UNIWERSYTET ŁÓDZKI

WYDZIAŁ EKONOMICZNO-SOCJOLOGICZNY

STUDIA STACJONARNE III STOPNIA

KIERUNEK: SOCJOLOGIA

MGR REMIGIUSZ ŻULICKI

NR ALBUMU: 5261

DATA SCIENCE W POLSCE. ETNOGRAFIA SPOŁECZNEGO ŚWIATA

Rozprawa doktorska
napisana w Katedrze Socjologii Kultury
pod kierunkiem
dra hab. Kazimierza Kowalewicza, prof. UŁ (emeritus)

ŁÓDŹ 2020

Spis treści

Rozdział 1. Uwagi wstępne

Czy sztuczna inteligencja pozbawia nas pracy? Algorytmy przejmują władzę nad światem? Czy big data sprawia, że jesteśmy bezustannie inwigilowani? Czy ogromna ilość danych zastępuje ekspertów i naukowców? Cokolwiek sądzimy na te tematy, jedno jest pewne – istnieje heterogeniczne środowisko ludzi, zajmujących się tzw. „sztuczną inteligencją" czy tzw. „big data" od strony technicznej i metodologicznej. To środowisko posługuje się rozmaitymi narzędziami matematycznymi i informatycznymi, wykonując operacje na danych cyfrowych za pomocą skryptowych języków programowania. Nie należy mylić ich z programistami (*software developers*). Pole ich działania nazywane jest data science (dalej DS), a oni data scientists. Nasza rozprawa poświęcona jest właśnie im, polskiemu środowisku DS, które traktujemy jako świat społeczny, wzorując się głównie na podejściu Adele E. Clarke. Celem głównym pracy jest opis etnograficzny polskiego społecznego świata data science.

Uzasadniamy wybór problematyki i celu pracy doktorskiej tym, że DS jest dynamicznym, zmiennym, nowym, bardzo słabo rozpoznanym i niejasnym zjawiskiem społecznym o poważnych konsekwencjach dla innych sfer życia. Niejasność spowodowana jest, naszym zdaniem, po pierwsze, wysokim poziomem zaawansowania matematyczno-informatycznego w substancjalnej sferze działalności data scientists; po drugie, omawianymi poniżej sporami definicyjnymi; po trzecie, ogromem szumu – nie tylko – medialnego i entuzjazmu wobec tej branży; po czwarte, charakterystyczną dla wysokobudżetowej branży nowoczesnych technologii kulturą tajemnicy:

> Tak jak czarnymi skrzynkami są często technologie AI, tak samo są nimi przemysłowe kultury, w których te technologie powstają. Rzadko możliwy jest wgląd w szczegóły systemów i ich otoczenia, właśnie z powodu korporacyjnych przepisów tajności (Crawford i in. 2018:11).

Z drugiej strony oczywistym jest to, że data scientists zajmują się formą komputerowej analizy danych cyfrowych. Skoro zatem pewni ludzie nazywają swoją działalność terminem *data science*, a siebie samych określają mianem *data scientists*; skoro inni ludzie chcą korzystać z usług tych pierwszych, albo dołączyć do nich i stać się data scientists; skoro wiadomo (choćby niezbyt precyzyjnie), co robią data scientists, to mówimy o DS jako o pewnej całości społecznej, wyróżnialnej z uwagi na działanie jej uczestników.

Pragniemy podkreślić, że rezultatem pracy data scientistów są elementy systemów technologicznych, nazywane potocznie „algorytmami" lub „sztuczną inteligencją", a bardziej techniczne – modelami uczenia maszynowego, z którymi na co dzień, często nieświadomie, w trakcie korzystania z takich narzędzi jak wyszukiwarka internetowa Google lub aparat fotograficzny w smartfonie, stykają się miliardy ludzi na całym świecie:

> Algorytmy w ostatnich latach stały się hasłem przewodnim (*catchworld*): ogniskiem publicznej fascynacji; rzadkim artefaktem, który wymaga nadzwyczaj wysokich wynagrodzeń dla tych, którzy je tworzą; piorunochronem dla tajemnicy biznesowej i „magiczną czarną skrzynką" dla tych, którzy z nich korzystają. Przy tym wszystkim, **bierzemy je za pewnik** [podkr. RŻ]. Kształtują dla nas kanały informacyjne (*newsfeeds*), personalizują nasze listy życzeń, włączają nam grzejniki i zoptymalizują nam ćwiczenia fizyczne. [algorytmy] To rzeczy codziennego użytku (Thomas, Nafus, i Sherman 2018:1).

Dojrzewanie i rozpowszechnianie się tzw. sztucznej inteligencji sprawia, że „algorytmy" (modele uczenia maszynowego) wtapiają się w ludzkie doświadczenie i otoczenie jako mediatorzy. Wpływają silnie, ale w sposób niemal niezauważalny – bo wygodny – na interakcje pomiędzy ludźmi i między ludźmi a otoczeniem (Taddeo i Floridi 2018a). Systemy te wdrażane są w rozmaitych dziedzinach życia, i tak jak w przypadku rozpoczętej kilkadziesiąt lat temu cyfryzacji (rozumianej jako stosowanie komputerów i internetu), stanowią one kolejne technologie ogólnego zastosowania. Tym samym jasne jest, że rezultaty pracy data scientistów wdrażane są nie tylko w sferze dóbr i usług konsumpcyjnych, ale także w bardziej wrażliwych obszarach, np. zatrudnienia, edukacji, ochronie zdrowia, opiece społecznej, transporcie, polityce, administracji publicznej, wymiarze sprawiedliwości.

Z uwagi na narosły wokół DS entuzjazm, dynamiczny rozwój tej dziedziny, a przy tym poważne konsekwencje społeczne – które omawiamy w części 2.2.3. – badanie etnograficzne, nastawione głównie na ogląd świata DS „od wewnątrz", uznajemy za istotny problem dla socjologicznej pracy doktorskiej.

1.1. Krótka historia terminów związanych ze światem społecznym data science

Big data, sztuczna inteligencja, uczenie maszynowe i data science. Występują one w mediach głównego nurtu jak i publikacjach akademickich. Są stosowane przez uczestników społecznego świata data science. W ostatnich dwunastu latach każdy z nich stał się coraz bardziej popularny, biorąc pod uwagę względną liczbę zapytań na te tematy, wpisywanych w

wyszukiwarkę internetową Google – por. rys. 1 – (Google Trends 2020a, 2020b). Różni autorzy nadają tym czterem terminom odmienne znaczenia i na wiele sposobów określają relacje zachodzące między nimi. Jeszcze większy chaos w tym zakresie panuje w dyskursie publicznym, w mediach i generalnie w gronie nie-technicznym (mamy na myśli także socjologię i inne nauki społeczne).

Przyglądamy się ewolucji analizowanych terminów i niejasnościom, które ich dotyczą. Ich znajomość jest drobnym, ale niezbędnym elementem wiedzy uczestnika świata społecznego data science (DS). Tym samym bez znajomości kilku znaczeń czterech wyróżnionych tutaj terminów nie moglibyśmy mówić o świece społecznym DS z perspektywy jego uczestników, a głównie taką perspektywę przyjmujemy w niniejszej rozprawie. Co do stosowanej konwencji zapisu, to termin big data zapisujemy zawsze małymi literami, chyba że cytujemy tekst w którym konwencja jest inna. Nie tłumaczymy na język polski terminów big data (z małymi wyjątkami, gdzie nie ma to negatywnego wpływu na zrozumiałość, a pozytywny na elegancję wywodu) ani data science (bez wyjątków), ponieważ tak postępują uczestnicy świata społecznego DS w Polsce. Stosujemy zamiennie określenia polsko- i anglojęzyczne dla terminów sztuczna inteligencja (*artiffcial intelligence,* AI) i uczenie maszynowe (*machine learning,* ML) z tego samego powodu.

Poniżej chcemy dać ogólne pojęcie o znaczeniu tych, jak sądzimy kluczowych w badanym świecie społecznym, terminów. W dalszych częściach pracy pokażemy m.in. co oznacza to, że uczestnicy społecznego świata DS kojarzą termin „big data" z narzędziem Spark, a termin „sztuczna inteligencja" z prezentacjami w Power Pointcie.

1.1.1. Big data

Termin „big data" prawdopodobnie powstał podczas rozmów w przerwie na lunch w firmie Silicon Graphics Inc. w połowie lat 90. Przyczynił się do tego John R. Mashey, jeden z głównych pracujących tam informatyków. Pierwsze referencje akademickie w dziedzinie informatyki to praca S. M. Weissa i N. Indurkhya („Predictive Data Mining. A practical guide" z 1998), w statystyce/ekonometrii zaś F. X. Diebolda („Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting", 2000 r.). Powyższe fakty przytaczamy za jednym z „ojców chrzestnych" terminu (Diebold 2012). Diebold użył go przypadkiem, nie słysząc terminu wcześniej, a samo określenie uznał za „trafne, dźwięczne i

intrygująco orwellowskie, szczególnie gdy zapisze się je z wielkich liter”[1] (2012:2). Big data zostało zdefiniowane w 2001 r. przez Douglasa Laneya z firmy doradczej Gartner jako dane, które charakteryzuje: wielkość (*Volume*), różnorodność (*Variety*), szybkość (*Velocity*) (Laney 2001). Ten sposób definicji był później wielokrotnie powtarzany jako „3 Vs of big data” (Berman 2013; Chen, Chiang, i Storey 2012; Cukier i Mayer-Schönberger 2014; Kwon, Lee, i Shin 2014; Ohlhorst 2013; Soubra 2012; TechAmerica 2012). Cechy te opisywane są następująco:

- Wielkość (*Volume*): rozmiar danych; bywa rozumiany jako zbiory większe niż 1 terabajt, a także takie, których nie da się przetwarzać za pomocą tradycyjnych narzędzi (relacyjnych baz danych i arkuszy kalkulacyjnych) (Gandomi i Haider 2015).
- Różnorodność (*Variety*): zróżnicowanie; różne struktury, formaty i charakter danych (np. teksty i fotografie pochodzące z blogów, stron www, mediów społecznościowych).
- Szybkość (*Velocity*): szybkość pojawiania się nowych danych; ciągły ich napływ, co powoduje, że potrzebne są metody umożliwiające wydobywanie informacji z danych w czasie rzeczywistym (Cukier i Mayer-Schönberger 2014; Gandomi i Haider 2015).

Powyższe wymiary pozostają w zależności – zmiana jednego powoduje zmianę drugiego. Twórca definicji „3Vs”, D. Laney, zobrazował to jako trzy różne osie w trójwymiarowym układzie współrzędnych (2001). Nie ma jednak konkretnych kryteriów dotyczących tego, gdzie „zaczyna się” big data (Gandomi i Haider 2015), choć można znaleźć żartobliwe stwierdzenia, jak „Big Data jest wtedy, gdy dane nie mieszczą się w Excelu”[2], czy poważniejsze „dane są wielkie, gdy wielkość danych zaczyna być problemem”[3] (Dutcher 2014). Filozof Luciano Floridi zauważył, że tego rodzaju definicje wielkości danych mówią nie o wielkości bezwzględnej, a o relacji w stosunku do dysponowanej mocy obliczeniowej, co prowadzi do rozwoju sprzętu i oprogramowania, które ma tę wielkość uczynić możliwą do skonsumowania (Floridi 2012:435–36). Pojawiają się także propozycje kolejnych cech czy wymiarów big data, również określane jako „Vs”. Są to (Gandomi i Haider 2015):

---

1 Oryginalnie: “To that end, I wanted a concise term that conjured a stark image. I came up with ‘Big Data’ which seemed apt and resonant and intriguingly Orwellian (especially when capitalized), and which helped to promote both goals.” [w całej pracy wykorzystuję własne tłumaczenia zródeł anglojęzycznych – RŻ].

2 Połączono oryginalne wyrażenia: “too big to fit in an Excel spreadsheet” oraz “the joke is that big data is data that breaks Excel” (tłumaczenie własne).

3 Oryginalnie: “data is big when data size becomes part of the problem” (tłumaczenie własne).

- wiarygodność (*Veracity*): dotyczy ona zagadnienia błędów w danych i ich prawdziwości;
- zróżnicowanie (*Variability and Complexity*): chodzi o duże zróżnicowanie wartości zmiennych i złożoność danych, które spowodowane są m.in. tym, że analizie poddawane są wszystkie dostępne dane, tzn. cała populacja a nie próba, co określa się podejściem N = all (Cukier i Mayer-Schönberger 2014);
- wartość (*Value*): rozumiana jako potencjał biznesowy, możliwości generowania zysków dzięki informacjom wydobytym z danych.

Zdaniem Karola Przanowskiego, fizyka pracującego w Szkole Głównej Handlowej, big data to nie tylko dane, jak wskazuje definicja „3Vs":

> (...) powinno się zatem definiować Big Data jako układ składający się z: danych opisanych własnościami 3Vs (5Vs), metod składowania i przetwarzania danych, technik zaawansowanej analizy danych oraz wreszcie całego środowiska sprzętu informatycznego. Jest to zatem połączenie nowoczesnej technologii i teorii analitycznych, które pomagają optymalizować masowe procesy związanie z dużą liczbą klientów, czy użytkowników.[4] (Przanowski 2014:13).

Dodać należy, że zasadniczym celem tego rodzaju analiz jest zazwyczaj stworzenie modelu matematycznego, wykonującego jakieś zadanie na podstawie danych.

Z punktu widzenia socjologa szczególnie ciekawe jest to, że zdaniem pewnych entuzjastów big data doprowadziło do przełomu cywilizacyjnego porównywalnego z wynalezieniem internetu, maszyny parowej czy druku (Cukier i Mayer-Schönberger 2014; Minelli, Dhiraj, i Chambers 2013). Autorzy jednej z najbardziej entuzjastycznych w tym temacie książki uważają, że big data to „rewolucja, która zmieni nasze myślenie, życie i pracę", (Cukier i Mayer-Schönberger [2013] 2014). Inni sądzą, że „big data może poznać nas lepiej niż sami siebie znamy"[5] (Gardner 2012). Wypowiedzi np. Drew Conway'a (firma Project Florida): „big data to ruch kulturowy, za pomocą którego kontynuujemy odkrywanie tego, jak ludzkość i świat oddziałują na siebie nawzajem"[6], czy też Daniela Gillicka (firma Google): „reprezentuje zmianę kulturową (…), coraz więcej decyzji podejmowanych jest za pomocą algorytmów działających na podstawie udokumentowanych, niezmiennych

4 Cytaty dłuższe niż 164 znaki ze spacjami wyróżniamy takim formatowaniem, bez cudzysłowu. Krótsze są ujęte w cudzysłów i nie wyróżnione w tekście.

5 Oryginalnie: „In a very real sense big data could know us better than we know ourselves." (tłumaczenie własne).

6 Oryginalnie: "Big data is a cultural movement by which we continue to discover how humanity interacts with the world." (tłumaczenie własne).

dowodów”[7] (Dutcher 2014) również wskazują na zmianę. Termin „big data” ogromną popularność zdobył w szeroko pojętym biznesie:

> Big data ma szanse zostać słowem roku. Chyba trzeba spędzić ostatnie miesiące na Księżycu, żeby nie myśleć o wdrożeniu big data w swojej branży. (...) Big data zmieniło historię: chodzi mi o wybór Donalda Trumpa i Brexit [mówi się o istotnej roli analizy danych „big data” w wywieraniu wpływu na głosujących – przyp. RŻ, por część 2.2.3.4. tej pracy]. Big data jest buzzwordem, nie ma konferencji biznesowej bez tego tematu[8] {notatka badacza, obserwacja 10}

Mówi się o końcu zapotrzebowania na ekspertów dziedzinowych – big data ma zapewniać lepsze decyzje, bez odwoływania się do jakoby ograniczających teorii czy intuicji; wystarczy „pozwolić przemówić danym” (Cukier i Mayer-Schönberger [2013] 2014: 19, 35, 185).

Pojawiają się teksty mówiące o światopoglądzie, filozofii czy religijnej wierze w dane – dataizmie (*dataism*), który zakładać ma, że każdy układ czy zjawisko o dowolnej naturze sprowadza się do przetwarzania danych. „Dataiści są sceptyczni wobec ludzkiej wiedzy czy mądrości, skłaniając się do zaufania w Big Data i algorytmy komputerowe” (Harari 2017:213–14). Istnieje strona internetowa dataists.com, na której opublikowano m.in. bardzo popularny diagram, prezentujący czym jest „nauka o danych” (*data science*) (Conway 2010), o czym piszemy dalej (por. rys. 2). Entuzjazm wokół big data był przedmiotem żartów już w 2013 r. Szeroko powtarzany i komentowany był wpis w mediach społecznościowych Dana Ariely (psychologa i ekonomisty): „big data jest jak seks u nastolatków: wszyscy o tym rozmawiają; nikt naprawdę nie wie, jak się to robi; każdy sądzi, że robią to inni, więc wszyscy twierdzą, że również to robią”[9] (Ariely 2013).

Powstały różnego rodzaju prace akcentujące zarówno pozytywne (Eagle i Greene 2014; Gardner 2012) jak i negatywne skutki big data (O’Neil 2013, 2017a; Surma 2017). Wśród pierwszych mówi się np. o zdrowszym i łatwiejszym życiu jednostek (Eagle i Greene 2014:31–50), optymalizacji ruchu ulicznego czy wykrywaniu aktywności przestępczych w przestrzeni miejskiej (Eagle i Greene 2014:85–98), ogólnie o zmienianiu świata na lepsze

7 Oryginalnie: “Big data represents the cultural shift in which more and more decisions are made by algorithms with transparent logic, operating on documented immutable evidence.” (tłumaczenie własne).

8 Otwierająca spotkanie wypowiedź Kamili Łepkowskiej, prowadzącej panel „Cyfrowe ślady, big data i polityka” w ramach wydarzenia „Igrzyska Wolności 2017 RE:wolucja / Cyfrowa Rzeczywistość” w Łodzi, zanotowana przez autora rozprawy (RŻ) podczas obserwacji uczestniczącej w dniu 21.10.2017 r. około godziny 18:15. Cytowania materiałów pozyskanych w ramach badań własnych oznaczamy w tekście jako {technika badawcza numer} np.{obserwacja 8}, i w odróżnieniu od cytatów z literatury zaznaczamy graficznie pionową, szarą linią z lewej strony tekstu. Informacje o dacie i innych cechach każdego aktu badawczego podajemy w Aneksie.

9 Oryginalnie: „Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it... ” (tłumaczenie własne).

dzięki rozwiązywaniu wcześniej nierozwiązywalnych problemów (Gardner 2012:14–15). Drugie poruszają np. kwestię niepewności i niedokładności wyników analiz dużych ilości danych (O'Neil 2013; Surma 2017), przedkładaniu skuteczności i szybkości działania algorytmów nad sprawiedliwość, co skutkuje m.in. kryminalizacją biednych dzielnic (O'Neil 2017a:134–39), czy zamienianiu nieświadomych tego klientów w produkty, poprzez sprzedawanie ich uwagi reklamodawcom (Surma 2017:66–69).

Temat „big data" był od lutego 2012 do marca 2017 popularniejszy w wyszukiwarce Google niż „data science" i uczenie „maszynowe" (rys. 1). Później jego popularność spadła nieznacznie, ale silnie wzrosła dla dwóch wymienionych tematów. Niezmiennie najbardziej popularnym z analizowanych był jednak temat „sztuczna inteligencja" (por. rys. 1A). W maju 2020 r., gdy jego popularność[10] osiągnęła maksymalną wartość 100, dla tematów „big data" popularność była równa 10, dla „uczenia maszynowego" 21, zaś „data science" 16.

1.1.2. Sztuczna inteligencja

Sztuczna inteligencja (*artificial intelligence*, AI) to, jak zaraz wskażemy, powiązany z big data termin. Za twórcę określenia „sztuczna inteligencja" uznawany jest matematyk John McCarthy, który użył go w opublikowanym dnia 31 sierpnia 1955 r. zaproszeniu do udziału w letnich warsztatach badawczych w Darmouth (McCarthy i in. 1955). Współcześnie AI definiowana jest zazwyczaj na cztery sposoby. Mówi się o systemach komputerowych, które: myślą w ludzki sposób (*thinking humanly*) lub myślą racjonalnie (*thinking rationally*) lub działają w ludzki sposób (*acting humanly*) lub też działają racjonalnie (*acting rationally*) (Russel i Norvig 2009:1–2). AI służy wielu zastosowaniom praktycznym: „automatyzacji różnego rodzaju rutynowej pracy, rozpoznawaniu obrazów, rozpoznawaniu mowy, wsparciu diagnoz medycznych czy badań naukowych". Początkowo AI rozwiązywała problemy „intelektualnie trudne dla ludzi, ale stosunkowo proste dla komputerów", czyli takie, które można opisać w sposób formalny (jak gra w szachy). Obecnie „wyzwaniem dla AI stały się problemy łatwe do wykonania dla ludzi, ale trudne do opisania formalnie" – takie, które ludzie rozwiązują intuicyjnie (rozpoznawanie mowy czy twarzy na obrazach) (Goodfellow, Bengio, i Courville 2016:1–2). Popularność tematu „sztuczna inteligencja" w wyszukiwarce

10 Miara nazywana przez Google popularnością (oś y na rys. 1.) reprezentuje względną częstość wyszukiwania . Przyjmuje wartości od 0 do 100 z dokładnością do 1; pojawia się też wartość „< 1". Miara uwzględnia skalę korzystania z wyszukiwarki Google ogółem, przedstawia wartość w liczbach względnych. Liczba 100 reprezentuje więc maksymalną popularność zanotowaną dla danego okresu i zakresu geograficznego. Wartość 50 oznacza, że dany temat był o połowę mniej popularny. Zero oznacza, że nie było wystarczającej ilości danych dla tego tematu (Google Trends 2020a).

Google była wielokrotnie wyższa, niż pozostałych trzech analizowanych (por. rys. 1). Temat był co najmniej czterokrotnie bardziej popularny, niż najbardziej popularny z pozostałych trzech tematów – w styczniu 2020 popularność „sztucznej inteligencji" była równa 91, podczas gdy dla „uczenia maszynowego" było to 21; to najniższa różnica popularności (rys. 1A). W analizowanym okresie średnia popularność tematu „sztuczna inteligencja" wynosi 76, dla „uczenia maszynowego" jest to 8, dla „big data" 7 i dla „data science" 5 (Google Trends 2020a).

Apokalipsą ludzkości wywołaną przez AI grozili m.in. słynny przedsiębiorca technologiczny Elon Musk[11] i astrofizyk Stephen Hawkins[12]. To z ich udziałem powstały „23 reguły AI z Asilomar". We wstępie napisano:

> AI dostarczyła już korzystnych narzędzi, które są używane codziennie przez ludzi na całym świecie. Jej dalszy rozwój, kierowany następującymi zasadami, będzie oferował niesamowite możliwości pomocy i wsparcia ludzi w nadchodzących dziesięcioleciach i wiekach (Future of Life Institute 2017).

Podobnie jak w przypadku entuzjastów big data, tam również mówi się o rewolucji czy przełomie. Andrew Ng, cieszący się ogromnym autorytetem badacz (Uniwersytet Stanforda) i praktyk (wcześniej Google, później chiński konkurent Google, firma Baidu) sztucznej inteligencji stwierdził, że „AI to nowa energia elektryczna", i zmieni całkowicie oblicze biznesu, tak jak przed stu laty zrobiła to elektryfikacja (Ng 2017). Temat AI chętnie podejmowany jest też w medialnych dyskusjach o tym, czy „roboty zabiorą nam pracę". Od sierpnia 2017 pisały o tym m.in. magazyny: „Wired" (Surowiecki 2017), „Forbes" (Ozimek 2017), „Independent" (Barnet 2017), „The New York Times" (Williams 2017) czy „The Guardian" (Elliott 2018). Istnieje strona internetowa willrobotstakemyjob.com, na której można wygodnie sprawdzić ryzyko automatyzacji dla konkretnego zawodu. Strona powstała w oparciu o prognozę naukowców z Uniwersytetu Oksfordzkiego, którzy podjęli ten temat z uwagi na rozwój technologii uczenia maszynowego (*machine learning*) będących subdyscypliną AI, oraz rozwój big data (Frey i Osborne 2017). Inny zespół z tej samej uczelni przeprowadził badanie ankietowe na próbie 352 naukowców z dziedziny uczenia maszynowego. Badani naukowcy przewidują, że sztuczna inteligencja będzie przewyższać ludzi w ciągu najbliższych dziesięciu lat. Mowa o zadaniach takich jak: tłumaczenie z języków obcych (do 2024 r.), pisanie esejów na poziomie szkoły średniej (do 2026 r.),

11 „AI jest naszym największym zagrożeniem egzystencjalnym", „za pomocą AI przywołujemy demona" (UNECE 2014).

12 „Myślę, że rozwój pełnej sztucznej inteligencji może oznaczać koniec rasy ludzkiej" (Clark 2014).

prowadzenie ciężarówki (do 2027 r.), praca w handlu detalicznym (do 2031 r.), napisanie bestsellerowej książki (do 2049 r.) i praca chirurga (do 2053 r.). Respondenci sądzą, że istnieje 50% szans na AI przewyższającą ludzi we wszystkich zadaniach w ciągu 45 lat od roku 2016, a na automatyzację wszystkich ludzkich stanowisk pracy w ciągu 120 lat (Grace i in. 2017:1–2).

Hiperentuzjastą sztucznej inteligencji jest Ray Kurzweil, którego cechuje religijna wiara w rozwój technologiczny. Jego pojęcie Osobliwości (*Singularity*) oznacza chwilę w rozwoju ludzkości, w której człowiek przekroczy granice biologiczne – ograniczeń swojego umysłu i ciała, z osiągnięciem nieśmiertelności włącznie (Kurzweil 2016:23–24). Twierdzi on, że:

> Najpoważniejszą z trzech głównych rewolucji leżących u podstaw Osobliwości (G, N oraz R [Genetyka, Nanotechnologia, Robotyka]) jest rewolucja R, która odnosi się do tworzenia inteligencji niebiologicznej przekraczającej inteligencję nieulepszonych ludzi. Bardziej inteligentny proces w naturalny sposób wygra z procesem mniej inteligentnym, dzięki czemu inteligencja stanie się najpotężniejszą siłą we wszechświecie. Choć R w GNR oznacza robotykę, w rzeczywistości chodzi o potężną SI (sztuczną inteligencję przekraczającą inteligencję ludzką). W tym wzorze zazwyczaj podkreśla się robotykę, ponieważ inteligencja potrzebuje ucieleśnienia, fizycznej obecności, aby wywrzeć wpływ na świat. Nie zgadzam się z tym dążeniem do fizycznej obecności, bo moim zdaniem najważniejszym problemem jest inteligencja. Inteligencja w naturalny sposób znajdzie metodę, by wpłynąć na świat, być może tworząc własne środki w celu pozyskania bytu i fizycznej manipulacji (Kurzweil 2016:251).

Pojęcie sztucznej inteligencji obecne jest w popkulturze. Jak wskazuje kulturoznawczyni Rozalia Knapik, teksty w których nadejście tzw. Osobliwości jest traktowane jako istotne zagrożenie dla ludzkości pojawiają się w kulturze popularnej o wiele częściej niż te, pokazujące inteligentne maszyny z podziwem, pokorą czy pobłażaniem dla ich służebnej roli. Autorka uważa, że popularność negatywnych wizji sztucznej inteligencji świadczy o trudnościach ze skonstruowaniem przekonującej wizji afirmatywnej. Dodaje jednak, wskazując m.in. na dzieła Stanisława Lema, Philipa K. Dicka i braci Wachowskich, że obawa przed Osobliwością jest niejednoznaczna i obok lęku zawiera też element fascynacji tym wytworem człowieka (Knapik 2018:11). Knapik twierdzi, że szczególnie sci-fi bazuje na obawie wobec nieprzygotowania ludzkości do korzystania ze zdobyczy technologii i zbyt ufnej symbiozie człowieka z technologią, wskazując na jej zdaniem kanoniczne książki: *Nowy Wspaniały Świat* A. Huxley'a i *Rok 1984* G. Orwella. Za współczesną grę z tymi dziełami uznaje brytyjski serial „Czarne Lustro" Charlie'go Brookera (ibidem:29). Autorka koncentruje się na związkach przedstawień AI z obrazowaniem religijnym i twierdzi, że rolę Boga „odgrywają na przemian konstruktorzy myślących maszyn i same sztuczne inteligencje" (Knapik 2018:13). Bezcielesność sztucznej inteligencji ma stawiać ją w „drabinie bytów"

wyżej niż gdyby posiadała ciało. Przy tym, z uwagi na trudność w zrozumieniu AI, jej działanie stanowi dla laika pokaz magiczny; laik wierzy w AI „na słowo". W tekstach kultury – filmach „Odyseja kosmiczna", „Terminator", „Ja, Robot", „Colossus" – pojawia się wątek skrajnie racjonalnego zbawienia ludzkości w postaci unicestwienia ludzi (Knapik 2018:113-129). To prowadzi do konkluzji, że takie przedstawienia stanowią „nową realizację prastarego zjawiska, jakim jest tworzenie sobie Boga", a „ludzie obawiają się Boga, ale uparcie wypatrują jego nadejścia" (2018:130; 188).

1.1.3. Uczenie maszynowe

Uczenie maszynowe (*machine learning*) jest jedną z najsilniej powiązanych z big data subdyscyplin AI (Goodfellow i in. 2016:9; Russel i Norvig 2009:2).

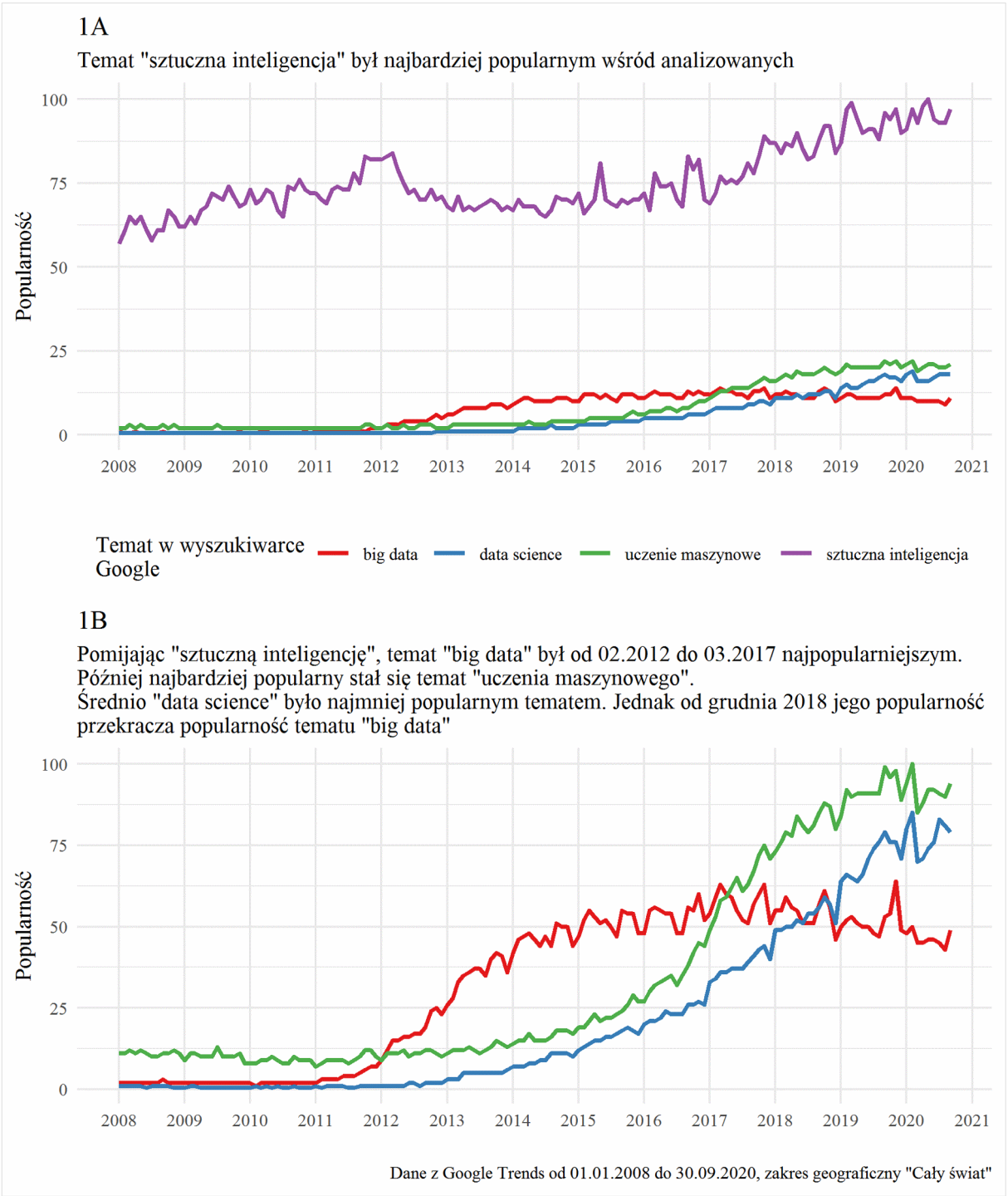
> W drugiej połowie XX w. uczenie maszynowe wyewoluowało z badań nad sztuczną inteligencją, w których projektowano samouczące się algorytmy, zdolne do pozyskiwania wiedzy z informacji oraz tworzenia na ich podstawie prognoz. Dzięki uczeniu maszynowemu nie trzeba zatrudniać ludzi do ręcznego określania reguł oraz tworzenia modeli poprzez analizowanie olbrzymich pokładów danych; omawiana dziedzina wiedzy oferuje efektywniejsze rozwiązanie polegające na stopniowym poprawianiu skuteczności modeli predykcyjnych oraz podejmowaniu decyzji na podstawie analizowanych danych (Raschka 2018:26).

Techniki uczenia maszynowego to sposoby na budowanie tzw. modeli, o których mowa w związku z big data: od znanych również w naukach społeczno-ekonomicznych modeli statystycznych, np. regresja liniowa, regresja logistyczna, po złożone od strony matematycznej i informatycznej modele, np. sztucznych sieci neuronowych (*artificial neural networks*) czy maszyny wektorów nośnych (*support vector machine*, SVM) (Larose 2005, 2012; Raschka 2018). Pojawiają się opinie, że big data – w sensie ilości danych, ich dostępności i możliwości przetwarzania, ale również innego nastawienia[13] do analizy danych – umożliwia rozwój uczenia maszynowego (Canton 2016; Hansen 2017; Kerzel 2017; Marr 2016). Przykładem tego jest Tłumacz Google. Dostęp do ogromnej ilości danych oraz przesunięcie akcentu z dokładności na szybkość obliczeń pozwoliło już na etapie prototypowej sieci neuronowej uzyskać tłumaczenie o 7 punktów BLEU lepsze[14] niż wcześniejsza wersja tłumacza, gdzie reguły programowano (Lewis-Kraus 2017). Zainteresowanie tematem „uczenie maszynowe" w wyszukiwarce Google przekroczyło

13 Mówi się o trzech zmianach: analizowaniu całej populacji zamiast próby, zmniejszeniu dokładności obliczeń na rzecz ich szybkości i zmianę pytania będącego podstawą budowania modeli z „dlaczego?" na „co?" (por. Cukier i Mayer-Schönberger [2013] 2014).

14 Test dla tłumaczenia z języka angielskiego na francuski; „poprawa o 1 punkt była postrzegana jako spory sukces, poprawę o 2 punkty postrzegano jako wybitną" (Lewis-Kraus [2016] 2017:153).

zainteresowanie „big data” w maju 2017 i utrzymuje się na poziomie wyższym, niż dla tematów „big data” i „data science” do września 2020 (rys. 1). Wyłączając popularność dla „sztucznej inteligencji”, w analizowanym okresie średnia popularność tematu „uczenie maszynowe” wynosi 35, dla „big data” jest to 33, zaś 22 dla „data science” (Google Trends 2020b). W momentach maksymalnej popularności – wrześniu 2019 i lutym 2020 – temat „uczenie maszynowe” był dwa razy bardziej popularny niż „big data” (por. rys. 1B).



Rysunek 1: Popularność tematów: "big data", "uczenie maszynowe", "data science", "sztuczna inteligencja" w wyszukiwarce Google

Źródło: 1A – popularność dla czterech tematów (Google Trends 2020a), 1B – popularność z wyłączeniem tematu „sztuczna inteligencja" (Google Trends 2020b); opracowanie własne za pomocą GNU R

Uwaga 1: Poprawiony system gromadzenia danych Google od 01.01.2016

Uwaga 2: Nazwę tematu „data science" zmieniono z tłumaczenia Google „danologia", aby zachować przyjętą konwencję

Uwaga 3: Google udostępnia dane z dokładnością do 1, więc wartość zmiennej Popularność równą „< 1" zmieniono na „0,5", aby zachować ilościowy charakter zmiennej

Przedstawione na rysunku 1. linie obrazują popularność jako względną częstość wyszukiwania tematów (nie jest ważna wielkość liter) w wyszukiwarce Google na całym świecie. Na osi y przedstawiono czas. Wykorzystano dane w interwale miesiąca od początku stycznia 2008 do końca września 2020.

1.1.4. Data science

Data science to dziedzina określana jako integrująca i wdrażająca w praktyce wszystkie omawiane powyżej pola: big data, sztuczną inteligencję i uczenie maszynowe. Nie znaczy to jednak, że każdy człowiek zajmujący się DS wykorzystuje np. big data, czy każdy pracujący z big data lub ML zajmuje się DS. W ogóle z tak ujętą integracją czy wdrażaniem AI, big data i ML przez DS niekoniecznie zgadzają się uczestnicy świata społecznego DS w Polsce, o czym więcej powiemy omawiając wyniki badań własnych.

Osoby zajmujące się DS nazywane są data scientists. Termin ten zazwyczaj nie jest tłumaczony na język polski, choć bywa prezentowany jako inżynier danych, analityk, mistrz danych (Przanowski 2014; Wikipedia b.d.). W Polsce wśród uczestników świata społecznego DS termin nie jest tłumaczony, ale bywa odmieniany zgodnie z regułami języka polskiego, np. „kilku data scientistów", „data scientistka", „data scientiści", „[kogo?] data scientista / scientisty", [jestem] „data scientistą / scientistem / scientistką", co wiemy z badań własnych. Pojęcia DS po raz pierwszy użyto w 1974:

> Data science to nauka zajmująca się danymi po ich pozyskaniu i zapisaniu, podczas gdy problematyka związku danych z tym, co mają one reprezentować, jest pozostawiona innym polom i dziedzinom nauki (Naur 1974).

Pierwsze użycie pojęcia we współczesnym sensie datowane jest jednak na rok 2001 (Andrus, Cook, i Sood 2017:1; Donoho 2015:13). Chodzi o pracę „Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics" (Cleveland 2001). Autor, wieloletni współpracownik Johna Tuckeya w laboratorium badawczym firmy Bell i profesor statystyki, postulował rozszerzenie technicznych obszarów statystyki uniwersyteckiej tak, by lepiej wspierać analityków danych (*data analyst*) pracujących dla rządowych lub komercyjnych podmiotów (Donoho 2015:13). Wiliam Cleveland zaproponował sześć obszarów aktywności dla uniwersytetów, sugerując rozkład pracy: badania multidyscyplinarne (25%), metody i modele analizy danych (20%), problemy obliczeniowe dla danych (15%), nauczanie (15%), ewaluacja narzędzi analitycznych (5%), teoria statystyczna (20%). David Donoho wskazuje, że czternaście lat po publikacji artykułu

Clevelanda zdecydowana większość statystyki akademickiej w USA wciąż poświęca 100% swojej pracy na teorię (2015:13). Nieco dalej Donoho powołuje się na pracę Leo Breimana (2001), również praktyka analiz w biznesie i profesora statystyki, wskazującego na istnienie dwóch kultur modelowania statystycznego. Jedna z nich, kultura modelowania generatywnego (*generative modeling*) ma być nastawiona na wnioskowanie (*inference*) o naturze zjawiska stojącego za związkiem statystycznym pomiędzy zmiennymi; podejście to ma przejawiać 98% statystyków akademickich. Inna, kultura modelowania algorytmicznego, ma na celu wyłącznie uzyskanie wyniku o jak najwyższym dopasowaniu do danych, przy całkowitym przemilczeniu rzeczywistych zjawisk przyrodniczych czy społecznych „stojących za" danymi. To podejście charakterystyczne jest dla 2% statystyków akademickich, ale także dla informatyków (*computer scientists*) i praktyków statystyki w biznesie (*industrial statisticians*). Tutaj właśnie stosuje się uczenie maszynowe, nazywane „epicentrum kultury modelowania algorytmicznego". Donoho dodaje, że uczeniem maszynowym od strony akademickiej zazwyczaj zajmują się wydziały informatyczne (*computer science departments*), nie statystyczne (2015:14-15). Zwracamy uwagę, że w entuzjastycznej pracy *Big Data: rewolucja, która zmieni nasze myślenie, życie i pracę* autorzy powtarzają wielokrotnie, że w erze big data obowiązuje pytanie „co?" zamiast pytania „dlaczego?" (por. Cukier i Mayer-Schönberger [2013] 2014:19, 35, 185). Pytanie „co?" jest zatem charakterystyczne dla kultury modelowania algorytmicznego (a więc raczej także dla DS), zaś „dlaczego?" dla modelowania generatywnego.

Do spopularyzowania terminu DS przyczyniły się głównie dwa artykuły prasowe: „What is Data Science" (Loukides 2010) i „Data Scientist: The Sexiest Job of the 21st Century" (Davenport i Patil 2012). Na artykuły te wskazują prace poświęcone historii DS (Andrus i in. 2017; Cao 2017a; Donoho 2015). W obu artykułach przytoczono słowa Hala Variana, głównego ekonomisty Google z wywiadu dla „The New York Times" w 2009. Powiedział on, że:

> W ciągu najbliższej dekady seksownym zajęciem będzie statystyka. Ludzie myślą, że żartuję, ale kto odgadłby, że seksownym zajęciem lat 90. będzie inżynieria informatyczna?[15] (Lohr 2009).

W obu artykułach mowa o możliwościach, jakie dają big data i DS, oraz o rosnącym w niesamowitym tempie zapotrzebowaniu na specjalistów z dziedziny DS. Wyraźnie widać, że

15 Oryginalnie: „The sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?" (tłumaczenie własne).

choć H. Varian użył słowa „statystyk", to Loukides oraz Davenport i Patij używają już terminu „data scientist". Szczególnie w tym drugim tekście przeinaczono oryginalną wypowiedź (poza nazwą profesji zmieniono najbliższą dekadę na całe stulecie), a dodatkowo wrażenie na czytelniku wywiera zawarcie wszystkich kluczowych słów w tytule *Data Scientist: The Sexiest Job of the 21st Century*.

W wyszukiwarce Google temat „data science" był w analizowanym okresie zawsze mniej popularny niż „sztuczna inteligencja" i „uczenie maszynowe" (rys. 1). Od grudnia 2018 jego popularność przekroczyła popularność tematu „big data" i utrzymuje się na wyższym niż dla niego poziomie. Średnio „data science" było najmniej popularnym tematem spośród omawianych (Google Trends 2020a, 2020b). Niemniej od czerwca 2012 r. jego popularność w wyszukiwarce Google ma trend rosnący. Jest silnie, dodatnio skorelowana[16] z popularnością tematu „uczenie maszynowe".

Data science jest często definiowane jako nowa dyscyplina nauki i praktyki, rodząca się na styku bardziej dojrzałych pól. Andrus z zespołem użyli metafory „dziecka" powstałego z „rodziców": ogólnej metodologii nauk, inżynierii danych, inżynierii oprogramowania, statystyki, wizualizacji danych i hakerstwa (jak podkreślają „w pozytywnym tego słowa znaczeniu") (Andrus i in. 2017). Ponieważ ich książka jest zasadniczo podręcznikiem wprowadzającym do DS, autorzy zaznaczają złożoność tej nowej dyscypliny mówiąc, że zapoznanie się z ich podręcznikiem nie czyni ze studenta eksperta, bo „prawdziwe DS jest bardziej odpowiednio nauczane na poziomie magisterskim i doktoranckim", i ponieważ DS powstaje „w świecie big data", to wymaga m.in. wiedzy i umiejętności z dziedziny rozproszonego przetwarzania petabajtowych zbiorów danych pochodzących z baz nierelacyjnych, znajomości uczenia maszynowego i zaawansowanej statystyki (ibidem). Donoho idzie jeszcze dalej, wskazując, że zdobycie umiejętności potrzebnych do pracy jako data scientist wymagać może lat doświadczenia zawodowego, już z dyplomem magistra (2017: 8-9). Z drugiej strony są autorzy sugerujący, że pracować w DS mogą z powodzeniem zarówno absolwenci informatyki (*computer science major*) – tych nazwano hakerami, jak i statystycy (*statistics major*) – których nazwano piszącymi skrypty, oraz absolwenci kierunków managerskich (*MBA*) – użytkownicy oprogramowania statystycznego (Barlow 2013:8–9).

16 Dla danych o trzech opisywanych tematach (Google Trends 2020b) współczynnik korelacji popularności tematów „data science" i „uczenie maszynowe" metodą rang Spearmana wynosi $r_s = 0{,}96$.

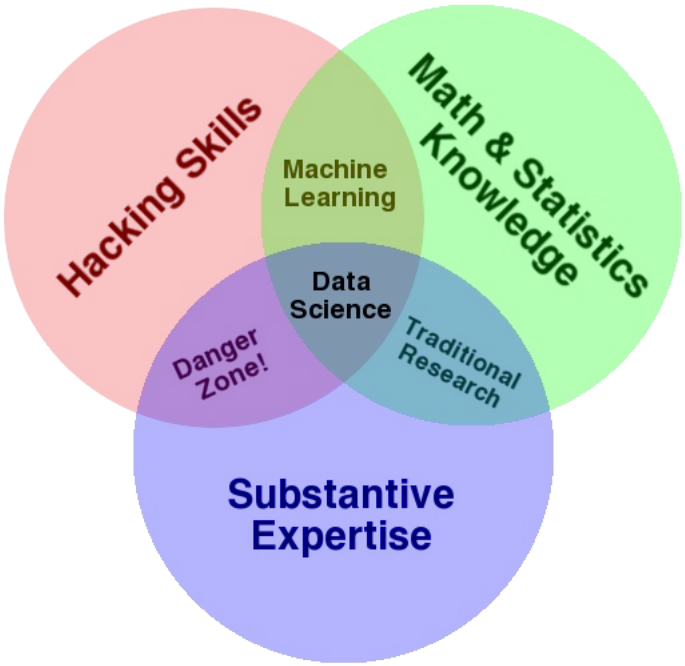
Już w 2013 r. dostrzeżono, że niejednoznaczność połączona z coraz większym rozgłosem i entuzjazmem wokół terminów „big data" i „data science" prowadzić może m.in. do problemów w zatrudnianiu data scientistów (Harris, Harlan, Murphy, Sean, i Vaisman 2013). Trudno dobrać właściwą osobę do projektu, kiedy pracodawca i kandydaci nadają tym terminom różne znaczenie. Te popularne terminy nazwano brutalnie „maszynką do mielenia modnych powiedzonek" (*buzzword meat grinder*). Autorzy podają przykład rozmowy kwalifikacyjnej, na której rekruter zapytany, jakiego pracownika konkretnie szuka powiedział:

> Chcę BOGA! Chcę gwiazdora rocka w programowaniu, który stworzył najbardziej wyszukane algorytmy uczenia maszynowego, zbudował platformę big data do rozproszonych obliczeń i założył swoją firmę! (Harris, Harlan i in. 2013:3–6).

Jerzy Surma we wstępie do swojej krytycznej pracy ujął rzecz następująco:

> (…) ta książka to rodzaj repulsji na to, co głoszą media głównego nurtu na temat Big Data. (…) Metody te niejednokrotnie są znane i stosowane od lat, ale w narracji nieświadomych dziennikarzy są emanacją inteligencji porównywalnej z ludzką. Ten **bezkrytyczny opis metod sztucznej inteligencji, irytująca antropomorfizacja, przy jednoczesnym braku podstawowej wiedzy na temat funkcjonowania tego typu algorytmów, przekroczyły – moim zdaniem – granicę akceptowalnej kuriozalności** [podkr. RŻ] (2017: 7-8).

W konsekwencji spory, niejednoznaczność i dość wysoki poziom emocji, towarzyszą próbom zdefiniowania czym jest DS i data scientist. Szeroko powtarzaną jest definicja Drew Conway'a (Conway 2010), gdzie dyscyplinę ukazano na przecięciu wiedzy matematyczno-statystycznej, umiejętności hakerskich i doświadczenia dziedzinowego. Niewątpliwie do popularności tego ujęcia przyczyniło się zwizualizowanie go w formie diagramu Venna (rys. 2).



Rysunek 2: Definiowanie data science – popularny diagram

Źródło: http://www.dataists.com/2010/09/the-data-science-venn-diagram/ (Conway 2010)

Sześć lat później na portalu Kdnuggets[17] opublikowano utrzymany w żartobliwym tonie artykuł *Battle of the Data Science Venn Diagrams* (Taylor 2016), w którym omówiono ukazany powyżej, najwcześniejszy diagram Conway'a oraz dwanaście (sic!) innych, powstałych w kolejnych latach. Autor we wstępie zaznacza, że postanowił uniknąć kontrowersji związanych z terminem „data scientist" i nazywa siebie „grotołazem danych (*data spelunker*)". W kolejnym zdaniu napisał: „data minerzy[18] i tak wyszli już z mody (*data 'miners' are out of vogue anyway*) (ibidem).

W nieco mniej żartobliwym tonie utrzymane są definicje DS (14 różnych) zebrane przez portal bigdata-madesimple.com (Baitu 2014). Mówi się m.in. o DS jako o „rzadkiej hybrydzie", a o data scientiście jako „Kolumbie i Colombo w jednym – wygłodniałym badaczu i sceptycznym detektywie". Większość zebranych tam wypowiedzi stosuje podobną metaforykę. Pojawiają się definicje mówiące o data scientistach jako „ludziach renesansu"; często powtarzana jest wypowiedź Josha Willsa: „data scientist to osoba lepsza w statystyce niż każdy programista, i lepsza w programowaniu od każdego statystyka" (Delapenha 2017). Popularny żart definicyjny (piszący te słowa słyszał go wielokrotnie podczas obserwacji) mówi, że „data scientist to analityk danych, który mieszka w Kaliforni" (Baitu 2014; Jarvis 2014). W podobnym tonie utrzymanych jest wiele wypowiedzi, np.: „data scientist to statystyk pracujący na MacBooku" (Big Data Borat 2013), „data science to domena ludzi, którzy decydują się wydrukować 'data scientist' na swoich wizytówkach i dostać podwyżkę" (Taylor 2016). Dyskusje toczą się wokół tego, czy DS to tylko nowa nazwa (*rebranding*) statystyki, czy też, że statystyka jest tylko mało znaczącą częścią DS (Donoho 2015:5–6; Hyndman 2014). Josh Bloom, profesor astronomii i pracownik Berkeley Institute for Data Science, zasłynął wypowiedzią: „pierwszą zasadą data science jest: nie pytaj o definicję data science"[19] (Azam 2014). W artykule mowa o tym, że w dzisiejszych czasach analiza danych jest tak popularna, że właściwie każdy badacz zajmuje się DS. Sam termin nie ma zatem żadnego konkretnego znaczenia i oznacza „różne rzeczy dla różnych ludzi"[20](Azam 2014). Pytanie „czym jest data science" jest „motywem przewodnim, centralnym zagadnieniem i mantrą" pracy statystyczki Rachel Schutt i matematyczki Cathy O'Neil (2015:303). Choć, jak

17 Wielokrotnie nagradzanym w branżowych konkursach „wiodącym portalu internetowym poświęconym analityce biznesowej, big data, data mining, data science i uczeniu maszynowemu"; https://www.kdnuggets.com/about/index.html, dostęp 5.01.2018 r.

18 Określenie „data mining" było powszechnie używane w przywoływanych już podręcznikach poświęconych analizie danych i uczeniu maszynowemu (por. Larose 2005, Larose [2006] 2012).

19 Oryginalnie: "The first rule of data science is: don't ask how to define data science" (tłumaczenie własne).

20 Oryginalnie: „data science is a meaningless term that means different things to different people" (tłumaczenie własne).

zaznaczyły one już na wstępie, ogrom szumu medialnego wokół DS spowodował ich sceptyczne podejście, to ostatecznie doszły do wniosku:

> Data science[21] jest zbiorem najlepszych praktyk stosowanych w firmach technologicznych operujących w szerokiej przestrzeni problemów, które można rozwiązywać za pomocą danych, i być może niekiedy zasługuje na miano nauki. Mimo to czasami nie jest niczym więcej niż szumnym określeniem, którego powinniśmy się wystrzegać i którego dodawania na siłę powinniśmy unikać (Schutt i O'Niel [2014] 2015:305).

To „rozwiązywanie problemów" było dla nas szczególnie ważną kategorią analizy jakościowej, zarówno w odniesieniu do działania podstawowego DS jako świata społecznego (por. część 5.2. rozprawy), jak i wartości tego świata (por. 5.4). Te same autorki ciekawie stwierdziły, że choć może nie istnieje coś takiego jak DS, to niewątpliwie istnieje praca w DS i wielki popyt na data scientistów (Schutt i O'Niel [2014] 2015:26). Nie można odmówić im racji, a właściwie trafności przewidywań: praca / posada (*job*) data scientist została uznana przez Glassdoor (jeden z największych amerykańskich portali z ofertami pracy) za najlepszą spośród około 1 700 badanych posad trzykrotnie w latach 2016, 2017 i 2018 (Piatetsky 2018a). O gwałtownie rosnącym popycie na pracowników zespołów DS pisze się wiele. Mowa np. o tym, że kariera w DS staje się bardziej atrakcyjna niż prawo lub medycyna (Burtch 2014); o bardzo silnym w tej branży rynku pracownika i doradza się pracodawcom strategie rekrutacyjne (PwC 2015); o prognozie popytu na data scientistów dla USA w 2018 r. rzędu 440 – 490 tys. stanowisk przy podaży około 300 tys. osób – tym samym popyt przewyższa podaż o około 50%, tzn. prawie 1/3 oferowanych stanowisk ma być w 2018 nie obsadzona (James i in. 2011); o najszybszym tempie wzrostu liczby stanowisk pracy na portalu LinkedIn dla posad „Machine Learning Engineer" (9,8 razy więcej osób z tym tytułem w 2017 niż w 2012) i „Data Scientist" (6,5 razy) (Economic Graph Team 2017).

Popyt na pracowników pociąga za sobą popyt na edukację w dziedzinie DS. Dość powiedzieć, że na pierwszej stronie wyników wyszukiwarki Google znajdujemy np. listę 368 kursów online wraz z ich recenzjami - stronę śledzi 127 100 osób (Class Central 2018); listę najlepszych kursów w tej dziedzinie - 52 pozycje (LearnDataSci 2018); listę najlepszych bootcampów[22] DS - 38 pozycji (switchup 2018); listę 23 najlepszych studiów magisterskich, gdzie nie zabrakło znakomitych uczelni, jak University of California-Berkeley i Carnegie Mellon University (masterindatascience 2018). O rynku pracy, edukacji i innych ilościowych charakterystykach polskiego świata DS mówimy więcej w części 5.1. tej rozprawy.

21 Przetłumaczone jako „badanie danych"; cytat zmieniono, by zachować przyjętą konwencję.

22 Bardzo intensywny kurs, stacjonarny, zaoczny lub online; słowo przeszło do biznesu z wojska i oznacza dosłownie „obóz rekrutów".

1.2. Cel pracy

Celem głównym pracy jest **opis etnograficzny polskiego społecznego świata data science**.

W pracy stawiamy sobie pięć celów szczegółowych. Wyznaczamy je na podstawie wybranych pojęć i procesów zaczerpniętych z teorii społecznych światów, w tym społecznych światów/aren Adele E. Clarke. Są to:

1. Rekonstrukcja tworzonych przez uczestników społecznego świata DS definicji działania podstawowego.

2. Zrozumienie roli technologii umożliwiających wykonanie działania podstawowego w odniesieniu do zachodzących w społecznym świecie procesów – przede wszystkim:

   a. wyznaczania granic społecznego świata;

   b. legitymizacji i zaświadczania o autentyczności;

   c. segmentacji – profesjonalizacji.

3. Rozpoznanie świata wartości uczestników, które rozumiane są jako „jako punkty orientacyjne, pozwalające dokonać oceny działania, a jednoczenie wskazujące, jak ma ono wyglądać i w jakich granicach przebiegać" (Kacperczyk 2016:44).

4. Określenie i opis głównych aren sporów.

5. Uchwycenie pełnej sytuacji badania.

Powyższe cele staną się zrozumiałe dla czytelniczki / czytelnika nie zaznajomionego z koncepcjami społecznych światów po lekturze 3. i 4. rozdziału niniejszej pracy.

Odnotujmy jednocześnie, że samo „środowisko" DS wykonuje różnego rodzaju akty samowiedzy. Przykładowo: od wspomnień, polemik i wywiadów (Alekseichenko 2018; Barlow 2013; Brooks 2014; Cao 2017a, 2017b; Chang 2015; Drejewicz 2017; Gutierrez 2014; Loukides 2010; O'Neil 2017a; O'Neil i Schutt 2015; Provost i Fawcett 2013), po zestawienia, sondaże i analizy danych (Barlow 2013; CrowdFlower 2017; Delapenha 2017; Fox i Leanage 2016; Kaggle 2017; King i Magoulas 2016; Onalytica 2017, 2018; Pascnau, Patraucean, i Precup 2018; Prokulski 2017; Rexer Analytics 2018; Sopyła 2017; Więcko, Słomczyński, i

Biecek 2016; Zhang, Muller, i Wang 2020). Dotychczas nie ma jednak pracy socjologicznej, w której podjęto by zadanie szerokiego oglądu tzw. „środowiska" / „społeczności" DS oczami jego uczestników. Niniejsza rozprawa zamierza wypełnić powstałą lukę.

1.3. Prezentacja struktury pracy

Pierwszy rozdział wprowadza w badany świat społeczny, w problematykę rozprawy i konstrukcję pracy, zaprezentowano też cel pracy. Rozdziały drugi i trzeci to przegląd literatury. Przyglądamy się najpierw literaturze akademickiej poświęconej DS, koncentrując się na problematyce konsekwencji społecznych i na – nielicznych – autorach polskich. Następnie rysujemy ramy teoretyczne rozprawy śledząc głównie prace poświęcone teorii światów społecznych.

Rozdział czwarty dotyczy metodologii. Prezentujemy w nim krótko założenia przyjętego paradygmatu interpretatywnego, ramy metodologiczne teorii ugruntowanej i ugruntowanego teoretyzowania, analizę sytuacyjną, zastosowane techniki badawcze i historię realizacji badań własnych, wraz z usytuowaniem badacza i problemami w postępowaniu badawczym.

Rozdziały 5. i 6. to właściwy raport etnograficzny, narracyjna prezentacja i synteza wyników badań własnych. Ustalenia są ilustrowane fragmentami zebranych materiałów jakościowych (wypowiedziami, obserwacjami, materiałami wizualnymi), rzadziej wynikami analiz ilościowych. Syntezy prezentujemy także w formie wizualnej (m.in. map według propozycji Adele E. Clarke). Wywód organizujemy kolejno wokół: opisu elementów społecznego świata DS (rozdział 5.), opisu jego dynamiki (6.): procesów zachodzących w badanym świecie społecznym (część 6.1.) oraz opisu aren tego świata (6.2.). Część 6.2. domyka ujęcie pełnej sytuacji badania, na którą łącznie składają się rozdziały od 2. do 6.

W zakończeniu rozprawy przedstawiamy podsumowanie i wnioski płynące z badań własnych. Prezentujemy także konkluzje dla socjologii. Ostatnia część pracy to aneks, zawierający spisy zrealizowanych badań oraz indeksy rysunków i tabel.

Rozdział 2. Data science w literaturze naukowej

W rozdziale dokonujemy przeglądu literatury naukowej – monografii, artykułów, podręczników – związanych z DS. Wyróżniamy dwa obszary: nauk ścisłych i technicznych (2.1) oraz nauk społecznych i humanistycznych (2.2). W pierwszym obszarze zagadnienia związane z DS rozwijane są głównie w wymiarze substancjalnym, w drugim dokonywana jest metaanaliza. W części 2.1 koncentrujemy się na autorach polskich, niemniej jednak z uwagi na cel rozprawy – opis świata społecznego - pokazujemy znaczące dla tego, czyli również polskiego świata czasopisma, książki i postaci z całego globu. Tym samym w obszarach ścisłym i technicznym przeglądamy literaturę podmiotu, a w obszarach społecznym i humanistycznym – przedmiotu. Zatem w części 2.1 jest to przegląd etnograficzny, nie systematyczny, nie sięgamy do źródeł już użytych w rozdziale pierwszym. W części 2.2 dokonujemy przeglądu systematycznego literatury w zakresie społecznych i humanistycznych prac poświęconych zagadnieniom związanym z DS.

2.1. Obszar nauk ścisłych i technicznych

Jednym z najbardziej rozpoznawalnych w badanym świecie społecznym naukowcem, zajmującym się zagadnieniami związanymi z DS, jest Andrew Yan-Tak Ng, znany jako Andrew Ng. Piszący te słowa słyszał wielokrotnie o Ng podczas prowadzonych wywiadów i obserwacji, widział jego wizerunek lub cytaty z jego wypowiedzi podczas prezentacji, w różnych materiałach powstających w ramach społecznego świata, a także na tapetach pulpitów komputerowych osób uczestniczących w warsztatach czy w postaci memów internetowych. Ng jest profesorem (*Adjunct Professor*) Stanford University, dyrektorem Stanford Artificial Intelligence Lab (Ng 2018a, b.d.). Jego badania dotyczyły m.in. rozpoznawania tematów w tekście metodą LDA – *Latent Dirichlet Allocation* – (Blei, Ng, i Jordan 2003; Ng, Zheng, i Jordan 2001), rozpoznawania obrazu i dźwięku za pomocą głębokich sieci neuronowych (Lee i in. 2009; Saxena, Sun, i Ng 2009), czy wsparciem diagnostyki medycznej z użyciem takich sieci (Kallenberg i in. 2016; Rajpurkar i in. 2018). Jest też współzałożycielem Google Brain Deep Learning Project, gdzie pracował w latach 2011-2012. Google Brain stworzyło wówczas sieć neuronową, rozpoznającą obiekty na filmach z należącego do Google portalu YouTube (Le i in. 2011). Rozpoznawanie odbywało się na zasadzie uczenia nienadzorowanego – pracowano na danych nieoznaczonych, zatem

wynikiem działania modelu było nadanie etykiet obiektom podobnym. W prasie określano odkrycie na zasadzie „komputer sam nauczył się rozpoznawać koty" (Clark 2012; Clarke 2012). Jesteśmy przekonani, że rozpoznawalność Ng jest spowodowana przez prowadzony przez niego kurs machine learningu dla początkujących na platformie Coursera. W roku 2012, po sukcesach otwartych kursów internetowych oferowanych przez Stanford w 2011, Ng założył platformę Coursera i opublikował bezpłatny, ogólnodostępny kurs. Wzorował się m.in. na zasadzie działania najbardziej popularnego forum internetowego dla osób programujących Stack Overflow (o którym piszemy w tej pracy wielokrotnie), a którego sam często używał (Ng i Widom b.d.). Był to jeden z pierwszych w historii kursów MOOC (*Massive Open Online Course*), oferujący wykłady z instruktorem, automatycznie sprawdzane prace domowe, bieżące wsparcie ucznia przez prowadzących i innych uczących się oraz certyfikat ukończenia (ibidem). Kurs jest przez cały czas dostępny. Wzięło w nim udział ponad 86 tys. osób, co wiadomo z liczby udzielonych przez uczestników ocen – dodajmy, że średnia to 4,9 na 5 (Ng 2012). Z kursu można skorzystać bezpłatnie, jednak możliwość uzyskania certyfikatu kosztuje 79 USD. W latach 2014-2017 Ng pracował w chińskim odpowiedniku Google – Baidu na pozycji Chief Scientist (Knight 2017). Obecnie – pod cytowanym wyżej hasłem „sztuczna inteligencja to nowa elektryczność" wspiera początkujące firmy AI (AI Fund 2018), prowadzi własną firmę (Landing AI 2018), założył też kolejną platformę edukacyjną w ramach której opublikował bezpłatny szkic podręcznika uczenia maszynowego (Ng 2018b).

Na Stanford University – od 2010 do 2011 – pracował również Peter Norvig. W latach 80. i wczesnych latach 90. był pracownikiem naukowym University of Southern California i University of California, Berkeley. Później pracował jako badacz i szef działów badawczych w Sun Microsystems Labs i NASA (Norvig 2018). Od 2001 pracuje w Google, gdzie obecnie jest szefem działu badań (*Director of Research*) (Google 2018; Spector, Norvig, i Petrov 2012). Na swojej stronie www w części „życiorys" napisał:

> Dla rekruterów: proszę, nie proponujcie mi pracy. Mam już najlepszą pracę na świecie w najlepszej na świecie firmie. Dla inżynierów i badaczy: zobaczcie dlaczego [link do careers.google.com] (Norvig 2018).

Norvig w pracy akademickiej zajmował się głównie problemami ilościowej analizy tekstów (Halevy, Norvig, i Pereira 2009; Michel i in. 2011; Norvig 1987). Już w 1994 roku opublikował książkę *Artificial Intelligence: A Modern Approach*, która w 2009 została wydana po raz trzeci (pracę cytowaliśmy w części 1.1) (Russel i Norvig 2009). Ma być ona

„wiodącym podręcznikiem w dziedzinie sztucznej inteligencji" (*the leading textbook in AI*), używanym na ponad 1,3 tys. uczelni w 110 krajach i czwartą najczęściej cytowaną pracą XXI wieku (ibidem). Norvig jest jednym ze współzałożycieli The Berkeley Institute for Data Science (2013). Podobnie jak Ng zajmuje się też nauczaniem AI w ramach MOOC. Wraz z Sebastianem Thrunem (o którym poniżej) prowadzi bezpłatny kurs „Wstęp do sztucznej inteligencji" (*Intro to Artificial Intelligence*) na platformie Udacity (2012). Poza pracą dydaktyczną Norvig także publikuje na temat edukacji w dziedzinie informatyki. Już w 2001 napisał tekst dotyczący nauki programowania, w którym krytykował pojawiające się wówczas w USA kursy typu „naucz się programować w 21 dni" (Norvig 2001). Najnowszą jego publikacja pokazuje wyniki badania ankietowego dotyczącego uczenia podstaw AI (Wollowski i in. 2016). Norvig wszedł też w polemikę z Noamem Chomsky'm na temat dwóch podejść do matematycznego modelowania zjawisk (Norvig 2012), o czym jako pierwszy wypowiadał się Breiman, cytowany już w części 1.1 tej rozprawy (Breiman 2001). Norvig zaangażowany jest w założony przez Petera Diamandisa i Raya Kurzweila Uniwersytet Osobliwości (*Singularity University*), organizację typu think-tank, która zgodnie z opisywaną wyżej wizją Kurzweila przygotowuje ludzi na przyszłość poza granicami ludzkiej biologii (Singularity University 2018).

Sebastian Thrun to naukowiec także związany z Uniwersytetem Stanforda – z przerwami pracował na tej uczelni od 2003 r, obecnie na stanowisku profesora (*Adjunct professor*) (Stanford University 2018). Wraz z Norvigiem przygotował dotyczący AI kurs MOOC dla początkujących, a sukces tego przedsięwzięcia zaowocował w 2012 r. założeniem przez Thruna platformy edukacyjnej Udacity (Udacity 2018). Thrun, jak Ng i Norvig, również pracował w Google. Był założycielem pierwszego zespołu zajmującego się w tej firmie samochodami autonomicznymi, a także zespołu Google X, zajmującego się m.in. projektem Google Glass – sterowanymi głosem okularami o rozszerzonej rzeczywistości (Stanford University 2018). W pracy badawczej zajmował się on m.in. przetwarzaniem obrazów w czasie rzeczywistym na potrzeby sterowania pojazdami (Anguelov i in. 2004, 2005; Diebel, Thrun, i Thrun 2005; Stavens, Hoffmann, i Thrun 2007). Otrzymał wiele wyróżnień, w tym nagrodę Insytutów Maxa Plancka w München (Stanford University 2018).

Yann LeCun to badacz znany z uwagi na wkład w rozwój sieci neuronowych oraz pracę dla innego niż opisani powyżej giganta technologicznego – Facebooka. LeCun po uzyskaniu stopnia doktora we Francji był badaczem (*postdoctoral research associate*) na Uniwersytecie w Toronto od 1987 do 1988 (LeCun b.d.). Pracował tam w laboratorium Geoffrey'a Hintona,

który nazywany jest ojcem chrzestnym głębokich sieci neuronowych (o czym dalej). Od roku 2003 pracuje na New York University (ibidem), tam w roku 2013 założył NYU Center for Data Science (2013). W tym samym czasie został pierwszym dyrektorem badań nad sztuczną inteligencją (*Director of AI Research*) w firmie Facebook (LeCun b.d.), gdzie obecnie jest zatrudniony jako Chief AI Scientist (Facebook 2018). LeCun jest autorem rozwiązania obecnie powszechnie wykorzystywanego do rozpoznawania obiektów na zdjęciach – konwolucyjnych sieci neuronowych (*convolutional neural networks,* CNN). Pierwszy artykuł dotyczył problemu rozpoznawania pisanych odręcznie kodów pocztowych (LeCun i in. 1989). Pracując w AT&T Research LeCun stworzył CNN o nazwie LeNet-5, służącą do rozpoznawania pisma odręcznego i pisma z wydruków (LeCun i in. 1998). Była to piąta wersja, a w cytowanym tutaj artykule porównano wyniki wcześniejszych iteracji. Sieć uzyskuje bardzo dobre wyniki niezależnie od zróżnicowania pisma: grubości linii, wielkości znaku, pochylenia, pozycji znaku na zdjęciu, szerokości znaku – na stronie internetowej autora dostępne są wizualizacje jej działania (LeCun b.d.).

> CNN jest podstawą wielu produktów i usług wdrożonych przez firmy takie jak Facebook, Google, Microsoft, Baidu, IBM, NEC, AT&T i inne do zadań zrozumienia obrazu i wideo, rozpoznawania dokumentów, interakcji człowiek-komputer i rozpoznawania mowy (LeCun b.d.)

Przeprowadzony przez Andrew Ng wywiad z LeCunem dotyczący podstaw CNN zawarto w materiałach kursowych specjalizacji głębokich sieci neuronowych (*deep learning*) na portalu Coursera (Ng 2018c).

Geoffrey Hinton karierę akademicką zaczynał w latach 70. z dyplomem studiów pierwszego stopnia z psychologii eksperymentalnej Uniwersytetu Cambridge i doktoratem ze sztucznej inteligencji obronionym na Uniwersytecie w Edynburgu. Jak wspomnieliśmy, pod koniec lat 80. pracował on na Uniwersytecie w Toronto, a w jego zespole był m.in. Yann LeCun. Z tą uczelnią Hinton związany jest zawodowo nieprzerwanie od 2001 r. (CS Department Toronto University b.d.). Miano ojca chrzestnego głębokich sieci neuronowych (Lee 2016; Mannes 2017; Somers 2017; Sorensen 2017) przyniosła mu praca poświęcona zastosowaniu w wielowarstwowych sieciach neuronowych metody propagacji wstecznej (*backpropagation*) (Rumelhart, Hinton, i Williams 1986). Metoda ta umożliwia korektę wag sieci neuronowej w sposób optymalizacyjny, czyli to, co nazywane jest uczeniem sieci neuronowej. Weźmy przykład uczenia nadzorowanego, gdzie mamy dane zaetykietowanie (przyjmijmy, że są to fotografie psów i kotów wraz z etykietami, „pies” albo „kot”). Ogólnie

rzecz działa następująco: wagi wszystkich sztucznych neuronów w sieci są losowane i sprawdzany jest błąd (koty zaklasyfikowane jako psy lub psy jako koty), następnie wagi są korygowane i ponownie sprawdzany jest błąd, procedura jest powtarzana wielokrotnie. Kierunek korekcji wag – ich zmniejszanie lub zwiększanie – w celu zminimalizowania błędu ustalany jest za pomocą reguły największego spadku (*gradient descent*) (Larose 2005:135–38). Hinton i współpracownicy spopularyzowali użycie propagacji wstecznej w sieciach neuronowych. Prace tego rodzaju podejmowano już od początku lat 60. (Schmidhuber 2015). Podstawa propagacji wstecznej, czyli reguła największego spadku, to mająca korzenie w siedemnastowiecznych pracach Leibniza i osiemnastowiecznej koncepcji równań Eluera-Lagrange'a, metoda redukowania błędów (ibidem).

Hadley Wickham jest naukowcem znanym wśród data scientistów z powodu stworzonych przez niego narzędzi. Jak sam pisze: „buduję narzędzia (obliczeniowe i poznawcze), które czynią data science łatwiejszym, szybszym i bardziej zabawnym". Pracuje on jako profesor statystyki (*Adjunct Professor of Statistics*) w University of Auckland, Stanford University i Rice University (Wickham 2018b). Narzędzia Wickhama są przeznaczone do pracy w języku R, a on sam pełni funkcję Chief Scientist w RStudio (RStudio 2018a). Dwa najpopularniejsze w DS języki programowania[23] to R i Python (DeZyre 2015; Hale 2018; Kromme 2017; Muenchen 2019; Piatetsky 2017b; Sharp Sight 2018; Silaparasetty 2018). Dodajmy, że język R powstał właśnie na uczelni w Auckland (Ihaka i Gentleman 1996), a Wickham uczył się go od twórców podczas swoich studiów magisterskich na tej uczelni (Kopf 2015). RStudio to firma, oferująca oprogramowanie umożliwiające użytkownikom efektywniejszą pracę z językiem R (RStudio 2018a). Zdecydowanie najpopularniejszym produktem firmy jest IDE (*Integrated Development Environment*) o nazwie RStudio Desktop (RStudio Team 2016). Narzędzie to jest wygodnym edytorem programistycznym, dostępnym bezpłatnie do większości zastosowań (Biecek 2015a). Istnieją wersje płatne o rozszerzonej funkcjonalności (RStudio 2018b). RStudio, w tym Hadley Wickham, tworzą w języku R narzędzia w formie pakietów (*packages*). Pakiet jest zbiorem funkcji, wykonujących określone zadania: „funkcjami możemy analizować dane, rysować dane, odsłuchiwać dane, wykonywać obliczenia, wysyłać maile, robić najróżniejsze rzeczy" (Biecek 2015a). Przykładowo podstawowy, domyślnie zainstalowany w R pakiet „base" zawiera funkcję „seq()", za pomocą której można stworzyć sekwencję liczb,

23 O technologiach świata społecznego data science traktuje osobna część w rozdziale 5. niniejszej rozprawy

Sekwencję stu jeden liczb od 0 do 10 z krokiem 0,1 utworzy i przypisze do zmiennej „sekwencja” następująca instrukcja (ibidem):

```
sekwencja <- seq(0, 10, 0.1)
```[24]

Wickham jest autorem rodziny pakietów „tidyverse” (Wickham 2017). W jej skład wchodzą pakiety: "ggplot2" – do wizualizacji danych (Wickham 2016), "tibble" – format danych tabelarycznych (Müller i Wickham 2018), "tidyr" – porządkowanie danych (Wickham i Henry 2018), "readr" – wczytywanie danych do R (Wickham, Hester, i François 2017), "purrr" – programowanie funkcjonalne (Henry i Wickham 2018), "dplyr" – manipulacje i transformacje na danych (Wickham i in. 2018), "stringr" – praca z ciągami znaków (*strings*) (Wickham 2018e), "forcats" – praca z danymi kategorialnymi (Wickham 2018a). Sam autor wskazuje, że pakiety te pozwalają na wykonanie 80% pracy w każdym projekcie DS, choć zazwyczaj nie będą wystarczające do realizacji całości zadania (Wickham i Grolemund 2017). Wickham jest autorem pakietów R o największej łącznie liczbie pobrań – ponad 113 milionów – i największej łącznie liczbie gwiazdek (*star*), będących wyrazem uznania na portalu z repozytorium kodu GitHub (Mortimer 2018). Przy tym jego pakiet do wizualizacji danych „ggplot2” znajduje się na drugim miejscu w rankingu pobrań pakietów, a na pierwszym w ranking gwiazdek (ibidem).

Wickham jest zarówno autorem podręczników do nauki R i DS (Wickham 2014a; Wickham i Grolemund 2017), jak i prac naukowych, w których prezentował podstawy teoretyczne i praktyczne tworzonych przez siebie narzędzi (Wickham 2010, 2014c). Wickham publikuje także polemiki, traktujące o pracy data scientistów czy o branży DS (Bryan i Wickham 2017; Wickham 2014b, 2018f). Podręcznik napisany wraz z Grolemundem został przetłumaczony m.in. na język polski (Wickham i Grolemund 2018), ale jedynie wersja anglojęzyczna dostępna jest w całości bezpłatnie na http://r4ds.had.co.nz/ (dostęp 03.06.2020). Wickham, naszym zdaniem, ma w środowisku użytkowników języka R (co nie jest tożsame ze środowiskiem DS) status autorytetu i celebryty. W mediach społecznościowych pojawił się np. żart w postaci przeróbki katolickiej modlitwy „Ojcze nasz” na modlitwę do Hadleya Wickhama, odmawianą przed uruchomieniem kodu w R (Votta 2018), a także fikcyjna okładka magazynu lifestyle’owego[25], na której obok różnych

24 Prezentowane w naszej rozprawie fragmenty kodu wyeksportowano do edytora tekstu Libre Office Writer za pomocą R Markdown (Allaire i in. 2019; Xie, Allaire, i Grolemund 2018).

25 https://twitter.com/W_R_Chase/status/1155212225621221376, dostęp 12.08.2019 r.

żartów dotyczących wizualizacji danych widnieje m.in. zapowiedź wywiadu z Wickhamem, który „ujawnia, dlaczego *nigdy* nie zmieni domyślnego stylu grafiki w ggplot".

W języku Python nie ma jednego twórcy narzędzi, którego dorobek byłby tożsamy z dorobkiem Hadley'a Wickhama. Przedstawmy krótko twórców pięciu najpopularniejszych pakietów Pythona do zastosowań w DS. Są to pakiety: „NumPy", „SciPy", „pandas", „matplotlib" i „scikit-learn" (Bobriakov 2017; Li i Paczuski 2017; Satyaseel 2018).

Pakiety „NumPy" i „SciPy" to narzędzia umożliwiające pracę z danymi w formie tabelarycznej i wykonywanie obliczeń naukowych lub inżynieryjnych. Choć w ich tworzeniu i rozwoju brało udział wiele osób, to chcemy zaznaczyć wkład Travisa Oliphanta. Oliphant jest jednym z inicjatorów powstania obu pakietów. Zaczynał pracę z językiem Python w 1998 r. jako doktorant w dziedzinie obrazowania biomedycznego. Z uwagi na ograniczenia dostępnych wówczas w Pythonie narzędzi, na potrzeby swojej pracy zaczął sam przepisywać do Pythona metody dostępne w językach C lub Fortran (Oliphant 2006, 2007). Ma wykształcenie matematyczne i z dziedziny inżynierii elektrycznej, w 2001 uzyskał stopień naukowy doktora w Mayo Clinic Research (Mayo Clinic b.d.). Publikował w czasopismach medycznych (Oliphant i in. 2001), obecnie nie jest aktywny naukowo. Zajmuje się rozwojem związanego z Pythonem otwartego oprogramowania, jest jednym z założycieli Anacondy, dyrektorem Python Software Foundation i dyrektorem NumFOCUS (Anaconda 2018b). Anaconda jest bezpłatną i wolną (*open source*) – a także jak napisano na oficjalnej stronie internetowej: najpopularniejszą i najbardziej zaufaną - platformą do programowania i zarządzania pakietami (ponad 1400 preinstalowanych) oraz środowiskiem pracy DS dla języka Python, choć dodatkowo możliwa jest praca w R (Anaconda 2018). NumFOCUS jest organizacją non-profit (*501(c)3 non-profit organization*), której motto brzmi „Otwarty Kod = Lepsza Nauka" (*Open Code = Better Science*). Założył ją Oliphant w roku 2011. Jej misją jest wspieranie wolnego oprogramowania na potrzeby nauki. Anaconda jest sponsorem NumFOCUS od 2012 (Anaconda b.d.; NumFOCUS 2018). NumFOCUS prowadzi międzynarodowy, dotyczący pracy z danymi program edukacyjny o nazwie PyData. Jedna z ostatnich konferencji PyData odbyła się 18.-21.11.2018 w Warszawie (PyData 2018). Pod nazwą „SciPy" kryje się też ekosystem pakietów do zastosowań naukowych, w skład którego wchodzą m.in. wymienione już „SciPy" i „NumPy", a także „pandas", „scikit-learn" i „Matplotlib" (SciPy.org b.d.).

„Matplotlib" to pakiet służący wizualizacji danych; dodajmy, że prace nad jego rozwojem są sponsorowane przez NumFOCUS (Matplotlib 2018). Głównym autorem pakietu

był nieżyjący już John Hunter, który opublikował artykuł przedstawiający ten pakiet w czasopiśmie „Computing in Science & Engineering" (Hunter 2007). Hunter był absolwentem Princeton i doktorem neurobiologii Universytetu Chicago (Princeton Alumni Weekly 2012). Podobnie jak Oliphant, Hunter zaczął pracę nad pakietem z uwagi na ograniczenia dostępnych narzędzi – jego środowisko naukowe korzystało głównie z komercyjnego oprogramowania MATLAB. Początkowo chodziło jedynie o możliwość wizualizacji wyników elektrokortykografii i elektroencefalografii pacjentów chorych na epilepsję, ponieważ dane z takich badań autor analizował na co dzień (Hunter i Droettboom 2012).

Pakiet „pandas" dostarcza – jak napisano na oficjalnej stronie internetowej – wysokowydajnych i prostych w użyciu struktur danych oraz narzędzi do analizy danych w języku Python (The pandas project 2018). Rozwój „pandas" także finansowany jest przez NumFOCUS (ibidem). Autorem narzędzia jest Wes McKinney (McKinney 2010, 2011). Jest on z wykształcenia matematykiem, podjął studia doktoranckie w dziedzinie statystyki na Duke University, gdzie przebywa na urlopie od 2011 roku (McKinney 2018a). Opublikował podręcznik do analizy danych z użyciem pakietów „pandas" i „NumPy" (McKinney 2012); drugie wydanie książki dostępne jest także w języku polskim (McKinney 2018b). McKinney współpracował z omawianym wyżej Hadley'em Wickhamem z RStudio, tworząc narzędzia dla DS niezależne od języka programowania – można używać ich w Pythonie, R i kilku innych (Dar 2018; McKinney 2018a).

Pakiet „scikit-learn" to jedyne z omawianych dotychczas narzędzi służące uczeniu maszynowemu:

> Scikit-learn to moduł Pythona integrujący szeroką gamę najnowocześniejszych [org. *state-of-the-art*] algorytmów uczenia maszynowego dla problemów nadzorowanych i nienadzorowanych w średniej skali. Pakiet ten koncentruje się na wprowadzaniu uczenia maszynowego dla osób niebędących specjalistami, używających wysokopoziomowego języka programowania ogólnego zastosowania[26]. (Pedregosa i in. 2011:2826)

Prace nad pakietem rozpoczęły się w 2007 r., ale to Fabian Pedregosa jest osobą, która przy udziale finansowym INRIA, francuskiego instytutu badawczego nauk cyfrowych, doprowadziła do publikacji pakietu (scikit-learn 2018). Brał on udział także w rozwijaniu m.in. omawianego powyżej pakietu „SciPy" (Pedregosa 2018). Pedregosa uzyskał tytuł doktora w dziedzinie informatyki (*computer science*), pracując nad neuroobrazowaniem z wykorzystaniem statystyki i uczenia maszynowego (Pedregosa 2015). W pracy naukowej

26 Wysokopoziomowy język programowania ogólnego zastosowania to m.in. język Python. W rozdziale 5. objaśniamy to zagadnienie szerzej, opisując technologie społecznego świata.

zajmował się zbliżonymi zagadnieniami (Pedregosa, Bach, i Gramfort 2014; Pedregosa, Leblond, i Lacoste-Julien 2017). Obecnie jest pracownikiem działu badań (*Research Scientist*) w Google (Pedregosa 2018).

Dorobek polskich naukowców związanych z DS zaczynamy od przedstawienia prac Ryszarda Tadeusiewicza. Jest on profesorem nauk technicznych związanym z Akademią Górniczo-Hutniczą w Krakowie, gdzie pracuje od lat 70. Zajmuje się systemami wizyjnymi robotów przemysłowych, systemami sensorycznymi; sieciami neuronowymi i biocybernetyką. Swoje prace sytuuje w dziedzinach informatyki, automatyki i robotyki oraz biocybernetyki i inżynierii biomedycznej (Tadeusiewicz 2018). Pośród ponad 1200 opublikowanych przez niego artykułów naukowych pierwsza praca dotycząca sieci neuronowych – techniki modelowania bardzo szeroko wykorzystywanej w DS i obecnie często utożsamianej ze sztuczną inteligencją – ukazała się pod koniec lat 70. (Tadeusiewicz i Mikrut 1978). Tadeusiewicz kontynuował badania dotyczące sieci neuronowych i innych metod uczenia maszynowego, głównie w zastosowaniach dla rozpoznawania obrazu lub dźwięku (Tadeusiewicz 1998, 2007; Tadeusiewicz i Flasiński 1991). W najnowszej pracy, której jest współautorem, wykorzystano metody uczenia maszynowego do klasyfikowania żołędzi na zdrowe, chore i wątpliwe. Klasyfikator ten jest częścią maszyny wykonującej automatyczną skaryfikację[27] żołędzi dębu szypułkowego (Adamczyk i in. 2018). Tadeusiewicz jest także autorem licznych opracowań popularnonaukowych (Tadeusiewicz 1989; Tadeusiewicz i Chrząstowski 1994), w tym książek dla dzieci (Tadeusiewicz 2015a, 2015b). Wielokrotnie wypowiadał się w mediach, prowadził cykliczne audycje radiowe, udzielił ponad 250 wywiadów prasowych (Tadeusiewicz 2018), z czego wiele w ostatnich latach poświęconych jest konsekwencjom społecznym wdrażania tzw. sztucznej inteligencji. Tadeusiewicz prezentuje w tej kwestii stanowisko optymistyczne, a wizję przejęcia władzy nad ludźmi przez roboty uważa za nierealną. Przykładowo prezentujemy wypowiedź dla tygodnika „Neewsweek", zwracając przy tym uwagę na odmienne stanowisko prowadzącej wywiad:

> Maszyny wciąż nie mają świadomości własnego istnienia, osobowości ani własnych dążeń - uspokaja prof. Tadeusiewicz. Ale ewolucja robotów już się zaczęła. Nie przebiega ona w pełni pod kontrolą człowieka i daje sztucznej inteligencji coraz więcej ludzkich cech. (…) Czy sztuczna inteligencja zyska kiedyś tę świadomość i chęć realizacji nie naszych, lecz własnych celów? (Burda 2018:67).

27 Polega to na odcinaniu 1/3 długości żołędzia z odpowiedniej strony. Pracownik ocenia wzrokowo zdrowie żołędzia. Sadzonki z żołędzi skaryfikowanych rosną szybciej niż z nieskaryfikowanych (Adamczyk i in. 2018).

Naukowcem, którego w przeprowadzonych w ramach niniejszej pracy wywiadach swobodnych z uczestnikami świata społecznego DS określano „polskim Hadleyem Wickhamem" jest Przemysław Biecek. Jest on profesorem Politechniki Warszawskiej i Uniwersytetu Warszawskiego, magistrem matematyki i informatyki Politechniki Wrocławskiej. Doktorat z biostatystyki obronił we Wrocławiu, habilitował się zaś w Instytucie Biocybernetyki i Inżynierii Biomedycznej PAN w dziedzinie statystyki medycznej (Biecek 2018c). Posiada on bogaty dorobek publikacyjny związany tematycznie ze swoimi rozprawami (Baurska i in. 2014; Biecek i Cebrat 2008; Bogdan i in. 2008; Dobrzynska i in. 2016; Donizy i in. 2018; Drozd-Sokołowska i in. 2018; Kondratiuk i in. 2017; Rudzki, Biecek, i Kaza 2017; Szczurek i in. 2011; Szynglarewicz i in. 2017). Współpracował m.in. z Rossem Ihaka, jednym z dwóch autorów języka R (Biecek 2018c; Ihaka i Gentleman 1996). Jest autorem narzędzi – pakietów dla języka R (Biecek 2018d; Biecek i Kosinski 2017; Caro i Biecek 2017; Molnar 2018; Staniak i Biecek 2018), w tym pakietu przygotowanego na potrzeby prowadzonego przez siebie kursu MOOC „Pogromcy Danych" (Biecek 2015c). Jest to jedyny bezpłatny, polskojęzyczny kurs MOOC z podstaw analizy danych w języku R (w ramach badań etnograficznych piszący te słowa ukończył obie jego części) (Biecek 2015a, 2015b). Biecek jest autorem podręczników do nauki analizy danych z programem R. W 2008 r. wydał pierwszą w Polsce książkę poświęconą w całości nauce R; obecnie doczekała się ona czwartego wydania (Biecek 2017a). W 2011 opublikował podręcznik poświęcony modelom liniowym i mieszanym z użyciem R; dostępne jest drugie wydanie (Biecek 2014). Nakładem prowadzonej przez siebie fundacji naukowej SmarterPoland P. Biecek wydał podręcznik – „Zbiór esejów o sztuce wizualizacji danych", także z użyciem R. Książka dostępna jest bezpłatnie na licencji Creative Common BY & SA, a kod w R napisany jest głównie z użyciem pakietu „ggplot2" autorstwa Wickhama. Dostępna jest wersja drukowana, pierwsze wydanie miało miejsce w 2014 r., obecnie dostępne jest drugie (Biecek 2016). Wcześniej pracował nad podręcznikiem „Na przełaj przez data mining. Zbiór przykładów użycia wybranych funkcji statystycznych i Data Mining dostępnych w programie R" (Biecek i Trajkowski 2011), który nie został ukończony i jest dostępny do wglądu online w wersji roboczej. Wspomniana fundacja SmarterPoland ma na celu popularyzację nauki oraz edukację dotyczącą metod analizy danych i czytania danych, gromadzenie danych dotyczących Polski oraz tworzenie opracowań tych danych. Autorem wpisów na blogu fundacji jest głównie P. Biecek. Przykładowo w czasie wejścia w życie przepisów RODO pisał o tzw. XAI czyli wyjaśnianiu działania modeli uczenia maszynowego w związku m.in. z tzw. prawem do

wyjaśnienia (Biecek 2018f, 2018b). Biecek jest obecnie jednym z najbardziej aktywnych polskich badaczy w obszarze XAI. Prace nad wspomnianym wyżej narzędziem (pakietem DALEX w R) służącym wyjaśnianiu działania modeli (Biecek 2018d) prowadzone są dzięki grantowi NCN w ramach grupy MI^2, łączącej studentów UW i PW w tzw. DataLab – laboratorium analizy danych prowadzone przez Biecka (MI^2 2018). Profesor Biecek jest także autorem projektu rozwijającego umiejętności analizy danych (*data literacy*) dla najmłodszych „Beta i Bit". Na projekt składają się opowiadania, komiks i gry komputerowe, te ostatnie do uruchomienia w środowisku R i grania za pomocą pisania kodu (Biecek 2018a). Pracuje on także w komercyjnym DS jako Principal Data Scientist w firmie Samsung SRPOL, ma doświadczenie w pracy dla działów R&D firm Netezza, IBM, iQor i Disney (Warsaw.AI 2020). P. Biecek jest aktywnym członkiem społeczności DS w Polsce, uczestniczy w wydarzeniach tego świata społecznego i w trakcie badań terenowych mieliśmy okazję go spotkać.

Istnieją liczne prace polskich autorów o charakterze technicznym czy podręcznikowym (nie powtarzamy prac Tadeusiewicza i Biecka). Mówi się zarówno o eksploracji danych czy data mining – raczej w starszych publikacjach (Lasek 2002, 2007; Morzy 2013; Osowski 2013; Szupiluk 2013), sieciach neuronowych, uczeniu maszynowym lub sztucznej inteligencji (Białko 2005; Krawiec 2003; Osowski 2006; Witkowska 2002), zaś w nowszych o analizie danych lub DS, z użyciem języków R lub Python (Gągolewski 2016; Gągolewski, Bartoszuk, i Cena 2016; Szeliga 2017).

Zwracamy uwagę na polemiczne teksty autorów z dziedzin technicznych i ścisłych. Już w 2000 r. wydano zbiór esejów „Granice sztucznej inteligencji", gdzie autorzy zdecydowanie obstają przy stanowisku, że AI inteligenta na miarę człowieka jest i będzie niemożliwa (Szumakowicz 2000). Profesor statystyki Mirosław Szreder wypowiadał się o społecznych konsekwencjach stosowania zaawansowanej analityki dużych zbiorów danych, zarówno w periodykach naukowych (Szreder 2015a, 2018a), jak i prasie (Szreder 2015b, 2016, 2018b). Konsekwencje społeczne omawianych zjawisk są także głównym tematem przeglądowej pracy Jerzego Surmy (2017). Autor zajmuje się sztuczną inteligencją od końca lat 80., i jego zdaniem

> świadomość zagrożeń z tym [big data] związanych jest tak niska, że w poczuciu elementarnej obywatelskiej odpowiedzialności postanowiłem wprost napisać o wszystkich znanych mi reperkusjach tego zjawiska w odniesieniu do publicznie dostępnych badań naukowych (Surma 2017:7).

Interesująca z punktu widzenia celu naszej rozprawy jest praca „Statystyka a data science" pod redakcją Grażyny Trzpiot. Choć książkę przygotowano jako podręcznik, to zauważmy, że tytułowe dyskusje o relacji między DS a statystyką trwają od końca lat 90. Jak wskazuje Cao (Cao 2017a:7) postulowano przemianowanie statystyki na DS w celu zorientowania tej dziedziny bardziej empirycznie (Wu 1997), rozszerzenie statystyki o zainteresowanie problemami obliczeniowymi i współpracą z informatykami (*computer scientists*) (Cleveland 2001), zaadoptowaniu do statystyki odmiennego podejścia do modelowania tzw. kultury modelowania algorytmicznego (Breiman 2001), uznania statystyki i uczenia maszynowego za odgrywające centralną rolę w DS (van Dyk i in. 2015). Mówi się zarówno, że DS to tylko przemianowanie (*rebranding*) statystyki, jak i przeciwnie – że statystyka jest najmniej ważną częścią DS, a DS bez statystyki jest nie tylko możliwe, ale i pożądane (Donoho 2015:4–7). Podkreśla się, że kształcenie na potrzeby DS jest odmienne od oferowanego obecnie nauczania statystki – zauważmy, że napisali to pracujący w RStudio profesorowie statystyki (Bryan i Wickham 2017). Dyskusje, rozważania i żarty o relacji statystyki i DS podejmowano na blogach i w mediach społecznościowych (Big Data Borat 2013; Hyndman 2014; Jarvis 2014; Taylor 2016; Wills 2012; Yau 2009). Zagadnienie to analizujemy w rozdziale 6., opisując świat społeczny DS w kontekście procesów segmentacji – pączkowania, odszczepienia i przecinania się światów społecznych (Kacperczyk 2016; Strauss 1984), m.in. statystyki i informatyki. G. Trzpiot definiuje DS następująco: „Data Science jest procesem zamieniania/przekształcania danych w konkretne działania" (2017:9). Statystyka ma zastosowanie w jednym z czterech obszarów działań DS. Są to: „uzyskaj" – pozyskanie danych; „przygotuj" – przygotowanie danych; „analizuj" – tutaj wymagana jest znajomość statystyki i innych metod analizy danych; „działaj" – wdrożenie wyników (Trzpiot 2017:11–15).

2.2. Obszar nauk społecznych i humanistycznych

Prace naukowe związane z DS dzielimy na cztery kategorie, wyróżnione z uwagi na problematykę:

- epistemologiczną i metodologiczną;
- włączającą charakterystyczne dla DS metody do badań własnych;
- dotyczącą konsekwencji społecznych stosowania DS / big data / sztucznej inteligencji (nazewnictwo jest niespójne);

- badającą środowisko ludzi zajmujących się DS jako wycinek rzeczywistości społecznej – takich tekstów jest bardzo niewiele, a właśnie taka jest problematyka naszej rozprawy.

Podział ten jest nierozłączny, co znaczy, że konkretne teksty można przyporządkować do więcej niż jednej kategorii. Szczególnie w części trzeciej poza tekstami naukowymi przywołujemy także publikacje popularnonaukowe, prasowe i inne.

2.2.1. Problematyka epistemologiczno – metodologiczna

Zaczynamy od problematyki epistemologicznej i metodologicznej, ponieważ działalność związania z DS zawiera elementy działalności poznawczej. Nazywamy tutaj DS działalnością paranaukową. Zdecydowanie zgadzamy się ze stanowiskiem, które wskazują Danah Boyd i Kate Crawford:

> Era Big Data właśnie się zaczęła, ale już teraz ważne jest, byśmy zaczęli kwestionować założenia, wartości i obciążenia (*biases*) tej nowej fali działalności badawczej. Jako naukowcy zajmujemy się wytwarzaniem wiedzy, zatem takie dociekania są najważniejszym elementem naszej działalności (Boyd i Crawford 2011:13).

Omawiane problemy podejmowano w obszarze socjologii i filozofii wiedzy w odpowiedzi na popularne wśród wczesnych entuzjastów tzw. big data podejście, które streszczają słowa „liczby / dane mówią same za siebie". W jednej z pierwszych w tej kategorii prac socjologicznych (Boyd i Crawford 2012) przywołano niezwykle wyraźny przykład tego podejścia. W 2008 r. w magazynie „Wired", piórem ówczesnego redaktora naczelnego, napisano wielokrotnie cytowane później słowa:

> Jest to [Epoka Petabajtów (*The Petabyte Age*)] świat, w którym ogromne ilości danych i matematyki stosowanej zastępują każde inne narzędzie, które można wykorzystać. Każdą teorię ludzkiego zachowania, od językoznawstwa po socjologię. Zapomnijmy o taksonomii, ontologii i psychologii. Kto wie, dlaczego ludzie robią to, co robią? Chodzi o to, że oni to robią i możemy śledzić i mierzyć to z niespotykaną wiernością. **Przy wystarczającej ilości danych**, **liczby mówią same za siebie** [podkr. RŻ] (Anderson 2008).

Analizy bibliometryczne wskazują, że w badaniach społecznych poświęconych big data w latach 2011-2015 ten tekst Andersona był najczęściej cytowaną pracą (Youtie, Porter, i Huang 2016:8). Boyd i Crawford nie zgadzają się z tezą o mówiących za siebie liczbach, podkreślając, że takie podejście uznaje za nieważne każdą teorię z każdej dziedziny nauki. Zauważają one, że owo „aroganckie" podejście entuzjastów big data marginalizuje inne sposoby zdobywania wiedzy (Boyd i Crawford 2012:666). Savage i Halford piszą jasno:

„liczby nie mówią same za siebie. To my umożliwiamy te wypowiedzi - za pomocą metod, które stosujemy i interpretacji, które czynimy" (2017:11). Z kolei entuzjaści mówią o końcu epoki ekspertów dziedzinowych dysponujących teoretyczną wiedzą substancjalną, ponieważ w myśl charakterystycznej dla Epoki Petabajtów zasady „co? zamiast dlaczego?" interesujące są jedynie relacje między zmiennymi, a nie teorie te relacje wyjaśniające. Kwestię tę analizowaliśmy w tekście *Potencjał Big Data w badaniach społecznych* (Żulicki 2017). Wskazaliśmy tam, że zdaniem entuzjastów w erze wielkich danych wiedza substancjalna jest więcej niż zbędna: ona przeszkadza. Opisując to zagadnienie Cukier i Mayer-Schönberger (których pracę uważamy za hiperentuzjastyczną) powołali się m.in. na zilustrowaną w filmie *Moneyball* historię drużyny baseballowej Oakland Athletics. Trener Billy Bean doprowadził tę, dotychczas słabą, drużynę do pierwszego miejsca w lidzie, uzyskując 20 zwycięstw pod rząd. Dokonał tego odrzucając wiedzę i doświadczenie emerytowanych zawodników i trenerów, a polegając na ilościowej analizie danych. Podejmował niepopularne decyzje – zrezygnował zupełnie m.in. z pewnego efektownego, ale jak wskazywała analiza danych nieefektywnego, elementu gry (tzw. kradzieży bazy). Podsumowując:

> Najważniejszym efektem *big data* będzie to, że decyzje oparte na danych ulepszą jakość ocen dokonywanych przez ludzi lub sprawią, że całkowicie stracą one na znaczeniu. (...) Ekspert czy specjalista w danej dziedzinie straci część swojego znaczenia na rzeczy statystyka czy analityka danych, którzy są nieskrępowani starymi metodami rozwiązywania problemów i pozwalają przemawiać danym. (Cukier i Mayer-Schönberger 2014:185).

Jasno i boleśnie ten koniec ery ekspertów ilustruje żart inżynierów zajmujących się tłumaczeniem maszynowym w firmie Microsoft: mówią oni, że „jakość przekładu rośnie za każdym razem, gdy z ich zespołu odejdzie jeden lingwista" (Cukier i Mayer-Schönberger 2014:186). Tym samym DS jest przez entuzjastów przedstawiana jako nowa nauka wiodąca czy uniwersalna (*Leitwissenschaft*) (Rath 2018). Boyd i Crawford, nie zgadzając się z podobnymi tezami wskazują jednak, że big data / DS zmieniają sposób, w jaki myślimy o badaniach w ogóle, i jest to zmiana o charakterze epistemologicznym i etycznym jednocześnie. Przeramowaniu podlegają kluczowe pytania o to, jaka jest konstytucja wiedzy, proces badania, obcowanie badacza z informacjami czy natura rzeczywistości. Akty pomiaru – czyli tutaj zapisywania dużych ilości danych w czasie rzeczywistym – kształtują mierzony świat (Boyd i Crawford 2012:665).

Na łamach czasopisma naukowego „Big Data & Society" temat ten podjął Rob Kitchin. Zwracamy szczególną uwagę na tę publikację, ponieważ syntetyzuje ona wiele wcześniejszych prac, a dodatkowo pojawiła się w okresie szczytowego zainteresowania

terminem „big data” (por. rys. 1). Autor artykułu stawia problem następująco: big data / DS mają wartość poznawczą dla wielu dziedzin nauki, ale nie bez odwołania do teorii dziedzinowych. Kitchin uważa, że mówienie o zmianie paradygmatów, końcu ery ekspertów i tym podobne, entuzjastyczne stwierdzenia są bezzasadne, a także niebezpieczne w sensie poznawczym. Wskazał on na cztery epistemologiczne obietnice big data: po pierwsze, pozwala ująć całość problemu (w myśl strategii N = all) i zapewnić pełne wsparcie decyzją; po drugie, do uzyskania wartościowych wyników nie potrzebne są przedmiotowe teorie, ani stawianie hipotez; po trzecie, ponieważ dane mówią same za siebie, nieobciążone zbędą teorią, to wyniki analiz są znaczące i zgodne z prawdą o świecie; po czwarte, wyniki analiz, niezależnie od przedmiotu, może interpretować każdy posiadający rozeznanie w statystyce (Kitchin 2014:4). Wskazał on dwie potencjalne ścieżki rozwoju współczesnej nauki. Tę zdecydowanie bardziej pożądaną nazywa nauką opartą na danych (*data-driven science*) – ma być to przeformułowanie sposobu jej uprawiania, w którym zmieszają się abdukcja, dedukcja i indukcja. Inną, naszym zdaniem niepokojącą, ścieżką rozwoju nauki może być skrajny empiryzm, czyli wspomniane „dane mówią same za siebie” (*data can speak for themselves free of theory*) jako zasada główna. (Kitchin 2014:10). Wystosowana przez Kitchina krytyka takiego podejścia (2014:5-6) odnosi się do wymienionych czterech epistemologicznych obietnic:

1. Strategia „N = all” jest taką tylko z pozoru. Zawsze badana jest próba, a nie populacja, chociażby z uwagi na ramy czasowe, a przy nieznanych jej obciążeniach wyciąganie wniosków o całej populacji może prowadzić do poważnych błędów. Poza tym dane nie są czystą reprezentacją jakiegoś wycinka rzeczywistości - zbierane są zawsze z pewnego punktu widzenia. Pomiary są społecznie konstruowane, silną i nieprzekraczalną ramę tworzą decyzje o tym, co zapisywać i przechowywać;

2. Big data nie wzięło się znikąd, zatem to podejście nie jest wolne od założeń filozoficznych i ontologicznych. Reprezentacjami tych założeń są technologie magazynowania, przetwarzania i modelowania danych. Zatem iluzją jest nie stawianie hipotez, i uzyskiwanie wartościowych informacji bez zadawania pytań – one zostały postawione wcześniej i gdzie indziej niż sądzą entuzjaści;

3. Dane nigdy nie przemówią same za siebie. Za wynikami analiz stoją zarówno zastosowane technologie, aparat matematyczny, jak i wiedza potoczna analityków (także wtedy, gdy analizy są zautomatyzowane). Wyniki same w sobie nie mają

żadnego znaczenia – interpretację np. dopasowania modelu predykcyjnego wykonują zawsze ludzie. Tym samym analiza danych zawsze odbywa się wewnątrz pewnej nieuświadomionej ramy, obciążającej uzyskiwane rezultaty;

4. Ignorowanie teorii substancjalnych, szczególnie w przypadku, gdy badane są zachowania ludzi, prowadzi do bardzo ograniczonych wniosków, nie uwzględniających mi.in. kontekstu kulturowego czy politycznego. Koncentracja na szukaniu w zbiorze danych wszelkich zależności prowadzi zazwyczaj do wniosków powierzchownych, trywialnych, bądź bezsensownych, będących skutkiem odkrycia związków pozornych.

Nie sugerujemy przy tym, że świat społeczny DS składa się wyłącznie ze skrajnych empirystów. Na ograniczenia epistemologiczne zwrócił uwagę chociażby mający w środowisku status gwiazdy i autorytetu[28] Nate Silver w popularnonaukowej pracy *Sygnał i szum*. Zdaniem tego autora większa ilość danych oznacza głównie więcej szumu. Odkrycie korelacji pomiędzy zmiennymi może zarówno odzwierciedlać pewien sposób funkcjonowania świata, jak i być korelacją pozorną w sensie takim, że związek zmiennych jest przypadkowy (Silver 2014). Twierdzi on, że:

> Liczba istotnych relacji między elementami zbioru danych (…) jest o całe rzędy wielkości mniejsza [niż relacji pozornych]. Nie rośnie też tak szybko jak ilość dostępnych informacji: ilość prawdy na świecie nie zmieniła się tak bardzo od czasu wynalezienia Internetu, a nawet prasy drukarskiej. Większość danych to zwykły **szum**, podobnie jak większość wszechświata stanowi pusta przestrzeń. (podkr. org) (Silver 2014:234-235).

Można także znaleźć branżowe artykuły nawołujące do współpracy data scientistów z ekspertami w projektach biznesowych (Viaene 2013). Teksty jednoznacznie krytyczne wobec epistemologii i metodologii big data / DS stanowią margines publikacji w tej dziedzinie. Autorzy badający publikacje związane z zastosowaniem DS w ochronie zdrowia twierdzą, że prace krytyczne stanowią około 5% tekstów (11 w próbie 206) (Stevens, Wehrens, i de Bont 2018:4–5). Wyróżnili oni pięć typów idealnych dyskursów o big data, z czego trzy są entuzjastyczne – modernistyczny, instrumentalny, pragmatyczny – zaś dwa krytyczne – naukowy i krytyczno-interpretacyjny, a wspomniane 5% prac przyporządkowano do tego ostatniego (ibidem).

28 Znany jest z sukcesu w prognozowaniu wyników wyborów i wyników sportowych w USA. Prowadzi popularny blog FiveThirtyEight.

Na szereg problemów epistemologicznych i metodologicznych związanych z badaniami zbliżonymi do big data / DS – szczególnie, gdy badania te dotyczą ludzi – wskazywano już przed Kitchinem, bądź zwracano uwagę na inne ich aspekty. Zagadnienie społecznej konstrukcji pomiarów podejmowano mówiąc o kontekstowości czasowej, przestrzennej, materialnej tego, co uznajemy za dane. Dane są zawsze wygenerowane, nigdy nie są one nam dane (od czasownika „dać"). Nie można zatem traktować danych jako neutralnej, przejrzystej, pewnej, autonomicznej, prawdziwej reprezentacji rzeczywistości (Bowker 2005; Gitelman i Jackson 2013). Tym samym „surowe dane (*raw data*) to jednocześnie oksymoron i zły pomysł; przeciwnie, dane należy troskliwie przyrządzać (*should be cooked with care*)" (Bowker 2005:184), czyli podchodzić krytycznie do założeń stojących za mechanizmem generowania danych. „Nie może umykać nam fakt, że nie są one [dane] neutralne i nie wyrażają konieczności, lecz są mniej lub bardziej arbitralną konstrukcją tworzoną przez ludzi" (Iwasiński 2017:126). Wszyscy badacze interpretują dane, a pierwszym krokiem takiej interpretacji jest uznanie czegoś za dane właśnie (Boyd i Crawford 2011:5). Pretensje do obiektywizmu i dokładności tylko dlatego, że danych jest dużo są więc całkowicie nieuzasadnione i mylące. Używa się w tym kontekście znanego już nam terminu dataizm (*dataism*) (Dijck van 2014; Harari 2017; Smith 2018). Ideologia dataizmu zasadza się na wierze w obiektywność kwantyfikacji i zaufaniu do agentów zbierających dane (Dijck van 2014:198). Surma wskazuje, jak niebezpieczne jest to stanowisko:

> Jest oczywiste, że nasza pamięć jest ułomna i niemal zawsze nie ma całkowitej pewności co do odtwarzania z pamięci faktów. Natomiast pamięć cyfrowa wydaje się w tym kontekście nieskazitelnie perfekcyjna, co powoduje efekt silnego do niej zaufania. Ma to dwa istotne skutki uboczne: [1.] pamięć cyfrowa rejestruje tylko to, co można w formie cyfrowej zapisać i jednocześnie to, co zostało zarejestrowane. (...) Jest zatem oczywiste, że cyfrowy obraz świata jest niepełny. Niemniej przy tak dużym zaufaniu do zapisów cyfrowych można budować fałszywe interpretacje historii. Ten efekt fałszu jest spotęgowany tym, że pamięć cyfrowa rejestruje gigabajty przypadkowych i nieistotnych zdarzeń, bez kontekstu i priorytetów (...). [2.] Zarejestrowane dane mogą podlegać zmianie (Surma 2017:30–31).

Surma dostrzega możliwości poprawiania, zmiany i celowego zniekształcania, przywołując orwellowską wizję panowania nad przeszłością. Powtarza, że „raz wprowadzona do sieci informacja zaczyna żyć własnym, niekoniecznie zgodnym z wolą autora, życiem", i przestrzega: „jeśli zaufanie do własnej pamięci zastąpimy pochopnym zaufaniem do pamięci cyfrowej, to władza totalitarna nie będzie już potrzebowała kontrolowania naszych umysłów" (ibidem). Zauważmy, że w tym miejscu epistemologia staje blisko takich problemów jak prywatność pojedynczych ludzi, co z kolei jest przedmiotem regulacji prawnych. Filozof L.

Floridi pokazuje te problemy w odniesieniu do konkretnego przypadku firmy Google i tzw. prawa do bycia zapomnianym (Floridi 2014a, 2015).

Można posunąć się jeszcze dalej w krytyce epistemologicznej i czerpiąc z myśli Jeana Baudrillarda spojrzeć na dane każdego rodzaju jako iluzję, reprodukowaną przez naukowców i analityków różnej maści, co prowadzi do konstatacji, że dane nie odzwierciedlają żadnej rzeczywistości, a są jedynie obrazem rzeczywistości wyobrażonej przez badacza (Koro-Ljungberg 2013).

Szczególnie w przypadku pozyskania dużej ilości nieuporządkowanych danych pierwszym etapem umożliwiającym dalszą analizę jest ich przygotowanie, tzw. czyszczenie danych. Jak wskażemy w rozdziale 5. czyszczenie to fundamentalna część pracy data scientista, określana zwyczajowo jako pracochłonna, zajmująca około 80% czasu realizacji projektów. Etap ten jest jednak kolejną interpretacją danych (Boyd i Crawford 2011:5). To, że ilość danych nie świadczy o ich jakości, a zatem np. nie uprawnia do uogólniania wyników z próby na populację jest uznane w metodologii badań społecznych od dziesięcioleci (Cain i Finch 1981). Ma to odzwierciedlenie w metodologii badań etnograficznych, sondażowych czy eksperymentalnych. Ponadto dane zbierane automatycznie w dużych ilościach mogą zawierać liczne błędy i braki (Pink i in. 2018). Przekonanie, że więcej danych oznacza lepsze analizy to zatem kwestia etosu (Boyd i Crawford 2011:6). Duża ilość danych, to także większa ilość szumu, mówiąc językiem cytowanego wyżej Nate'a Silvera. Prowadzi to błędu zwanego apofenią (*apophenia*): dostrzegania nieistniejących wzorców, które pojawiają się w danych przypadkowo tylko ze względu na ich ilość (Boyd i Crawford 2011:2). Zatem małe dane mogą z powodzeniem prowadzić do bardziej wartościowych poznawczo rezultatów niż dane wielkie (Bornakke i Due 2018; Faraway i Augustin 2018; Jifa i Lingling 2014; Kitchin i Carrigan 2014; Polskie Towarzystwo Badaczy Rynku i Opinii 2017). Znany jest przykład pierwszego sondażu przedwyborczego Gallupa. Przyniósł on w latach 30. widocznie lepsze rezultaty niż badanie na większej, ale obciążonej próbie, gdzie respondentami byli wyłącznie prenumeratorzy „Literary Digest" (Babbie 2006). Z drugiej strony metody uczenia maszynowego nie zakładają tego, że dane stanowią reprezentatywną próbę losową z określonej populacji, a techniką służącą generalizowaniu uzyskanych wyników jest testowanie wyuczonego modelu na innym, nieznanym mu zbiorze danych (Larose 2005). Badacze społeczni są świadomi tej utraty monopolu na dostarczanie wiedzy o świecie społecznym (Afeltowicz i Pietrowicz 2008; Savage i Burrows 2007) i wskazują, że obecnie do zastosowań biznesowych korzystanie z eksploracji danych behawioralnych /

niedeklaratywnych – czyli analizy typu big data / DS – daje rezultaty lepsze z punktu widzenia praktyki rynkowej niż sondaże, głównie dzięki posiadaniu większej liczby bardziej szczegółowych danych, a co za tym idzie możliwości np. bardziej szczegółowej segmentacji klientów (Savage i Burrows 2007). Nie zgadzamy się jednak ze stawianymi przez entuzjastów big data tezami o śmierci ekspertów, w tym badaczy społecznych (Żulicki 2017).

Także matematycy, na przykładzie modelowania zjawisk przyrodniczych, argumentowali za brakiem możliwości uprawiania nauki na podstawie wyłącznie „surowych danych", bez sięgania do teorii dziedzinowych i stawiania hipotez (Hosin i Vulpiani 2018). Autorzy odwołali się bezpośrednio do cytowanych wyżej słów Ch. Andersona o mówiących same za siebie liczbach. Ich zdaniem, stały się one manifestem ideologicznym danocentrycznego (*datacentric*) entuzjazmu, i można w nich wskazać dwa charakterystyczne dla tego entuzjazmu poglądy. Po pierwsze, wygoda – za przekonaniem o mówiących za siebie liczbach stoi obietnica uzyskania prostej odpowiedzi na pytanie dotyczące bardzo skomplikowanej materii. Zdaniem autorów, poprzez korzystanie np. z wyszukiwarki Google jesteśmy przyzwyczajeni do takich odpowiedzi; nie rozumiemy przecież mechanizmu działania algorytmu PageRank, po prostu zadajemy pytanie i korzystamy z wyniku. Po drugie, odrzucenie przestarzałych poglądów – ostrożne i pracochłonne przyrodnicze podejście do zdobywania wiedzy o świecie jest nieefektywne, ograniczające i hamujące rozwój nowoczesnych technologii, zatem nie ma już dla niego miejsca i powinno być odrzucone (Hosni i Vulpiani 2018:121). N. Chomsky w słynnym wystąpieniu na Uniwersytecie Michigan także mówił (od 54. minuty nagrania), że praca z big data wydaje się łatwa, bo obiecuje duży efekt bez dużego wysiłku umysłowego. Jego zdaniem duża ilość danych nie ma znaczenia, jeśli nie wiemy, czego szukać. Analitykę danych pozbawioną podstaw teoretycznych porównał do sytuacji, w której człowiek chciałby zostać biologiem i dostaję poradę: to proste – idź do biblioteki biologicznej Harvarda, tam mają wszystko (Chomsky 2013).

Chcemy powiedzieć, że big data / DS jest innym, (ale nie zawsze, bezwzględnie i do-każdego-celu lepszym) sposobem dostarczania kwantyfikowalnej wiedzy niż badania ilościowe wywodzące się z tradycji nauk przyrodniczych, w tym sondaże. To w tym kontekście mówi się o dwóch kulturach modelowania zjawisk czy dwóch podejściach do nauki w ogóle (Breiman 2001; Ceri 2018; Manhart 1996). Przeciwstawia się naukę opartą na teoriach (*formal science model-driven*) nauce opartej na danych (*data science data-driven*). Pierwsza ma polegać na wyjściu od teorii, postawieniu hipotez i testowaniu ich za pomocą

danych zebranych w badaniu, druga zaś na analizie danych bez wcześniejszych pytań i założeń (przy czym jedno podejście nie wyklucza drugiego) (Ceri 2018). Tym samym DS to niewątpliwie inne narzędzia służące poznawaniu świata, lecz nie „tylko" narzędzia, a „aż" narzędzia. Przypomina to słowa, którymi zreferowano jedną z tez Marshalla McLuhana: „kształtujemy nasze narzędzia, a następnie one kształtują nas" (*We shape our tools and thereafter they shape us*) (Culkin 1967:70). Zgadzamy zatem się z tezą autorek *Six Provocations for Big Data*, czerpiących z myśli Latoura (2009), że te nowe narzędzia zmieniają sposób myślenia o badaniach społecznych i o teoriach społecznych (Boyd i Crawford 2011:3). Ten nowy sposób myślenia określano krytycznie jako „uwodzicielski pseudopozytywizm" (Dalton, Taylor, i Thatcher 2016:6), czy łagodniej „wywodzący się z idei pozytywistycznej" (Kitchin 2014:7), a także „odnowiony naturalizm", w którym społeczeństwo traktowane jest jak fenomeny przyrody – lawiny śnieżne, ruchy ławic ryb, interakcje komórek (Törnberg i Törnberg 2018:2). Przyjmujący taki naturalistyczny sposób myślenia badacze zapominają – poza szeregiem wymienionych wyżej problemów – o dyskusjach toczonych w naukach społecznych we wczesnej fazie rozwoju internetu. Mówiono wówczas, że cyfryzacja życia jest nieodłączną częścią procesu przejścia od modernizmu do postmodernizmu, co czyni życie społeczne bardziej otwartym, płynnym czy mozaikowym. Tym samym cyfryzacja może powodować, że życie społeczne jest coraz mniej, a nie coraz bardziej uchwytne, szczególnie za pomocą ilościowej analizy dużych ilości danych (ibidem). Kazimierz Krzysztofek poszedł jeszcze dalej i nazwał zjawisko „fragteracją", czyli równoczesnym rozpadaniem się, ale i integrowaniem świata społecznego (Krzysztofek 2010).

2.2.2. Data science jako strategia badań w naukach społecznych i humanistyce

Krytyka epistemologiczna i metodologiczna nie oznacza, że nowe narzędzia, metody czy podejście do badań społecznych nie są wdrażane przez wielu współczesnych naukowców. Na pograniczu humanistyki, nauk społecznych i DS wyrastają dziedziny aplikujące charakterystyczne dla DS metody do badań własnych: humanistyka cyfrowa (*digital humanities*) i obliczeniowe nauki społeczne (*computational social sciences*). Mówi się o zwrocie obliczeniowym (*computational turn*) dotyczącym myślenia i prowadzenia badań we współczesnej nauce. Boyd i Crawford porównują ten zwrot do zmiany społecznej, która zaczęła się od wprowadzenia produkcji masowej w fabrykach Forda. Tak jak wprowadzony tam system taśmowej produkcji samochodów zmienił rozumienie pracy, związku człowieka z

pracą i społeczeństwo w szerokim sensie w XX w., tak w wieku XXI zmiany społecznej ma dokonać zwrot ku obliczeniowemu rozumieniu świata dokonujący się za pomocą dużych i nieustrukturyzowanych zbiorów danych i narzędzi DS, umożliwiających pracę z tymi danymi (Boyd i Crawford 2011:3, 2012:665).

Jerzy Surma (2017:23) manifestem obliczeniowych nauk społecznych nazwał pracę Davida Lazera z zespołem (2009). Zauważyli oni, że nowe możliwości zbierania i analizy danych dotyczących ludzi są już przedmiotem pracy badawczej firm (wymieniono Google i Yahoo) i agencji rządowych (wskazano U.S. NSA[29]), zatem jeżeli akademia nie chce pozostawić innym podmiotom tego rodzaju prac – musi działać (Lazer i in. 2009:721). Artykuł rozpoczyna się słowami „żyjemy w sieci (*we live life in the network*)" (ibidem). Zauważmy, że badania nad sieciami społecznymi (*social network analysis,* SNA) w naukach społecznych sięgają lat 30. Jak wskazuje Linton C. Freeman, pierwsze użycia tej metafory w filozofii sięgają XIII w., jednak za początki metodycznego stosowania metafory sieci w naukach społecznych uznaje on socjometrię J. L. Moreno i H. Jenningsa (2014:26). Wypracowano wtedy założenia analizy sieci społecznych: uznano, że powiązania między ludźmi tworzą ważną strukturę o charakterze społecznym, zatem analiza sieciowa to nie badanie jednostki, ale relacji między nimi. Do badania takiej struktury korzystano z danych o charakterze relacyjnym; prezentowano model takiej struktury graficznie; rozwijano matematyczne metody opisu i wyjaśnienia modelu. Moreno zaproponował tzw. socjogramy, przygotowywane na podstawie badań kwestionariuszowych, gdzie pytano respondentów z danej grupy o ich wybory towarzyskie w obrębie tejże grupy. Na tej podstawie tworzono macierz „wybierający*wybierani", i diagram – socjogram, czyli wizualne przedstawienie struktury tych wyborów. „Moreno wprowadził niektóre kluczowe elementy do współczesnej analizy sieciowej: rysowanie mapy relacji między aktorami w ujęciu przestrzennym, aby przedstawić strukturę tych relacji" (Turner 2004:565). Wśród inspiracji teoretycznych i metodologicznych socjometrii, a zatem i teorii sieci społecznych w ogóle, Turner wskazuje także na dorobek antropologów: Johna A. Barnesa, Jamesa Clyde Mitchella i Elizabeth Bott. Uznaje on, że ci antropologowie połączyli idee badania konfiguracji powiązań między aktorami z bardziej metodycznym konceptualizowaniem własności sieci i nowatorskimi wówczas badaniami empirycznymi (Turner 2004:562-563).

29 Przypomnijmy, że w 2013 Edward Snowden ujawnił tajne dokumenty dotyczące zbierania cyfrowych danych o ludziach właśnie przez NSA (Dijck van 2014).

Badania nastawione na analizę sieci kontynuowali m.in. R. K. Merton czy C. Lévi-Strauss. Później, w latach 70., analizę sieci społecznych rozwinęła i ujednoliciła tzw. szkoła harvardzka H. C. White'a. Dopiero w późnych latach 90. zagadnieniem zainteresowali się m.in Albert-Laszlo Barábasi i Albert Réka (fizyk i biolożka). Pojawiła się także praca kanadyjskiego naukowca proponującego analizę sieci komputerowych jako sieci społecznych (Wellman 2001). Barábasi jest jednym z autorów cytowanego wyżej manifestu obliczeniowych nauk społecznych (Lazer i in. 2009). Od lat 90., jak uważa Freeman (2014), prace środowiska badaczy społecznych i fizyków przenikają się, jednak środowiska są odseparowane. Co ciekawe badanie Barábasiego jest przywołane w entuzjastycznej pracy *Big Data. Rewolucja…*, gdzie autorzy uzasadniają – w swoim przekonaniu kluczowe i nowatorskie - podejście „N = all". Barábasi dokonał analizy sieci na podstawie danych uzyskanych od europejskich operatorów sieci komórkowych. Były to wszystkie anonimowe logi telefonów z okresu czterech miesięcy; obejmować miały one około 1/5 mieszkańców Europy. Autorzy twierdzą, że odkryto zależność, której nie ujawniały mniejsze zbiory danych: dla stabilności sieci ważniejsze są osoby z niewielką ilością odległych powiązań niż te z dużą liczbą bliskich relacji (Cukier i Mayer-Schönberger 2014:49-50). Sam Barábasi w popularnonaukowej pracy *Linked: The New Science of Networks* powołuje się zarówno na Leonarda Eulera, osiemnastowiecznego matematyka, który stworzył podstawy teorii grafów[30], jak i na Stanley'a Milgrama i jego słynny opis sześciu stopni oddalenia poszczególnych ludzi w sieciach społecznych, nazwany problemem „małego świata" (Barabási 2002). Nie sugerujemy, że reprezentowana m.in. przez Barábasiego nowa nauka sieci (*New Science of Network,* NSN) zawdzięcza wszelkie swoje dokonania naukom społecznym. Przecież socjometrycy z lat 30. nie napisali pojęcia sieci społecznych na czystej tablicy: korzystali z zaplecza filozoficznego, aparatu matematycznego a także z nauk medycznych - J. L. Moreno był psychiatrą (Freeman 2014). Z kolei jak uważa Fran Webster, M. Castells i jego *Społeczeństwo sieci* wydane po raz pierwszy w 1996 r. niewątpliwie zainspirowane zostało nie tylko szkołą harvardzką, ale także (a może głównie?) zjawiskami społecznymi związanymi z rewolucją informacyjną, nowymi mediami i siecią Word Wide Web; co ciekawe Webster widzi u Castellsa postmarksizm (Maciąg 2014:165).

Kończąc myśl dotyczącą przecinania się światów nauk społecznych, przyrodniczych i biznesu zauważmy, że algorytm wyszukiwarki Google – PageRank – jest implementacją zainspirowanego przez socjologów i rozwiniętego przez fizyków pojęcia centralności sieci.

30 Pojęcia wierzchołków (*nodes*) i krawędzi (*edges*) z matematycznej teorii grafów pojawiają się także w kontekście architektur baz danych czy systemów rozpraszania obliczeń, o czym w części 5.3. naszej pracy.

Dotyczy ono tego, że węzły posiadające większą liczbę połączeń stają się istotnymi centrami (Freeman 2014). Przy tym autorzy polscy zwracają uwagę na to, że nowa nauka sieci (NSN) i badanie sieci społecznych (SNA) pozostają odrębnymi dyscyplinami, choć SNA rozwija się w oderwaniu od tradycji humanistycznej (Afeltowicz i Pietrowicz 2008:57–59).

Ośrodkiem rozwoju obliczeniowych nauk społecznych w początkach XXI w. był Massachusetts Institute of Technology. Jedne z pierwszych prac koncentrowały się na analizach behawioralnych z uwzględnieniem sieci społecznych i lokalizacji osób, początkowo dzięki specjalnie skonstruowanemu przez badaczy urządzeniu do noszenia na ciele (Choudhury, Pentl, i Pentland 2002), później z użyciem telefonów komórkowych i sensorów (zapisywano m.in. dane z połączeń bluetooth, WiFi, GPS, sieci komórkowych) (Eagle 2005; Eagle i Pentland 2006). W tym czasie pojawiły się liczne prace autorstwa badaczy z innych uczelni. Podejmowano m.in. tematy związane zarówno z zarządzaniem, polityką jak i zdrowiem publicznym: od mierzenia produktywności pracy (Brynjolfsson, Aral, i Alstyne 2007), rozprzestrzeniania się palenia papierosów w sieciach społecznych (Fowler i Christakis 2008b), sponsorowania działalności politycznej (Fowler 2006), do rozprzestrzeniania się szczęścia (Fowler i Christakis 2008a). W Polsce pionierem tego rodzaju badań jest Dominik Batorski, który już w 2004 obronił na Uniwersytecie Warszawskim pracę doktorską w dziedzinie socjologii pt. *Sieci społeczne: Charakterystyka, uwarunkowania i konsekwencje struktur relacji społecznych na przykładzie komunikacji internetowej.* Na potrzeby tej rozprawy analizował duże zbiory danych z popularnego w owym czasie komunikatora internetowego Gadu Gadu z użyciem języka programowania GNU R (Batorski 2004). W pracy naukowej Batorski prowadził także badania użytkowania internetu w ramach Diagnozy Społecznej i wypowiadał się o internecie jako polu badań i źródle danych dla badań społecznych (Batorski i Olechnicki 2007). D. Batorski jest aktualnie organizatorem otwartych spotkań, tzw. meetupów, grupy Data Science Warsaw (Data Science Warsaw 2018b), był członkiem rady programowej „największego wydarzenia data science w Polsce" (Data Science Warsaw 2018a), współorganizował lub promował liczne imprezy tego świata społecznego (ChallengeRocket.com 2018; Kaggle 2018; PyData 2018; WDI18 2018). Jest jednym z założycieli i Chief Scientist w firmie Sotrender oferującej narzędzie do analizy marketingu w mediach społecznościowych (Sotrender 2018). Obecnie pojawiają się artykuły określane jako obliczeniowe nauki społeczne, gdzie oprócz analizy sieci społecznych używane są inne metody analiz ilościowych, w tym analizy tekstów (Anderson, Kleinberg, i Mullainathan 2016; Barrio, Goldstein, i Hofman 2016; Hofman, Sharma, i Watts 2017). Od

stycznia 2018 ukazuje się czasopismo *Journal of Computational Social Science,* w którym publikowane są:

> przełomowe badania z dziedzin nauk społecznych (socjologii, ekonomii, nauk politycznych, psychologii, lingwistyki i innych), fizyki, biologii, nauk o zarządzaniu, informatyki (*computer science*) i data science (Springer 2018).

W ostatnim wydaniu ukazały się artykuły dotyczące m.in. prognozowania miejsca i czasu wystąpienia przestępstwa (Kupilik i Witmer 2018), symulacji ruchu ulicznego (Umemoto i Ito 2018) czy rozprzestrzeniania się propagandy Jihadu (Badawy i Ferrara 2018).

Cyfrowa humanistyka (*digital humanities*) posługuje się często metodami ilościowej analizy tekstów, w tym ilościową eksploracja tekstów (*text mining*). Teksty są jednym z typów danych nieustrukturyzowanych. Jak wskazywaliśmy w części 1.1.1., zdaniem niektórych autorów big data to właśnie dane zarówno duże, jak i nieustrukturyzowane. Mogą być to np. wpisy na portalach społecznościowych, komentarze na forach internetowych, treści stron WWW czy blogów, zdigitalizowane bądź wydawane wyłącznie w formie cyfrowej artykuły, książki, dokumenty. Text mining to komputerowa analiza tekstów, traktowanych jako dane. Używany jest także termin obliczeniowa analiza tekstu (*computational text analysis*) (O'Connor, Bamman, i Smith 2011). Jak wskazaliśmy (Żulicki 2017), analiza jakościowa z użyciem oprogramowania CAQDAS, gdzie człowiek z pomocą komputera i np. programu Atlas czy NVivo koduje transkrypcje wywiadu, nie jest analizą text mining.

> (…) text mining w porównaniu do analizy jakościowej wykonywanej zazwyczaj przez człowieka wydaje się atrakcyjny pod kątem stuprocentowej powtarzalności wyników, złożoności czasowej metody, natomiast może jej ustępować pod kątem poprawności wyników (Dzieciątko i Spinczyk 2016:11).

Text mining to zatem analiza ilościowa, także statystyczna, umożliwiająca różnego rodzaju pomiary tekstów / dokumentów. W uproszczeniu służy ona odkrywaniu dominujących wzorów użycia słów i wzorów powiązań między słowami lub dokumentami (O'Connor, Bamman i Smith 2011). Metody text mining wywodzą się z eksploracji danych ustrukturyzowanych (*data mining*) i dziedziny zwanej NLP (*natural language processing*), której przedmiotem jest przetwarzanie informacji zawartej w języku naturalnym - np. w celu umożliwienia sterowania urządzeniem za pomocą głosu, bądź w celu automatycznej zamiany mowy na tekst czy odwrotnie (Dzieciątko i Spinczyk 2016). Zastosowania text mining to m.in. takie zadania, jak: identyfikacja słów / zdań kluczowych, kategoryzacja dokumentów według wzorca albo grupowanie bezwzorcowo (wyestymowanie kategorii), wykrywanie

określonych treści w dokumentach, czy wspomniana jako ogólne zadanie text mining identyfikacja wzorów i powiązań między słowami / dokumentami (ibidem). Wspomniana już metoda LDA (*Latent Dirichlet Allocation*) jest przykładem bezwzorcowego grupowania dokumentów – rozpoznawaniem tematów (Blei i in. 2003).

Cyfrowy humanista ma dwa znaczenia: jako osoba, która za pomocą narzędzi cyfrowych, obliczeniowych, programistycznych pracuje w dowolnej dziedzinie humanistyki oraz jako osoba, która za przedmiot swoich badań humanistycznych bierze przestrzeń cyfrową, czyli np. gry komputerowe, portale społecznościowe, strony internetowe (Szpunar 2016a:364). Można łączyć te znaczenia, co robi chociażby Aleksandra Przegalińska, o której powiemy dalej, czy Stefania Druga, badająca m.in. interakcje dzieci ze sztuczną inteligencją (Druga i in. 2018; Williams i in. 2018). Obie badaczki można, naszym zdaniem, równie dobrze zaklasyfikować do obliczeniowych nauk społecznych.

Wskazuje się, że humanistyka cyfrowa charakteryzuje się podejściem do uprawiania nauki, na które składają się otwartość rozumiana jako nielimitowany dostęp do publikacji, do danych i do kodu, zatem występują przecięcia z ruchami wolnego oprogramowania (*open source*) i otwartej nauki (*open access*). Cyfrowy zwrot w humanistyce ma być także odejściem od właściwej dla kultury druku linearności w kierunku współpracy i otwartości o charakterze wikinomicznym, czyli „raczej modelu bazaru, ułatwiającego twórczy ferment niż elitarnego i zamkniętego modelu katedry", co jest metaforą zaczerpniętą od Erica Raymonda, libertanina, hakera i filozofa open source (ibidem:362-64). Ponadto

> Otwartość i gotowość do dzielenia się wiedzą, a także nieustanne uczenie się – także od innych, niekoniecznie profesjonalistów i specjalistów w danej dziedzinie. Odejście w tym przypadku od elitarnej działalności wiedzotwórczej na rzecz pluralistycznego bazaru stać się może szansą na innowacyjność i niekonwencjonalność rozwiązań niedostrzeganych przez zanurzonych w problemie specjalistów (Szpunar 2016a:364).

Za początki humanistyki cyfrowej uznawany jest rok 1949, ale wskazuje się, że do lat 70. komputerowe przetwarzanie tekstów służyło głównie tworzeniu skorowidzów. Później rozwinęły się zastosowania analityczne w lingwistyce czy stylometrii, a czasy jej rozkwitu określa się jako okres rozpowszechniania internetu i WWW – początek lat 90. (Hockey 2004). Pierwsze czasopisma naukowe z dziedziny humanistyki cyfrowej sięgają właśnie lat 90. Odnaleźć można prace poświęcone ilościowym analizom literatury pięknej (Burrows 1996; Olsen 1993; Tompa 1996). Nieco później pojawiały się czasopisma, gdzie publikowano artykuły poświęcone nie tylko analizie tekstów (Wolff 2007), ale np. analizie obrazów (zarówno malarstwa jak i fotografii) nastawionej na znajdowanie podobieństw (Nauta

2008) czy wielowymiarowemu cyfrowemu zapisowi sutr tybetańskich (Patrik 2007). Jednym z bardziej znanych autorów z dziedziny cyfrowej humanistyki jest badacz kultury i mediów Lev Manovich, który realizuje tego typu prace od 2007 roku (Manovich 2012b:462). Wśród współczesnych badań wymienić można książkę Manovicha *AI Aesthetics*, w której częściowo zastosowano metody obliczeniowe, a obszarem badawczym jest rola sztucznej inteligencji w mediach i sztuce. Chodzi zarówno o rekomendowanie treści – książek, obrazów, muzyki – z użyciem AI, jak i komputerowe generowanie takich treści (Manovich 2018). Powstały formalne grupy zajmujące się humanistyką cyfrową. Istnieje międzynarodowa wspólnota nazywająca się akronimem DARIAH – Digital Research Infrastructure for the Arts and Humanities. Jej członkami są badacze z siedemnastu krajów europejskich. Organizację DARIAH opisano jako infrastrukturę wspierającą badania i nauczanie metod humanistyki cyfrowej, w tym text mining. Zapewnia ona m.in. możliwość magazynowania danych; narzędzia przetwarzania i analizy danych; oraz procedury mające zapewnić interoperacyjność w różnych lokalizacjach, dyscyplinach naukowych, różnych kontekstach akademickich i kulturowych, a także różnych językach (DARIAH 2018). W Polsce na zbliżonym polu działa Laboratorium Cyfrowe Humanistyki Uniwersytetu Warszawskiego. Powstało z inicjatywy wydziałów humanistycznych UW, na których prowadzi się badania z użyciem narzędzi cyfrowych oraz Wydziału Matematyki, Informatyki i Mechaniki i Interdyscyplinarnego Centrum Modelowania Matematycznego i Komputerowego. W misji jednostki napisano:

> LaCH UW włącza się w rozwój społeczeństwa informacyjnego, daje wsparcie priorytetowym obecnie multidyscyplinarnym kierunkom badań, które zakładają wykorzystanie na szeroką skalę technologii informatycznych w humanistyce. (…) Wspierając badania z zakresu humanistyki cyfrowej LaCH UW promuje jednocześnie całą humanistykę, wierząc, że narzędzia cyfrowe przyczyniają się do usprawnienia i przyspieszenia transferu wiedzy, otwierają przed humanistyką nowe możliwości badawcze i edukacyjne (Laboratorium Cyfrowe Humanistyki UW 2017).

W innym wymiarze francuski filozof Bernard Stiegler postuluje wprowadzenie studiów cyfrowych (*digital studies*), których dział stanowi humanistyka cyfrowa (*digital humanities*) w ramach kształcenia we wszystkich uniwersyteckich dyscyplinach. Uważa on, że obecnie trwa

> (…) kryzys uniwersytetu, który jest efektem radykalnego przekształcenia nowoczesnego świata. Przekształcenia wywołanego najpierw przez pojawienie się technologii analogowych, a następnie technologii cyfrowych (Stiegler 2017).

Ten kryzys „nauczania pomyślanego na bazie pisma" jest bezpośrednio związany z rozwojem technologii analogowych w latach 60. Ich hegemoniczne panowanie Horkheimer i

Adorno nazwali przemysłem dóbr kulturowych. Od lat 90. rozpoczęło się panowanie technologii cyfrowych. Ma to związek z całościowym projektem uniwersytetu, co Stiegler sprowadza do „technologii ducha" (ibidem:22-23). Stiegler uważa, że rozwój uniwersytetu zgodny z obecnym trendem, będzie opierał się na bez-myśli (*impensé*) i wyparciu. Polega to na błędnym odrzuceniu roli techniki w formułowaniu wiedzy, „konstytuowaniu się 'duszy noetycznej' w sensie ogólnym". Stiegler przekonuje, że pismo jako mnemotechnika jest warunkiem teoretycznego działania rozumu, a zatem w konsekwencji formułowania wiedzy. Dalej wskazuje, że technologia analogowa, a później cyfrowa jest nowym stadium pisma i lektury. Przekształca to warunki prowadzenia badań, nauczania, relacji uczelni z otoczeniem. Otoczenie przenika do uczelni poprzez technologie (np. smartfony, internet). Autor wątpi w stawienie oporu i utrzymanie kształtu uniwersytetu niczym z Aten 4 wieku p.n.e., stąd każde studia uniwersyteckie mają być cyfrowymi (Stiegler 2017).

Współcześnie polscy humaniści i badacze społeczni pracują w różnych obszarach humanistyki cyfrowej czy obliczeniowych nauk społecznych. Kulturoznawca i socjolog Radosław Bomba raczej nie korzysta z metod obliczeniowych, ale z pewnością z cyfrowych i prowadzi analizy cyfrowej kultury – jest zarówno badaczem gier komputerowych (Bomba 2015), jak i propagatorem humanistyki cyfrowej (Bomba 2017; Radomski i Bomba 2013). Podobnie można scharakteryzować dorobek cytowanej wyżej Magdaleny Szpunar. Sama siebie nazywa socjolożką internetu, zatem jest „cyfrowa" w sensie przedmiotu badań, a jej prace dotyczyły np. lęku wobec technologii (Szpunar 2005, 2006, 2018b), specyfiki komunikowania się za pomocą internetu (Szpunar 2012, 2015, 2016b), czy nierówności technologicznych, w tym algorytmicznych (Szpunar 2013, 2018a). Maciej Eder, filolog i dyrektor Instytutu Języka Polskiego PAN w Krakowie zajmuje się ilościową analizą literatury, m.in. do zadań atrybucji autorskiej i stylometrii. Poza licznymi publikacjami naukowymi (Eder 2013, 2015, 2017; Rybicki i Eder 2011), jest on autorem narzędzia stylometrycznego – pakietu „stylo" w języku R (Eder, Rybicki, i Kestemont 2016). Był on prelegentem na ogólnopolskich konferencjach użytkowników R w 2017 i 2018 r. (Why R? Foundation 2018). Socjolog Marek Troszczyński wykorzystał metody uczenia maszynowego i text mining do wykrywania mowy nienawiści na forach internetowych (Troszczyński i Wawer 2017). Krzysztof Tomanek i Grzegorz Bryda stosowali podobne narzędzia do badania postaw nauczycieli akademickich w opiniach studentów (2015), pisali także o metodyce prowadzenia podobnych badań (Tomanek i Bryda 2017), zaś Agnieszka Kwiatkowska użyła text miningu do analizy polskich debat parlamentarnych (2017). Metody ilościowej analizy

tekstu zostały zauważone przez polskich badaczy z kręgu socjologii jakościowej i są wdrażane do praktyki badawczej obok oprogramowania CAQDAS. Wspomniani Bryda i Tomanek już w 2014 r. pisali o zastosowaniach text mining w książce Jakuba Niedbalskiego (2014). W roku 2016 na XVI zjeździe Polskiego Towarzystwa Socjologicznego problematyce poświęcono panel „Big Data, CAQDAS i nowe technologie w polu socjologii jakościowej" (PTS) z udziałem wyżej wymienionych socjologów. W roku 2017 ukazał się „Przegląd Socjologii Jakościowej" (tom XII nr 2) pod tytułem „Big Data i CAQDAS w badaniach jakościowych". Artykuł wprowadzający redaktorów tomu zawiera rozważania na temat stosowania wspomnianych metod, jak i poruszonej wyżej specyfiki epistemologicznej tego rodzaju analiz (ateoretyczność - zmianę pytania będącego podstawą budowania modeli z „dlaczego?" na „co?" i „dane mówią same za siebie", analizowaniu całej populacji zamiast próby) (Brosz, Bryda, i Siuda 2017). W „Studiach Socjologicznych" pojawiły się także tematycznie pokrewne teksty pokazujące możliwości korzystania z nowych źródeł danych w naukach społecznych: „Google big data: charakterystyka i zastosowanie w naukach społecznych" (Turner, Zieliński, i Słomczyński 2018) i „Twitter jako przedmiot badań socjologicznych i źródło danych społecznych: perspektywa konstruktywistyczna" (Rodak 2017). Według Z. Tufekci (autorka ta nazywa siebie technosocjolożką) dane z Twittera są bardzo popularnym polem badań z uwagi na względną łatwość w ich pozyskaniu (np. w porównaniu z danymi z Facebooka), ale badacze niewiele uwagi poświęcają problematyzowaniu tego źródła – głównie mowa o nieznanym obciążeniu badanej próby (Tufekci 2014:1–2). Dane z tego portalu społecznościowego analizowano np. we wspomnianym artykule dotyczącym sieci społecznych jihadystów (Badawy i Ferrara 2018), czy dążąc do stworzenia kontekstualnego wykrywacza sarkazmu (*contextualized sarcasm detection*) – modelu klasyfikującego twitty (Bamman i Smith 2015). Zauważmy jednak, że w drugim z wymienionych artykułów autorzy należą do nauk informatycznych (*computer science*), nie społecznych. W Polsce badanie przekazywania informacji między ludźmi za pomocą Twittera realizowane jest przez fizyków. Projektem RENOIR – Reverse EngiNeering of sOcial Information pRocessing (RENOIR 2018) finansowanym w ramach programu EU Horyzont 2020 kieruje pracownia Fizyki w Ekonomii i Naukach Społecznych Wydziału Fizyki Politechniki Warszawskiej. Opublikowano tutaj artykuły dotyczące np. mierzenia entropii w dynamice komunikacji międzyludzkiej (Kulisiewicz i in. 2018), wykrywania źródła informacji w sieci społecznej (Paluch i in. 2018) czy kwantyfikacji podobieństwa warstw sieci (Bródka i in. 2018).

Badania z użyciem metod ilościowych, a dodatkowo dotyczące sztucznej inteligencji, prowadzi Aleksandra Przegalińska. Zajmowała się ona interakcją między człowiekiem a chatbotem (Ciechanowski i in. 2018), czy możliwością mierzenia uwagi człowieka za pomocą ubieralnego gadżetu, sprzedawanego jako pomoc w treningu medytacji (Przegalińska i in. 2018). Przegalińska jest doktorem filozofii, jej rozprawa doktorska pt. *Fenomenologia istot wirtualnych* osadzona była w polu fenomenologii techniki i technologii. Cel pracy autorka określiła następująco: „chodzi mi zatem o wykorzystanie kategorii fenomenologicznych w opisie pierwszoosobowego, żywego doświadczenia rzeczywistości wirtualnej" (Przegalińska 2013:10). Tutaj jeszcze nie korzystała z podejścia ilościowego, ale dokonywała analizy „określonej grupy wytworów sztucznej inteligencji, czyli chatbotów i awatarów" (Przegalińska 2013:13). Przegalińska wypowiadała się wielokrotnie w mediach i pracach popularnonaukowych o przyszłości ludzi i sztucznej inteligencji (Brzezińska i Przegalińska 2015; Przegalińska 2014, 2015, 2016a, 2016c, 2016b), była też jedną z głównych prelegentek na warszawskiej konferencji branżowej DS (PyData 2018).

W pracach obliczeniowych nauk społecznych widoczny jest entuzjazm wobec stosowanej metodologii i epistemologii. Pojawiają się pretensje do prawdziwości, obiektywności, pewności, słuszności i użyteczności uzyskiwanych wyników. Pisano o „eksploracji rzeczywistości" (Eagle i Pentland 2006), „sygnałach prawdy" jako „twardych miarach" zachowań społecznych (Arena, Pentland, i Price 2010; Pentland i Heibeck 2008), analizach o bezprecedensowym zasięgu, głębi i skali (Lazer i in. 2009:722), inżynierii społecznej dla lepszego świata (Eagle i Greene 2014), uzyskaniu „punktu widzenia Boga na nas samych" (*God's eye view of ourselves*) (Pentland 2009:80). Kazimierz Krzysztofek wskazuje, że Barabási – opisywany wyżej fizyk, jeden z twórców NSN – wyrażał nadzieję odkrycia ścisłych, matematycznych praw opisujących i pozwalających prognozować ludzkie zachowanie dzięki analizie obiektywnych danych (Krzysztofek 2011:126). Prezentowaliśmy wcześniej krytyków tego rodzaju pretensji, należy jednak zauważyć, że naukowcy używający metod obliczeniowych w badaniach społecznych także poddawali w wątpliwość własne założenia i metody. Pisano o sprzeciwie wobec głosów o końcu teorii i danych mówiących samych za siebie (Chang, Kauffman, i Kwon 2014), czy o pułapce pozornego obiektywizmu (Eder 2014).

Powstają także prace postulujące włączanie big data / DS do arsenału badawczego nauk społecznych, obok innych metod i technik. Językiem przywołanych wcześniej autorów badających dyskursy o big data w ochronie zdrowia, można by przyporządkować te teksty do

kategorii „naukowych" - krytycznych, zwracających uwagę na ograniczenia, wskazujących na potencjał, jeśli chodzi o eksplorację danych (Stevens i in. 2018:4–5; 8–9). W naukach społecznych mówi się o integracji wielkich danych i grubych danych (*big data with thick data*). Autorką dźwięcznego określenia „grube dane" (*thick data*) jest Tricia Wang, socjolożka stosująca metody etnograficzne w badaniach marketingowych (Wang 2013, 2016). Określenie zainspirowane zostało terminem big data jak i zaproponowanym w latach 70. „opisem gęstym" (*thick decription*) C. Geertza. Oznacza ono materiały zbierane za pomocą metod i technik badań jakościowych. Łączenie wglądów (*insights*), jakie dają analiza danych ilościowych typu big data i analiza materiałów jakościowych ma prowadzić do osadzenia wyników analizy ilościowej w kontekście, zatem ułatwić ich wyjaśnianie i rozumienie (Bornakke i Due 2018:1–4). Jako przykład autorzy podają m.in. badania Wendy F. Hsu, która zajmowała się niezależną sceną muzyki rockowej osób pochodzących z Azji w USA. Hsu używała typowych technik etnograficznych: wywiadów, obserwacji, w tym obserwacji mediów społecznościowych i stron www. Obserwując szeroko wówczas wykorzystywany przez muzyków portal MySpace zauważyła, że ulokowane w USA zespoły mają fanów z wielu azjatyckich – i nie tylko – krajów. Skorzystała z metod ściągania danych (*webscraping*) o lokalizacji fanów z MySpace, co pozwoliło poszerzyć wnioski i pokazać zasięg geograficzny społeczności fanów badanej muzyki (Hsu 2014). Nawet w popularnonaukowym badaniu dotyczącym przekraczania granic intymności kobiet bawiących się w nocnych klubach w Brazylii wyraźnie widać, że dopiero osadzenie analizy ilościowej w kontekście prowadzi do zrozumienia problemu. Badaczki ubrane w sukienkę wyposażoną w niewidoczne sensory dotknięć uczestniczyły w imprezach, niejawnie nagrywano także materiał audio i wideo. Choć w materiale z tego badania podkreślona jest rola sukienki z sensorami, to nie wykresy obrazujące ilość dotknięć, a nagrane zachowania i wypowiedzi prezentujące interakcje badaczek z mężczyznami pozwalają zrozumieć problem, o czym świadczą zaprezentowane w materiale reakcje odbiorców badania (Ogilvy 2018). O komplementarności badań etnograficznych i DS przekonuje Heather Ford, która we współpracy z data scientistami badała powstawanie artykułów na Wikipedii (Ford 2014, 2016). Podobne badania wykonywano nawet wcześniej, niż wszedł do użycia termin „data science" – w czasach popularności określenia „data mining" (Anderson i in. 2009). Powstała już pierwsza dotycząca omawianego zagadnienia monografia o tytule *Ethnography for a data-saturated world*. Przy współpracy data scientistów i badaczy społecznych zawarto tam głównie porady

metodologiczne i praktyczne oraz przykłady badań, ale i próby badań etnograficznych środowiska data scientistów, do których jeszcze powrócimy (Knox i Nafus 2018).

Także w Polsce pisze się o potrzebie łączenia etnografii i etnografii cyfrowej z metodami ilościowymi typu big data na zasadzie triangulacji metod. „Bogactwo analityki danych [ilościowych] zwiększa, a nie zmniejsza zapotrzebowanie na uzupełnianie ich o wyniki badań etnograficznych" (Jemielniak 2018:17). Ten autor opublikował także pierwszą w Polsce, bogatą monografię „Socjologia Internetu" (2019), gdzie pokazał liczne metody ilościowe i jakościowe dla tego medium i obszaru badań. Zwracamy uwagę, że także Rob Kitchin w omawianym wyżej krytycznym artykule postulował ostrożne korzystanie z możliwości, jakie daje big data / DS, obok tradycyjnego podejścia do badań ilościowych i jakościowych w naukach społecznych. Mówił on o możliwości przyjęcia nowej, epistemologicznej strategii, która miałaby łączyć w sobie indukcję, dedukcję i abdukcję. Chodzi o naukę opartą na danych (*data-driven science*) w tym sensie, że dzięki eksploracyjnej analizie dużych lub nieustrukturyzowanych danych można generować pytania i hipotezy dla kolejnych etapów osadzonego w ramach teorii dziedzinowej badania (Kitchin 2014:5–6). Bazując m.in. na jego myśli pozwoliliśmy sobie we wcześniejszej pracy postulować włączanie big data / DS do badań społecznych w ramach triangulacji metod i technik, zarówno w badaniach prowadzonych w paradygmacie pozytywistycznym (strukturalnym), jak i humanistycznym (interpretatywnym) (Żulicki 2017:200). Inni autorzy także zainspirowani m.in. omawianym artykułem Kitchina mówiąc o „symfonicznych naukach społecznych" postulują podobnie jak on abdukcyjne podejście, upatrując w nim przyszłość tych nauk (Halford i Savage 2017:1145). W ramach nauk społecznych powstawały już prace łączące dane ilościowe z różnych źródeł z teorią społeczną. Wymienia się wydane w 2000 *Bowling Alone* Roberta Putnama*, The Spirit Level* autorstwa Richarda Wilkinsona i Kate Pickett z 2009 oraz Thomasa Picketty'ego *Capital* z 2014 r. (Halford i Savage 2017:1132–34). Jeszcze innym, według nas ważnym i słabo zagospodarowanym polem współpracy DS / specjalistów od uczenia maszynowego z przedstawicielami nauk społecznych jest wyjaśnianie działania gotowych modeli (Miller 2019). Nie chodzi o uczenie ludzi teorii stojącej za modelowaniem ani o uczenie umiejętności modelowania. Chodzi o wyjaśnianie człowiekowi skąd biorą się konkretne odpowiedzi modelu, ważności zmiennych w modelu itd., szczególnie dla metod o bardzo wysokiej złożoności (jak *deep learning* lub zespoły modeli ML tzw. *model ensemble*). Ten obszar badań nad uczeniem maszynowym nazywany jest XAI (*eXplainable AI*); w Polsce zajmuje się nim m.in. Przemysław Biecek (por. część 2.1).

Zgadzamy się z Boyd i Crawford, które twierdzą, że źródłem zwrotu obliczeniowego (*computational turn*) jest nie tylko akademia. Mówią one o silnym dążeniu firm i rządów do wydobywania z danych maksymalnej użyteczności, w postaci np. profilowania reklamy, projektowania produktów, zarządzania ruchem ulicznym czy przeciwdziałania przestępczości (Boyd i Crawford 2011:13). Van Dijck zwraca uwagę na podnoszone już kwestie epistemologiczno-metodologiczne. Zarówno rządy, biznes jak i nauka traktują dane jak Świętego Graala, prawdziwe odzwierciedlenie ludzkiego zachowania zebrane za pomocą przezroczystych termometrów, zatem najbardziej wiarygodne, a co za tym idzie użyteczne źródło informacji (Dijck van 2014:199). Krzysztofek mówi w tym kontekście o „logice cywilizacji numerycznej" – istnieje to, co jest kwantyfikowalne, a musi być takie przynajmniej po to, aby możliwa była wycena w pieniądzach (Krzysztofek 2012:225–26). Inni autorzy mówią o kulturze algorytmicznej (Striphas 2015). Z powodu nastawienia na użyteczność nazywamy w tej pracy DS działalnością paranaukową.

> Myślenie współczesne jest pod przemożnym wpływem nauki, a naukowość – rzeczywista lub pozorowana – traktowana jest jako najwyższa nobilitacja idei. Nic dziwnego, że legitymizację naukową usiłują pozyskać poglądy nie mające z nauką nic wspólnego. (…) Obserwujemy dziś rozrost sektora **paranukowego** [podkr. RŻ], zwłaszcza w tych dziedzinach, w których nauka odpowiada na jakieś palące ludzkie potrzeby i nadzieje (Sztompka 2007:298).

Wśród krytyków pojawia się określenie „pseudonauka" (Crawford i in. 2018:4, 8, 14), ale mowa nie o całej dziedzinie, a konkretnie o technikach maszynowego rozpoznawania emocji na podstawie obrazu twarzy. To nastawienie na użyteczność niesie wiele konsekwencji etycznych i społecznych, które obok krytyki metodologiczno-epistemologicznej są przedmiotem zainteresowania tzw. krytycznych studiów danych (*critical data studies*) (Dalton i in. 2016).

Istnieje jeden tekst, w którym autorzy postulują łączenie nauk społecznych i data science, ale nie w sensie opisywanej wyżej triangulacji, a włączania powstałej na gruncie nauko społecznych krytyki metodologiczno-epistemologicznej i dotyczącej konsekwencji społecznych do badań i praktyki data science (Neff i in. 2017). Artykuł przygotowano częściowo na podstawie badań etnograficznych przeprowadzonych w zajmującym się data science środowisku akademickim w USA, jednak nie opisujemy tego jako przykładu badań społecznych poświęconych data science (2.2.4.), ponieważ tekst ma charakter interwencyjny – jest przede wszystkim wezwaniem do działania na rzecz bardziej etycznego data science.

2.2.3. Konsekwencje społeczne data science

Kate Crawford i Danah Boyd[31] oraz związane z nimi instytuty AI Now oraz Data & Society (AI Now Institute 2018; Data & Society 2018) są obecnie najbardziej aktywne w polu krytycznych studiów danych. Działalność Boyd i jej grupy została pozytywnie oceniona przez osoby zajmujące się DS w USA, gdyż „stawiają oni [instytut Data & Society] pytania o antropologiczne implikacje big data" (Gutierrez 2014:320). Boyd i Crawford prowadzą działania na pograniczu nauki i interwencji społecznej – np. współtworzyły bezpłatnie dostępny spersonalizowany serial dokumentalny o prywatności w sieci i gospodarce internetowej „Do Not Track" (UPIAN 2015) – co naszym zdaniem jest wartościowe. Zauważmy także, że Boyd i Crawford pracują w firmie Microsoft na stanowiskach dyrektorek badań (*Principal Researcher*) (Microsoft Research 2019a, 2019b). Choć krytyka praktyk związanych ze zbieraniem informacji o ludziach, inwigilacją czy nierównościami związanymi z technologią jest znacznie starsza niż DS, big data czy tzw. sztuczna inteligencja, to omawiamy aktualną krytykę autorstwa Crawford i Boyd, oraz osób z nimi współpracujących.

AI Now Institute już po raz trzeci opublikował raport poświęcony konsekwencjom społecznym najnowszych wydarzeń w omawianych dziedzinach. Autorzy posługują się terminem sztuczna inteligencja – AI – mając na myśli szeroką gamę technologii, od bazujących na zaprogramowanych regułach, przez te korzystające ze statystyki czy uczenia maszynowego, do najnowszych technik uczenia głębokiego (*deep learning*), które wykonują zadania korzystając z danych i obliczeń. Chodzi o zadania rozpoznawania mowy, tłumaczenia, rozpoznawania obrazów, prognozowania, podejmowania decyzji (*determination*) (Crawford i in. 2018:44). Zdaniem autorów w centrum skandali, jakie zaszły w AI w 2018 r. jest pytanie o odpowiedzialność / rozliczalność (*accountability*) (Crawford i in. 2018:7) - kto jest odpowiedzialny za to, że system AI robi człowiekowi krzywdę? Autorzy raportu posługują się nierzadko publicystyczną retoryką i dyrektywnymi hasłami, ale warto przyjrzeć się dokładniej omawianej publikacji z uwagi na liczne przykłady tzw. skandali z systemami sztucznej inteligencji w roli głównej. AI Now Institute twierdzi, że w 2018 roku systemy AI gwałtownie rozwijały się w kolejnych dziedzinach życia społecznego, stwarzając ryzyko dla coraz większej liczby ludzi. Choć AI wciąż oferuje warte rozważenia obietnice, to szybkie wdrażanie tych systemów bez odpowiedniej oceny (*assessment*), odpowiedzialności (*accountability*) i kontroli (*oversight*) może powodować poważne zagrożenia. Według

31 Właściwie danah boyd (z małych liter), zgodnie z objaśnieniem na prywatnej stronie internetowej socjolożki (Boyd b.d.). Przyjmujemy pisownię nazwiska tej autorki wielką literą, zgodnie z polską konwencją pisania rozprawy doktorskiej.

autorów, potrzeba niezwłocznych regulacji prawnych systemów AI, sektor po sektorze, ze szczególną uwagą zwróconą na technologie rozpoznawania twarzy i emocji, oraz rygorystycznych badań na potrzeby wsparcia tych regulacji. Autorzy stawiają problem jednoznacznie: pytanie nie brzmi „czy w systemach AI są obciążenia (*biases*) i czy są one krzywdzące" – dyskusja jest zamknięta, bowiem w wyniku wydarzeń roku 2018 nie ma żadnych wątpliwości, że tak właśnie się dzieje. Następnym zadaniem jest zaadresowanie tych krzywd, co jest szczególnie pilne z uwagi na skalę działania systemów AI, ich funkcjonowanie centralizujące wiedzę i władzę w rękach nielicznych, oraz powiększającą się nierówność dystrybucji kosztów i benefitów, która tej centralizacji towarzyszy (Crawford i in. 2018:42). Podkreślmy, że słowo *bias* jest powszechnie używanym określeniem o charakterze technicznym, także w polskim świecie społecznym DS, a rozmaite rozwiązania mające adresować problem takich obciążeń w systemach uczenia maszynowego proponowali zarówno naukowcy, aktywiści, jak i są one oferowane jako usługa komercyjna – firmę audytującą algorytmy prowadzi np. wielokrotnie przez nas wspominana C. O'Neil (por. Courtland 2018).

Wraz ze wzrostem wszechobecności, złożoności i skali systemów AI brak znaczącej odpowiedzialności / rozliczalności (*accountability*) i nadzoru, na który ma się składać podstawowe zabezpieczenie odpowiedzialności, odpowiedzialność prawna i właściwy proces, jest zdaniem autorów coraz bardziej palącym problemem. Wskazują oni pięć obszarów problemowych, zaznaczając, że dotyczą one głównie USA, ale przypominają zarazem, że to w USA ulokowane są największe firmy tworzące systemy AI o zastosowaniu globalnym (Crawford i in. 2018:7). Przedstawiamy te obszary, a następnie omawiamy je kolejno:

1. rosnąca przepaść odpowiedzialności (*accountability gap*) w AI, gdzie faworyzowani są ludzie tworzący i wdrażający te technologie, a obciążani ci najbardziej dotykani ich skutkami;
2. używanie AI do zwiększania i wzmacniania inwigilacji, szczególnie w połączeniu z maszynowym rozpoznawaniem twarzy i emocji powiększa możliwości scentralizowanej kontroli i opresji;
3. zwiększanie rządowego użycia automatycznych systemów decyzyjnych, które bezpośrednio wpływają na jednostki ludzkie i społeczności, przy braku ustalonych struktur odpowiedzialności / rozliczalności (*accountability*);
4. nieregulowane i niemonitorowane eksperymenty z użyciem AI na populacjach ludzkich;

5. ograniczenia rozwiązań technologicznych w problemach dotyczących uczciwości, obciążeń (*bias*) i dyskryminacji ludzi przez systemy oparte na AI.

2.2.3.1. Twórcy a użytkownicy

Autorzy raportu AI Now Institute uważają, że rośnie przepaść odpowiedzialności między twórcami czy wdrażającymi systemy bazujące na AI, a ich użytkownikami. Jednocześnie starania o odpowiedzialność / rozliczalność (*accountability*) sytemów AI nie oznaczają po prostu szukania winnego za uczynioną szkodę (Boyd 2017). Autorka wskazuje, że przede wszystkim chodzi o odpowiedzialne czy rozliczalne działanie systemów AI, a zatem o sprawdzanie i ocenianie działania tych systemów już po ich wdrożeniu, z uwzględnieniem występowania efektów ubocznych. Boyd mówi wyraźnie – firmy takie, jak Google czy Facebook nie projektują swoich systemów AI tak, by były one dyskryminujące. Te systemy są dyskryminujące w sposób niezaplanowany i niekontrolowany. Problemem jest brak wychwytywania tego rodzaju skutków ubocznych przez firmy (ibidem). Zdaniem autorów omawianego raportu, podmioty tworzące i wdrażające systemy AI nie są zainteresowane badaniem tego, czy systemy te działają w sposób odpowiedzialny. Przepaść odpowiedzialności (*accountability gap*) jest zatem, ich zdaniem, spowodowana m.in. szybkością i tempem pracy oraz nastawieniem na innowacyjność i to zarówno w badaniach naukowych nad AI jak i w samych firmach wdrażających systemy komercyjnie. Nie ma tam miejsca na rozważania etyczne. Panuje kultura wyrażająca się w hasłach: „wdrażaj i poprawiaj później" (*launch and iterate*), oraz „działaj szybko i psuj" (*move fast and break things*) (Crawford i in. 2018:12). Wgląd w tę kulturę oraz w omawianą przepaść między osobami tworzącymi i wdrażającymi AI, a tymi ludźmi na których systemy AI mają największy wpływ, daje popularnonaukowa praca matematyczki Cathy O'Neil. Autorka była profesorem matematyki, odeszła z uczelni, by zatrudnić się jako analityk na Wall Street, a później pracowała jako data scientist:

> [na uczelni] dokuczało mi jednak ślimacze tempo prac. Chciałam być częścią postępującego szybkim krokiem prawdziwego świata. (...) Byłam dumna z tego, że idę do Shaw, znanego jako Harvard wśród funduszy hedgingowych[32], i będę mogła pokazać ludziom, że mój intelekt może przekładać się na pieniądze. W dodatku zarabiałam trzy razy tyle, ile na stanowisku profesora (O'Neil 2017a:64).

32 Fundusz hedgingowy (*hedge fund*) to fundusz inwestycyjny, stosujący różnorodne strategie inwestycyjne (m.in. zarówno krótki jak i długi horyzont czasowy) i instrumenty inwestycyjne, co ma zapewniać elastyczność i zyski w różnych warunkach koniunktury. Słowo hedge oznacza zabezpieczenie, jednak fundusze hedgingowe należą do ryzykownych. Nazywane są alternatywnymi instrumentami finansowymi, w odróżnieniu od klasycznych. „W środowisku finansowym uważa się, że m.in. działalność funduszy hedgingowych przyczyniła się do powstania kryzysu finansowego, dlatego same fundusze mają swoich zwolenników i przeciwników" (Gierałowska 2013).

Autorka ta sugeruje ponadto, że środowisko osób zajmujących się DS postrzega siebie jako elitę intelektualną i finansową, a jednocześnie zwraca uwagę raczej na wymierne wskaźniki ilościowej produktywności, nie myśląc kategoriami odpowiedzialności czy etyki systemów, nad którymi pracuje:

> Dostrzegałam cały szereg podobieństw między światem finansów a światem Big Data. Obydwie gałęzie ekonomii czerpią z tego samego zasobu talentów, z których większość wywodzi się z (...) MIT, Princeton i Stanford. Ci nowo zatrudnieni są żądni sukcesu i koncentrują się na zewnętrznych wskaźnikach takich jak punktacja SAT [odpowiednik matury w USA] czy testy wstępne na uczelnię. To stanowi całe ich życie. Niezależnie, czy trafią do świata finansów, czy technologii, mówi im się to samo: będą bogaci i będą rządzić światem. Ich produktywność stanowi dowód na to, że podążają we właściwym kierunku, a to przekłada się na zarobione dolary. Prowadzi to do mylnego wniosku, że cokolwiek zrobią, by zarobić jeszcze więcej pieniędzy, jest dobre. (...) W obydwu biznesowych kulturach majątek przestał już być środkiem do życia. Stał się natomiast bezpośrednio powiązany z osobistą wartością danego człowieka. (...) Obydwie te dziedziny są dość odległe od rzeczywistego świata z całym jego bałaganem. Widoczne jest dążenie do zastępowania ludzi śladami zostawianych przez nich danych, tak aby osiągnąć określone cele i przekształcić ich w skuteczniejszych kupujących, wyborców czy pracowników. Łatwo to zrobić i usprawiedliwić w sytuacji, gdy sukces pojawia się w postaci anonimowej punktacji, a ludzie, których te działania dotykają, pozostają tak samo abstrakcyjni jak liczby tańczące na ekranie (O'Neil 2017a:80).

Wypowiedzi O'Neil traktujemy jako wspomnienia i opinie osoby ze środowiska, zatem materiał do naszych badań etnograficznych, nie publikację naukową. Dodajmy, że autorka ta pozostaje aktywna w nagłaśnianiu konsekwencji społecznych big data / sztucznej inteligencji, i wyraziła się krytycznie na temat bezczynności naukowców w tej dziedzinie (O'Neil 2017b). Na zarzuty natychmiast odpowiedziano, krytykując ignorancję autorki i wskazując na liczne prace w dziedzinie krytycznych studiów danych (*critical data studies*), czy studiów nad nauką i techniką (*Science and Technology Studies,* STS) w ogóle (Vinsel 2017).

Zauważmy, że nastawienie na szybkość wdrożenia, przy braku obaw o konsekwencje porażki, nie jest charakterystyczne dla DS, a dla liderów branży nowoczesnych technologii informatycznych w ogóle. Za przykład może służyć książka autorstwa prezesa wykonawczego Google'a i doradcy współzałożyciela tej firmy Larry'ego Page'a:

> Proces rozwoju stał się szybszy i bardziej elastyczny, a diametralnie lepsze produkty nie powstają już na podstawie poprzednich hitów, tylko w wyniku dziesiątek iteracji. Podstawą sukcesu i jedynym sposobem na utrzymanie najwyższej jakości produktów jest więc szybkie działanie (Schmidt, Rosenberg, i Eagle 2014:39).
> (...) zarządzanie produktami w Dolinie [Krzemowej] przypomina lot F-16 z prędkością Mach 2 nad ziemią, nad terenem pokrytym wielkimi głazami. Do tego, **jeśli się rozbijecie, to konsekwencje są mniej więcej takie jak w grze telewizyjnej, a nas stać na mnóstwo żetonów. Super!** [podkr. RŻ] Najlepsze branże to właśnie te, w których pędzi się F-16 starając się nie rozbić, ale mając świadomość, że w razie czego możecie po prostu wyciągnąć z kieszeni kolejny żeton (Schmidt i in. 2014:209–10).

Analiza dyskursu na podstawie wypowiedzi publicznych CEO Facebooka Mark Zuckerberga z lat 2004-2014 wskazała, że szybkie wprowadzanie zmian i oczekiwanie na informację zwrotną od użytkowników, a następnie dalsze iteracyjne zmiany stanowiły modus operandi działania Facebooka (Hoffmann, Proferes, i Zimmer 2018:211–12). Niemniej Zuckerberg już w 2014 r. oświadczył, że „działaj szybko i psuj" (*move fast and break things*) nie jest już mottem pracy w Facebooku, i że odchodzi on od swoich hakerskich początków. Firmie miało przyświecać bardziej dojrzałe hasło stabilności rozwiązań (Murphy 2014; Statt 2014). Instytut AI Now wskazuje jednak na szereg nieetycznych działań Facebooka w roku 2018 (Crawford i in. 2018:10), z tzw. skandalem Cambridge Analytica na czele, o czym później. Problem nastawienia na szybkość i iteracyjność pracy oraz podejścia do ludzi, których dotyczą analizowane przez data scientistów dane i którzy są użytkownikami wypracowywanych przez nich rozwiązań będziemy omawiać szczegółowo w dalszej części rozprawy, opisując wartości świata społecznego DS w świetle badań własnych. Zauważmy, że w naukach prawniczych już pod koniec lat 90. dostrzegano problem naruszania prywatności i pozbawiania ludzi indywidualności w wyniku działania systemów bazujących na analizie danych, choć nie używano terminów „AI" ani „data science" (Vedder 1999). Gdzie indziej stwierdzano, że ponieważ analitycy pracują z danymi, to definiują się jako mający styczność z matematyką, nie z ludźmi (Metcalf i Crawford 2016:3).

Nie sugerujemy, jak zdają się to czynić autorzy z AI Now Institute (Crawford i in. 2018:12), że firmowa kultura nastawienia na szybkość i innowacyjność technologiczną jest tożsama z całkowitym brakiem zainteresowania kwestiami etycznymi wśród ich pracowników, czy w ogóle wśród data scientistów[33]. W badanym przez nas świecie społecznym DS pojawiają się wypowiedzi takie, jak choćby Piotra Migdała, jednego z bardziej znanych w Polsce data scientistów, aktywnego członka społeczności – m.in. założyciela facebookowego fanpage'u Data Science PL[34] i współorganizatora konferencji PyData Warsaw:

> [pytanie - Bagiński:] (...) ta praca [data scientista] wydaje się dość przydatna społecznie. Faktycznie można zmienić kawałek świata na lepsze? [odpowiedź - Migdał:] Jest wiele rzeczy, których kilka lat temu albo nie dałoby się zrobić, albo wysiłek byłby zbyt duży. Część tej pracy widać – gdy maszyna poprawnie zanalizuje zdjęcie albo podpowie komuś produkt do kupienia. Oczywiście są też sprawy kontrowersyjne, choćby analiza danych z Facebooka. Są przecież algorytmy pokazujące treści w taki sposób, by użytkownik stawał się od nich uzależniony.

33 Później wskażemy, że data scientiści pracują nie tylko w tzw. firmach technologicznych. Poza tym forma współpracy bywa inna niż zatrudnienie na etacie, więc nie zawsze można mówić o pracowniku. Pokażemy także, że uczestnicy społecznego świata DS w Polsce to nie tylko osoby zajmujące się data science zarobkowo.

34 https://www.facebook.com/groups/datasciencepl/, dostęp 22.04.2019 r.

Takimi rzeczami nie chciałbym się zajmować, uważam to za nieetyczne (Bagiński 2018).

Zdaniem informatyka i artysty Jonathana Harrisa żyjemy w czasach, w których grupa ludzi, licząca niewiele ponad kilkaset osób, zamieszkująca głównie takie miasta jak San Francisco i Nowy Jork, składająca się głównie z mężczyzn w wieku 22-35 lat ma ogromny wpływ na całą resztę gatunku ludzkiego. Dawniej życie tak ogromnych grup ludzi mogła zmienić wojna, głód, zaraza czy religia. Dziś dzieje się to po cichu, za pomocą oprogramowania używanego codziennie przez owych ludzi. Oprogramowanie można, jego zdaniem, postrzegać jak nowy rodzaj leków. Jednak inaczej niż lekarstwa, działa ono na wzorce zachowania całego społeczeństwa, nie na pojedyncze ciało (Harris 2012).

Myśl ta jest bliska Bernardowi Stieglerowi, który mówi o farmakonie. Ten orędownik humanistyki cyfrowej i uczeń J. Derridy twierdzi, że:

> [4.1.] (…) technologie cyfrowe są współczesnym wcieleniem tego, co starożytni Grecy określali mianem *hypomnémata*, tj. mnemotechnik. Albowiem owe mnemotechniki wciąż stanowią, zgodnie z terminologią platońską, farmaka, tzn. że są *jednocześnie trucizną i lekarstwem*. [4.2.] Stawiamy ogólniejszą hipotezę, że 1) każda technika jest „farmakologiczna" jako czynnik, który może generować zarówno złe jak i pozytywne skutki; 2) w sytuacji kiedy brakuje jasno zdefiniowanej postawy „terapeutycznej", co Grecy określali terminami *mélétè* i *epimeleia* (dyscyplina, troska, opieka), która wymaga technicznego podejścia do siebie samego, *farmakon* staje się nieuchronnie toksyczny [podkr. org.] (Stiegler 2010).

Harris tym samym uznaje, że inżynierowie oprogramowania (*software engineers*) to raczej inżynierowie społeczni (*social engineers*), jednak niewielu z nich zdaje sobie z tego sprawę, a jeszcze mniej przejmuje się etycznymi następstwami takiego rodzaju władzy. Nasze badania wskazują jednoznacznie, że data scientiści i inżynierowie oprogramowania (czy tzw. programiści) to przecinające się, ale zdecydowanie różne światy społeczne. Według autora w Facebooku pyta się projektantów „gdzie tu jest serotonina?", co oznacza projektowanie oprogramowania działającego nawet nie jak lek, a jak narkotyk – użytkownik ma czuć ciągłą potrzebę wracania po więcej. Technologia staje się wirusowo popularna (*viral*), kiedy rozszerza i wzmacnia coś, co już jest w ludziach, ale było zablokowane. Facebook zyskał 500 mln użytkowników w niecałe 5 lat, ponieważ doprowadził do usunięcia blokady w potrzebie bycia połączonym i dzielenia się (*our need to share and connect*). Zdaniem Harrisa znaczna część oprogramowania projektowana jest tak, by była uzależniająca. W Dolinie Krzemowej takie właściwości uzależniające uznawane są za zaletę (*asset*). Harris (2012) wyróżnił dwa typy firm działających w internecie: Plac targowy (*marketplace*) i Ekonomia uwagi (*attention economies*). Kontynuując metaforę leków Harris uznaje pierwsze – place targowe – za rzeczywiście uzdrawiające: takie firmy łączą jedną grupę ludzi z drugą, umożliwiając

rozwiązanie jakiegoś problemu (np. witryny aukcyjne, randkowe, noclegowe). Tu liczy się, by użytkownik zrealizował swoją potrzebę jak najszybciej. Te drugie – ekonomie uwagi – przypominają narkotyki, kryjąc się za hasłami komunikacji czy łączenia ludzi. Zatem „(...) uwaga stała się walutą. Dane także. Jedno i drugie można było łatwo zmonetyzować, kiedy reklamy i sklepy połączyło jedno proste kliknięcie" (Vaidhyanathan 2018:335). Według Harrisa w rzeczywistości twórcom zależy nie na ułatwieniu komunikacji, ale by użytkownik spędził w serwisie jak najwięcej czasu, doświadczając proponowanych treści i co najważniejsze, reklam. W internecie ludzie nauczyli się nie nadawać rzeczom bezpośredniej wartości, bo za zdecydowaną większość usług nie płacą. Tym samym najpopularniejszym modelem biznesowym jest dystrybucja stworzonego produktu za darmo, osiągnięcie dużej bazy użytkowników i zamiana tych użytkowników na produkt - sprzedanie ich uwagi reklamodawcom i ich danych marketingowcom. Autor twierdzi, że jest to groźne nie w sensie ekonomicznym, a w ludzkim. Użytkownicy tracą w oczach firm swoją indywidualność, liczą się wyłącznie jako agregat towarów niskiej jakości (Harris 2012:200–202). To samo napisał Jerzy Surma, zaczynając od pytania „dlaczego dla Google poczucie bezpieczeństwa jego użytkowników jest tak istotne?". W roku 2015 około 70% zysków tej firmy - ponad 52 mln dolarów - pochodziło z reklam zamieszczonych na stronie wyszukiwarki i na YouTube (który rzecz jasna należy do Google). Zatem:

> (…) użytkownik wyszukiwarki jest produktem, który Google oferuje swoim klientom płacącym za reklamy. Z tego względu Google bardzo dba o swoje „produkty" [tzn. użytkowników], oferując im całe portfolio różnego rodzaju darmowych serwisów, w tym najpopularniejszą stronę WWW w całym Internecie, jaką jest wysokiej jakości wyszukiwarka (Surma 2017:66–67).

Fundacja Panoptykon przygotowała infografikę, obrazującą proces od wejścia użytkownika na stronę internetową do wyświetlenia mu spersonalizowanej reklamy. Użytkownik jest towarem, którego uwagę mogą sprzedać dostawcy treści, a dla reklamodawcy jest potencjalnym klientem, którego uwagę warto kupić (Zespół Fundacji Panoptykon 2017).

Autorzy omawianego raportu AI Now twierdzą, że powodami pogłębiania się przepaści odpowiedzialności są (obok kultury i dodajmy modeli biznesowych firm technologicznych): asymetria władzy pomiędzy firmami a użytkownikami oraz brak regulacji prawnych i wewnętrznych regulacji w firmach (Crawford i in. 2018:7). W świetle pracy ekonomistów Brynjolfssona i McAffe można pracowników firm technologicznych, a szczególnie data scientistów, określić jako wygranych w wyścigu z maszynami, co wspiera tezę o asymetrii

władzy. Są to pracownicy wysoko wykwalifikowani, a dodatkowo elitarni, pożądani, posiadający status określany przez autorów mianem „supergwiazd". Autorzy wskazują na pogłębiające się różnice w poziomie zarobków między najbardziej a najmniej zamożnymi, przy ogólnym wzroście średniej produktywności i zamożności, bowiem „zyski zwycięzców mogą być wyższe niż straty przegranych, ale to małe pocieszenie dla tych, którzy wychodzą na postępie jak Zabłocki na mydle" (Brynjolfsson i McAfee 2015:50–51). W tej samej pracy wskazano, że prawo, działanie instytucji i tzw. mentalność większości społeczeństwa nie nadąża za szybkim tempem rozwoju technologii – nadążają wyłącznie wygrani (ibidem 16-17; 63-64).

Podsumowując wątek zwiększania się przepaści odpowiedzialności między twórcami a odbiorcami, użytkownikami czy podmiotami działania systemów, bazujących na AI stwierdzamy, że choć niewątpliwie istnieje między nimi szereg różnic i kolosalnych nierówności kulturowych, materialnych, oraz dotyczących wiedzy i władzy, a brak odpowiedzialności za krzywdę wyrządzoną komukolwiek uznajemy bez najmniejszego wątpienia za poważny problem społeczny, to nie możemy jednoznacznie zgodzić się z tezą o „zwiększaniu się przepaści". Nie widzimy podstaw do tak uproszczonego postrzegania zachodzących procesów i zwracamy uwagę, że np. w świetle przemian prywatności oraz procesu zwanego danetyzacją – o czym poniżej – opozycja „złe firmy" versus „pokrzywdzeni użytkownicy" nie ma sensu.

Asymetria władzy widoczna jest także w obszarze niekoniecznie dotyczącym użytkowników, a wyłącznie pracowników czy quasipracowników firm technologicznych. Jak się okazuje do tych (quasi) pracowników zaliczają się nie tylko inżynierskie, naukowe, managerskie czy inne wysoko wykwalifikowane stanowiska, ale także osoby wykonujące zadania rutynowe, powtarzalne niczym proste prace manualne na taśmie produkcyjnej, choć odbywają się w środowisku cyfrowym. Mowa o dwóch różnych kategoriach relatywnie niskopłatnych i nie wymagających wysokich kwalifikacji zajęć: umożliwiających działanie AI, oraz wspierających / zastępujących deklarowane działanie tejże. Pierwsza kategoria to przede wszystkim zadania ręcznego etykietowania danych. Powtórzmy, że algorytmy nadzorowanego uczenia maszynowego (*supervised learning*) ogólnie działają na zasadzie przybliżania funkcji, łączącej zmienne niezależne (*features*) ze zmienną zależną (*target*). Żeby nauczyć model i ocenić poprawność jego działania, wartości zmiennej zależnej muszą być znane. Podręcznikowymi przykładami najprostszych algorytmów uczenia nadzorowanego jest regresja liniowa i logistyczna. Łatwiej o etykiety, czy ogólnie o znaną wartość zmiennej

celu przypadku takich danych, jak dane o transakcjach kartą płatniczą, a trudniej gdy chodzi np. o fotografie. W związku z tym zdarza się, że nadanie etykiety milionom zdjęć jest zlecane poza firmę technologiczną np. do Azji Południowo-Wschodniej, gdzie „klikacze" (*clickworkers*) oznaczają je, by można było później zasilić algorytm (Crawford i in. 2018:33–34). Przykład podobnego pozyskania danych dzięki „klikaczom" za pomocą serwisu Amazona „Mechanical Turk" przedstawimy w części 2.2.3.4., omawiając tzw. aferę Cambridge Analytica[35]. W drugiej kategorii mówimy o bardziej złożonych i wyżej wynagradzanych zadaniach, z którymi, kolokwialnie mówiąc, AI sobie nie radzi. Stąd bierze się wyczerpująca psychicznie praca moderatorów treści (*content moderator*), np. w należącym do Google serwisie wideo YouTube czy na Facebooku, którzy zapewniają rodzaj dodatkowego filtru nałożonego obok lub po AI czy innych systemów informatycznych na filmy/zdjęcia/teksty napływające od użytkowników (Chemaly i Buni 2016; Newton 2019). Stąd biorą się sytuacje, w których „człowiek udaje robota udającego człowieka (*humans pretending to be robots pretending to be humans*)". Sprzedawany jest jakoby bazujący na AI system, chatbot, inteligentny robot-asystent w postaci aplikacji na urządzenia mobilne, dzięki której szybko i wygodnie zorganizujemy podróż, randkę czy przyjęcie. Faktycznie „niemal wszystkie (*almost every*)", jakoby magicznie wykonywane przez AI zgłoszenia, obsługują pracujący 24 godziny na dobę przez 365 dni w roku ludzie. Niewidoczni, ukryci ludzie, pracujący w imieniu cyfrowego konsjerża, sekretarki i pomocnika, w opisanym startupie X.ai byli zatrudnieni na stanowiskach trenerów sztucznej inteligencji (*AI trainer*) (Huet 2016).

2.2.3.2. Inwigilacja, kontrola, opresja?

Co do drugiego obszaru problemowego – zwiększania i wzmacniania inwigilacji z użyciem AI, co prowadzi do zwiększania scentralizowanej kontroli i opresji – to autorzy raportu zwracają szczególną uwagę na techniki umożliwiające jakoby automatyczne rozpoznawanie emocji na podstawie obrazu twarzy. Wskazano, że technologie te mają pozwalać też na rozpoznawanie nastroju (*mood*), kondycji zdrowia psychicznego, poziomu zaangażowania, winy bądź niewinności (*guilt or innocence*). Powtórzmy, że zdaniem autorów takie systemy bazują na dawno zdyskredytowanej pseudonauce. Zaznaczają oni, że takie technologie są obecnie w użyciu bez wiedzy osób, wobec których są stosowane. Wskazują, że używanie takich systemów w procesie rekrutacji i zatrudnieniu, edukacji czy wymiarze

35 Dodajmy, że trudno odróżnić „klikacza" od użytkownika serwisu zbierającego dane, co zwane jest zamianą klienta w produkt. Pracę „klikacza" użytkownik może wykonywać nieświadomie, rozwiązując CAPTCHA. Por. https://www.kdnuggets.com/2017/06/acquiring-quality-labeled-training-data.html, dostęp 05.02.2019 r.

sprawiedliwości stanowi poważne zagrożenie dla praw człowieka i wolności obywatelskich (Crawford i in. 2018:11; 13–17). Zauważmy, że o zagrożeniach związanych z maszynowym rozpoznawaniem twarzy wypowiadano się w ten sposób znacznie wcześniej (Introna i Wood 2002). Inwigilacja z użyciem AI jest w raporcie określana jako niebezpieczna z trzech powodów. Po pierwsze, działania są automatyzowane, co umożliwia inwigilację przekraczającą możliwości ludzkiej kontroli. Po drugie, jest skalowalna, co oznacza, że może łatwo zwiększać skalę działania – przetwarzać i analizować coraz większe ilości danych znajdując powiązania trudne do wykrycia innymi metodami. Po trzecie, metody AI pozwalają na prognozowanie, co używane jest do kontrolowania zachowań jednostek ludzkich i grup (Crawford i in. 2018:12). Autorzy podają szereg przykładów. W Chinach system rozpoznawania twarzy jest używany do kontroli na granicy Hong Kongu i Shenzhen, w pięciu prowincjach działa monitorowanie mieszkańców z użyciem dronów, w będącym pod wyjątkowo natężoną kontrolą autonomicznym regionie Xinjiang[36], metody uczenia maszynowego są stosowane do wskazywania osób niepoprawnych politycznie, które wysyłane są do „obozów reedukacyjnych" (*re-education camps*), zaś w całym kraju od 2013 działa mający do 2020 objąć wszystkich obywateli Chin system punktacji społecznej (*social credit*) (Crawford i in. 2018:13). System ten wzbudził uwagę zachodnich mediów. Wzorowany jest na bankowych systemach oceny zdolności kredytowej, jednak ma oceniać solidność (*trustworthiness*) obywatela, przyznając każdemu konkretną liczbę punktów. Brane pod uwagę są m.in. osiągnięcia szkolne, zakupy, historia zatrudnienia, konflikty z prawem, kontakt z obcokrajowcami czy prywatne finanse (Associated Press 2015; Botsman 2017; Creemers 2015; Denyer 2016; Gostkiewicz 2017; Kędzierski 2018; Nguyen 2016; Wong i Chin 2016). System budowany jest we współpracy z chińskimi firmami, m.in. gigantyczną platformą handlu internetowego Alibaba. Nie można powiedzieć, żeby system był jednoznacznie odrzucany przez obywateli Chin – np. największy, gromadzący w 2015 r. 90 milionów osób chiński portal randkowy Baihe umożliwił użytkownikom prezentowanie liczby swoich punktów na profilu, z czego większość skorzystała (Hatton 2015). W tym samym artykule zaznaczono, że odpowiedzialna za algorytm obliczający tę „solidność" obywatela firma odmówiła rozmowy z BBC oświadczając, że algorytm jest złożony, a przekazanie szczegółów zachodnim mediom byłoby złamaniem umowy z rządem (ibidem). O podobnych praktykach tajności i nieprzejrzystość powiemy więcej poniżej, wprowadzając termin „broń matematycznej zagłady". Wracając do przykładów z raportu, w Wenezueli

36 Graniczy m.in. z Tybetem, a część terytorium jest przedmiotem sporu z Indiami.

wprowadzono inteligentny dowód tożsamości, który pozwala na integrację i przeglądanie przez władze danych konkretnego obywatela o jego prywatnych finansach, historii medycznej i głosowaniu w wyborach. W USA rządowa agencja ds. imigracji i ceł (*Immigration and Customs Enforcement,* ICE) we współpracy z firmą Amazon zbudowała system integrujący dane o przybywających do kraju imigrantach. Łączone są dane z systemów publicznych z danymi kupowanymi od firm prywatnych za pośrednictwem brokerów danych, a uczenie maszynowe służy do kategoryzowania imigrantów do dalszych ścieżek postępowania (w tym deportacji). Również we współpracy z firmą Amazon policja w miastach Orlando i Waszyngton korzystała z bazujących na AI systemach rozpoznawania twarzy o nazwie Rekognition, wyszukując podejrzane osoby na zdjęciach tłumu. Pracownicy Amazona protestowali przeciwko tego rodzaju współpracy z policją i ICE, jednak firma oficjalnie zapewniła o tym, że oprogramowanie Rekognition jest wartościowe, jedynie może obciążać wyniki, gdy baza danych nie jest odpowiednio reprezentatywna. Do problemu reprezentatywności, czyli obciążenia (*bias*) zbiorów danych, na których uczone są modele będziemy powracać wielokrotnie. W Nowym Jorku policja korzystała z podobnego rozwiązania firmy IBM. Badania wykazały niedokładność i obciążenie (*bias*) działania tych systemów – system Amazona wskazał 28 członków Kongresu USA jako osoby z bazy 25 tys. osób aresztowanych, co jest niedorzecznością, przy czym błędnie rozpoznawanych było około 5% zdjęć białych członków Kongresu, a 40% czarnoskórych. W innym badaniu podobnych systemów AI stwierdzono bardziej precyzyjne działanie rozpoznawania twarzy dla zdjęć osób o jasnej skórze niż ciemnej, a także dla mężczyzn niż kobiet, co przypisuje się właśnie charakterystyce używanych do uczenia zbiorów danych. Autorzy raportu dodają, że podobnych systemów inwigilacji może być znacznie więcej, ponieważ informacje o nich rzadko podawane są publicznie (Crawford i in. 2018:12–13; 15–16). Znanym przykładem omawianego obciążenia systemu rozpoznawania obrazów jest działanie aplikacji Google Photos, która umożliwia automatyczne etykietowanie zdjęć dzięki AI. Dwie czarnoskóre osoby zostały oznaczone na fotografiach jako goryle, za co firma przepraszała (Grush 2015). W raporcie podano także przykłady zastosowania i prac nad rozpoznawaniem twarzy na uczelniach. Na Uniwersytecie Św. Tomasza w Minnesocie studenci są podczas zajęć obserwowani przez kamerę bazującą na systemie AI firmy Microsoft, który ma w czasie rzeczywistym oceniać ich stan emocjonalny, pomagając nauczycielowi na bieżąco w prowadzeniu zajęć (Crawford i in. 2018:15). Na Uniwersytecie Stanforda powstał zaś model z użyciem sieci neuronowej, która ma rozpoznawać orientację seksualną osoby na podstawie

fotografii twarzy. Choć Wang i Kosinski, autorzy tego modelu wskazują, że ich odkrycia demaskują zagrożenia prywatności i bezpieczeństwa osób homoseksualnych, to spotkały się z krytyką amerykańskiego środowiska naukowego (Crawford i in. 2018:17; Wang i Kosinski 2018).

Powyższe przykłady wskazują zarówno na zbieranie danych, ich analizę, jak i wdrażanie wyników. Jednak bez wątpienia etap zbierania danych jest konieczny do zaistnienia kolejnych – bez danych nie ma analizy danych. W entuzjastycznej pracy *Big data. Rewolucja…* ukuto termin danetyzacja (*datafication*), który oznacza zamianę w dane różnych rodzajów ludzkiej aktywności (Cukier i Mayer-Schönberger 2014). Biznes i rządy korzystając z danych i metadanych serwisów, np. wyszukiwarki Google, portalu Facebook, Twitter, LinkedIn, Tumblra, iTunes, komunikatora Skype, WhatsApp, serwisu YouTube, i darmowych platform poczty elektronicznej jak Gmail czy Hotmail mają dostęp do zapisu pewnego wycinka ludzkich zachowań, które jeszcze 20 lat temu były rejestrowane w bardzo ograniczonej skali, bądź wcale.

> Dane i metadane, nie tak dawno traktowane jako bezwartościowy produkt uboczny serwisów internetowych, zostały stopniowo zmienione w cenny zasób, który można wydobywać, wzbogacać i wykorzystywać ponownie w dochodowych produktach (Dijck van 2014:199).

Danetyzacja jest zatem ogólnym „dążeniem do zawężania obszarów, które nie podlegają ewidencji" (Iwasiński 2016:137). Łukasz Iwasiński przywołuje w tym kontekście słowa Lva Manovicha (2012a:335) który uważa, że efektem danetyzacji będzie zamiana świata w jedną wielką bazę danych. Entuzjaści piszą w tym kontekście o „świadomości big data", a więc „przekonaniu, że istnieje mierzalny komponent wszystkiego" i że da się „przekształcić niezliczone wymiary rzeczywistości w dane" (Cukier i Mayer-Schönberger 2014:132). Te zdania można traktować jako wyraz wspomnianego wcześniej dataizmu – wiary i zaufania w dane. Zdaniem badaczki komunikacji i nowych mediów José van Dijck, danetyzacja jest przejawem ideologii dataizmu. Wyznając ją uznaje się, że dane są najważniejszym elementem dla zrozumienia rzeczywistości (Dijck 2014). Zwracamy uwagę na widoczną w powyższym cytacie kwestię ponownego wykorzystania danych. W myśl zasady „zachowaj wszystko — przeanalizujesz, co zechcesz" (Iwasiński 2017:122), gromadzona jest jak największa ilość danych, nawet gdy nie ma obecnie możliwości czy pomysłu na to, co z nimi zrobić. Dane wykorzystywane powtórnie, zapisywane w celu wykonania jakiejś innej czynności bądź przy okazji jej wykonywania mogą być chaotyczne i niekompletne. Tak zwane „dane resztkowe" (Cukier i Mayer- Schönberger 2014:151), które mają być charakterystyczne dla big data to np.

wpisy w wyszukiwarce Google, z których już skorzystaliśmy dzięki Google Trends (por. część 1.1.3.). Powstaje pytanie, na ile inwigilacją jest zapisywanie czy korzystanie z takich resztkowych danych do celów innych niż dostarczenie usługi. Wskazaliśmy, że ludzie sami chętnie generują dane o sobie, oraz na swoje potrzeby. Nie ma mowy o jakimkolwiek przymusie, może co najwyżej niewielka jest świadomość użytkowników, co dzieje się z danymi o nich. Powołaliśmy się na ruch zwany *quantified self* (skwantyfikowany ja), polegający na dążeniu do nieustannego monitorowania parametrów swojego ciała, pracy czy właściwie dowolnych aktywności celem ich usprawniania (Żulicki 2017:184). Odbywa się to za pomocą łatwo dostępnych gadżetów, jak smartwatch lub opaska fitness sparowana ze smartfonem. Dziś jesteśmy przekonani, że problem ten ma dwa wymiary: zachodzą przemiany w postrzeganiu swojej prywatności przez jednostki ludzkie – czyli ludzie chcą generować dane, i ma miejsce intencjonalna inwigilacja ludzi przez firmy w celu zarabiania pieniędzy na wygenerowanych przez nich danych. Jak podaliśmy, m.in. za Surmą (2017), użytkownik usług Google czy portalu Facebook w rzeczywistości nie jest ich klientem, choć firmom tym przeświecają jakoby nastawione na użytkownika misje „uporządkowania światowych zasobów informacji tak, by stały się powszechnie dostępne i użyteczne dla każdego" (Google 2019b) i „wspierania ludzi w budowaniu wspólnoty i łączeniu świata ze sobą" (Facebook 2019). W ich modelu biznesowym użytkownik nie płaci za usługi, bo to on jest produktem, który te firmy sprzedają reklamodawcom i inny, rozmaitym podmiotom zainteresowanym uwagą użytkowników czy danymi o nich (Surma 2017:66–67; 70). Przy tym nie ma mowy o działaniu wbrew użytkownikom, co prowadzi do konstatacji, że ideologia danetyzacji jest, jak stwierdziła van Dijck, oparta na powszechnych normach społecznych. Użytkownicy zapewniają firmom dostęp do danych o sobie w zamian za usługi – jest to forma barteru. Taki model biznesowy w usługach komunikacji elektronicznej van Dijck uznaje za normalny, wskazując na badania sondażowe z których wynika, że mniej niż 1/3 użytkowników byłaby skłonna płacić za owe usługi, gdyby firma je dostarczająca miała przestać korzystać z jego danych na potrzeby marketingowe. Tym samym ideologia danetyzacji jest najczęściej przyjmowana bezkrytycznie przez wszystkie zainteresowane strony (Dijck van 2014:200–201). Dziennikarka Julia Angwin we wspomnianym serialu dokumentalnym „Do Not Track" (od 6 min. 17 s., odcinek 2.) przedstawiła ten problem w kontekście historycznym, na przykładzie znanych każdemu korzystającemu ze stron www „ciasteczek", czyli plików cookies:

> To wszystko [bycie śledzonym w Internecie przez cookies] mogło wyglądać inaczej - to mogło

> nigdy nie zaistnieć, gdybyśmy umówili się np. na niewielką opłatę za korzystanie z emaili. Niestety w latach 2000 powstała niechęć użytkowników do płacenia za cokolwiek w Internecie. Więc teraz tak płacimy - za pomocą naszych danych (UPIAN 2015).

Jednocześnie od kilkunastu lat mówi się o społeczeństwie postprywatnym, które ma charakteryzować się transparentnością stosunków społecznych. Transparentność ta ma na celu zapewnienie ładu publicznego i ochrony osobistej, a uzyskana jest dzięki bezustannej inwigilacji (Dopierała 2013:229). Renata Dopierała stwierdza, że trwa proces w którym to, co prywatne staje się upubliczniane, zatem prywatność jest sukcesywnie redukowana. Składa się na to z jednej strony zwiększanie się otwarcia w kwestiach uważanych za prywatne, z drugiej strony zaś poprzez ingerencje technologii zmniejsza się strefa prywatności (ibidem). Mówi się także o postprywatności jako postawie leseferyzmu wobec cyfrowej inwigilacji. Postawa ta bazuje na przekonaniu, że prywatność w świece cyfrowym jest zarówno nieosiągalna, jak i nieopłacalna (*socially unrewarded*). Zdaniem Andersona i Burkarta postawa ta jest odpowiedzią na poczucie utraty kontroli w sferach autonomii, bycia widocznym (*self visibility*) oraz autoprezentacji (Burkart i Andersson Schwarz 2014:218). Giganci technologiczni legitymizowali swoje działania za pomocą uproszczonej wersji powyższej argumentacji. Założyciel i CEO Facebooka Mark Zuckerberg stwierdził publicznie, że prywatność nie jest już obowiązującą normą społeczną (Johnson 2010). Wskazuje się, że Zuckerberg używał retoryki zachęcającej w istocie do przekazywania Facebookowi jak największej ilości danych. Stosował określenia „dzielenie się" (*sharing*), „łączenie" (*connection*) i „otwartość" (*openness*) w odniesieniu do aktywności ludzi na Facebooku w połączeniu z zapewnieniami o tym, że użytkownicy mają kontrolę nad udostępnianymi treściami, co prowadzić ma do rekonceptualizacji prywatności (Hoffmann i in. 2018:201–2; 206). W świetle badań wskazujących na to, że wyższe zainteresowanie kontrolą swojej prywatności online jest związane z wyższą aktywnością użytkownika Facebooka (Jordaan i Van Heerden 2017), retoryka ta wydaje się nam skuteczna.

Całość zjawiska big data można postrzegać także jako „szereg intencjonalnych działań będących wyrazem nowej logiki akumulacji" – kapitalizmu inwigilacji (*surveillance capitalism*) (Zuboff 2015:75–76). Kapitalizm inwigilacji jest, zdaniem Zuboff, nastawiony na przewidywanie i modyfikowanie ludzkiego zachowania w celu uzyskania dochodu i kontroli rynku (ibidem). Mówi ona o danych jako aktywach inwigilacji (*surveillance assets*), zaznaczając, że można traktować je jako kontrabandę w sensie dobra będącego przedmiotem kradzieży, bowiem są one zabrane od kogoś (*taken*), a nie przez tego kogoś oddane (*given*),

ani nie wyprodukowane (ibidem:81). Autorka uważa, że uzyskanie stałej inwigilacji usuwa potrzebę istnienia zaufania. Zaufanie do drugiego człowieka w sytuacji np. zawarcia umowy nie ma sensu, gdy możliwe jest stałe monitorowanie tego, jak druga osoba wypełnia kontrakt i reagowanie na zachowanie tej osoby w czasie rzeczywistym. Zauważmy, że jest to tożsame z transparencją stosunków społecznych, która ma być charakterystyczna dla społeczeństwa postprywatnego omawianego wyżej. Byt, który ma zapewnić taki poziom transparencji, a zdaniem Zuboff inwigilacji, nazywa ona – w nawiązaniu zarówno do big data, Big Brothera jak i meadowskiego znaczącego innego (*siginificant other*) – Big Other i określa jako uniwersalną architekturę, gdzieś pomiędzy Bogiem a naturą. Ten „Wielki Inny" to „nowy reżim niezależnych i niezależnie konstruowanych faktów", który „zapisuje, modyfikuje i porządkuje codzienne doświadczenia (…), a wszystko po to, aby wytyczyć nowe ścieżki do monetyzacji i zdobywania profitów" (ibidem:81-82). W wielu publikacjach naukowych, medialnych czy popularnonaukowych poświęconych big data lub AI pojawiały się nawiązania do pojęć orwellowskiego Wielkiego Brata, panoptykonu Jeremy'ego Benthama i Michela Foucaulta czy Nowego Wspaniałego Świata Aldousa Huxleya (Biedrzycki 2017b, 2017a; Botsman 2017; Ćwiklak 2017; Diebold 2012; Dijck van 2014; Frankowski 2018; Gostkiewicz 2017; Iwasiński 2017; Kędzierski 2018; Kulesza 2018a, 2018b; Morzy 2014; Ożóg 2009; Zespół Fundacji Panoptykon 2016). Zuboff uważa, że te charakterystyczne dla XX w. metafory władzy totalitarnej nie przystają do „Wielkiego Innego". Kapitalizm inwigilacji ma spowodować sytuację, w której brak jest jakiegoś scentralizowanego ośrodka nadzoru czy władzy, przed którego oczyma można by próbować się schronić czy wobec którego można by przyjąć postawę konformisty albo opozycjonisty. Wielki Inny jest wszędzie, jest rozproszoną i w dużej mierze niekwestionowaną formą władzy, która przemienia wszystko w towar, rozdziela kary i nagrody za pomocą nowego rodzaju niewidzialnej ręki (Zuboff 2015:82). Na nieadekwatność totalitarnych metafor w odniesieniu do inwigilacji cyfrowej – choć w ich pracach nie było terminów AI, big data, DS – wskazywali już wcześniej autorzy polscy. Decentralizacja spojrzenia ma zaczynać się wraz z powstaniem Web 2.0, rozumianego jako prywatyzacja technologii umożliwiająca wzajemną obserwację wszystkich, tzn. każdy może być zarówno obserwującym jak i obserwowanym. W wyniku tego powstały heterogeniczne systemy inwigilacji rządowej i prywatnej nazywane asamblażami nadzoru (Dopierała 2013:230–31; Haggerty i V. Ericson 2000; Ożóg 2009:18–19), w wyniku proliferacji których „powstają zdelokalizowane, zdematerializowane, 'subtelnie' opresyjne, nieuświadamiane, ale zarazem wszechobecne, silniej oddziałujące i zniewalające mechanizmy sprawowania

władzy" (Dopierała 2013:231). Taki postpanoptyczny nadzór jest nieopresyjny (Iwasiński 2017:131), co ma sens także w odniesieniu do przywołanej wyżej do tezy van Djik o powszechnej akceptacji idelologii danetyzacji. Zauważmy także, że Zuboff mówi o „internecie wszystkiego", w którym cały świat jest połączony. Tego rodzaju wizję, w której swoje miejsce mają zarówno podłączone do internetu urządzenia gospodarstwa domowego jak i pływające w ludzkich naczyniach krwionośnych zbierające dane nanoroboty, a cała Ziemia staje się wielkim, inteligentnym, cyfrowym systemem nerwowym rysuje np. zarówno hiperentuzjastyczny Kurzweil (2016:23–24; 415), jak i ostrożny Harris (2012:202). Zuboff wskazuje także, że kapitalizm inwigilacji stanowi poważne zagrożenie dla demokracji. W tym modelu nie ma już wzajemności pomiędzy firmami a populacjami ludzi. W tradycyjnym kapitalizmie występowała wzajemność w sferach zatrudnienia i konsumpcji, a w kapitalizmie inwigilacji firmy mają traktować ludzi wyłącznie jako cel, którego dane chcą wykorzystać. Autorka nazywa to „radykalną rezygnacją z życia społecznego" i uważa, że kapitalizm inwigilacji jest tym samym antydemokratyczny. Dodatkowo jest antydemokratyczny także w tym sensie, że demokracja zagraża uzyskiwaniu dochodów z inwigilacji ludzi (Zuboff 2015:86). Tezy autorki zostały rozwinięte w jej ostatniej książce, gdzie wskazuje, że kapitalizm nie jest już – jak pisał Marks – wampirem żywiącym się pracą proletariuszy, a żywi się każdym aspektem ludzkiego doświadczenia (Zuboff 2019).

Podsumowując, problem zwiększania i wzmacniania inwigilacji dzięki używaniu dużych ilości danych i AI uznajemy z jednej strony za istotny i trafnie sformowany. Sądzimy tak, bowiem coraz większe możliwości techniczne (dokładność działania algorytmów, wydajność, skalowalność), coraz niższe koszty, a co za tym idzie dostępność tych technologii, połączone są z wiarą – a może nawet są wyrazem tej wiary – w „dobre" rozwiązywanie rozmaitych problemów firm, rządów, grup i jednostek ludzkich właśnie za pomocą takich technologii, co niewątpliwie prowadzi do zwiększania i wzmacniania „inwigilacji". Słowo to zapisujemy celowo w cudzysłowie, bowiem z drugiej strony mamy poważne wątpliwości wobec adekwatności jego użycia, i jesteśmy znacznie bardziej przekonani do określenia „danetyzacja". Skłaniamy się zatem ku opisowemu, nie-bezwzględnie-negatywnemu ujęciu zidentyfikowanych procesów jako zwiększającej się i wzmacniającej danetyzacji.

Przy tym teza o wzmacnianiu scentralizowanej kontroli jest według nas uzasadniona, ponieważ o ile kontrola ta stała się nieopresyjna, niewidoczna i milcząco akceptowana, to niewątpliwie nowymi ośrodkami władzy – już nie totalitarnej, a m.in. ekonomicznej, informacyjnej, oraz władzy nad uwagą – stają się quazi-monopole, czyli najpotężniejsze firmy

gospodarki cyfrowej. Określa się je jako: GAFA – Google, Apple, Facebook, Amazon (Rogers 2018:11) oraz szerzej GFGC – Google, Facebook, Amazon, Samsung, Apple, Huawei, Alibaba, Yandex (Surma 2017:62-63), a ich sukces oparty jest na efektach sieciowych, wykorzystujących dodatnie sprzężenie zwrotne. Bezpośredni efekt sieciowy polega na tym, że wartość przyłączenia się do sieci zależy od liczby już obecnych w sieci użytkowników. Tym samy firmy silne stają się jeszcze silniejsze (ibidem); im więcej znajomych osoby A jest np. na Twitterze, tym większe prawdopodobieństwo założenia tam konta przez A.

2.2.3.3. Automatyczne systemy decyzyjne

Trzeci obszar problemowy to zwiększanie rządowego użycia automatycznych systemów decyzyjnych, które bezpośrednio wpływają na jednostki i społeczności przy braku ustalonych struktur odpowiedzialności / rozliczalności (*accountability*). Autorzy raportu AI Now Institute wskazują na znaczny wzrost liczby automatycznych, bazujących na AI systemów w obszarach takich jak wymiar sprawiedliwość, edukacja, imigracja czy opieka społeczna dzieci (*child welfare*). System te, pod hasłami zwiększania efektywności i zmniejszania kosztów, wspomagają albo zastępują procesy decyzyjne. Osoby, których dotyczą decyzje przeważnie nie mają ani możliwości wglądu i zrozumienia procesu decyzyjnego, ani odwołania się od decyzji. Podano przykład publicznej opieki nad osobami niepełnosprawnymi w Arkansas, gdzie po wprowadzeniu w 2016 r. automatycznego systemu decyzyjnego określającego tygodniową liczbę godzin domowej opieki pielęgniarskiej setki osób z dnia na dzień i bez wyjaśnienia doświadczyły redukcji opieki, np. z 56 do 32 godzin (Crawford i in. 2018:18). Autorzy wskazują, że dzięki pozwowi przygotowanemu przez organizację pozarządową przeciw stanowi Arkansas ów system został uznany przez sąd za obarczony błędami (*erroneous*) i niekonstytucyjny (ibidem). Zdaniem autorów instytucje finansowane publicznie w USA są obecnie pod ogromną presją dotyczącą cięcia kosztów. Legitymizują swoje działania hasłami, o których mówiliśmy podnosząc wątek krytyki metodologiczno-epistemologicznej big data / DS – bezstronności, obiektywności i sprawiedliwości decyzji zapadających na podstawie wyników systemów bazujących na analizie danych. Tym samym instytucje przenoszą na systemy AI część odpowiedzialności za pozbawianie ludzi dostępu do usług publicznych. Systemy są zatem projektowane i oceniane głównie pod kątem tego, czy ich wdrożenie będzie skutkowało redukcją kosztów, tak w wyniku zmniejszenia kosztu procesu jak w wyniku zmniejszenia ilości udzielonych usług (Crawford i in. 2018:18–19).

Abstrahując od konkretnego problemu zjawisko takie opisał K. Krzysztofek. Jego zdaniem ludzie przekazują władzę zbierania danych i ich analizy technologiom z dwóch powodów. Ludzki mózg nie wystarcza do wykonania tak dużych i złożonych analiz, a co ważniejsze zawierzenie technologiom jest wygodne, zwalnia z wysiłku poznawczego, i przynajmniej częściowo z odpowiedzialności za skutek decyzji podjętych na podstawie uzyskanych wyników (Krzysztofek 2012). W innym tekście stwierdził on, że „w przeważającej większości współczesny człowiek nie rozumie technologii, jakimi się posługuje, są one dlań czarną skrzynką" (Krzysztofek 2015:11). Technologia analizy danych może być czarną skrzynką z dwóch powodów: po pierwsze, z uwagi na zastosowaną metodę[37]; po drugie, z powodu braku dostępu do informacji o zasadach działania algorytmu z uwagi na tajemnice firmowe czy handlowe. Zauważmy, że powszechnie używaną czarną skrzynką jest wyszukiwarka internetowa Google. Algorytm PageRank, decydujący o ważności wyników wyszukiwania czyli kolejności tych wyników, którą widzi użytkownik został w wersji wzorcowej opublikowany przez autorów i założycieli firmy (Page i in. 1999). Mechanizm nadawania rangi nazwano demokratycznym (Page i in. 1999:11), bowiem zgodnie z wzorcem jego działania ranga strony www jest tym wyższa, im więcej stron zawiera link do tej strony i im wyższą rangę mają owe linkujące strony. Google jednak nie podaje publicznie, jak dokładnie działa tworzenie rankingu, a ponadto w 2016 usunęła z wyszukiwarki informację o wyniku PageRank konkretnej strony (Sullivan 2016). Tajemniczość tworzenia rankingu i to, że nie bazuje on tylko na PageRank podkreśla się w branży SEO (*search engine optimisation*), czyli usługach polegających na poprawianiu pozycji w ranking (Bannister 2017). O tej tajemniczości piszą też autorzy techniczni, uprawiający odwrotną inżynierię wyszukiwarki Google, tworzący modele mające na celu odtworzyć wyniki wyszukiwania (Brinkmeier 2006; Langville i Meyer 2004). Dodajmy, że w krytycznej pracy *Tajemnice Google'a* autor twierdzi, że firma ta od początku działalności pilnie strzegła tajemnic rozwijanych technologii, co nazywa podwójnymi standardami – pretenduje do uporządkowania światowych zasobów informacji tak, by stały się powszechnie dostępne, o ile nie dotyczą one tej firmy (Sanchez-Ocana 2013:27–28).

Wracając do sfery publicznej – jaskrawym przykładem czarnej skrzynki jest system oceny nauczycieli szkolnych w Houston, który bazując na wynikach testów uczniów rekomendował decyzje dotyczące m.in. kontynuacji zatrudnienia i awansu konkretnego

37 Przykładowo w tzw. deep learningu czyli głębokich sieciach neuronowych ilość wykonanych obliczeń sprawia, że nie ma możliwości, by człowiek zrozumiał dokładnie zasadę działania modelu – nie ma możliwości prześledzenia równania przez człowieka, jak w np. modelu regresji liniowej.

nauczyciela. Organizacja zawodowa nauczycieli powołując się na prawo pracy, konstytucję i prawa człowieka pozwała stosujące ów system władze tzw. dystryktu szkolnego (*Houston Independent School District*). W toku śledztwa okazało się, że żaden pracownik dystryktu, przy posiadaniu kompletu danych na których system działa, nie potrafi wyjaśnić ani nawet powtórzyć wyniku działania systemu. Skarżący się na zapadające decyzje nauczyciele twierdzą, że jeszcze przed śledztwem pracownicy powiedzieli im, że „ten system jest po to, żeby w niego wierzyć, a nie kwestionować". Firma, która dostarczyła dystryktowi system powołując się na tajemnicę handlową odmówiła prawnikom dostępu do zasad działania systemu. Sąd zawyrokował, że używanie systemu było nielegalne z uwagi brak możliwości wyjaśnienia decyzji nauczycielowi, dystrykt zaś ostatecznie zrezygnował z jego użycia (Crawford i in. 2018:18-20). Brak możliwości uzyskania wyjaśnienia, odwołania się od decyzji czy poprawienia danych jest, zdaniem O'Neil, charakterystyczne dla tego, co nazywa ekonomią danych. „Osoby takie jak Helen Stokes [osoba skrzywdzona] nie są ich klientami. Są produktem, a reagowanie na ich skargi kosztuje" (2017:208). Szereg podobnych przykładów tyczy się wymiaru sprawiedliwości (por. Richardson, Schultz, i Crawford 2019). Działają systemy: np. LSI-R, Level of Service Inventory-Revised (Flores i in. 2006) oraz COMPAS, na który składają się General Recidivism Risk and Violent Recidivism Risk (Northpointe Inc. 2011). W uproszczeniu mają one modelować ryzyko recydywy konkretnej osoby. Model LSI-R bierze pod uwagę m.in. ilość kontaktów z policją, co dyskryminuje mężczyzn o innym niż biały kolorze skóry. W Nowym Jorku czarnoskórzy i Latynosi byli obiektem prawie 41% zatrzymań i przeszukań, stanowiąc niecałe 5% mieszkańców. 9 na 10 tych mężczyzn okazało się niewinnych (O'Neil 2017a:51–53). Do podobnych wniosków doprowadziła analiza systemu COMPAS. Algorytmy dostarczone przez firmę Northpointe przyznają wyższe oceny ryzyka osobom czarnoskórym niż białym, przy kontroli innych zmiennych (Julia i in. 2016; Larson i in. 2016). Dochodzi zatem do reprodukowania i wzmacniania istniejących w wymiarze sprawiedliwości – i nie tylko – uprzedzeń, bowiem jeżeli dane, na których opracowywano modele były „rasistowskie", to model również będzie „rasistowski", bo w taki sposób uzyskuje się dopasowanie modelu do danych. O'Neil mówi w tym kontekście o negatywnym sprzężeniu zwrotnym. Pętla sprzężenia zwrotnego polega na coraz gorszej sytuacji prawnej i majątkowej grup już dyskryminowanych, i kolejno silniejszej dyskryminacji, i tak dalej w sposób zapętlony (O'Neil 2017:58). Tym samym założenia epistemologiczno-metodologiczne i ideologiczne stojące za systemami AI mają bezpośrednie przełożenie na praktykę społeczną. O'Neil, matematyczka i analityczka danych, podkreśla, że

każdy model jest zawsze uproszczeniem rzeczywistości, więc musi pomijać pewne informacje. Autorka zaznacza, że te luki odzwierciedlają intencje, założenia i osobiste przekonania twórców. W tym sensie wyraźnie widoczne jest, że wbrew opiniom laików i entuzjastów analiza danych nie jest obiektywna, bezstronna czy sprawiedliwa – jest narzędziem w ludzkich rękach, a „modele są opiniami opisanymi w języku matematyki" (O'Neil 2017:46-48). Systemy, które są nieprzejrzyste, oparte na silnych założeniach, statyczne (tzn. nie ewaluowane i nie aktualizowane), wykorzystujące zmienne pośrednie[38], a jednocześnie działające na dużą skalę wobec ludzi, autorka określa publicystycznie jako broń matematycznej zagłady (*weapon of math destruction*), co w polskim tłumaczeniu oddano jako Beemzet (ibidem:27). Jej zdaniem sprawiedliwość nie idzie w parze ze skutecznością, a Beemzety stawiają zawsze na tę drugą (ibidem:138-139). Nie działają one rzecz jasna tylko w podmiotach finansowanych publicznie, ale z drugiej strony nie każde oparte na AI czy w ogóle analizie danych rozwiązanie jest uznawane przez nią za Beemzet. Wśród pozytywnych przykładów wymienia ona m.in. analitykę sportową (ponieważ modele działają na wycinku rzeczywistości o znanych regułach, wykorzystywane są w nich pomiary zmiennych bezpośrednich, jak liczba podań czy trafień, zatem ryzyko uwzględniania związków pozornych jest niewielkie, a modele są na bieżąco oceniane i poprawiane z meczu na mecz), system rekomendacyjny Amazona (z uwagi na stałe monitorowanie i ulepszanie), czy systemy bankowej oceny zdolności kredytowej (bowiem dziedzina została uregulowana prawnie, co skutkuje przejrzystością, np. możliwością odwołania się konsumenta od niekorzystnej dla niego decyzji, sprawdzenie jej powodów, poprawę błędnych danych) (ibidem:44-46; 142-143; 197).

Podsumowując, mamy do czynienia z mniej abstrakcyjnym problemem niż w przypadku pierwszych dwóch i zgadzamy się z tym, że przyjmowanie „na wiarę" w połączeniu z zaprojektowaną, założoną z góry nieprzejrzystością systemów wspomagających lub zastępujących procesy decyzyjne ma negatywne konsekwencje społeczne. Sądzimy, że ma to zastosowanie nie tylko do systemów bazujących na AI i nie tylko w odniesieniu do instytucji publicznych. Odczytujemy te procesy jako kolejny przykład powszechności wiary w dataizm i danetyzację.

38 O'Neil podaje m.in. przykład kodu pocztowego w modelu oceniającym kandydata do pracy. Takie zmienne pośrednie mogą być powiązane pozornie ze zmienną celu, a ich użycie jest potencjalnie dyskryminujące, np. z uwagi na miejsce zamieszkania w ubogiej dzielnicy osoba kompetentna nie przechodzi do dalszych etapów rekrutacji (2017a:44-45).

2.2.3.4. Eksperymenty na populacjach ludzkich

Zacznijmy od głośnych, drastycznych przykładów. Śmierć osoby potrąconej przez kontrolowany z użyciem AI samochód autonomiczny (*self-diving*) firmy Uber miała miejsce w marcu 2018, w tym samym miesiącu zginął człowiek siedzący na pokładzie podobnego pojazdu firmy Tesla. Żadna z firm nie poniosła konsekwencji tych wydarzeń. Autorzy raportu zwracają uwagę, że do obu wypadków doszło w Phoenix w stanie Arizona. Arizona, zdaniem autorów, wiedziona perspektywą pojawienia się tam nowych miejsc pracy i napływu kapitału firm technologicznych, w 2015 roku przyjęła prawo zezwalające na ruch samochodów autonomicznych bez człowieka za kierownicą po drogach publicznych. Jest zatem obecnie poligonem dla testowania tego rodzaju technologii w realnym ruchu ulicznym (Crawford i in. 2018:23). W innych badaniach wskazano, że w przypadku spowodowania wypadku przez pojazd autonomiczny kierowcy-ludzie będący na pokładzie są silniej obwiniani przez opinię publiczną niż kierowcy-maszyny (Awad i in. 2020).

Eksperymenty na nomen omen żywym organizmie (*in the wild*) (Crawford i in. 2018:8) mają miejsce także w sferze ochrony zdrowia. Przykładowo bazujący na AI system wsparcia diagnostyki i leczenia nowotworów Watson Oncology firmy IBM jest testowany w szpitalach na całym świecie, przy czym śledztwo dziennikarskie ujawniło dokumenty wewnętrzne firmy, w których mowa o licznych przykładach niebezpiecznych i błędnych rekomendacji (Ross i Swetlitz 2018). Autorzy raportu mówią o zbyt szybkim wdrażaniu technologii pod, omawianą już wyżej, presją dotrzymania kroku innowacjom, czemu podlegać ma także administracja publiczna. Urząd odpowiedzialny za dopuszczanie do obrotu żywności i leków (*U.S. Food and Drug Administration*) uznał za bezpieczny dla konsumentów i zezwolił na sprzedaż inteligentnego zegarka Apple Watch z funkcją EKG, która ma informować o nieregularnym biciu serca, co wzbudziło kontrowersje – tak z uwagi na poprawność działania urządzenia, jak i prywatność oraz bezpieczeństwo danych o charakterze medycznym osób je użytkujących (Crawford i in. 2018:23-24). Jako przykład eksperymentu psychopedagogicznego na dzieciach i młodzieży pokazano działanie firmy Pearson, dostarczającej oprogramowanie i produkty edukacyjne. Firma na grupie około 9000 uczących się sprawdzała, jaka jest różnica w ilości podejmowanych i rozwiązywanych ćwiczeń pomiędzy grupami osób otrzymujących a nieotrzymujących wiadomości, mające budować nastawienie do samorozwoju (*growth-mindset*). Autorzy wskazują, że firma wykonała eksperyment na potrzeby własne, bez zgody i bez wiedzy uczniów, rodziców i nauczycieli (ibidem). Firma Pearson, promująca się hasłem „always learning", prowadzi zespół DS także w Polsce. Była aktywnym uczestnikiem i

jednym ze sponsorów ogólnopolskich konferencji WhyR? – spotkań użytkowników języka R w roku 2017 i 2018, czyli wszystkich dotychczasowych (Why R? Foundation 2018). Eksperymenty, o których mowa, są określane testami A/B, i są odpowiednikiem znanych w medycynie czy psychologii badań z grupą eksperymentalną i kontrolną. Surma omawia kontrowersyjne eksperymenty Facebooka. Jeden dotyczył wpływu wyświetlania określonych treści w newsfeedzie na frekwencję wyborczą – okazuje się, że komunikat miał niewielki, ale istotny statystycznie wpływ mobilizujący do udziału w wyborach. Inny dotyczył transferu stanów emocjonalnych – komunikaty negatywne miały większy zasięg niż pozytywne (Surma 2017: 64; 77-80). Autorzy badania dotyczącego frekwencji wyborczej opublikowali swoje wyniki w czasopiśmie „Nature". Wyświetlane w newsfeedzie wiadomości bezpośrednio wpłynęły na zachowania – polityczne wyrażanie siebie (*political self-expression*), poszukiwanie informacji i głosowanie w wyborach do senatu USA 61 milionów ludzi. Ponadto wiadomości wpływały nie tylko na użytkowników, którzy je otrzymali, ale także na ich znajomych i znajomych ich znajomych. Wpływ tej transmisji społecznej na głosowanie w świecie rzeczywistym był większy niż bezpośredni efekt samych wiadomości, a prawie cała transmisja odbywała się między bliskimi przyjaciółmi, którzy częściej kontaktowali się twarzą w twarz (Bond i in. 2012). Kontrowersyjne jest nie tylko to, że starano się wpłynąć na zachowania ludzi, ale i to, że podobnie jak w przykładzie z firmą Pearson, działano bez zgody i świadomości uczestników tych testów (Surma 2017:64; 77-80).

Z Facebookiem związany jest też tzw. **skandal Cambridge Analytica**, który zdaniem AI Now Institute „w każdym normalnym roku" – a roku 2018 najwyraźniej nie zaliczają oni do takich – byłby największym z możliwych skandali, ponieważ działania firmy Cambridge Analytica miały na celu wpływ na demokratyczne głosowania w USA i Wielkiej Brytanii, z zastosowaniem danych z mediów społecznościowych i algorytmicznego targetowania reklamy politycznej (Crawford i in. 2018:10). Choć autorzy raportu AI Now Institute nie omawiają szczegółowo tego skandalu, to warto to zrobić w kontekście problemu nieregulowanych i niemonitorowanych eksperymentów z użyciem AI na populacjach ludzkich. Były pracownik Cambridge Analytica, tzw. sygnalista (*whistleblower*) Christopher Wylie w rozmowie z dziennikarką The Guardian, nazwał działania firmy „rażąco nieetycznym eksperymentem, ponieważ grano na psychice całego kraju, bez zgody i świadomości ludzi[39]" i przyznał, że czuje się za to odpowiedzialny oraz żałuje tego, co się wydarzyło – m.in. z tej przyczyny zdecydował się ujawnić sprawę w prasie (Cadwalladr

39 Oryginalnie: „It was a grossly unethical experiment because you are playing with an entire country, the psychology of an entire country, without their consent or awareness".

2018a). Cambridge Analytica wspierała sztab zwycięskiego kandydata na prezydenta USA Donalda Trumpa podczas kampanii wyborczej w 2016, o czym dla magazynu Newsweek oficjalnie i w kategorii sukcesu mówił jej prezes, Alexander Nix: „gdy podano wyniki wyborów, wpadłem w ekstazę" (Ćwiklak 2017:73). Nix dowodził skuteczności usług Cambridge Analytica, które nazywa konsultingiem politycznym. Zauważmy, że wywiad miał miejsce prawie rok przed przełomowym dla sprawy wywiadem sygnalisty Wylie'a. Jak pisze Ćwiklak:

> [prezes firmy] chwali się, że jego firma Cambridge Analytica zna profil osobowości każdego dorosłego Amerykanina i że wykorzystała tę wiedzę, by zapewnić zwycięstwo Donaldowi Trumpowi (2017:73).

Dzięki wykonanym przez firmę analizom sztab Trumpa zdecydował np. o większych nakładach na kampanię w stanie Wisconsin. Choć w Wisconsin po raz ostatni głosowano na republikanina w roku 1984, Trump odwiedził ten stan pięciokrotnie, podczas gry Hillary Clinton nie zrobiła tego ani razu (Ćwiklak 2017:74). Analiza danych służyła do targetowania reklamy politycznej, bo

> cała sztuka polegała na tym, by mówić tylko do tych, których można przekonać do głosowania na Trumpa. I tak personalizować reklamy, by każdemu mówić to, co chce usłyszeć. Czy to manipulacja? Nix obrusza się. Staramy się dostarczyć odbiorcy najistotniejsze dla niego treści, by mógł podjąć decyzję uzbrojony w jak najwięcej faktów – tłumaczy. Chodzi o to, by tak zniuansować przekaz, aby jak najlepiej trafił do odbiorcy (Ćwiklak 2017:75).

Były pracownik Cambridge Analytica, sygnalista (*whistleblower*) Christopher Wylie, mówił o firmie jako machinie propagandowej, która nie tylko budowała profile osób na podstawie analizy danych, ale tworzyła sprofilowane treści na stronach internetowych czy blogach i dbała o to, by dotarły one do odpowiedniego profilu odbiorców. (Cadwallard 2018). Wcześniej Nix zapewniał jednak, że firma nie miała nic wspólnego z tzw. fake news (Ćwiklak 2017). Wylie nie twierdził, że podawano nieprawdziwe informacje. Mówił, że ich praca polegała na tym, by zamiast mówić o określonej sprawie do tłumu, stojąc na środku miejskiego rynku szeptać o niej do ucha każdemu z osobna wiedząc, jak sformułować komunikat dla tej konkretnej osoby (Cadwallard 2018). Firma chwaliła się swoją skutecznością w wywieraniu wpływu na wyborców:

> „Cambridge Analytica pomogła zidentyfikować zwolenników, przekonać nieprzekonanych i doprowadzić ich do urn. Firma dostarczała informacje, na których sztab wyborczy opierał swoje kluczowe decyzje dotyczące kierunków kampanii, komunikacji i rozłożenia zasobów. W ramach kampanii prowadzonej w cyfrowych mediach Cambridge Analytica była w stanie na bieżąco testować i dostosowywać komunikaty oraz kanały dotarcia do ludzi. Dzięki cyfrowemu zespołowi Cambridge Analytica była to pierwsza kampania, w której udało się wykorzystać

wiele nowych trendów w przemyśle reklamowym, takich jak reklama natywna". Takie oświadczenie – wprost sugerujące kluczową rolę analizy danych i nieprzejrzystych technik reklamowych – znalazło się na stronie firmy zaraz po zwycięstwie Trumpa (Szymielewicz 2017).

Targetowanie reklamy politycznej wykonywano m.in. dzięki wykorzystaniu pracy specjalizujących się w psychometrii psychologów z Uniwersytetu w Cambridge i użyciu danych z portalu Facebook. Michał Kosiński, David Stillwell i Thore Graepel pokazali, że łatwo dostępne cyfrowe zapisy zachowań, czyli tzw. lajki (*likes*) facebookowe (od funkcji „like" – „polub"), mogą być wykorzystywane do automatycznego i dokładnego przewidywania wielu wrażliwych cech osobistych, takich jak: orientacja seksualna, pochodzenie etniczne, poglądy religijne i polityczne, cechy osobowości, inteligencja, szczęście, uzależnienie, separacja rodziców, wiek i płeć. Przedstawiono analizę opartą na danych ponad 58 000 ochotników, którzy wzięli udział w badaniu za pomocą przygotowanej przez psychologów aplikacji MyPersonality i udostępnili swoje profile facebookowe, szczegółowe dane demograficzne i wyniki kilku testów psychometrycznych. Proponowany model wykorzystuje redukcję wymiarów do wstępnego przetwarzania lajków, a następnie wykonuje prognozę psychodemograficzną za pomocą metod regresji logistycznej i regresji liniowej. Przykładowo ich model prawidłowo rozróżnia homoseksualnych i heteroseksualnych mężczyzn w 88% przypadków, zaś między demokratami i republikanami rozróżnienie było poprawne w 85% (Kosinski, Stillwell, i Graepel 2013). Dodajmy, że autorzy pracowali w języku programowania R. Zauważmy też, że zastosowano wręcz podręcznikowe, powszechnie wykorzystywane w naukach społecznych i przyrodniczych techniki statystyczne, a nie uczenie maszynowe. Tym samym należy zwrócić uwagę przede wszystkim na to, jaki jest potencjał danych z mediów społecznościowych – że w ogóle możliwa jest taka analiza lajków, która daje wyniki wysoce zbieżne z testami psychometrycznymi. Autorzy podkreślili ten wątek, wskazując na potencjalne implikacje uzyskanych wyników dotyczące personalizacji i prywatności w internecie (ibidem). Michał Kosiński wypowiedział się w tym kontekście dla wspominanego wcześniej serialu dokumentalnego Do Not Track (od 2 min. 8 s., odcinek 3.) mówiąc, że dzięki danym z mediów społecznościowych „komputer jest lepszy w ocenie Twojej osobowości od Twojego męża czy żony" (UPIAN 2015). Christopher Wylie jako początkujący doktorant zainteresowany prognozowaniem zachowań politycznych przeczytał omawiany artykuł, nie będąc jeszcze pracownikiem Alexandra Nixa. Poznał go w czasie, gdy zainspirowany pracą Kosińskiego i innych proponował swoje usługi identyfikowania wyborców partiom

politycznym. Nix ponoć przekonał go do pracy w Cambridge Analytica obietnicą całkowitej wolności w eksperymentowaniu i testowaniu swoich najbardziej szalonych pomysłów (Cadwallard 2018). Prasa zaznaczała, że Kosiński z zespołem odmówili współpracy z firmą Cambridge Analytica, a właściwie Strategic Communication Laboratory, bo pod taką nazwą funkcjonowała, realizując wcześniej kilkaset zleceń dla rządów, wojska i innych organizacji z całego świata (Cadwalladr 2018a; Ćwiklak 2017; Czubkowska 2018).

Współpracy z firmą podjął się inny psycholog z Uniwersytetu w Cambridge, Aleksandr Kogan, który odtworzył aplikację o nazwie thisisyourdigitallife, na wzór MyPersonality i modelu predykcyjnego Kosińskiego, Stillwella i Graepela. Uczestników, wyłącznie Amerykanów, rekrutował on za pomocą portalu z ogłoszeniami o proste, powtarzalne prace Mechanical Turk na Amazonie, oferując 1 – 2 dolary za udział w ankietach (Ćwiklak 2017; Schwartz 2017). Ogłoszenie zamieściła w 2014 r. należąca do Kogana firma Global Science Research, zaś w 2015 Amazon usunął je z powodu powtarzających się skarg, że udział w ankietach wymaga wyrażenia zgody na dostęp do danych profilu Facebook – uczestnika ankiety i wszystkich jego znajomych, co było naruszeniem regulaminu Mechanical Turk (Davies 2015; Schwartz 2017). Właśnie od tego momentu pojawia się problem nie etyki, a potencjalnego złamania prawa. Podawane są zarówno informacje, że Global Science Research Kogana informowała uczestników, iż zbierane dane posłużą wyłącznie do celów badawczych i nie miała prawa przekazywać ich Cambridge Analytica, jak i to, że zgoda dotyczyła celów dowolnych (Davies 2015).

Tak czy inaczej Global Science Research pozyskała i przekazała Cambridge Analytica dane kilkudziesięciu milionów użytkowników Facebooka. Podawane są liczby od 30 (Schwartz 2017) przez 50 (Cadwalladr 2018) do 87 milionów (Ghosh 2018; Isaak i Hanna 2018:56), przy czym podkreślmy, że w ankiecie wzięło udział znacznie mniej osób, ponieważ pobierane były także dane niczego nieświadomych znajomych uczestnika ankiety. Podawana jest średnia liczba 340 znajomych na jednego użytkownika Facebooka w 2014 r. (Davies 2015), co przy prawoskośnym rozkładzie liczby znajomych wskazywałoby na około 200 – 300 tysięcy respondentów przy 50 milionach pobranych punktów danych; Facebook podał informację o około 270 tys. respondentach (Grewal 2018). Nie ma informacji, by naruszeniem prawa było udzielenie dostępu przez Facebook dla aplikacji firmy Kogana. Media nazywały całą tę sytuację naruszeniem danych (*data breach*) – w języku polskim używano słów „wyciek danych” (Gersz 2018) – z Facebooka i oskarżały go o niedopilnowanie bezpieczeństwa swoich użytkowników (Cadwalladr i Graham-Harrison 2018a; Hockney i

Green 2018). Dostęp do danych użytkowników Facebooka i ich znajomych był możliwy dzięki pierwszej wersji technologii Facebook Graph API[40], która funkcjonowała od 2010 do 2015 r. Facebook za jej pomocą sprzedawał dane swoich użytkowników rozmaitym podmiotom: marketingowym, biznesowym, badawczym i prawnym (Albright 2018; Symeonidis, Tsormpatzoudi, i Preneel 2015). Nie jest zatem dziwne, iż Facebook zaprzeczał, że działanie aplikacji Kogana było naruszeniem (*data breach*), jednocześnie informował o zawieszeniu kont Cambridge Analytica, Kogana i Wylie'a na portalu (Grewal 2018). Dziennikarze wskazują na słuszność tego zaprzeczenia, ponieważ zbieranie danych przez aplikację Kogana (również od nieświadomych niczego znajomych osób wyrażających zgodę na udział w badaniu) było zgodne z ówczesną polityką prywatności Facebooka. Kogan zatem działał legalnie, a o naruszeniu (*data breach*) mówi się w przypadku ataków hakerskich, co w tym wypadku nie zachodzi (Franceschi-Bicchierai 2018). Taka linia obrony nie zatrzymała fali krytyki wobec Facebooka. Wspominana socjolożka Zeynepf Tufekci zwracała uwagę, że skoro takie wykorzystanie danych Facebooka jest legalne, to świadczy to jak najgorzej o jego modelu biznesowym i stawia firmę w jeszcze gorszym świetle niż gdyby dane uzyskano naruszając jej przepisy (Tufekci 2018). Tę jej wypowiedź znaleźliśmy udostępnioną na facebookowym fanpage'u Data Science PL ze słowami „najlepszy komentarz". Autorka już wcześniej pisała krytycznie o działaniach m.in. Facebooka i Google, mówiąc o „algorytmicznych krzywdach" w kontekście naruszania prywatności i wpływu na wybory polityczne (Tufekci 2015:209–12; 215). O modelu biznesowym Facebooka w podobnie krytycznym tonie wypowiadał się polski data scientist i socjolog (Batorski 2018). Powstał ruch na portalu Twitter pod tagiem #deletefacebook, gdzie nawoływano do usuwania kont na portalu i należących do niego Instagramie i Whatsapp w ramach sprzeciwu wobec modelu, w którym użytkownik jest faktycznie produktem (Aron 2018; Bhardwaj 2018). W dniach od 16.03. (piątek) do 19.03.2018 (poniedziałek) cena akcji Facebooka spadła o 8% (Bhardwaj 2018) – pierwsze materiały z udziałem sygnalisty Wylie'a opublikował The Guardian w sobotę 17.03. (Cadwalladr i Graham-Harrison 2018b). Ten spadek cen akcji był poważny, a dodatkowo po nim wzrosła zmienność (*volatility*) cen. Ogółem wzrosła zmienność cen akcji spółek indeksu NASDAQ-100, co badacze określają jako niekorzystne dla inwestorów z uwagi na zwiększone ryzyko (Peruzzi i in. 2018). W odniesieniu do sprawy pisano o konieczności uregulowania prawnego kwestii dostępu do danych z mediów

40 API oznacza *application programming interface*, co w uproszczeniu oznacza zaprogramowany mechanizm pozwalający na komunikację pomiędzy dwoma aplikacjami, np. Facebookiem a aplikacją innego twórcy i może służyć np. do pobierania danych wykorzystywanych później do analizy.

społecznościowych (Isaak i Hanna 2018). W tej sprawie, po wywiadach prasowych sygnalisty, Mark Zuckerberg dwukrotnie zeznawał przed Kongresem USA (Wąsowski 2018b). Facebook oficjalnie przepraszał za zaistniałą sytuację i całą winą obarczał Kogana wskazując, że nielegalne było przekazanie przez niego danych do innych podmiotów (Grewal 2018; Wong 2018). Wskazuje się, że Facebook wiedział o sytuacji w 2015 r. – firma informowała, że wtedy wyłączyła dostęp aplikacji Kogana do portalu (Grewal 2018) – jednak przeprosiny i poważniejsze zmiany w polityce prywatności przyniosło nagłośnienie sprawy w mediach w marcu 2018 (Davies 2018; Wąsowski 2018a; Wong 2018). Dodatkowo Kogan (posługujący się w działalności naukowej także nazwiskiem Spectre) oficjalnie współpracował z Facebookiem w latach 2014-2015 (Dwoskin, Harwell, i Timberg 2018), m.in. publikując artykuł naukowy, bazujący na danych z tego portalu we współautorstwie z pracownikami firmy. Dane zbierano w identyczny sposób jak te, które później przekazano CA (Spectre i in. 2015), ale zasadniczo nie świadczy to o legalności przekazania innych przecież danych do innych celów. Kogan twierdził, że był przekonany, że jego działanie w ramach firmy Global Science Research, tzn. zbieranie danych osób i ich znajomych za pomocą ankiety połączonej z portalem Facebook oraz przekazanie ich CA do celów marketingu politycznego, było w pełni legalne i zgodne z polityką Facebooka (Cadwalladr i Graham-Harrison 2018b). Wskazywał on także, że z uwagi na to, że urodził się pod koniec lat 80. na terenie ówczesnego ZSRR jest wdzięcznym kozłem ofiarnym (Mac 2018), a po sugestiach Christophera Wylie'a nazywano go rosyjskim szpiegiem (Cadwalladr i Graham-Harrison 2018b; Mac 2018). Prasa nazywała sygnalistę (*whistleblower*) Christophera Wylie'a terminem data scientist, a on sam wypowiadał się jako osoba biegła w kwestiach dotyczących analizy danych (Cadwalladr 2018a). Kogan, również nazwany data scientistem twierdzi, że Wylie nie ma z tą dziedziną wiele wspólnego i „nazywać go data scientist to jak mnie nazywać ikoną mody, a zazwyczaj noszę dres" (Mac 2018). Ma być to argument za tym, że Wylie wyolbrzymiał możliwości targetowania, jakie miała CA i stworzył nieprawdziwe wrażenie znaczącego wpływu działań firmy na wyniki wyborów prezydenckich (ibidem). Odpowiedzialny za media cyfrowe w kampanii Trumpa Brad Parscale (Diamond i Bash 2018; Lapowsky 2016) twierdził, że choć CA dostarczała mu użytecznych analiz, to w kampanii w ogóle nie stosowali psychologicznego targetowania reklamy (Lapowsky 2017).

Zwracamy uwagę, że CA pracowała nie tylko dla sztabu Donalda Trumpa. Wcześniej na etapie prawyborów wspomagała sztab republikanina Teda Cruza (Davies 2015), zaś z Republikanami w ogóle łączą firmę osoby Steve'a Bannona oraz Roberta Mercera

(Cadwalladr i Graham-Harrison 2018b; Ćwiklak 2017). Pierwszy był członkiem zarządu CA, redaktorem naczelnym prawicowego amerykańskiego medium Breitbart News, następnie szefem kampanii Trumpa i jego doradcą, a drugi z wymienionych ufundował założenie firmy, a wcześniej był zajmującym się sztuczną inteligencją inżynierem i opisywany jest jako wspierający finansowo kampanię na rzecz Brexitu i Breitbart News. (Cadwalladr i Graham-Harrison 2018b; Ćwiklak 2017). Christopher Wylie nazywał swoją pracę w CA budowaniem kulturowej broni (*weapon to fight a culture war*) dla S. Bannona (Cadwalladr 2018a; Solon 2018). Mówi się o ścisłej współpracy Cambridge Analytica z firmą AggregateIQ, która świadczyła usługi podobne do tych oferowanych przez CA dla organizacji wspierających wyjście Wielkiej Brytanii z UE: Vote Leave, BeLeave, Veterans for Britain, DUP oraz Leave.EU (Cadwalladr 2018b; Cadwalladr i Townsend 2018). Należy pamiętać, że CA korzystała nie tylko z danych pozyskanych z Facebooka, a także z innych mediów społecznościowych, przeglądarek internetowych, danych o zakupach online, wynikach głosowań i innych, pobieranych samodzielnie lub kupowanych od brokerów danych (Isaak i Hanna 2018:57; Szymielewicz 2017). Kogan twierdził, że tego rodzaju działalności i sposobów pozyskania danych jest na całym świecie bardzo, bardzo wiele, a akurat skandal CA został wyjątkowo nagłośniony, co ma pozytywny wydźwięk w postaci dostrzeżenia przez opinię publiczną problemu dostępu i wykorzystywania danych z serwisów społecznościowych (Mac 2018). Przykładowo w Polsce Fundacja Panoptykon pisała o tym skandalu kilkakrotnie, zachęcając do korzystania z przygotowanych przez siebie materiałów edukujących w zakresie prywatności w serwisach społecznościowych (Obem 2018). Różne media wskazywały, że choć nie ma mowy o działaniach potencjalnie nielegalnych, to CA lub podobne do niej firmy analityczne pracowały w wielu kampaniach wyborczych (Szymielewicz 2017). Wymienia się np. kampanie Baracka Obamy w 2008 i 2012 roku (Rutenberg 2013), a także Andrzeja Dudy w 2015 (Gersz 2018; Matuszewski 2018). W rocznicę publikacji wywiadu z sygnalistą Ch. Wylie ponownie pisano o wspomnianym, pozytywnym wymiarze afery jako rozbudzającej debatę publiczną o dostępie i wykorzystywaniu danych oraz prywatności ludzi, szczególnie użytkowników Facebooka. Zaznaczono, że choć zaszły pewne istotne zmiany na korzyść prywatności jednostek, jak np. wdrożenie przez Facebook szyfrowania wiadomości dla serwisów Messenger i Instagram, to problemów jest bez porównania więcej niż rozwiązań. Problemów, tak po stronie firm, podmiotów stanowiących prawo (*regulators*), jak i użytkowników, wymagających jednocześnie – a co wydaje nam się sprzeczne – wygodnego i darmowego internetu oraz kontroli nad danymi o sobie (Lapowsky 2019).

Tego rodzaju eksperymenty na nieświadomych ludziach w marketingu politycznym są postrzegane inaczej niż kiedy chodzi o marketing dóbr i usług konsumpcyjnych, co wyraźnie ilustrują słowa prezeski Fundacji Panoptykon:

> Zasada działania amerykańskich kampanii wyborczych nie różni się od profesjonalnie przygotowanej kampanii jakiegokolwiek produktu. Najważniejsze to dobrze określić i poznać target. W drugim kroku trzeba dopasować przekaz, złapać klienta wyborcę we właściwym momencie i sprowokować do interakcji. Kampania działa, jeśli dość duży odsetek złapanych kupi produkt: poda swój numer karty kredytowej i kliknie albo pójdzie do urny. W przypadku nowego produktu dobra reklama to konieczność – klient musi go poznać, a przynajmniej zapragnąć, zanim kupi. Na pozór w tym świecie rządzi jednak klient wyborca, a nie agencja reklamowa: zarówno marka nowego mydła, jak i profil nowego prezydenta muszą być dobrane do jego emocji i potrzeb. A kiedy te się zmieniają, stara linia produktów już nie przejdzie. Demokraci musieli ustąpić republikanom. Czyżby zatem nowy paradygmat politycznego marketingu zbliżał nas do ideału demokracji bezpośredniej, w której programy rządzących dynamicznie się zmieniają w zależności od tego, co w danej chwili pragną usłyszeć lub zobaczyć ich wyborcy? Takie spojrzenie promują sami politycy, twierdząc, że dzięki zbieranym danym i mikrotargetowaniu przekazu mogą być bliżej swoich wyborców. Po to, by lepiej odczytać ich potrzeby, czy raczej po to, by je skuteczniej wykreować? Niestety, granica między badaniem społecznych nastrojów a polityczną manipulacją jest płynna. Ilu swoich wyborców Trump wykreował, ilu zmanipulował, a do ilu po prostu skutecznie dotarł? Żeby odnaleźć swoje miejsce na tej osi, **wyborca musi przede wszystkim wiedzieć, że stał się królikiem doświadczalnym w laboratorium politycznego marketingu** [podkr. RŻ]. Większość Amerykanów nie miała tej świadomości w czasie kampanii prezydenckiej: nikt ich nie pytał o zgodę na śledzenie w sieci, profilowanie i wykorzystywanie danych z całkiem z polityką niezwiązanych sfer życia. Być może zaczynają sobie z tego zdawać sprawę teraz, czytając butne wypowiedzi „cudotwórców" z Cambridge Analitica. Jednak ani ich świadomość, ani ewentualne oburzenie, nie mają w tej grze większego znaczenia. (…) Realnego wpływu na to, w oparciu o jakie zachowania i cechy jesteśmy profilowani i dlaczego dociera do nas taki, a nie inny przekaz, nie mamy ani przy wyborze mydła, ani prezydenta. Jednak **ze względu na konsekwencje politycznych wyborów – dotyczące nie tylko osoby, która dała się „złowić" na sprofilowany przekaz – nieprzejrzystość technik politycznego marketingu to o wiele poważniejszy problem niż reklama behawioralna** [podkr. RŻ] (Szymielewicz 2017).

Wspominana wyżej Zuboff już przed omawianym skandalem twierdziła, że demokracja zagraża uzyskiwaniu dochodów z inwigilacji ludzi, zatem tzw. kapitalizm inwigilacji jest antydemokratyczny (Zuboff 2015:86). Siva Vaidhyanathan, medioznawca i historyk kultury, uważa, że media społecznościowe – a przede wszystkim Facebook – generalnie zagrażają demokracji. Nie mówi on konkretnie o opisywanym skandalu. Jego zdaniem Facebook jest „wydestylowanym obrazem Doliny Krzemowej", która ma charakteryzować się przywiązaniem do logicznego myślenia, podejmowania decyzji w oparciu o dane, kosmopolityzmem i tolerancyjnością, oraz misjonaryzmem. Misjonaryzm ten głosi połączenie (*connectivity*) ludzi, a zatem rozprzestrzenianie wiedzy, co ma zmieniać ich życie na lepsze. Zdaniem autora:

> Żadna inna firma [niż Facebook] nie przedstawia tak dobrze marzenia o w pełni połączonym

> świecie. (…) Żadna inna firma nie przyczyniła się bardziej do paradoksalnego załamania podstawowych dogmatów debaty i demokracji (Vaidhyanathan 2018:12–13).

Autor mówi o trzech problemach w budowie Facebooka. Wszystkie one mają być antydemokratyczne, a przynajmniej nie sprzyjać demokratyzmowi. Po pierwsze, w neews feedzie facebookowym trudno jest zauważyć źródło oglądanej informacji, ponieważ wszystkie mają jednakową ramkę, czcionkę i formatowanie. Ma się to przyczyniać do łatwego i szybkiego rozprzestrzeniania się informacji nieprawdziwych i wprowadzających w błąd, w tym tzw. fake newsów. Po drugie, mechanizm reagowania użytkowników na widziane treści – kiedyś wyłącznie lajkowania, obecnie siedem reakcji do wyboru – powoduje promowanie tego, co wywołuje silne emocje. Przypomnijmy o pojęciu ekonomii uwagi, czyli serwisach dążących do zwiększania czasu spędzonego na nich przez użytkowników. Emocje są tym, co trzyma użytkownika w zalogowaniu. Tym samym opinie wyważone, słabo nacechowane emocjonalnie nie są promowane na Facebooku. Przeciwnie, treści najbardziej ekstremalne i polaryzujące wywołują największe zaangażowanie, której miarą jest liczba reakcji, a w konsekwencji mają największy zasięg. Po trzecie, na Facebooku powstają bańki filtrujące (*filter bubbles*). Jest to także wywołane przez ekonomię uwagi zjawisko, polegające na tym, że użytkownikom dostarcza się tylko te treści, które są zgodne z ich zainteresowaniami i poglądami. To właśnie oznacza bycie w bańce – brak treści sprzecznych. Jest to nazywane personalizacją treści, a umożliwia ją rzecz jasna zbieranie danych o użytkowniku i dalej wybór najbardziej angażujących emocjonalnie treści z użyciem m.in. metod uczenia maszynowego. „W ten sposób Facebook utrudnia różnym grupom spotkanie się w celu przeprowadzenia spokojnej, merytorycznej i konstruktywnej rozmowy" (Vaidhyanathan 2018:14–18).

W skandalu CA widać także ogromną asymetrię władzy, o której także mówi Zuboff jako cesze charakterystycznej kapitalizmu inwigilacji, na linii wyborcy – kandydaci z całym swoim zapleczem. O'Neil określiła to następująco:

> Marketerzy polityczni tworzą obszerne kartoteki na nasz temat, dostarczają nam skrawków informacji i sprawdzają, jak na nie reagujemy. (…) Taka asymetria informacyjna uniemożliwia poszczególnym stronom łączenie sił, a przecież taka jest właśnie idea rządów demokratycznych (2017a:264).

Z uwagi na omówioną strategię „szeptania każdemu do ucha zamiast przemawiania do tłumu na rynku" (Cadwalladr 2018) należy nazwać to nierównowagą na linii wyborca (nie „wyborcy") – kandydat z zapleczem. Surma, abstrahując od marketingu politycznego i

odnosząc się do ogólnej roli analizy danych w biznesie nazywa to przejściem od segmentacji do personalizacji. Pierwsza to tradycyjne podejście, gdzie relacja przebiega na linii firma – wielu klientów w możliwie jednorodnej grupie. Personalizacja to relacja firma – pojedynczy klient, co ma być charakterystyczne dla marketingu z użyciem big data:

> Klient jest traktowany podmiotowo jako niezależna jednostka o specyficznych potrzebach i preferencjach. Firma pragnie poznać tę charakterystykę poprzez 'rozmowy', utrzymywanie relacji oraz poprzez zbieranie i analizę śladów cyfrowych, jakie klient pozostawia w ramach różnych interakcji (Surma 2017:48-49).

Ma być to przejście od konsumeryzmu do prosumeryzmu, bo skoro produkt jest personalizowany dla pojedynczego klienta na podstawie danych o nim, to klient jest niejako współproducentem tego produktu (Nowacki 2014:14–17; Surma 2017:39). Iwasiński wskazuje, że zwolennicy marketingu bazującego na big data twierdzą, że działają wyłącznie na korzyść konsumenta, bowiem dzięki personalizacji oferta rynkowa będzie coraz lepiej odzwierciedlać ludzkie preferencje, a konsumenci będą mogli poznać i zrealizować swoje głębokie czy nawet nieuświadamiane potrzeby (Iwasiński 2017:125). Widać echa omawianych już pretensji do obiektywizmu i prawdziwości rekomendacji na podstawie danych, a także zwalniania jednostki z odpowiedzialności za podejmowane decyzje. Iwasiński wskazuje, że suwerenność konsumenta wobec marketing big data jest ambiwalentna. Pozornie podmiotowe, indywidualne jego traktowanie co prawda sprzyja poczuciu komfortu i autonomii, ale arbitralność bazujących na danych systemów i ciągła inwigilacja zapewniająca dostęp do danych prowadzi do licznych wątpliwości, co do zakresu wolnej woli i odpowiedzialności konsumenta (Iwasiński 2017:130-131). Mówi on o dobrach i usługach konsumpcyjnych, nie o demokratycznych wyborach. W skandalu Cambridge Analytica uderzające jest to, że firma wspierała marketing polityczny tak, jak byłby to marketing dobra konsumpcyjnego lub usługi. Zdaniem Yonatana Zungera, byłego inżyniera Google w dziedzinie prywatności i bezpieczeństwa, skandal CA wskazuje na kryzys etyczny w informatyce (*computer science*). Zunger jako fizyk twierdzi, że informatyka (*computer science*) w odróżnieniu od innych nauki i praktyki ani razu nie poniosła poważnych, negatywnych konsekwencji działania swoich przedstawicieli. Podaje on przykłady wynalazków czy idei wykorzystywanych do krzywdzenia ludzi, jak dynamit wynaleziony przez chemików, bomba atomowa stworzona przez matematyków i fizyków czy eugenika na gruncie biologii człowieka, czy tych mogących skrzywdzić, jak błędnie zbudowany przez inżynierów drogownictwa i zawalony most. Każda z tych dziedzin chciała zmienić świat na lepszy i poniosła poważne tego konsekwencje w postaci regulacji prawnych, procedur

walidacji wyników badań i zdobywania uprawnień do wprowadzania ich do praktyki (jak w przypadku leków, pozwoleń na budowę czy uprawnień do wykonywania zawodów technicznych i medycznych), czy wprowadzenia powszechnego kształcenia w dziedzinie etyki i kodeksów etycznego postępowania. Autor zwraca uwagę na omawiane już podejście „działaj szybko i psuj", będące jego zdaniem wyrazem presji na szybkość wdrożenia ze strony liderów biznesowych przekonanych o tym, że w razie błędu lub porażki nie poniosą poważniejszych konsekwencji. Zauważa też, że sami informatycy (*software enigineers*) postrzegają problemy bezpieczeństwa i etyki jako dodatek do projektowanych rozwiązań, nie jako ich podstawę. Kształcą się w przekonaniu, że żeby zmienić świat wystarczy nauczyć się programować. Autor uważa, że systemy informatyczne są testowane tylko w sensie konsekwencji technicznych, tzn. czy nie zepsują się przy konkretnych scenariuszach, a nie konsekwencji ludzkich. Jego zdaniem gdyby powszechną praktyką było dostrzeganie konsekwencji w wymiarze ludzkim, w ogóle by nie wdrożono wspomnianej technologii Facebook Graph API, umożliwiającej pobieranie danych z Facebooka przez rozmaite zewnętrzne podmioty, w tym takie aplikacje, jak thisismydigitallife Aleksandra Kogana (Zunger 2018).

Podsumowując z właściwego nam punktu nauk społecznych bazujących na badaniach terenowych, uważamy, że praktyki zbierania informacji o ludziach – kiedy ci nie zgadzają się na to, bądź nie są tego świadomi – budzą wątpliwości etyczne, nie mówiąc o próbach wpływu na zachowania tych ludzi. Podkreślamy więc ponownie, że uznajemy DS za działalność paranaukową, w której cele poznawcze są zawsze podrzędne wobec utylitarnych, przy czym nie oznacza to w żadnym wypadku, że ta działalność jest immanentnie wątpliwa etycznie. Wątpliwe etycznie są cele utylitarne, realizowane z użyciem DS, a właściwie z użyciem technologii w ogóle. Powtórzyć należy spostrzeżenia dotyczące wyraźnie widocznych przemian w sferze prywatności, których wyrazem jest zarówno legalne jak i nielegalne zbieranie i korzystanie z danych o ludziach do celów, których nie są oni – nieważne czy z woli swojej czy zewnętrznej – świadomi, jak i to, że ludzie w ogóle zostawiają w (upraszczając i uogólniając) internecie ogromną ilość danych o sobie.

Krytyka w omawianym obszarze płynie także pod adresem kolonialnego charakteru tzw. big data/data science. Na kolonialny charakter ma wskazywać m.in. fakt, że choć firma CA stosowała swoje antydemokratyczne praktyki marketingu politycznego w 2017 r. w Kenii oraz 2015 r. w Nigerii, to jej działalność stała się skandalem dopiero w roku 2018, po ujawnieniu działalności w wyborach prezydenckich w USA i w kampanii dotyczącej Brexitu

(Mohamed, Png, i Isaac 2020:11). Działalność CA w Kenii i Nigerii opisano jako znane z czasów kolonialnych, a stosowane np. przez naukowców brytyjskich wyeksportowanie testowania (*beta-testing*) praktyk potencjalnie nieetycznych i prawnie ryzykownych we własnym kraju, na grupach odległych geograficznie i słabo chronionych przez prawo, słabych ekonomicznie, co określane jest jako tzw. zrzut etyczny (*ethics dumping*). Testowanie ryzykownych praktyk na bezpiecznym, pozbawionym konsekwencji gruncie, przed „realnym" zastosowaniem ich na gruncie USA i UK (ibidem).

2.2.3.5. Rozwiązywanie problemów etycznych za pomocą technologii

W piątym, ostatnim z wyróżnionych przez Instytut AI Now obszarów problemowych wskazano na ograniczenia technologii w rozwiązywaniu problemów uczciwości, obciążeń (*bias*) i dyskryminacji. Autorzy wskazują na wzmożone w 2018 r. prace firm tworzących i wdrażających systemy AI w sferze nastawionej na przeciwdziałanie powyższym problemom. Ich zdaniem podejmowano takie prace z uwagi na liczne skandale roku 2018, a we wcześniejszych latach były one marginalne (Crawford i in. 2018:29; 32). Zasadniczo mowa o dwóch strategiach: rozwiązaniach technologicznych oraz deklaratywnych. Te pierwsze autorzy nazywają próbami zapewnienia algorytmicznej sprawiedliwości (*algorithmic fairness*) i uznają je za charakterystyczne dla osób i środowisk zajmujących się AI od strony technicznej i akademickiej. Rozwiązania technologiczne zasilane są nowymi metodami algorytmicznymi i statystycznymi, które mają na celu diagnozowanie i łagodzenie obciążeń (*bias*). Sukces w tego rodzaju podejściu definiowany jest w odniesieniu do, bazujących na matematycznych regułach, ilościowych definicjach sprawiedliwości (*fairness*) (Crawford i in. 2018:24–25). Dominujące definicje omawiają Skirpan i Gorelick (Skirpan i Gorelick 2017). Autorzy wskazują na istotną rolę społeczności konferencji Fairness, Accountability, and Transparency in Machine Learning (FAT ML 2019) w rozwoju tych rozwiązań, w tym definicji i metryk sprawiedliwości, i wymieniają szereg rozwiązań już wdrożonych do biznesu, z których większość ma status open-source. Są to m.in.: „AI Fairness 360", zestaw narzędzi autorstwa IBM, zawierający dziewięć różnych algorytmów, które mają łagodzić obciążenia na każdym etapie budowania systemu AI (od przygotowania danych do uczenia modelu); „What-If" autorstwa Google's People + AI Research group (PAIR), który pozwala na wizualizacje efektu zastosowania różnych metod i metryk łagodzenia obciążeń na konkretnym, wyuczonym modelu; pakiet w języku Python „fairlearn.py" autorstwa firmy Microsoft, który ma zapewniać sprawiedliwość w zadaniach klasyfikacji binarnej (takich jak np. przekazać / nie przekazać CV do drugiego etapu rekrutacji) na podstawie zadanych przez

człowieka ograniczeń (*constrains*); wskazano też podobne prace firm Facebook i Accenture (Crawford i in. 2018:24; 28-29). Wspominaliśmy już o tego rodzaju pracach polskiego naukowca i jednego z najbardziej znanych uczestników świata DS w Polsce – chodzi m.in. o pakiet „DALEX" w języku R, który ma służyć wyjaśnianiu działania złożonych modeli uczenia maszynowego, także tzw. czarnych skrzynek (Biecek 2018d). P. Biecek zwracał też uwagę, jak takie wyjaśnianie ma się do potencjalnych, nowych regulacji prawnych w Europie, tzw. prawa do wyjaśniania, które może wynikać z RODO (Biecek 2018f). Generalnie w badaniach nad uczeniem maszynowym istnieje dziedzina zajmująca się tzw. wyjaśnialną sztuczną inteligencją (XAI, por. 2.2.2 i 6.2.1). Choć autorzy omawianego raportu oceniają tego rodzaju algorytmiczne rozwiązania pozytywnie, ponieważ świadczą one o zainteresowaniu technicznego środowiska akademickiego zaadresowaniem problemów uczciwości, obciążeń (*bias*) i dyskryminacji, to wskazują na ich trojakie ograniczenia. Po pierwsze, są to rozwiązania o charakterze szybkiej, nie prowadzącej do żadnych zasadniczych zmian naprawy (*fix*), mające „rozwiązać problem", „wyeliminować obciążenia", zatem wpisują się one w charakterystyczne dla firm technologicznych podejście (Crawford i in. 2018:29). Po drugie, poleganie na takich naprawach (*fix*), które z kolei bazują na „godnych zaufania" pracach akademickich może prowadzić do fałszywego przekonania o rzeczywistym rozwiązaniu problemu i w konsekwencji uzyskaniu prawdziwie sprawiedliwego, nieobciążonego systemu AI (Crawford i in. 2018:26). Po trzecie, autorzy powołując się na badania (Chouldechova 2017; Kleinberg 2018) wskazują, że różne kwantyfikowalne kryteria sprawiedliwości bazują na odmiennych, arbitralnych założeniach o tym, co jest, a co nie jest sprawiedliwe, a więc są wzajemnie sprzeczne. Zatem poprawa metryki jednego może powodować spadek innego (Crawford i in. 2018:26-27). Podsumowując, rozwiązania te projektowane są z pominięciem kontekstów społecznych, historycznych i politycznych, które sprawiają, że systemy AI, a szczególnie zbiory danych, na których są one uczone, odzwierciedlają, powielają i wzmacniają już istniejące wzory społecznego defaworyzowania czy dyskryminacji (Crawford i in. 2018:29). Podano wyraźny przykład stworzonego przez Amazon systemu wspierającego decyzje w procesie rekrutacji pracowników do tej firmy. Odkryto, że system, tworzący ranking osób kandydujących na stanowisko pracy, dyskryminuje kobiety. Przy kontroli innych zmiennych niższe wyniki w rankingu otrzymywały osoby, które w CV wpisały ukończenie żeńskiej szkoły wyższej (*all-women's college*), a także w ogóle wpisały do niego słowo „kobieta". Firma odkrywszy to zastosowała techniczne rozwiązania oparte na algorytmicznej sprawiedliwości (*algorithmic fairness*).

Naprawa skończyła się niepowodzeniem, ponieważ pracowano na danych zebranych w trakcie wielu lat rekrutowania do Amazona, a dane te odzwierciedlały utrwalone praktyki defaworyzowania kobiet w procesie rekrutacji. Ostatecznie system usunięto (Crawford i in. 2018:38-39). Słowami prawnika Franka Pasquale (2018) autorzy wskazali, że algorytmy same nie zapewnią odpowiedzialnego działania innych algorytmów (Crawford i in. 2018:27).

O takim przekonaniu dotyczącym możliwości rozwiązywania złożonych problemów społecznych za pomocą technologii pisano w sposób krytyczny już w 1995 r. Daniel Fox określił to jako „technokratyczny solucjonizm", co rozwinął Evgeny Morozov na początku drugiej dekady XXI wieku w publicystycznej pracy „To Save Everything Click Here: Technology, Solutionism and the Urge to Fix Problems That Don't Exist". Mowa tutaj nie tylko o technologiach bazujących na tzw. big data czy sztucznej inteligencji, ale także innych, np. o internecie (Carr 2013).

Drugą grupę strategii stanowią rozwiązania deklaratywne. Należą do nich kodeksy etyczne, kodeksy zasad itp., oraz certyfikaty lub zezwolenia. Zdaniem autorów wiele z nich powstało w odpowiedzi na skandale i rosnące obawy opinii publicznej o konsekwencje społeczne rozwoju AI. Jednak mają one jakoby niewielki wpływ na praktykę, a szczególnie zmniejszenie przepaści odpowiedzialności / rozliczalności (*accountability gap*), o ile nie są bezpośrednio wbudowane w proces tworzenia systemów (Crawford i in. 2018:9). Przykładowo omówiono siedem zasad przewodnich (*guiding principles*) społecznej odpowiedzialności AI w firmie Google, wykazując, że jedna z nich jest bezpośrednią odpowiedzią na niedawny skandal. Chodzi o protest pracowników Google'a wobec realizowania projektu „Maven" – zamówionego przez Pentagon systemu AI służącego inwigilacji za pomocą dronów. Firma, po protestach i rzecz jasna ujawnieniu sprawy w mediach, nie przedłużyła umowy z Pentagonem. W zasadach, które powstały po tym fakcie, zapisano, że Google „nie realizuje systemów AI obejmujących broń i inne technologie, których głównym celem lub implementacją jest spowodowanie lub bezpośrednie umożliwienia obrażeń ludzi" (Crawford i in. 2018:29-30). Google jako kolejny krok po przyjęciu tych siedmiu zasad przewodnich powołało także radę etyki technolgii, w tym sztucznej inteligencji (*AI ethics council*) o nazwie Advanced Technology External Advisory Council, w skład której zaproszenie przyjął m.in. cytowany wcześniej filozof Luciano Floridi (Floridi 2019b). Zajmuje się on niemal wyłącznie etyką technologii cyfrowych (Allo i in. 2016; Cath i in. 2018; Floridi 2006, 2012, 2014b, 2015, 2018, 2019a; Floridi i Taddeo 2016; Taddeo i Floridi 2018b; Turilli i Floridi 2009). Rada została rozwiązana po tygodniu istnienia,

ponieważ pracownicy Google zebrali ponad 2,3 tys. podpisów osób sprzeciwiających się włączeniu w skład ośmioosobowej rady Kay Coles James, prezeski Heritage Foundation. James dała się poznać amerykańskiej opinii publicznej jako konserwatystka związana z administracją Tumpa, działająca w tzw. ruchu antyaborcyjnym, będąca orędowniczką budowy proponowanego przez Trumpa muru przeciw-imigracyjnego na granicy USA z Meksykiem, oraz przeciwniczką „Aktu Równościowego (*Equality Act*)", nastawionego w zamyśle na ochronę przed dyskryminacją z powodu płci, tożsamości płciowej (*gender identity*) i orientacji seksualnej (Cicilline 2019; Levin 2019). Floridi w poście facebookowym wyjaśnił, że w całej rozciągłości nie zgadza się z poglądami James, wolałby, by nigdy nie nastąpiło zaproszenie jej do rady Google – a skoro już nastąpiło i zostało oprotestowane, to oczekiwałby jej usunięcia ze składu – i choć cenione przez niego osoby namawiały go do takiego gestu protestu, to nie zrezygnuje z pracy w radzie. Jego zdaniem, cały incydent świadczy o potrzebie podwojenia wysiłku na rzecz etyki technologii cyfrowych (Floridi 2019b).

Używamy określenia „rozwiązania deklaratywne", ponieważ wskazano, że firmy Google, Facebook, Microsoft czy Axon, co prawda wprowadziły kodeksy a także rady etyki czy etycznych doradców do swojej działalności w zakresie AI, czego skutkiem jest dostarczenie języka do dyskusji o problemach etycznych, a także możliwość oceny rezultatów działania firm przez opinię publiczną w odniesieniu do zawartych w kodeksach ogólnych stwierdzeń. Jednak właściwie nie wiadomo, co i czy coś można dzięki temu wyegzekwować (Crawford i in. 2018:30). Nie oznacza to, że kodeksy są zupełnie niepotrzebne. Jedną z rekomendacji wobec odpowiedzialności big data, które przygotowywali m.in. Boyd, Crawford i Pasquale jest tworzenie takich kodeksów (Boyd i in. 2017:6). Są one jednak dalece niewystarczające. Taka strategia, w której korporacje mówią: oto kodeks – teraz nam ufajcie, może być próbą uniknięcia regulacji prawnych (Crawford i in. 2018:30). Autorzy powołują się na pracę na Bena Wagnera. Wagner podaje przykład projektu Google DeepMind, którego pracownicy zespołu etyki wielokrotnie zapewniali publicznie o etyczności swojej firmy i swojego działania w sytuacji, gdy nielegalnie udostępnili dane o charakterze medycznym ponad półtora miliona osób – określali tę sytuację jako błąd rządu brytyjskiego, odżegnując się od jakiejkolwiek odpowiedzialności (Wagner 2018:84). Wskazuje on, że i politycy są niezdolni lub niechętni do dostarczania właściwych regulacji prawnych, a etyka jest postrzegana i przez nich, i przez firmy jako „łatwa" lub „miękka" opcja, która może pomóc w ustrukturyzowaniu i nadaniu znaczenia istniejącym inicjatywom

samoregulacji. Etyka to nowa „samoregulacja przemysłu" (ibidem). Ponadto, bazując na badaniu treści kodeksów etycznych AI (Greene, Hoffman, i Stark 2019) wskazano na zawarte w nich założenia i wizję świata. Zwrócono uwagę na wąską, ekspercką wizję sprawczości (*agency*) etycznej – decyzje o tym, jakie ma być etyczne AI mogą podejmować tylko i wyłącznie wykwalifikowani eksperci. Ponadto kodeksy mówią o benefitach i ryzykach związanych z AI w kontekście ludzkości jako takiej, nie zauważając żadnych mniejszości, których sytuacja może drastycznie się różnić. Kodeksy te są zatem wyrazem niekwestionowanego przekonania o nieuniknionym rozwoju AI, wzywają one po prostu do budowania lepszych systemów (Crawford i in. 2018:31). Zauważmy, że o potrzebie kwestionowania ekspansji AI pisał m.in. Andrzej Zybertowicz z zespołem, pytając:

> Czy są takie procesy myślowe albo typy interakcji międzyludzkich, których nie wolno nam automatyzować? Takie, które stanowią rdzeń człowieczeństwa, których naruszenie, przekształcenie oznacza kres ludzkiego świata, takiego jaki znamy? (Zybertowicz i in. 2015:232).

Omawiane wyżej obserwacje autorów z AI Now Institute zostały określone jak mało odkrywcze i nie zaskakujące, a także „uwikłane" – o czym poniżej – przez Sebastiana Benthalla, pracującego na Uniwersytecie Berkeley interdyscyplinarnego badacza z pogranicza informatyki, DS, nauk społecznych i filozofii. Benthall we wpisie na swoim blogu, powołując się m.in. na pogląd przywołanej wcześniej Luke Stark wskazuje, że omawiane problemy z rozwiązywaniem problemów etycznych są tak stare, jak kapitalizm i można je do kapitalizmu sprowadzić. Konsekwencje społeczne osiągnięć inżynierów również uznaje za problem niezmienny od czasu dziewiętnastowiecznej rewolucji przemysłowej. Autor mówi o panowaniu hegemonii neoliberalizmu, której wyrazem są, jego zdaniem, także poglądy i postawy reprezentowane przez AI Now Institute. Twierdzi, że autorzy raportu nie zauważają, iż problemy defaworyzowania przez firmy tworzące systemy AI pewnych mniejszości są wyrazem małej siły nabywczej tychże, co sprowadza się do tego, że firmom nie opłaca się inwestować w mniejszości. Wytyka on autorom tworzenie dwóch opozycji oni – my. Po pierwsze: (oni) potężne firmy i inżynierowie przeciw (my) liczni i różnorodni, poddawani działaniu firm ludzie. Po drugie: (oni) informatyka (*computer science*) przeciw (my) społecznie zaangażowane, humanistyczne dyscypliny. Z ciekawej argumentacji autora rozumiemy, że opozycje te są dla autorów wygodne, bo podmiot taki jak AI Now Institute jest tak samo „uwikłany" w kapitalizm, jak firma technologiczna: neoliberalny, elitarny i broniący swojej pozycji. Autor krytykuje arbitralną formułę omawianego raportu AI Now Institute i postuluje badania wpływu systemów AI na życie społeczne za pomocą populacyjnych badań

ekonomicznych, nastawionych na relacje ludzi, dóbr i kapitału. Mówi przy tym o niepopularności takich badań, bo nie wpisują się one w wygodną opozycję inżynierowie vs. humaniści. Zdaniem autora badacze społeczni postrzegają problemy generowane przez technologie jako problemy społeczne, bo nie potrafią i jednocześnie nie chcą postrzegać ich jako problemów ekonomicznych. Twierdzi on, że mówienie o problemach ekonomicznych jest tematem tabu w hegemoni neoliberalizmu, a dodatkowo badacze społeczni nie chcą pokazać swojej niemocy w tym zakresie i przestać być postrzegani jako eksperci (Benthall 2018).

W argumentacji tego autora widzimy echa pretensji do obiektywności szeroko zakrojonych badań ilościowych, jednak postrzeganie problematyki społecznych konsekwencji AI / big data / DS jako – także – ekonomicznej uznajemy za cenne poznawczo. Staramy się w niniejszej pracy uwzględniać tę perspektywę, ale przyjmujemy za Pierrem Bourdieu, że pieniądze i szerzej zasoby materialne są tylko jednym z czterech rodzajów kapitału[41] i unikamy podejścia akcentującego racjonalność ekonomiczną. Bourdieu ukazywał, że pozornie – tj. z punktu widzenia racjonalności ekonomicznej – nieracjonalne praktyki nie są uwarunkowane np. tradycją czy działaniem pod wpływem emocji, a są właśnie racjonalne, zgodne z logiką interesów kulturowych czy symbolicznych (Bourdieu 1977, 1986). Kończąc wątek krytyki podjętej przez Benthalla, zauważamy, że dość trafnie wskazuje on niespójność argumentacji autorów omawianego raportu. Mówią oni o pracownikach firm technologicznych jako o oderwanej od problemów etycznych elicie, zaś w innym miejscu tekstu pracownicy owych firm protestują przeciwko realizacji nieetycznych projektów (Benthall 2018).

2.3.2.6. Pozytywne konsekwencje społeczne tzw. AI

W całym raporcie AI Now Institute nie znajdujemy żadnych przykładów pozytywnych konsekwencji rozwoju big data / DS / sztucznej inteligencji. Stwierdzono jedynie, że systemy bazujące na AI „wciąż mają wiele do zaoferowania" (*still offer considerable promise*) (Crawford i in. 2018:42). Na potencjał pozytywnych konsekwencji AI wskazują Floridi i Taddeo, jednak ich praca jest apelem o etyczność w projektowaniu, użyciu i regulacjach prawnych sztucznej inteligencji, a nie stricte analizą konsekwencji. Wyrażają oni wyłącznie pragmatycznie pojęte korzyści: obniżanie kosztów, zmniejszanie ryzyka, poprawę zwięzłości

41 Są to: 1) kapitał ekonomiczny – pieniądze i przedmioty materialne; 2) społeczny – pozycje i relacje w w grupach społecznych; 3) kulturowy – umiejętności, zwyczaje, nawyki, style językowe, rodzaj ukończonych szkół, styl życia i gust; 4) symboliczny – wykorzystywanie symboli do uprawomocnienia posiadania pozostałych typów kapitału (Bourdieu 1986).

i niezawodności (*consistency and reliability*), i ogólnie umożliwienie nowych sposobów rozwiązania złożonych problemów. Takie korzyści mogą, ich zdaniem, następować pod warunkiem wykonywania nastawionego na etyczność skoordynowanego wysiłku społeczeństwa obywatelskiego, polityków, biznesu i akademii (Taddeo i Floridi 2018a). Nie znajdujemy również innych prac z dziedziny nauk społecznych, które poświęcone byłyby pozytywnym konsekwencjom. Przypomina to o przywołanej w części 1.1.2. obserwacji Knapik, która wskazuje, że w popkulturze obrazy zagłady ludzkości spowodowanej przez AI zdecydowanie dominują nad innymi wizjami relacji człowieka ze sztuczną inteligencją (Knapik 2018:11). W naukach społecznych i humanistyce jest, naszym zdaniem, podobnie. Może dominuje następujące podejście:

> Ponieważ na ogół korporacje i przedsiębiorcy, którzy przodują w technicznej rewolucji, sami naturalnie wychwalają własne wytwory, zadanie bicia na alarm i wyjaśniania wszelkich możliwych czarnych scenariuszy przypada w udziale socjologom, filozofom i historykom – takim, jak ja (Harari 2018:10).

Szereg pozytywnych konsekwencji społecznych widoczny jest w pracach entuzjastów rozwoju omawianych dziedzin, co pokazaliśmy w częściach 2.2.1. i 2.2.2. Bardzo upraszczając, mówi się o możliwości uzyskania obiektywnej wiedzy, która pomoże rozwiązać najbardziej palące problemy ludzkości. Z kolei w części 2.2.3. zauważalna jest nadzieja podmiotów tworzących, wdrażających i kupujących systemy oparte na AI na pragmatycznie pojęte pozytywne konsekwencje, a właściwie korzyści: zwiększenie swojej skuteczności, efektywności, skali, szybkości działania, redukcji kosztów, zwiększeniu wygody. Tym samym w całej części 2.2.3. moglibyśmy do niemalże każdego argumentu krytycznego dostarczyć szereg kontrargumentów. Nie zrobiliśmy tego, ponieważ naszym celem było omówienie prac powstających na gruncie nauk społecznych. W analizie kodeksów etycznych w dziedzinach zajmujących się AI wskazano, że dominuje tam myślenie o ogromnym potencjale sprawczym sztucznej inteligencji. Wyraża się ono w przekonaniu, że zarówno pozytywny, jak i negatywny wpływ AI jest powszechnym, uniwersalnym zmartwieniem (*the positive and negative impacts of AI are a matter of universal concern*) (Greene i in. 2019).

Z naszych badań wynika, że w świecie społecznym DS podzielane jest takie podejście. Rozwój technologiczny postrzega się jako nieunikniony, jest on imperatywem, dotknie każdej dziedziny życia, a szczególnie technologie tzw. AI będą miały coraz to większe znaczenie. Technologie, w tym tzw. AI czy – technicznie rzecz ujmując – uczenie maszynowe,

postrzegane są raczej jako neutralne narzędzia w ludzkich rękach. Może to akcentować ludzką odpowiedzialność za cel używania technologii:

> Technologia jest **niewinna** (*innocent*, podkr. org.). To data scientist ustawia dane wejściowe (*input*) i zadaje modelowi zmienną celu. Pojawią się wzory, a niektóre z nich mogą być dla wielu ludzi krzywdzące. Musimy być świadomi ostatecznego celu, jak w każdej innej technologii (Casas 2018).

W polskim świecie DS pojawiają się porównania do noża / młotka, którymi można przygotować posiłek / wbić gwóźdź lub zabić człowieka. Rozwijamy to zagadnienie w częściach poświęconych wynikom badań własnych.

2.2.4. Data science jako badany wycinek rzeczywistości społecznej

W latach 2017 – 2018 powstały cztery artykuły z dziedziny nauk społecznych dotyczące – przyjemniej częściowo – DS jako badanego wycinka rzeczywistości społecznej. Choć, jak wskazywaliśmy wyżej, krytyka epistemologiczna, metodologiczna czy dotycząca konsekwencji społecznych rozwoju technologii związanych z DS zaczęła się około roku 2010, to do 2017 r. nie znajdujemy żadnych opublikowanych badań empirycznych skoncentrowanych na ludziach zajmujących się DS.

Mówimy o czterech tekstach:

- na problemie traktowania algorytmów jako fetyszy koncentruje się tekst bazujący na częściowo ustrukturyzowanych wywiadach z osobami zawodowo zajmującymi się wizją komputerową (*computer vision*) w USA i Azji, oraz trzyletnich badaniach etnograficznych osób należących do społeczności zajmującej się monitorowaniem i poprawą parametrów własnego ciała (*quantified self,* QS) (Thomas i in. 2018);
- inny tekst mówi o praktykach odróżniania DS od statystyki publicznej, której to DS ma być odłamem (*fraction*), oraz włączaniu praktyk DS / big data do statystyki publicznej. Bazowano na badaniach etnograficznych urzędów i organizacji statystyki publicznej w kilku krajach europejskich (Grommé, Ruppert, i Cakici 2018);
- w tej samej książce znalazł się najbardziej dla nas interesujący tekst poświęcony stawaniu się prawdziwym data scientistem (*becoming a real data scientist*), przygotowany na podstawie badań etnograficznych środowiska data scientistów w Rosji, związanych z nowo powstałą jednostką DS na jednej z uczelni wyższych, we współpracy z firmą Yandex (Lowrie 2018);

- na podstawie tych samych badań Ian Lowrie przygotował też tekst bliższy socjologii nauki i techniki, poświęcony relacjom osób zajmujących się DS z algorytmami i tworzenia tzw. algorytmicznej racjonalności (Lowrie 2017).

W pracach, które tutaj przedstawiamy, deklarowano podejście etnograficzne. Niemniej nigdzie nie znalazły się dokładniejsze wyjaśnienia dotyczące metodologii, ani podstawowej ramy teoretycznej. W żadnej prac nie ma odniesień do teorii społecznych światów, zatem nasza rozprawa może być pierwszym zastosowaniem tej ramy do badania DS.

Pierwszy z wymienionych tekstów (Thomas i in. 2018) podkreśla asymetrię pomiędzy osobami tworzącymi algorytmy (u nich są to specjaliści wizji komputerowej) a użytkującymi algorytmy (społeczność QS). Rozróżnienie na twórców i użytkowników może być rozmyte lub ostre. Mowa o asymetriach społecznych – status, eksperckość (*expertise*), społeczność (*community*), kulturowych – praktyki, rytuały, wiedza, oraz ekonomicznych – bowiem różna jest rola twórców i użytkowników wobec algorytmów i danych jako zmaterializowanych artefaktów (czyli jak rozumiemy tego, za co można otrzymać pieniądze) (Thomas i in. 2018:9).

O tym samym wielokrotnie pisze AI Now (Crawford i in. 2018:7, 33–34). Thomas z zespołem zauważają jednak, że asymetrie występują także po stronie twórców całych systemów zawierających algorytmy – pomiędzy specjalistami tworzącymi same algorytmy (*algorithm developers*) a twórcami oprogramowania (*software developers*), „obudowy" dla algorytmów, a także (o czym mówi Craford z zespołem) pomiędzy osobami zajmującymi się rutynowymi zadaniami przygotowania danych do dalszych czynności a specjalistami tworzącymi algorytmy (Thomas i in. 2018:5–6). Chociaż określenie „twórcy algorytmów" odnosi się jedynie do luźno traktowanego podzbioru tego, co nazywamy w naszej rozprawie światem społecznym DS, to zdecydowanie zgadzamy się z wyraźnym odróżnieniem osób zajmujących się DS od tych zajmujących się szeroko pojętą informatyką (*computer science*) czy tworzeniem oprogramowania (*software engineering / development*). Trudno nam jednoznacznie zgodzić się z asymetrią tych pozycji, niemniej badany przez nas świat chętnie definiuje się w opozycji do tzw. informatyków czy programistów, już na etapie określenia działania podstawowego (por. część 5.2. tej pracy).

Autorzy nazywają algorytmy fetyszami, ponieważ ludzie (trzymając się zaproponowanej asymetrii – tak samo twórcy jak i użytkownicy) przyznają algorytmom możliwości nie wynikające ani z ich podstaw matematycznych, ani z ich zapisu w kodzie.

Algorytmiczna obietnica jest zatem znacznie większa niż możliwości algorytmów (Thomas i in. 2018:9). Autorzy twierdzą, że ten sprawczy, demoniczo-boski status algorytmów przekłada się na praktyki wysokiego wyceniania i doceniania pracy twórców algorytmów (*algorithm developers*), przy nie dostrzeganiu pracy będących jakoby w opozycji użytkowników (ibidem). Słusznie sugerują oni, że ta opozycja jest o tyle fałszywa, że algorytmy działają i są ulepszane dzięki danym tzw. użytkowników końcowych. Przypomina to przywoływane przez nas ujęcie użytkowników jako współtworzących produkt prosumentów, a także ujęcie użytkowników jako produktu (Surma 2017:39, 66–67). W dalszej części pracy przedstawiamy, że w badanym przez nas świecie społecznym DS i na przecięciach z innym światami występuje zjawisko, dla którego opisu kategoria fetyszu jest użyteczna. Mowa o fetyszyzowaniu narzędzi lub metod – stają się one (z różnych powodów, do różnych celów i dla różnych osób) ważniejsze niż rezultat ich zastosowania.

W drugim z wymienionych tekstów znajdują się licznie odwołania do P. Bourdieu. Autorzy opisują tworzenie się figury data scientista jako tworzenie się pola profesji (*professional field*). Powstawanie tego pola wiążą oni ze zmianami sposobów oceny metod, technologii, ekspertyzy (*expertise*), umiejętności, wykształcenia i doświadczenia, szczególnie w porównaniu go do pola statystyki publicznej. Jako przedmiot walki (*struggle*) pokazują oni wysiłki ustalania ram (*framing*) dyskusji – czy DS zagraża istnieniu statystyki publicznej, czy też nie zagraża. Wytworzenie się nazw „data science" i „big data" uważają oni za formę materialno-semiotycznej praktyki, która jest, według nich, przemocą symboliczną, a demonstruje współzawodnictwo metod określania prawdy – jak rozumiemy chodzi z jednej strony o metody oficjalnej statystyki, z drugiej o DS (Grommé i in. 2018:37–38).

Statystycy poświęcają, zdaniem autorów, wiele wysiłku na definiowanie profesji data scientistów i na definiowanie swojej do nich relacji, co prowadzi do redefiniowania profesji statystyka i obiektów jej działania, czyli statystyki publicznej/oficjalnej. Mówią oni o podobnym, np. w wymiarze edukacji, habitusie statystyków i data scientistów, ale różnym w sensie kapitału kulturowego. Wskazano, że data scientiści pracują z mniej uporządkowanymi danymi niż statystycy (np. danymi z mediów społecznościowych); używają raczej metod algorytmicznych niż statystycznych; bardziej cenią i więcej uwagi poświęcają wizualizacji danych. W komunikacji wyników data scientiści mówią o wartości biznesowej, o eksperymentowaniu i pracy metodą prób i błędów. Statystycy koncentrują się na odpowiedzialności zawodowej, ponieważ dostarczają „oficjalnych" wyników dla władzy i opinii publicznej. Jednocześnie istnieje trend polegający na włączaniu do statystyki publicznej

źródeł danych, tzw. big data, metod algorytmicznych i narzędzi charakterystycznych dla DS. Autorzy uważają, że jest to także redefiniowaniem profesji statystyka publicznego, szczególne poprzez włączanie elementów habitusu DS do habitusu statystyka. Mówią także o tworzącej się nowej profesji, iStatistician (Grommé i in. 2018:45–50).

Wśród statystyków publicznych jakoby podkreślana jest rola źródeł danych, zwanych big data. Trwają spory, czy i w jakim zakresie te źródła danych mogą wzbogacić lub zastąpić spisy powszechne czy sondaże. Naszym zdaniem Grommé z zespołem przyjęli tę perspektywę od uczestników swoich badań. Prowadzi ich to do stwierdzenia, już na samym początku artykułu, że data scientiści to profesja postrzegana jako realizująca potencjał big data, czyli nowych źródeł danych (Grommé i in. 2018:33).

Z naszych badań wynika, że w polskim świecie DS używanie terminu big data w sensie innym niż techniczny (związany z technologiami typu Hadoop i Spark, o czym w części 5.3.) jest uznawane za przejaw słabej orientacji w temacie i jest jasnym wskaźnikiem bycia spoza świata DS. Nie krytykujemy autorów omawianego tekstu. Wskazujemy, że w świecie społecznym DS określenie „big data" ma inne niż u nich znaczenie. Przy tym w badanym przez nas świecie osoba zajmująca się DS w żadnym razie nie jest zawsze jednocześnie osobą zajmującą się big data (w sensie technicznym). Mówienie o ilości czy nowych źródłach danych jest jedną z praktyk legitymizacji działania badanego świata (por. 6.1.2. tej rozprawy).

Trzeci tekst, najbardziej dla nas interesujący, omówimy tutaj krótko, gdyż odnosimy się do niego kilkakrotnie w empirycznej części naszej rozprawy – rozdziałach 5. i 6. Lowrie w odróżnieniu od ww. autorów sytuuje swoje badanie w Moskwie (Lowrie 2018:63), jako konkretnym czasie, miejscu i kulturze, zaś Grommé z zespołem (2018:35-36) pisali o polach międzynarodowych (*transnational*), w oderwaniu od lokalnych kontekstów. Bardzo dużo w tekście Lowrie'go jest mowy o matematyce. Twierdzi on, że jest ona postrzegana jako podstawa dla innych umiejętności data scientistów, jako podstawa niezmienna (w odróżnieniu od konkretnych rozwiązań technologicznych), jako to, co uczy ścisłości myślenia i właściwego rozwiązywania problemów niezależnie od dziedziny, jako uniwersalny język porozumiewania się data scientistów. Posiadanie solidnego przygotowania matematycznego oraz znajomość najnowszych technologii do pracy z danymi są konieczne, by zajmować się DS, ale to nie czyni prawdziwego data scientista. W świetle wyników autora prawdziwy data scientist to m.in. taki, który jest bystry (*clever*) i ma dobre podejście do rozwiązywania problemów (*good approach to solving problems*) (Lowrie 2018:63). Innym kryterium bycia prawdziwym data scientistem jest rozwiązywanie prawdziwych problemów – to odróżnia data

scientistów od matematyków i informatyków (*computer scientist*) (Lowrie 2018:71-72). Niemniej prawdziwy data scientist może zarówno pracować komercyjnie, jak i akademicko – informator autora wskazał, że „wszyscy jesteśmy data scientistami, wszyscy myślimy tak samo" (Lowrie 2018:79).

Pojęcie prawdziwego czy autentycznego uczestnika jest typowe dla teorii społecznych światów, odwołujemy się do niego w części 6.

> Dyskusja dotyczy także tego, kto jest autentycznym uczestnikiem świata, jaki powinien być „prawdziwy uczestnik", jakie cechy powinien posiadać, jakie wartości „powinny" mu przyświecać? W ten sposób społeczny świat zaczyna pełnić funkcję grupy odniesienia normatywnego, która wyznacza jego członkom nie tylko to, co jest ważne i wartościowe, jak należy działać i w jaki sposób dokonywać oceny czyjegoś (i swojego własnego) działania, ale także, poprzez generowanie wspólnych perspektyw działających aktorów (Kacperczyk 2016:17–18).

Wyniki prezentowane w omawianym tekście są częściowo zbieżne z wynikami naszych badań, m.in. w kwestii przynależności do świata społecznego DS, tak samo uczestników pracujących komercyjnie, jak i naukowo. Twierdzimy, że sam fakt pracy w firmie lub na uczelni nie ma znaczenia w perspektywie uznawania się i bycia uznawanym za uczestnika badanego świata (choć określenie data scientist jest raczej nazwą stanowiska pracy w podmiocie komercyjnym niż nazwą każdego uczestnika badanego świata, por. część 5.2.2.). Lowrie zwraca także uwagę na uznawanie uczenia się przez całe życie za element określający prawdziwego data scientista. My proponujemy ujęcie tego problemu w kategoriach samorozwoju jako wartości badanego świata (część 5.4.). Podkreślmy, że zapoznaliśmy się z tekstami Iana Lowrie (jak i dwoma innymi tutaj omawianymi) już po zakończeniu badań i analiz własnych.

Inny tekst tego autora bazuje na powyższej tezie o braku zasadniczych różnic między data scientistami w komercji i akademii oraz ją rozwija. Zbiorowa praktyka intelektualna data scientistów tworzy, zdaniem Lowrie, spójną formę racjonalności algorytmicznej, nieredukowalną do konglomeratów praktyk lub skrawków teorii zapożyczonych z matematyki lub informatyki. Ta racjonalność ma być nierozerwalnie związana z pojawieniem się infrastruktury dużych zbiorów danych i głęboko zakorzeniona w obliczeniowej nowoczesności (*computational modernity*) (Lowrie 2017:1). Jak widać autor ten także jest pod wpływem tego, co wyżej nazwaliśmy praktyką legitymizacji działań świata DS (por. część 6.1.2.1. tej pracy – pojawienie się dużej ilości nowych typów danych i rozwój mocy obliczeniowej), co rzecz jasna nie znaczy, że go krytykujemy.

Jego zdaniem oddzielenie nauki od technologii (co jest tożsame z rozróżnieniem na zajmowanie się DS od strony akademickiej a od strony komercyjnej) w DS nie jest możliwe, bo algorytmy można badać tylko pośrednio, jako część systemów technologicznych. Algorytmy, w odróżnieniu od innych części matematyki nie są badane i rozwijane za pomocą dowodów, które można wykonywać wyłącznie za pomocą ludzkiego intelektu, z pomocą papieru i ołówka. Można badać je tylko pośrednio, za pomocą rozmaitych miar efektywności, i do tego muszą one być zaimplementowanie w języku programowania, na hardware z odpowiednią mocą obliczeniową, oraz uruchomione na zbiorze danych. Algorytm bez konkretnego zapisu w postaci kodu w języku programowania i bez konkretnego zbioru danych, na którym jest walidowany pozostaje neutralny (*inert*). Koncentracja na redukowaniu kosztów i uzyskaniu wyniku prowadzi do pozornie poprawnej konkluzji, że DS to dziedzina techniki czy inżynierii, a nie nauki. Lowrie, powołując się na Jean-Françoisa Lyotarda, wskazuje, że w inżynierii dominuje kryterium efektywności, zaś w nauce kryterium prawdy. Ma to pasować do DS w tym sensie, że działanie jest w DS oceniane jako dobre, kiedy daje lepszy wyniki w sensie utylitarnym. Nie ma mowy o ocenianiu wyniku w kategoriach prawdy. DS zmieniło technologiczne kryterium efektywności w standard epistemologiczny, w służbie nowej formy badań naukowych nad algorytmami. Połączenie nauki i inżynierii pod hasłem logiki efektywności jest jakoby najsilniej widoczne w biznesowych zastosowaniach DS. Optymalizacja działania algorytmu jest optymalizacją biznesu – zmniejszeniem odpływu klientów czy zwiększoną wartością koszyków zakupowych. Mogą być to optymalizacje rzędu 0,5 pp. (np. zwiększenie dopasowania modelu z 94% do 94,5%), dla dużych podmiotów komercyjnych przekłada się to nierzadko na miliony dolarów zysku (Lowrie 2017:3–9).

Za cenne w powyższym tekście uważamy zwrócenie uwagi na kategorię efektywności, którą opisujemy jako centralną wartość w badanym świecie społecznym (część 5.4. tej pracy). Niemniej nie zgadzamy się z ujęciem Lowrie'go w tym wymiarze, że traktuje on DS jako coś homogenicznego, i mówi o podejściu data scientistów jako takich do kolejnych, interesujących go zagadnień. Każde z tych zagadnień może być w naszej perspektywie areną sporu subświatów w DS i światów przecinających się z DS.

Zdecydowanie zgadzamy się jednak ze stwierdzeniem, że

> choć od 2013 r. **wiele pisze się w humanistyce i naukach społecznych o algorytmach i danych, to bardzo mało uwagi poświęcono osobom, które budują te epistemologiczne ramy** [podkr. RŻ] poprzez rozwój i wdrażanie określonych technologii (Lowrie 2017:2).

Te osoby to data scientists.

Rozdział 3. Ramy teoretyczne rozprawy doktorskiej

Jako najważniejsze teorie socjologiczne dostarczające sposobów konceptualizacji wybranego obszaru badania i aparatu pojęciowego do badania takiego rodzaju całości społecznej przyjmujemy tutaj teorie społecznych światów w ujęciach Tamotsu Shibutaniego, Anselma L. Straussa, Howarda S. Beckera, a przede wszystkim teorię społecznych światów/aren Adele E. Clarke. Koncepcja Clarke jest przez nas traktowana, zgodnie z propozycją tej autorki, jako części pakietu teorii/metod badań, co rozwijamy poniżej i dalej w rozdziale 4. Inspiracją są dla nas także studia nad nauką i techniką (STS), z klasycznym, oryginalnym podejściem Bruno Latoura na czele. Korzystamy w dużej mierze z publikacji przygotowanych przez pracowników Instytutu Socjologii UŁ, którzy wprowadzili do polskiej socjologii metodologię teorii ugruntowanej – Krzysztofa T. Koneckiego oraz teorie społecznych światów – Annę Kacperczyk.

Agnieszka Kretek-Kamińska wskazuje, że około 1/3 spośród analizowanych przez nią empirycznych, socjologicznych prac doktorskich z lat 1993-2013 nie udało się przyporządkować ani do paradygmaty normatywnego, ani do interpretatywnego (2016:227). Autorka stwierdza, że

> kategorie paradygmatu normatywnego i interpretatywnego w zasadnie nie były obecne w rozprawach. (…) Na ogół autorzy badanych prac nie odwoływali się bezpośrednio do tych właśnie sformułowań, choć w dalszej części rozprawy opisywali ich założenia jako podstawę prowadzonych działań badawczych (Kretek-Kamińska 2016:222).

W naszej rozprawie deklarujemy usytuowanie pracy w paradygmacie interpretatywnym, o czym więcej w części 4.

3.1. Teorie społecznych światów w ujęciach Tamotsu Shibutaniego, Anselma L. Straussa, Howarda S. Beckera i Adele E. Clarke

Za pierwsze użycie pojęcia „świat społeczny” (mniej więcej w omawianym sensie) traktuje się pracę Paula Cressey’a *The Taxi Dance Hall…* Cressey uznał środowisko określonego rodzaju klubów tanecznych za „odrębny społeczny świat”, z „własnymi sposobami działania, mówienia, myślenia”, posiadający „własny słownik, sobie właściwe działania i interesy, własną koncepcję tego, co jest znaczące w życiu, i do pewnego stopnia

własny schemat życia" (Cressey, 1932:31, za: Kacperczyk 2016:31; Konecki 2010:31). Sens bliski temu pojęciu dla Anselma L. Straussa czy Adele E. Clarke zauważyć można u Georga Simmela. W pracy *The Web of Group Affiliation* z 1955 r. wprowadził on określenie „krąg społeczny" (*social circle*) dla wielkomiejskich całości społecznych, charakteryzujących się koncentracją wokół jakiegoś specjalnego zainteresowania, woluntaryzmem, brakiem związku z bliskością przestrzenną. Było to ujęciem odmiennym niż u Cressey'a, który jednoznacznie łączył społeczny świat z bliskością przestrzenną ludzi, z konkretnymi miejscami na mapie miasta (Unruh 1980:273–74).

Rama teoretyczna społecznych światów ma swoje korzenie w tzw. szkole chicagowskiej, w której socjologowie bazując na myśli George'a H. Meada i Johna Dewey'a praktykowali badania terenowe niewielkich grup społecznych (jak enklawy etniczne) lub przestrzeni miejskich (jak wspomniane kluby taneczne). Badania takich całości społecznych (*social wholes*), jak nazwał je W. I. Thomas w 1914 r., nastawione były na wgląd w interakcje, w praktyki wspólnego wytwarzania znaczenia i działania na podstawie tego znaczenia, także wgląd w tożsamości. Stosowano podejście ekologicznie w sensie osadzenia badania w naturalnym środowisku badanej całości społecznej, a zatem w określonym miejscu i czasie. Kontynuacja i rozwój tego podejścia po II wojnie światowej doprowadziły do syntezy zainteresowań badaniem znaczenia z badaniem tożsamości, co zdaniem Clarke i Star stało się socjologią jednocześnie materialną i symboliczną, interaktywistyczną, procesualną i strukturalną. Powołując się na pracę Jamesa Carey'a (Carey 2002:202) wskazują one, że Anselm Strauss wynalazł wówczas socjologię strukturacji, lata przed wymyśleniem terminu „strukturacja" przez Anthonego Gidensa. Takie „strukturacyjne" ujęcie struktury jako wyłaniającej się (*emergent*) było, ich zdaniem, znaczące dla sformułowania teorii społecznych światów przez Straussa. Od końca lat 50. do końca 80. teorię społecznych światów rozwijali przedstawiciele szkoły chicagowskiej: A. Strauss, T. Shibutani, H. S. Becker. Autorki wskazują, że do lat 80. przedstawiciele symbolicznego interakcjonizmu, przyjmujący perspektywę społecznych światów, koncentrowali się na badaniach problemów społecznych, sztuki, medycyny, prac (*occupations*), zawodów (*professions*). Później wzrosło zainteresowanie badaniem nauki i techniki (*Science and Technology Studies*, STS) z zastosowaniem teorii społecznych światów, a jedną z pierwszych pracy były badania uczennicy Straussa – A. E. Clarke z 1987 r. W końcu XX w. badania tego rodzaju zaczęły uwzględniać infrastrukturę materialną i narzędzia świata społecznego, nazywane aktorami nie-ludzkimi (*nonhuman*), co czynili m.in. B. Latour i S. L. Star (Clarke i Star 2008:113–15).

Sądzimy, że zastosowanie w naszej rozprawie podejścia wzorowanego szczególnie na A. E. Clarke, czyli teorii społecznych światów/aren, analizy sytuacyjnej i strategii ugruntowanego teoretyzowania, będzie właściwe dla badania DS. Data science stanowi bowiem chaotyczne i dynamiczne połączenie m.in. nauki, techniki, pracy, zawodu. Jest przedmiotem sporów i dyskusji wewnątrz (jak np. wysiłki definiowania DS i jego specjalizacji; określanie działania podstawowego; zalety i wady technologii do wykonywania działania podstawowego i działań wspomagających) i na zewnątrz „środowiska" (np. praktyki rekrutowania, zatrudniania, wynajmowania, zlecania pracy jednostkom/zespołom/firmom DS; ostrzeżenia przed konsekwencjami społecznymi, regulacje prawne). Data science jest światem nierozerwalnie związanym z infrastrukturą techniczną i narzędziami pracy, dlatego zależy nam na ujęciu aktorów nie-ludzkich.

Zdaniem Anselma Straussa (1978:122) najważniejszą cechą definicyjną świata społecznego jest istnienie jednej uderzająco ewidentnej działalności podstawowej (*primary activity*). Ta działalność jest oczywista i staje się kryterium wyodrębnienia świata społecznego (Kacperczyk 2016:31). Skoro zatem działalność podstawowa data scientists jest dość oczywista – polega na komputerowych operacjach na danych cyfrowych[42] – oraz skoro niezwykle trudne jest wykazanie innych kryteriów pozwalających na wyróżnienie DS jako całości społecznej, to decydujemy się przyjąć założenie o możliwości wydzielenia takiej całości społecznej jak DS na podstawie jej działania podstawowego. Tym samym **zakładamy, że data science jest społecznym światem** w rozumieniu wyżej wymienionych autorów. Ponadto opisane w części 1.1.4. praktyki autodefiniowania DS jako całości – „dziecka" wyrastającej z „rodziców", tzn. bardziej dojrzałych dziedzin nauki i praktyki, lub całości znajdującej się na przecięciu tych dziedziny (przykładowo statystyki i matematyki, informatyki, biznesu; por. rys. 2) traktujemy jako wyraźne markery do zastosowania pojęcia społecznego świata. Zarówno pojęcie społecznego świata, jak i cały zestaw pojęć w ramach koncepcji społecznych światów traktujemy jako pojęcia uwrażliwiające (*sensitizing concepts*) w rozumieniu Herberta Blumera. Takie podejście proponują A. E. Clarke i S. L. Star (2008:118) oraz A. Kacperczyk (2016:608).

W myśl Straussa można mówić o świecie społecznym jako uniwersum dyskursu, ale nie jest to jedyna możliwość. Nie należy ograniczać się do patrzenia jedynie na formy komunikacji czy tworzenia symboli (*symbolization*). Badane mogą być też elementy

42 Jak wskażemy później, takie określenie działania podstawowego społecznego świata data science nie jest zupełnie poprawne – jest za szerokie i nie wskazuje na sposób wykonywania działania podstawowego, co jest w tym przypadku ważne. Zajmujemy się tą kwestią w części 5.2.

namacalne (*palpable*), jak podejmowana aktywność, członkostwo, miejsca, technologie czy organizacje charakterystyczne dla określonego świata społecznego (Strauss 1978:121). Tym samym jako świat społeczny można traktować różnorodne wycinki rzeczywistości społecznej:

> Z pozoru można dostrzec niezliczone światy [społeczne]: operę, baseball, surfing, zbieranie znaczków, muzykę country, homoseksualizm, politykę, medycynę, prawo, matematykę, naukę, katolicyzm.... Niektóre światy są małe, inne ogromne; niektóre są międzynarodowe, inne lokalne. Niektóre są nierozłączne z daną przestrzenią; inne są powiązane z miejscami, ale znacznie mniej identyfikowalne przestrzennie. Niektóre są publiczne i upubliczniane; inne ledwie widoczne. Niektóre są tak wyłaniające się (*emerging*), że ledwo możliwe do uchwycenia (*barely graspable*); inne są dobrze osadzone (*established*), a nawet dobrze zorganizowane. Niektóre mają stosunkowo szczelne granice; w innych granice są przepuszczalne. Niektóre są bardzo hierarchiczne; inne mniej lub wcale. Niektóre są wyraźnie związane z klasą społeczną, inne (jak baseball) biegną w poprzek klas. Należy pamiętać, że działania i komunikacja w tych światach koncentrują się wokół różnych spraw intelektualnych, zawodowych, politycznych, religijnych, artystycznych, seksualnych, rekreacyjnych, naukowych; to znaczy, że światy społeczne są charakterystyczne dla każdego obszaru merytorycznego (*substantive area*) [tłumaczenie własne RŻ] (Strauss 1978:121-22).

Korzystając z przystępnego wyjaśnienia Anny Kacperczyk, przybliżamy podstawowe pojęcia i procesy w teorii światów społecznych. Pojęciem kluczowym dla tej teorii jest działanie podstawowe.

> Serce świata społecznego stanowi podstawowe działanie (*primary activity*). Wokół niego koncentruje się cała energia uczestników. Stanowi ono także kryterium wyodrębnienia społecznego świata jako pewnej całości, bo jego uczestnikami są aktorzy podejmujący takie działanie (Kacperczyk 2016:34).

Za Straussem (1993:216) autorka podaje także określenie „działania podstawowego" jako *core activity* (Kacperczyk 2016:50).

Działaniu podstawowemu towarzyszą działania wspomagające w skali makro i mikro. „W perspektywie makrospołecznej, podstawowe działanie obrasta całą siecią działań pomocniczych, służących przetrwaniu i ekspansji społecznego świata jako całości" (Kacperczyk 2016:35). Strauss wyróżnił 14 subprocesów pomocniczych wobec działania podstawowego, m.in. finansowanie (*funding),* reklamowanie (*marketing*), budowanie organizacji (*organizational building*). Skala mikro to działania wspomagające działanie podstawowe pojedynczego aktora społecznego. Jego „działanie podstawowe zanurzone jest w innych podtrzymujących i umożliwiających je działaniach" (Kacperczyk 2016:35; Strauss i Corbin 1990:236).

Za istotną cechę definicyjną społecznego świata uznawana jest jego technologia – pojęcie szczególnie ważne dla niniejszej rozprawy. Oznacza ona:

> wykorzystanie określonych środków technicznych, umożliwiających specyficzny sposób wykonania działania. Bez względu na to, czy technologia zostaje 'odziedziczona', czy stanowi innowacyjny sposób prowadzenia podstawowego działania, zawsze jest uwikłana w budowanie i podtrzymywanie społecznego świata, wiedzę o stosownych miejscach i okolicznościach, w których można podejmować dane działanie, oraz o specjalistycznym sprzęcie, jakim 'należy' dysponować, aby wykonywać je prawidłowo (Kacperczyk 2016:36-37).

Na technologie społecznego świata niejako „powołują się" procesy legitymizacji i zaświadczania o autentyczności. Rozwój technologii społecznego świata może prowadzić do procesu segmentacji czy jej „odmiany" – profesjonalizacji pewnych subświatów, o czym za chwilę. Za pomocą opisu zmian technologii w czasie można dokonać także opisu tzw. areny świata społecznego. Zmiany technologii ujawniają interpretacyjne zmagania nad znaczeniami (Clarke i Casper 1996:615).

Pojęcie granic społecznego świata jest wielce problematyczne z punktu widzenia socjologów krytykujących te koncepcje. Według Kacperczyk (2016:37) brak wyraźnych granic społecznych światów jest uznawany za najpoważniejszy problem związany z ich konceptualizacją. Strauss (1993) uważa, że problem podnoszą badacze przyzwyczajeni do ujmowania całości społecznych (*social units*) jako odgraniczonych czy wydzielonych od innych. W przypadku światów społecznych nie da się takich granic dookreślić - ta relatywna płynność granic jest jedną z podstawowych właściwości światów społecznych (1993:212-213). Świat społeczny nie jest sprowadzalny do grupy czy organizacji; „przenika [on] liczne organizacje formalne i układy społeczne, krzyżuje się z nimi i przenika inne społeczne światy" (Kacperczyk 2016:38).

> W istocie jedną z najbardziej uderzających cech wielu społecznych światów jest nieustanna wewnętrzna dysputa oraz podejmowanie decyzji odnośnie do wyobrażeń o ich własnych granicach. Ujawnia się to w licznych pytaniach zadawanych przez członków i dyskusjach dotyczących tego, czy dana czynność, osoba czy produkt jest 'rzeczywiście' reprezentatywny dla nich i dla ich świata. Dzieje się tak, dlatego że nawet uczestnicy, członkowie społecznego świata, zazwyczaj nie znają ostatecznej odpowiedzi na pytanie o granice ich świata, a ich główną działalność przeplata się zazwyczaj z procesem permanentnego negocjowania własnego statusu, roli i granic własnych działań oraz nieustannego podejmowania wysiłku samookreślenia się (Strauss 1993:214, za: Kacperczyk 2016:38).

Działania we wszystkich społecznych światach i arenach obejmują ustanawianie i utrzymywanie granic między społecznymi światami oraz uzyskanie legitymizacji dla samego świata (Clarke i Star 2008:118).

Arena to pojęcie, na które szczególną uwagę zwraca A. E. Clarke. Jej zdaniem jest ono „polem działań i interakcji pomiędzy potencjalnie nieskończoną ilością rozmaitych zbiorowych całości" (Clarke 1991:128). Jest polem, na którym toczy się istotny dla świata

społecznego problem i według Straussa oznacza niezgodę na kierunek działania podejmowany przez świat społeczny lub jego segment (1993:227). Przykładowo w świecie opieki paliatywnej w Polsce arena skoncentrowała się wokół obiektu granicznego - morfiny i podobnych leków opioidowych (Kacperczyk 2006). Obiekt graniczny „staje się punktem zapalnym, jądrem krystalizacji dla areny, a pośrednio dla autodefinicji uczestników zamieszanych w dyskusję światów" (Kacperczyk 2016:41). Pojęciu temu przyjrzymy się szerzej w kolejnej części, omawiając ważną dla naszej rozprawy specyfikę podejścia A. E. Clarke do ujęcia społecznych światów / aren.

W uproszczony i zdroworozsądkowy sposób pomaga zrozumieć utożsamianie społecznego świata z arenami następujący przykład historyczny:

> Być Europejczykiem w 1618 roku oznaczało mieć obsesję na tle niewielkich różnic doktrynalnych między katolikami a protestantami albo między kalwinistami a luteranami i być gotowym z powodu tych różnic zabijać lub ginąć. Jeśli kogoś w 1618 roku te konflikty nie obchodziły, to ów człowiek był może Turkiem albo Hindusem, ale na pewno nie Europejczykiem (Harari 2018:147).

Tym samym zrozumienie, o co spierają się uczestnicy społecznego świata (czy światów) jest dla badaczy niezbędne dla zrozumienia społecznego świata w ogóle, np. zachodzących w nim procesów i wartości tego świata.

Znane w naukach społecznych pojęcie wartości, nabiera w koncepcjach światów społecznych bardziej konkretnego znaczenia. Jest ono bowiem odnoszone do działania podstawowego:

> w społecznym świecie działanie nigdy nie jest działaniem dowolnym. Obwarowane jest regułami, wyobrażeniami, jak powinno przebiegać, a przede wszystkim z jakich pobudek powinno wypływać. (...) O granicach pomiędzy subświatami decydują ostatecznie zasadnicze różnice w pojmowaniu istoty podstawowego działania, a te są tak naprawdę różnicami w sferze wartości, które są tutaj rozumiane jako punkty orientacyjne, pozwalające dokonać oceny działania, a jednocześnie wskazujące, jak ma ono wyglądać i w jakich granicach przebiegać (Kacperczyk 2016:44).

Również znane, szczególnie na gruncie symbolicznego interakcjonizmu, pojęcie tożsamości znajduje się w uniwersum elementów społecznego świata. Tożsamość uznaje się tutaj za sprzężoną czy splecioną z wartościami. „Jeśli tożsamość widzimy jako zobowiązanie jednostki wobec jej grupy odniesienia..." przypomnieć należy pojawiające się już u Beckera (1974, 1986) pojęcie *commitment,* rozumiane jako konstytuujące świat społeczny zobowiązanie do podejmowania działania podstawowego „...oznacza to, że jednostka jest związana ze swoją 'grupą' wartościami rozumianymi jako kryteria oceny czy perspektywy

poznawcze" (Kacperczyk 2016:48). Te perspektywy i kryteria są nieustannie poddawane „testowi rzeczywistości" (Shibutani 1955:569).

> Relacja pomiędzy wartościami a tożsamością jest wzajemnie sprzężona i spleciona wokół budujących autodefinicje narracji. Ponieważ aktorzy mają liczne wielokrotne i równoległe zobowiązania, które mogą pozostawać w konflikcie, najistotniejsze staje się to, co faktycznie robią, rzeczywiście podejmowane przez nich działania (Kacperczyk 2016:48).

Autorka, powołując się głównie na A. L. Straussa (1993), zwróciła uwagę na procesy charakterystyczne dla światów społecznych:

> społeczny świat rozwija się w specyficznym dla siebie tempie i posiada własną niepowtarzalną dynamikę, ale w każdym świecie odnaleźć można immanentnie wpisane w społeczną rzeczywistość procesy przemian, takie jak: segmentacja, przecinanie się i legitymizowanie działalności społecznego świata oraz wskazywanie na jej autentyczność (Kacperczyk 2016:49).

Wśród podstawowych procesów, zachodzących w społecznym świecie nieustannie dochodzi do legitymizacji i zaświadczania o autentyczności. Legitymizacja wiąże się ze wskazaniem na wyjątkowy sposób wykonywania działania podstawowego przez uczestników społecznego świata oraz na konkretną technologię, jaką dysponuje dany świat lub segment świata. W skali makro zachodzi proces „wyznaczania granic prawidłowej działalności". W skali mikro szereg objaśnień, uzasadnień, wymówek pozwalających pojedynczym aktorom kontynuować działanie podstawowe (Kacperczyk 2016:36). Procesy te realizowane są głównie poprzez zabiegi komunikacyjne o charakterze objaśniania, neutralizowania lub teoretyzowania. Strauss poświęcił procesom legitymizacji osobną pracę, wyróżniając (Strauss 1982:173):

- odkrywanie i deklarowanie wartości (*discovering and claiming worth*);
- dystansowanie (*distancing*);
- teoretyzowanie (*theorizing*);
- wyznaczanie standardów, wcielanie, ewaluacja (*standard setting, embodying, evaluating*);
- wyznaczanie granic i stawianie wyzwań wobec granic na arenach (*boundary setting, and boundary challenging in arenas*).

Odwołując się do Straussa (1993:215) oraz artykułu Klinga i Gersona (1978) Kacperczyk stwierdziła, że „jedną z najważniejszych cech społecznego świata jest jego nieuchronna segmentacja (*segmentation*), prowadząca do wewnętrznego różnicowania

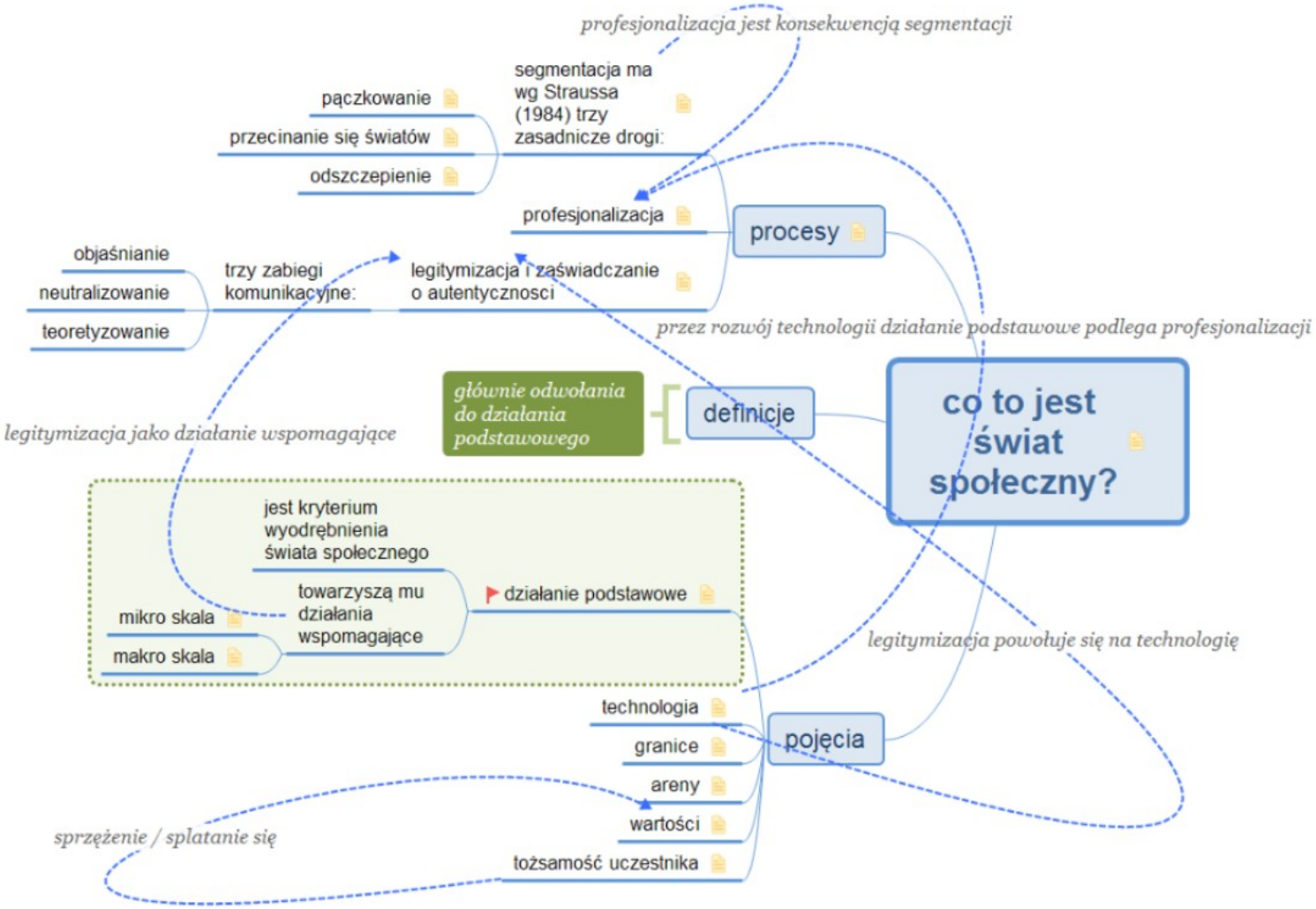
się i wydzielania w jego ramach subświatów", zaś jako przyczynę segmentacji przyjęła specjalizację zainteresowań i różnicowanie się interesów pewnych podgrup w świecie społecznym (2016:49). Segmentacja, zdaniem Straussa, ma trzy zasadnicze drogi: pączkowanie, odszczepienie i przecinanie się światów. Pączkowanie (*budding off*) (Strauss 1984:125) to „wydzielanie się nowej specjalizacji, opartej o inną (udoskonaloną lub uwstecznioną) technologię, oraz poczucie odmienności wykonywanego działania podstawowego" (Kacperczyk 2016:50). Zdaniem autorki

> pączkowanie jest niezwykle interesującym subprocesem, odsłaniającym niepohamowaną skłonność ludzi do **twórczych permutacji działania** [podkr. org.] (por. *Continual Permutations of Action*, Strauss 1993) i do wytwarzania coraz to nowych jego wariantów. W procesie tym ujawnia się niezliczona różnorodność rozwiązań i pomysłów na to, jak można zmienić, udoskonalić, udziwnić, przerobić czy sparodiować podstawowe działanie. Ta nieograniczona aktywność remodelowania działania ma w sobie wyraźny wymiar twórczy, a nawet artystyczny (Kacperczyk 2016:50).

Odszczepienie (*splitting off*) (Strauss 1984:125) to „radykalne oddzielenie się od wyjściowego świata społecznego", które jest uzasadniane różnicami ideologicznymi. Świat odszczepiony rozpoznaje sam siebie jako zasadniczo odmienny od świata wyjściowego, nie ma w nim żadnej ciągłości czy kontynuowania tradycji świata wyjściowego (Kacperczyk 2016:50). Przecinanie się (*intersection*) światów społecznych jest, zdaniem Straussa, procesem charakterystycznym dla współczesnych społeczeństw (1993:217). „Przecięcia światów łączą i dzielą jednocześnie" (Kacperczyk 2016:55). Chodzi nie o to, że płynne granice świata społecznego mieszają się i nakładają z innymi – „inne światy aktywnie go przenikają w każdym momencie jego istnienia" (Kacperczyk 2016:54). Formy interakcji między przecinającymi się światami bywają różnorodne, a kooperacja lub rywalizacja to najprostsze przykłady (Kacperczyk 2016:55).

> Proces przenikania się sięga mikroświata pojedynczego uczestnika, który doświadczać może przecinania się własnych wewnętrznych aren lub swoich własnych linii działania - jeśli jest członkiem kilku przenikających się światów. Oznacza to, że niekiedy owego 'innego' uczestnik spotyka w samym sobie (Kacperczyk 2016:55).

Powyższe wprowadzenie przedstawiamy w uproszczonej formie graficznej tzw. mapy myśli. Ta mapa (rys. 3) jest fragmentem naszych notatek sporządzonych w początkowej fazie zainteresowania skorzystaniem z teorii społecznych światów w tej rozprawie.



Rysunek 3: Podstawowe pojęcia i procesy w teorii społecznych światów – mapa myśli

Źródło: opracowanie własne na podstawie (Kacperczyk 2016:31-57) za pomocą XMind 8

3.1.1. Specyfika ujęcia społecznych światów / aren Adele E. Clarke

Adele E. Clarke (1991) używa metafory mozaiki, mówiąc o wielości światów i subświatów społecznych, w których uczestniczą aktorzy. Anna Kacperczyk ujęła rzecz następująco: „ludzie uczestniczą w wielu społecznych światach. Równocześnie i/lub równolegle podejmują działania wiążące ich z różnymi subświatami społecznymi" (2016:49). Clarke określa świat społeczny w odwołaniu do pojęcia środowiska społecznego, „świata czegoś", czyli „grupy wspólnie zaangażowane w pewną działalność", „grupy mającej wspólne kompetencje do wykonywania pewnych działań, dzielącej różnego rodzaju zasoby, by osiągać swoje cele i budującej wspólne ideologie dotyczące tego, jak realizować własne interesy". Charakter działań – a właściwie działania podstawowego – wokół którego koncentruje się się świat społeczny, może być różnorodny i związany np. ze sprawami tak artystycznymi czy religijnymi, jak i zawodowymi, rekreacyjnymi, komercyjnymi (Clarke 1991:131, za: Kacperczyk 2016:32). Takie mozaikowe, akceptujące chaos widzenie społecznych światów Clarke przełożyła na strategie badawcze i analityczne. Mówi ona o renowacji i regeneracji (*renovate and regenerate*) klasycznej teorii ugruntowanej na rzecz ugruntowanego teoretyzowania (Clarke 2003:558). Ugruntowane teoretyzowanie dokonywane jest na podstawie analizy sytuacyjnej, o czym więcej w rozdziale 4. rozprawy. Tę ramę ujmowania społecznych światów/aren nazwano pakietem[43] teorii/metod badań (*theory/methods package*), który jest zakorzeniony w konstruktywizmie społecznym, zawiera w sobie symboliczny interakcjonizm i filozofię pragmatyczną (Clarke i Star 2008:12).

> Zjawiska, z jakimi spotyka się badacz, są nieuporządkowane i rozmyte, wyślizgują się sztywnym ujęciom, są pełne rys i pęknięć a ich granice zanikają lub wymykają się definicjom. Clarke (1991) używa określeń: „zamazane (*blurred*), niechlujne (*sloppy*), rozmyte (*fuzzy*), zabałaganione (*messy*)". Te założenia na temat struktury rzeczywistości społecznej – zmiennej, fluidalnej, płynnej − nie mogą pozostać bez wpływu na proces badawczy i stawiane w nim pytania (Kacperczyk 2016:569).

Podejście Clarke cechuje zatem holizm metodologiczny (*methodologischen Holismus*), oraz pragmatyzm lub neopragmatyzm (Diaz-Bone 2013). Autorka ta uważa, że procesy zachodzące w społecznych światach badacz może zrozumieć tylko wtedy, kiedy założy, że nieład stanowi immanentną cechę obserwowanej przez siebie rzeczywistości społecznej (Kacperczyk 2016:568). Jest to założenie o charakterze ontologicznym. Rzeczywistość społeczna jest postrzegana jako pełna konfliktu i nieprzezwyciężalnych sprzeczności.

43 Przypomnijmy o pojęciu pakietu w świecie społecznym data science – pakiet to narzędzie, używane w języku programowania; jest zbiorem funkcji, wykonujących określone zadania (por. 2.1. i 5.3.1.).

Zdaniem Clarke znajduje to wyraz w tzw. arenie sporu. Areny są, w podejściu tej autorki, tożsame ze społecznymi światami. Nie istnieje badanie świata społecznego bez badania jego aren. Clarke wyraża to, posługując się określeniami „teoria społecznych światów/aren" lub „analiza społecznych światów/aren" (Kacperczyk 2016:572). Pojęcie aren w odniesieniu do społecznych światów zostało wprowadzone przez Straussa (1978), który podjął dyskusję z pracą Shibutani'ego (1955). Zwracamy uwagę na to, że areny występują na różnym poziomie rzeczywistości społecznej. Arena jako przestrzeń sporu może dotyczyć jednego lub kilku subświatów, przez całe światy społeczne czy kilka światów, do tzw. problemów publicznych, czyli zagadnień dziejących się pomiędzy wieloma światami społecznymi (*multiple world issues*). Strauss podkreślał, że występują przecięcia pomiędzy arenami na każdym z poziomów. Szczególnie w przypadku aren obejmujących wiele światów należy badać uwikłanie reprezentantów (tj. zabierających głos przedstawicieli światów biorących udział w sporze) w spory wewnątrz własnych światów czy subświatów.

> W każdym świecie społecznym różne sprawy (*issues*) są dyskutowane, negocjowane, zwalczane, podlegają wymuszaniu i manipulacji przez przedstawicieli uwikłanych subświatów. Obszary te obejmują działalność polityczną, ale niekoniecznie organy ustawodawcze i sądy. Sprawy (*issues*) są również zwalczane w subświatach przez ich członków. Przedstawiciele innych subświatów ([pochodzący z] tego samego i z innych światów), mogą również wejść do walki. Niektóre z tych spraw (*issues*) świata społecznego mogą stanowić wiadomości na pierwszych stronach, a inne są znane tylko członkom, lub innym zainteresowanym stronom. Media świata społecznego są pełne takich częściowo niewidocznych aren. Wszędzie tam, gdzie przecinają się światy i subświaty, możemy oczekiwać, że wraz z towarzyszącymi im procesami politycznymi powstają areny. (…) Jeśli chodzi o większe kwestie publiczne (np. co zrobić z zanieczyszczeniem środowiska lub alkoholizmem), socjolog musi zapytać nie tylko, które światy społeczne są reprezentowane na większej arenie, ale także które [to] segmenty których światów społecznych. Co więcej, [musi zapytać] z jakimi innymi arenami (wewnętrznych światów) powiązana jest ta reprezentacja wystawiona na arenę (pomiędzy wieloma światami)? Zagadnienia pomiędzy wieloma światami (*multiple world issues*) są nierozstrzygalne w oderwaniu od szerszego kontekstu wewnętrznych działań politycznych w świecie [wewnętrznym] [tłumaczenie własne RŻ] (Strauss 1978:124–25).

W powyższym fragmencie widać, że Strauss zwracał uwagę na unikanie nadmiernego upraszczania rzeczywistości, a szczególnie na docieranie do głosów „spoza pierwszych stron", czyli niszowych, słabo widocznych. Clarke rozwinęła ten kierunek poszukiwań, łącząc go z analizą relacji władzy, o czym powiemy dalej.

Analiza sytuacyjna, będąca metodologiczną konsekwencją przyjęcia omówionych wyżej założeń, jest dla badaczy „zaproszeniem do uwzględnienia chaotyczności licznych elementów sytuacji: ludzi, nie-ludzi (*nonhuman*), dyskursów, technologii i innych" (Morrison 2016:519). Kacperczyk, powołując się na Straussa (1978) i Beckera (1986) uznaje, że w

badaniu społecznych światów/aren jednostką analizy są nie pojedynczy ludzie, a właśnie społeczny świat/arena. Traktowana jest ona dwojako: jako zbiorowość (*social unit*) wytwarzająca znaczenia i podejmująca kolektywne działania oraz jako uniwersum dyskursu (Kacperczyk 2016:573). To samo mówią Clarke i Star. Społeczny świat/arena jest jednostką analizy szczególnie w badaniach nauki i techniki; może być to badanie kontrowersji (podano przykład pigułki wczesnoporonnej RU 468, którą zajęli się w 1993 r. Clarke i Montini) lub badanie dyscypliny (jak badanie S. L. Star z 1989 r. dotyczące neurofizjologii brytyjskiej w XIX w.) (2008:123-4). Pojęcie areny Clarke i Star wyjaśniają następująco:

> Jeśli i kiedy liczba światów społecznych staje się duża i poprzecinana konfliktami, różnymi rodzajami kariery, różnymi punktami widzenia, źródłami finansowania itd., to ta całość jest analizowana jako arena. Arena składa się zatem z wielu społecznych światów zorganizowanych ekologicznie wokół zagadnień, będących przedmiotem wspólnego zainteresowania i zaangażowania (*commitment*) w działanie [tłumaczenie własne RŻ] (Clarke i Star 2008:113).

Perspektywa społecznych światów/aren jest, jak zaznaczyliśmy wyżej, istotna dla naszej rozprawy i naszym zdaniem właściwa dla badanego podmiotu, czyli DS. Wzorujemy się i inspirujemy się całym pakietem teorii/metod badań proponowanym przez A. E. Clarke. Charakterystyczne dla Clarke pojęcie sytuacji i analizy sytuacyjnej oraz strategię ugruntowanego teoretyzowania prezentujemy w rozdziale 4. rozprawy.

3.1.2. Społeczne światy/areny w badaniach nauki i techniki (STS)

Poczynając od pracy A. E. Clarke poświęconej naukom reprodukcyjnym w USA z roku 1987 stosowano teorię społecznych światów/aren w badaniach nauki i techniki. Poza licznymi pracami Clarke i Star, tego rodzaju badania prowadzili m.in. (Clarke i Star 2008:124-6):

- Sara Shostak, która w pracach z 2003 i 2005 r. analizowała wyłonienie się dziedziny genetyki środowiskowej (*environmental genetics*) w latach 1950-2000, jako świata na przecięciu farmakogenetyki, epidemiologii molekularnej, epidemiologii genetycznej, ekogenetyki i toksykologii;
- Isabelle Baszanger w 1998 r. opublikował badanie dotyczące klinik leczenia bólu we Francji, gdzie pokazał praktyki współpracy bez zgody (*cooperation without consensus*);
- Monica Casper w tym samym czasie badała świat/arenę chirurgii płodów, wskazując na zmieniającą się wizję płodu ludzkiego jako pacjenta, zmiany postrzegania sprawczości

(*agency*) płodu, a także związków rozwoju chirurgii płodowej z działalnością ruchów antyaborcyjnych;

- Eswaran Subrahmanian w 2003 pokazał, że zmiany w składach zespołów inżynierskich, projektujących i wytwarzających oprogramowanie komputerowe zaburzały postrzeganie tworzonych przez nich prototypów. Działo się tak, ponieważ zmiany personalne naruszały ustalone zasady współpracy (i współpracy bez zgody), przez co otwierały debatę co do obiektów granicznych, jakimi są właśnie prototypy;
- Uri Gal badając w podobnym czasie przemysł AEC (*architecture, engineering and construction industry*) stwierdził, że zmiany stosowanych technologii informacyjnych, które traktowano jako obiekty graniczne, powodowały zmiany nie tylko wewnątrz tych światów, ale w ich tożsamości. Tym samym używane technologie informacyjne nie są tylko – mówiąc językiem Latoura – narzędziami translacji, ale źródłami formowania i wyrażania tożsamości zawodowej na zewnątrz świata społecznego.

Późniejsze badania typu STS z zastosowaniem teorii i metod społecznych światów/aren to m.in.:

- Case study informatyzacji wybranej uczelni. Badano procesy towarzyszące adaptacji systemu komputerowego w administracji uczelni – jako arenę na przecięciu światów społecznych urzędników i informatyków zatrudnionych na tym samym uniwersytecie (Vasconcelos 2007);
- Fenomen powstawania produktu – obiektu granicznego, który jest nie tylko wytwarzany, ale także prowadzony i rozwijany przez swoich użytkowników (użytkownicy to jednocześnie inżynierowie, architekci, projektanci, managerowie, zarząd i właściciele). Badanie wykonano na przykładzie mapy open-source OpenStreetMap. Produkt (mapa) traktowana jest jako obiekt graniczny, który stwarza okazję do współpracy aktorów z różnych światów społecznych. Aktorzy ci współtworzą mapę negocjując między sobą znaczenie procesu mapowania, danych na których pracują i samego OpenStreetMap (Lin 2011);
- Kontrowersje dotyczące zamieszkania w pobliżu miejsc składowania odpadów łatwopalnych, ze szczególnym uwzględnieniem konstruowania wiedzy w społeczności lokalnej na podstawie swojego odczytania badań naukowych (Andersen 2014);

- Ruchy nauki „zrób to sam” (*do-it-yourself science*), czyli badań wykonywanych poza opieką instytucjonalną uczelni przez hobbystów, majsterkowiczów, pasjonatów technologii, tzw. makerów. Wskazano na przemiany w dyskursie o tym ruchu, a także „obiektywizowanie” ruchu jako dyskursu przez to, że jest on wyrażany w dyskursie (Ferretti 2019).

Najnowszą polską pracą socjologiczną w dziedzinie STS jest rozprawa doktorska „Aktorzy-sieci w kolektywach hakerskich w Polsce” (Zaród 2018). Praca interesuje nas nie tylko jako przykład z dziedziny STS, ale także z uwagi na badany podmiot i rezultaty analiz. Mogą występować przecięcia pomiędzy światem DS a hakerami (por. rys. 2). Marcin Zaród przeprowadził szereg badań terenowych, głównie obserwacji etnograficznych w tzw. haker spejsach. Choć w jego pracy pojawiają się stosowane także przez Clarke, Star i ich naśladowców terminy – np. jako obiekt graniczny traktowana jest drukarka 3D (2018:307-13) – to nie ma w niej żadnych odniesień do społecznych światów. Jak sam tytuł wskazuje, autor pracuje w ramie Teorii Aktora-Sieci (*Actor-Network Theory,* ANT) Bruno Latoura. Swoją metodę badań nazywa żartobliwie „antografią laboratorium”, nawiązując jasno do latourowskiej etnografii laboratorium, „realizowanej ściśle w ramach ANT, która jednocześnie opisuje badany kolektyw i wychodzi poza jego granice” (Zaród 2018:199).

ANT jest specyficzną, odmienną od światów społecznych koncepcją. Dostarcza ona wyjątkowego ujęcia wiedzy, a właściwie konstruowania sieci jako procesu wiedzotwórczego. Niezbędne jest włączenie w ten proces aktorów pozaludzkich (*nonhumans*) (Abriszewski 2010). Sądzimy, że możliwość uwzględnienia takich aktorów jak: źródła danych, hardware umożliwiający składowanie danych lub wykonywanie obliczeń, software i koncepcje użyte do magazynowania danych czy do ich przetwarzania/eksploracji/modelowania/wizualizacji będzie użyteczna w opisie świata społecznego DS (por. Żulicki 2016:25). W ANT, z uwagi na porzucenie rozróżnień: kultura-natura, podmiot-przedmiot czy ludzie-rzeczy, nie zakłada się ograniczenia sprawczości bądź refleksyjności wyłącznie do aktorów ludzkich (Latour 2010). Przyglądamy się zagadnieniu sprawczości / refleksyjności pozaludzkich aktorów DS w badaniach własnych. Naszym zdaniem, uczestnicy świata DS antropomorfizują lub personifikują modele uczenia maszynowego, mówiąc np. że „on tego nie widzi”, „model myśli”, „mówimy mu, a on daje odpowiedź”, „nauczył się / nie nauczył się”, „jest na to za głupi”, „radzi sobie dobrze” itp., ale akcentują sprawczość ludzką (por. część 5.4.2.), zaś przecenianie sprawczości technologii uznają za charakterystyczne dla laików, tzw. osób nie-technicznych (por. 6.2.1.1.-2). Jak wspominaliśmy, Surma mówił o „irytującej

antropomorfizacji" systemów bazujących na uczeniu maszynowym (2017:8) w dyskursie medialnym.

Teksty B. Latoura przywoływaliśmy już kilkukrotnie, co wskazuje, że rozprawa socjologiczna bliska badaniom nauki i technik nie może pomijać tego wyjątkowego autora. Początkowym stadium tworzenia ANT były badania nazwane etnografią laboratorium. Latour wraz ze S. Woolgarem prowadzili badania etnograficzne w pracowni badań biologicznych i endokrynologicznych Salk Institute. Praca wydana w 1979 r. jako *Laboratory Life. The Social Construction of Scientific Facts* bliska jest konstruktywizmowi społecznemu i stanowi jedną z klasycznych pozycji w dziedzinie badań nad nauką i techniką (STS). Rodząca się wówczas ANT jest koncepcją sytuującą się na pograniczu różnych dyscyplin, które w uproszczeniu można nazwać łączącymi zainteresowanie badaniem nauki i wiedzy (Abriszewski 2010). Koncepcja powstała w międzynarodowym gronie uczonych zajmujących się STS. Za twórców ANT uznaje się trzy osoby: Bruno Latoura, Michela Callona i Johna Lawa. Koncepcja zaczęła rozwijać się pod koniec lat 70. XX w. Kontynuatorami myśli Latoura i współpracowników są obecnie m.in. Annemarie Mol i – cytowany w części 2.2.1. naszej pracy – Geofrey C. Bowker (Abriszewski 2012). Koncepcja łączy się z socjologią wiedzy, socjologią nauki, i socjologią ogólną (zwłaszcza etnometodologią H. Garfinkla) etnografią laboratorium i antropologią nauki, filozofią oraz semiotyką (Abriszewski 2010). Autor ten uznaje wersję ANT Latoura za „kanoniczną", najciekawszą filozoficznie, a stosowane przez niego liczne pojęcia za najbardziej interesujące, bo zaznaczające odmienność ANT od innych teorii czy tradycji refleksji nad nauką i wiedzą (Abriszewski 2012). ANT, szczególnie w wydaniu Latoura, nazywa myślą bliską Nietzschemu i de Saussere'owi, a będącą w opozycji do Kanta, Wittgensteina czy Jamesa. Oznacza to, że ANT niejako „zamazuje" rozróżnienie na ontologię i epistemologię, a także na podmiot i przedmiot. Blisko do koncepcji de Saussere'a, który uznawał świat za sieć obiektów, których tożsamość określona jest przez odróżnianie się od innych. Tym samym zainteresowania ANT uznać należy za ontologiczne, pytające o kształt świata czy zbiorowości, nie epistemologiczne, koncentrujące się na wiedzy. Kluczowe pytanie w ANT to zatem: „jak działa poznanie świata", rozumiane jako czynne tego świata przekształcanie. Zdaniem Abriszewskiego, ANT jest socjologią wiedzy bez pojęcia „wiedzy": w opozycji do kantowskiego rozróżnienia na nierozpoznawalny świat i jego konceptualizacje (które to są określane właśnie jako „wiedza"), w ANT mówi się metaforycznie o budowaniu sieci. Sieć wiąże ze sobą różne punkty, tworzy też nowe punkty przez połączenia „nitek" w węzłach. Stanowi ona rezultat relacji między różnymi bytami, zaś relacje ustabilizowane

tworzą aktora. Inaczej ujmując, w ANT badana jest dynamika procesów poznawczych („wiedzotwórczych"), nazywanych budowaniem sieci (Abriszewski 2010, 2012, Latour 2010). Proces wiedzotwórzy (poznawczy) w ANT jest więc jednocześnie procesem budowania sieci, łączenia aktorów, tworzenia aktorów i zmian relacji między aktorami. Pamiętajmy, że jest to Teoria Aktora-Sieci: aktor i sieć są równorzędni, a nawet więcej - są zasadniczo tym samym (Latour 2010). Nie ma mowy o żadnym aktorze poruszającym się po niciach sieci. Sieć jest dynamicznym zestawem relacji pomiędzy aktorami. Jeżeli relację utrwalają się i stają wystarczająco silne, to są aktorem. ANT była początkowo rozważana jako rama teoretyczna tej rozprawy (Żulicki 2016), jednak po kilku pierwszych miesiącach badań terenowych zdecydowaliśmy się traktować ją jako poboczną inspirację, koncentrując się na perspektywie społecznych światów (głównie światów/aren A. E. Clarke).

Koncepcje społecznych światów i ANT poza licznymi różnicami różnią się także kulturowo:

> Porównując ANT Latoura ze społecznymi światami / arenami, mówi się, że scentralizowany charakter władzy w ANT jest bardziej francuski, podczas gdy pluralizm perspektyw w koncepcji światów społecznych jest charakterystyczny dla Amerykanów. Bowker i Latour (1987)[44] stworzyli nieco podobny argument, porównując francuskie i angloamerykańskie STS. Argumentowali oni, że oś racjonalności / władzy jest tak naturalna dla francuskiej technokracji (i zbadanej między innymi u Foucault), że właśnie to musi być uwzględniane w pracach angloamerykańskich, bowiem w kontekście angloamerykańskim zwykle „zakłada się" (*assumed*), że nie ma związku między tymi dwoma [władzą i racjonalnością] [tłumaczenie własne RŻ] (Clarke i Star 2008:123).

Koncepcja światów społecznych jest bardziej odpowiednia do pracy nad opisem DS niż ANT, także dlatego, że DS ma angloamerykański rodowód. Jednak zgodnie z zakorzenioną w szkole chicagowskiej strategią badania społecznych światów/aren – analizą sytuacyjną – chcemy osadzić nasze badania w konkretnym kontekście przestrzennymi i czasowym. Data science w Polsce może mieć swój specyficzny charakter także dlatego, że technokracja jako idea historycznie odgrywała w Polsce znikomą rolę (Kurczewska 1997:XII). Od początków technokracji we Francji XIX w., kiedy to Henri de Saint-Simon przedstawił pomysł administrowania ludźmi tak jak rzeczami, poprzez kontynuatorów amerykańskich w okresie po I wojnie światowej – H. Scotta, W. H. Smytha, T. Veblena – była ona ideą wyrosłą na odmiennej niż dominujące w ówczesnej Polsce filozofia życiowa, polityczna czy etyka pracy. Jak wskazuje Joanna Kurczewska,

44 Mowa o artykule: Bowker, Geoffrey C. & Bruno Latour (1987) "A Booming Discipline Short of Discipline: (Social) Studies of Science in France," Social Studies of Science 17:715–48.

> przedstawiciele polskich elit intelektualnych częściej myśleli o Polsce i świecie w kategoriach politycznych czy narodowo-kulturowych aniżeli w kategoriach cywilizacyjno-gospodarczych czy też w takich, które politykę bezpośrednio uzależniały od kondycji cywilizacyjnej społeczeństwa (1997:XIII).

Nie znaczy to jednak, że nie było w polskiej myśli społecznej tradycji przygotowującej do recepcji zachodnioeuropejskich i amerykańskich idei technokratycznych. Autorka mówi o tradycji pozytywistów warszawskich „ceniących sobie naukę i technikę jako dźwignie postępu społecznego" jako nurcie konkurencyjnym i jednocześnie znacznie mniej popularnym niż romantyzm emancypacyjno-narodowy (Kurczewska 1997:XII). Sądzimy, że elementy idei technokratycznych można odnaleźć w świecie społecznym DS. Choć praca Kurczewskiej z pewnością nie jest osadzona w teorii społecznych światów ani w STS, to uważamy, że jej analizy będą cenne dla tej rozprawy. Tym samym nie redukujemy swoich horyzontów literaturowych do autorów posługujących się wyłącznie taką perspektywą teoretyczno-metodologiczną, jaka przyjęta jest w tej rozprawie.

Rozdział 4. Metodologia badań własnych

Jednoznaczną konsekwencją usytuowania rozprawy w wyżej wskazanej ramie teoretyczno-metodologicznej społecznych światów/aren jest jej osadzenie w szerszej perspektywie symbolicznego interakcjonizmu, oraz paradygmatu interpretatywnego w socjologii. Taka deklaracja jest, naszym zdaniem, niezbędna dla rozprawy empirycznej, ponieważ:

> Świat empiryczny możemy postrzegać jedynie za pośrednictwem pewnych schematów czy wyobrażeń o nim. Cały akt badania naukowego jest zorientowany i ukształtowany przez tkwiący u jego podłoża, przyjęty obraz świata empirycznego. Obraz ten rządzi selekcją i formułowaniem problemów, przesądza o tym, co będzie danymi, jakich środków się użyje, uzyskując owe dane, i w jakich twierdzeniach zostaną one przywołane (Blumer 2007).

Ów obraz świata, będący zbiorem założeń ontologicznych i epistemologicznych, określany jest jako model czy właśnie – paradygmat (Silverman 2010:136). Konecki określił paradygmat jako „różne sposoby ujmowania rzeczywistości w różnych orientacjach socjologicznych" (Konecki 2000:17). Paradygmatem nazywa się kanon, który jest przyjmowany i akceptowany jako oczywisty przez grono naukowców (Krzemiński 1986). Pojęcie paradygmatu zostało wprowadzone do filozofii nauki przez Thomasa Kuhna. Nie zostało ono zdefiniowane precyzyjnie, jednak Kuhn określa go m.in. jako: „powszechnie uznawane osiągnięcia naukowe, które w pewnym czasie dostarczają społeczności uczonych modelowych problemów i rozwiązań" (Kuhn 2001:10). Choć wskazuje się, że „pojęcie to [paradygmatu] zostało rozdęte do tego stopnia, że stało się mętne, niejednoznaczne, zbyt ogólne, tajemnicze i mylące" (Jodkowski 1990:139), to bazując na przedstawionej dalej argumentacji Koneckiego (2000:20-23) uznajemy, że socjologiczna rozprawa badawcza jest prowadzona w ramach wybranego paradygmatu.

W socjologii wyróżnia się paradygmat normatywny i interpretatywny (Hałas 2006:29; Wilson 1971, za: Konecki 2000:17; Piotrowski 1998:7–26). Ireneusz Krzemiński nazywa je odpowiednio socjologią pozytywistyczną i humanistyczną (1986:7-12). Pierwszy czy pierwsza z nich zakłada, że jednostka jest otoczona strukturami stabilnymi, zewnętrznymi wobec siebie, co warunkuje jej zachowanie poprzez konieczność podporządkowania się im. Do takiego zachowania zmuszają ją również czynniki wewnętrzne, psychiczne czy biologiczne, zatem sprawczość (*agency*) jednostki jest niewielka. W przypadku paradygmatu interpretatywnego rzeczywistość ujmowana jest jako konstruowana przez jednostki,

nazywane aktorami, w toku interakcji. Rzeczywistość postrzegana jest w kontekście nadawania znaczeń. Mówi się raczej o działaniu niż zachowaniu, a sprawczość (*agency*) jest głównie po stronie aktorów (Krzemiński 1986, Hałas 1987, Konecki 2000). Z paradygmatem normatywnym kojarzone są ilościowe, a z interpretatywnym jakościowe techniki badawcze: „luźne wywiady, analiza przypadków oraz obserwacja uczestnicząca" (Krzemiński 1986:92). Podkreślmy jednak, że stosowanie takich czy innych technik nie przesądza o przyjęciu danego paradygmatu (Konecki 2000:20). Paradygmat jest przede wszystkim rozstrzygnięciem ontologicznym, a dalej epistemologicznym i metodologicznym, zatem to decyzja o przyjętej wizji rzeczywistości społecznej prowadzi do wyboru strategii badawczej, nie odwrotnie. Tym samym metody, techniki i materiały pochodzące z badań ilościowych oraz jakościowych mogą być z powodzeniem stosowane w jednym projekcie badawczym, w ramach jednego, określonego paradygmatu. Wiąże się z tym pojęcie triangulacji: danych, badaczy, teorii oraz metod (Konecki 2000:86). Ma ona na celu szerokie uchwycenie interesującego nas wycinka rzeczywistość, ogląd różnych perspektyw, wzajemne wzbogacanie i weryfikację uzyskiwanych wniosków. Prowadzi do zminimalizowania błędów i ograniczeń związanych z podejściem jednostronnym. Używanie różnych technik, metod i źródeł danych jest strategią, która może uczynić wnioski „lepszymi" (o ile zależy badaczowi na szerokim obrazie) poprzez efekt synergii. Istnieje jednak postulat nie przekraczania granic paradygmatów (normatywnego i interpretatywnego), co jest uzasadnione ich odmiennymi założeniami ontologicznymi. Argumentacja Koneckiego – którą uważamy za słuszną – uwypukla różnice pomiędzy paradygmatem normatywnym a interpretatywnym, dotyczące podstawowych założeń o własności rzeczywistości społecznej. Można te różnice zredukować do wizji rzeczywistości „istniejącej" (co cechuje paradygmat normatywny) vs. „konstruowanej" (co charakteryzuje model interpretatywny). Konecki uznaje, że są to zasadniczo odmienne podejścia ontologiczne, więc w jednym projekcie badawczym należy jasno wybrać jedno z nich (Konecki 2000:20-23), co w naszej pracy czynimy.

Zwięzłego, choć uproszczonego porównania paradygmatu normatywnego (PN) i interpretatywnego (PI) na pięciu płaszczyznach dokonali Teddlie i Tashakkori (2009:86).

- Na płaszczyźnie epistemologii, rozumianej jako relacja między badaczem a rzeczywistością badaną, charakterem wiedzy i jej weryfikacją:
    - W PN badacz i badana rzeczywistość są od siebie niezależni, badacz jest zewnętrzny wobec obiektu badanego. Występuje pojęcie przedmiotu badania. Dalej powiemy o metaforze badacza-górnika S. Kvale.

- W PI badacz jest częścią badanej rzeczywistości, wchodzi z nią w interakcje, jest wewnętrzny także w sensie bycia częścią procesu badawczego. W tym kontekście mówi się, że badacz jest narzędziem. U Kvale pojawia się metafora badacza-podróżnika.

- Co do możliwości określania związków przyczynowo-skutkowych:
  - PN nastawiona jest na możliwość identyfikacji przyczyn zdarzeń, poszukuje następstw, wywołujących określone efekty w czasie.
  - PI z uwagi na zmienność badanej rzeczywistości uznaje, że możliwa jest sytuacja w której nie da się odróżnić przyczyny od skutku.
- Co do możliwości generalizacji wiedzy:
  - PN uznaje, że możliwa jest wiedza nomotetyczna – ogólna, niezależna od kontekstu czasowego i przestrzennego.
  - PI uznaje możliwość uzyskania wiedzy idiograficznej – opisowej, osadzonej w kontekście.
- Jeżeli chodzi o aksjologię, rozumianą jako rola wartości w procesie poznania:
  - PN postuluje poznanie empiryczne jako nie wartościujące.
  - PI uznaje, że poznanie empiryczne jest zawsze powiązane z wartościowaniem.
- W zakresie ontologii – założeń co do natury rzeczywistości:
  - PN zakłada jawność rzeczywistości, możliwość jej fragmentaryzacji, a także prostotę.
  - PI zakłada, że rzeczywistość jest ukryta, holistyczna (zatem niemożliwa do podziału na fragmenty), oraz złożona.

Istnieje pojęcie trzeciej drogi metodologicznej, trzeciego paradygmatu czy trzeciego ruchu metodologicznego, dążącego do integracji dwóch wyróżnionych paradygmatów na różnych poziomach (Teddlie i Tashakkori 2009), jednak nie zaproponowano nic innego niż triangulację metod i technik pod nazwą metod mieszanych (*mixed methods*). Zauważmy, że koncepcjami dążącymi do przezwyciężenia tego, co nazywane jest opozycją skrajnych wizji rzeczywistości społecznej: subiektywizmu/konstruktywizmu (charakterystycznego dla paradygmatu interpretatywnego) i obiektywizmu/strukturalizmu (odpowiednio dla paradygmaty normatywnego) (tzw. opozycja *agency and structure*) zajmowali się wybitni

teoretycy myśli socjologicznej, by wymienić przynajmniej Pierre'a Bourdieu i Anthony'ego Giddensa. W niniejszej rozprawie przyjmujemy paradygmat interpretatywny z jego założeniami ontologicznymi, epistemologicznymi i metodologicznymi, ponieważ zgadzamy się z tezą, że:

> (…) [problem zasadniczy] polega nie na tym, jak patrzeć na rzeczywistość społeczną i jaką jej stronę badać, lecz na tym, jaka ona *jest* [podkr. org.]. Rozumienie to nie tylko sposób jej poznawania, ale i zasada samego jej istnienia (Szacki 2002:872).

Głównych inspiracji metodologicznych dostarcza nam podejście badawcze Adele E. Clarke, które nazywa ona analizą sytuacyjną i strategia ugruntowanego teoretyzowania (Charmaz i Clarke 2013; Clarke 2003, 2005; Clarke i Friese 2007; Clarke i in. 2017). Decyzja o takim wyborze jest motywowana trojako. Po pierwsze, autorka proponuje „ugruntowane teoretyzowanie" zamiast „teorii ugruntowanej", co oznacza prowadzenie analizy nastawionej na tzw. sytuację, zrozumienie złożoności i heterogeniczności badanej rzeczywistości oraz uwzględnianie zanurzenia w lokalnych kontekstach. Za immanentną cechę badanej rzeczywistości przyjmując nieład, rozmycie czy pogmatwanie. Jak argumentowaliśmy wcześniej (por. 3.1.1.) traktujemy to jako adekwatne podejście do tak dynamicznie zmieniającej się dziedziny jak DS. Po drugie, Clarke uznaje aktorów pozaludzkich (*non-human actants / objects*) za pełnoprawne elementy badanego kontekstu (por. 3.1.2.), co w przypadku pracy dotyczącej DS jest niezbędne. Działanie tego świata, tak podstawowe jak i wspomagające, odbywa się prawie zawsze z użyciem sprzętu i oprogramowania komputerowego. Trzecim powodem jest to, że autorka stosowała i sugeruje owo podejście do badań typu *Science and Technology Studies* (STS), które są nam bliskie z uwagi na teorie, metody i badany podmiot (por. 3.1.2.). Argumentem nadrzędnym jest to, że taka metodologia jest konsekwencją przyjęcia – w naszym przekonaniu dobrze dopasowanej do DS – ramy teoretycznej społecznych światów, ze szczególnym naciskiem na społeczne światy/areny A. E. Clarke.

Cel niniejszej pracy – opis etnograficzny świata społecznego DS w Polsce – jest zgodny z podstawowym zadaniem etnografii: „opisowym objaśnieniem świata życia badanych" (Konecki 2010:20).

> **Badanie etnograficzne** (często nazywane jako badanie terenowe, badanie jakościowe, obserwacja uczestnicząca, interakcjonistyczne badanie, badanie konstrukcjonistyczne i studia naturalistyczne) **może być przede wszystkim zdefiniowane jako studium światów życia poszczególnych grup ludzi poprzez aktywne interakcje i wymianę z uczestnikami tych światów**. Przeprowadzając etnograficzne studium badacz przyjmuje do realizacji następujące zadania: a) przeniknięcie do społecznego świata 'innego' (uczestników), b) osiągnięcie

> bezpośredniej i bliskiej znajomości przeżyć innego, c) ostrożne i pełne zbieranie i zapisywanie informacji o świecie życia, i d) przekazywanie innym (outsiderom) wiedzy o świecie innego (uczestnika) w sposób zrozumiały dla outsiderów, a jednocześnie jak najbliższy przeżyć uczestników. **Pierwszym i podstawowym zadaniem etnografa jest: osiągnięcie całościowego i wyczulonego na szczegół opisowego objaśnienia świata życia innego** [podkr. RŻ] (Prus 1997:192, za: Konecki 2010:20).

Etnografia jako metoda (Angrosino 2010:45–46) jest naszym pierwszym, najwcześniejszym i „globalnym" podejściem do badań DS, a jednocześnie naszej rozprawy doktorskiej, ponieważ od początku byliśmy zdecydowani na pracę empiryczną. Pierwsze i najwcześniejsze oznacza, że przystępowaliśmy do badań terenowych kierując się jedynie przekonaniem, że podejście etnograficzne będzie odpowiednie dla naszych zainteresowań DS i ogólnym rozumieniem etnografii jako (ibidem):

- pracy badacza w terenie, czyli w naturalnym środowisku życia badanych, nie w laboratorium;
- badania spersonalizowanego, czyli takiego, w którym badacz i badani pozostają w bezpośrednim kontakcie, a badacz jest jednocześnie uczestnikiem i obserwatorem;
- badania wieloczynnikowego – używane też jest słowo „wielostanowiskowego" (Marcus 1995) – co oznacza korzystanie z więcej niż jednej techniki zbierania danych;
- badania wymagającego długotrwałego zaangażowania badacza w przebywanie i kontakt z badanymi;
- indukcyjnego podejścia do generowania wiedzy, zatem zbierania szczegółowych opisów badanego wycinka rzeczywistości w celu wskazania wzorów czy wyjaśnień na ich podstawie, a nie weryfikowania w badaniach wcześniej postawionych hipotez;
- dialogu z badanymi, czyli uzyskiwania na bieżąco komentarza od badanych co do wniosków i interpretacji badacza;
- podejścia holistycznego, zatem starania o jak najpełniejszy obraz badanego środowiska.

Poszukiwanie ram teoretycznych rozprawy, od ANT przez STS, do (ostatecznie) społecznych światów było późniejsze niż przyjęcie perspektywy właściwej etnografii. Etnografia jest zatem dla nas „globalnym" podejściem metodologicznym w tym sensie, że od niej zaczynaliśmy i jednoznacznie przy niej pozostajemy. Znajduje to wyraz w tytule rozprawy. Pozostajemy przy etnografii nie tylko w sensie metody, ale też rezultatu naszego badania – raportu etnograficznego. Raport taki jest narracją, opowieścią, „której głównym celem jest danie czytelnikowi możliwości pośredniego doświadczania społeczności, w której

badacz żył i z członkami której wchodził w interakcje" (Angrosino 2010:46). Przyjmujemy realistyczny styl naszej narracji, co oznacza kreślenie bezosobowych portretów w sposób emocjonalnie neutralny, nawet jeśli badacz był zaangażowanym emocjonalnie uczestnikiem życia badanych. Styl konfesyjny, w którym doświadczenie i punkt widzenia badacza przyjmuje w opowieści rolę centralną jest stosowany rzadko, z wyraźnym zaznaczeniem[45] (ibidem:46-47). Narrację uzupełniamy charakterystycznymi dla Clarke wizualizacjami analitycznymi, tzw. mapami, oraz innymi wizualizacjami, co rozwiniemy później. O ile narracja pozwala na pokazanie szczegółów, to mapy umożliwiają spojrzenie z lotu ptaka (Uri 2015:146).

Podejście etnograficzne, szczególnie wielostanowiskowe, jest wykorzystywane zarówno w badaniach społecznych światów, jak i STS.

> W ciągu kilku ostatnich dziesięcioleci interpretatywne studia przypadków, wielostanowiskowe (*multi-site*) lub multimodalne (*multi-modal*), stały się najbardziej popularnym podejściem metodologicznym w STS, zazwyczaj korzystając z wywiadów, danych archiwalnych, etnograficznych czy publikacji. (…) Ponieważ STS bada, ogólnie rzecz biorąc, produkcję wiedzy, to metody praktyków STS były zawsze lustrowane (*been under scrunity*) (...). Tym samym refleksyjność i triangulacja źródeł danych nie są dodatkiem, a są wbudowane do metodologicznej skrzynki z narzędziami STS [tłumaczenie własne RŻ] (Clarke, Friese, i Washburn 2015:46).

Etnografia wielostanowiskowa jest metodą analizy sytuacyjnej (Clarke 2005:xxxiii). Zaród wskazuje, że „koncepcja etnografii wielostanowiskowej Marcusa oparta była na badaniach etnograficznych prowadzonych w ramach STS", i że stosował on takie podejście w swoich badaniach hakerów z uwagi na „latourowski postulaty podążania za aktorami i rozpraszania tego, co lokalne" (Zaród 2018:156). Nasze badania również określamy jako etnografię wielostanowiskową. Zawierają one także elementy autoetnografii, netnografii i etnografii opartej na współpracy (*collaborative ethnography*), o czym więcej w części 4.2.

4.1. Od metodologii teorii ugruntowanej do analizy sytuacyjnej i ugruntowanego teoretyzowania

Jak wskazaliśmy wyżej, Clarke przełożyła mozaikowe, akceptujące chaos i sprzeczności widzenie społecznych światów/aren na strategie badawcze i analityczne. Mówi ona o renowacji i regeneracji (*renovate and regenerate*) teorii ugruntowanej na rzecz ugruntowanego teoretyzowania, dokonywanego na podstawie czy w ramach analizy sytuacyjnej (Clarke 2003:558). Jest to tzw. pakiet teorii/metod badań (*theory/methods*

45 Wyłącznie w prezentacji materiału autoetnograficznego.

package), zakorzeniony w konstruktywizmie społecznym, zawierający w sobie symboliczny interakcjonizm i filozofię pragmatyczną (Clarke i Star 2008:12). Traktuje się podejście Clarke m.in. jako o jedną czterech szkół teorii ugruntowanej, spośród: glaseriańskiej (*Glaserian* – Barney'a G. Glasera), straussowskiej (*Straussian* – A. L. Straussa), charmaziańskiej (*Charmazian* – Kathy Charmaz) i właśnie clarke'ańskiej (*Clarkeian*) (Apramian i in. 2017). W sensie historycznym propozycja Clarke jest szkołą najnowszą (Clarke i in. 2015:36–37; 43). Clarke deklarowała chęć popchnięcia teorii ugruntowanej w kierunki postmodernistycznego zwrotu (2005:xxiii). Istotnym źródłem inspiracji była dla niej myśl postmodernistyczna J. Derridy, B. Latoura i M. Foucaulta (Kacperczyk 2007:6). Później Clarke, kontynuując swój projekt renowacji i regeneracji, mówiła o teorii ugruntowanej po zwrocie interpretatywnym (*interpretive turn*) (Clarke i in. 2017).

Klasyczna, pierwsza wersja metodologii teorii ugruntowanej (dalej MTU) została przedstawiona w drugiej połowie lat 60. (Glaser i Strauss 1967). B. Glaser, A. L. Strauss, L. Schatzman i J. Corbin mieli za sobą głównie zaplecze teoretyczne w jednej z postaci symbolicznego interakcjonizmu, zwanego szkołą chicagowską. B. Glaser studiował też na Uniwersytecie Columbia, skąd w zmodyfikowanej formie przeniósł na grunt MTU pojęcia z metodologii ilościowej (Konecki 2000:32-34). To, jak wskażemy dalej, było nazywane pozytywizmem i stało się przedmiotem krytyki, tak ze strony K. Charmaz, jak i A. E. Clarke. MTU jest uznawana za metodologię naturalistyczną, a badacz stosujący MTU widzi świat nie mechanistycznie, a humanistycznie czyli interpretatywnie. Perspektywa ta jest skoncentrowana na „kształtowanych symbolicznie ludzkich procesach poznawczych" (Konecki 2000:34), których badanie jest drogą do objaśniania oraz zrozumienia rzeczywistości. Założenia dotyczące rzeczywistości według badacza humanistycznego można ująć w jej czterech cechach (Schatzman i Strauss 1973:5, za: Konecki 2000:34):

- Człowiek może przyjąć pewną postawę czy perspektywę wobec siebie i działać wobec siebie – co odpowiada podstawowemu twierdzeniu symbolicznego interakcjonizmu, szczególnie szkoły chicagowskiej, tj. że jednostka definiuje siebie oraz przedmioty, działania i cechy tych działań. Poniższe cechy rzeczywistości w podejściu humanistycznym są pochodną tego założenia.

- W różnych sytuacjach człowiek może mieć jednocześnie wiele różnych perspektyw (wobec siebie, innych rzeczy i zdarzeń); perspektywy mogą być sprzeczne. W nowych sytuacjach człowiek tworzy nowe perspektywy – co odnosi się do aspektu

kreatywnych działań ludzkich w odniesieniu do sytuacji, o czym mówił już G. H. Mead, a nieco później H. Becker.

- Perspektywy jednostkowe mają pochodzenie społeczne i pośrednio wywodzą się z otoczenia społecznego. Symboliczny interakcjonizm tłumaczy, że to otoczenie ma charakter symboliczny, gdzie symbol jest bodźcem. Bodziec nie powoduje jednak reakcji przez działanie fizyczne – dzieje się to raczej przez znaczenia i wartości, które to są internalizowanymi w procesie socjalizacji własnościami kultury społeczeństwa. O zagadnieniu pisali m. in. A. M. Rose, N. Herman, L. Reynolds, A. Lindesmith.

- Warunkami działań człowieka są przedstawiane przez niego perspektywy i definicje, które faktycznie są jego wytworami.

Przyjęcie określonych założeń o rzeczywistości jest ściśle związane z przyjęciem pewnej metodologii badań. Trzy istotne cechy MTU (Konecki 2000:35-36) odpowiadają charakterystycznej dla symbolicznego interakcjonizmu, założonej teoretycznie cesze „zmienności i kreatywności rzeczywistości społecznej" (ibidem). Po pierwsze, postulowane jest jak najsilniejsze ograniczenie prekonceptualizacji rzeczowej badania. Po drugie, MTU jest teorią dyskusyjną, bez stwierdzeń deklaratywnych. Jej budowanie nie jest nigdy zakończone, może zawsze podlegać modyfikacjom, gdyż rzeczywistość społeczna nie jest skończona, istnieje niezliczona ilość możliwości jej porządkowania. Po trzecie, w badaniu z użyciem MTU można dokonywać analiz sytuacyjnych, tzn. analizować ludzkie działania w ścisłym związku z ich kontekstem. Szczególnie ten aspekt rozwinęła Clarke.

W teorii symbolicznego interakcjonizmu rzeczywistość jest ujęta procesualnie, co wpływa na dyrektywy metodologiczne w MTU. Badacze pracujący w tym podejściu akcentują sekwencyjność i fazowość działań społecznych, zwraca się też uwagę na różne aspekty związane z czasem, jak np. jego społeczne konstruowanie, czy na tzw. ład temporalny (Konecki 2000:39-44). B. Glaser zalecał, by korzystając z MTU poszukiwać podstawowego procesu społecznego (*Basic Social Process*). Taki proces jest często tzw. kategorią centralną (*core category*). Wokół tejże kategorii integruje się cała teoria ugruntowana (ibidem). Poszukiwanie procesów jest strategią preferowaną w klasycznej MTU, ponieważ mają one być głównymi zmiennymi wyjaśniającymi zjawiska społeczne. Teoria ugruntowana ma wskazywać badany proces jako stadia i warunki ich pojawiania się (ibidem). Symboliczny interakcjonizm i MTU cechuje też tzw. empiryzm metodologiczny. Odkrywanie teorii czy

hipotez osadza się w badaniu empirycznym. W MTU teoretyzowanie jest w głównej mierze indukcyjne; teoria wyłania się z danych po rozpoczęciu badań empirycznych (Konecki 2000: 45-46).

Glaser i Strauss stworzyli podstawy MTU prowadząc badania dotyczące temporalnego procesu umierania pacjentów w szpitalach. Badania wykonano w Stanach Zjednoczonych, w pierwszej połowie lat 60. Praca dotyczyła m. in. tego, jak lekarze i chorzy przekazują informacje o umieraniu pacjentów, jaka jest organizacja pracy personelu szpitala w związku z tym procesem, jaki jest temporalny porządek umierania w organizacji szpitala. W trakcie tejże pracy Glaser i Strauss stworzyli kompleksowe podejście metodologiczne, które może być stosowane w badaniach jakościowych przeróżnych zagadnień. Strategia ta jest systematyczna i koncentruje się na praktycznym aspekcie badań jakościowych. Można jakoby dzięki temu rozwinąć teorię, która ugruntowana jest w badaniach, zamiast wyprowadzać z istniejących już teorii hipotezy do empirycznego sprawdzenia. Po publikacjach związanych z procesem umierania pacjentów w szpitalach Glaser i Strauss wydali w 1967 wskazaną wyżej pracę *The Discovery of Grounded Theory: Strategies for Qualitative Research* (wydana w 2009 r. Polsce jako *Odkrywanie teorii ugruntowanej*), opisującą ich strategię badawczą (Charmaz 2009:10–13).

> [MTU] polega na budowaniu teorii średniego zasięgu na podstawie systematycznie zbieranych danych empirycznych. (…) Teoria wyłania się tutaj, w trakcie systematycznie prowadzonych badań terenowych, z danych empirycznych, które bezpośrednio odnoszą się do obserwowanej części rzeczywistości społecznej (Konecki 2000:26).

Metodologia ta była próbą zmiany popularnego w połowie XX w. podejścia do gabinetowego budowania teorii w naukach społecznych. „Wielkie Teorie" – termin C. Wrighta Millsa dotyczący wysoce abstrakcyjnych i mało użytecznych w badaniach teorii (Silverman 2010:438) – są zazwyczaj trudne do zoperacjonalizowania ze względu na niejasne odniesienia empiryczne. Teoria ugruntowana jako wyprowadzona z empirycznych badań określonego wycinka rzeczywistości ma być łatwa do zoperacjonalizowania (Konecki 2000:26).

Według Glasera i Straussa teorię ugruntowaną w jej aspekcie praktycznym definiują następujące elementy (Charmaz 2009:13):

1. jednoczesne zbieranie danych i analiza tychże;

2. konstruowanie analitycznych kategorii indukcyjnie z danych, nie dedukcyjnie na podstawie przyjętych hipotez;

3. stosowanie metody ciągłego porównywania na każdym etapie analizy;

4. tworzenie not w celu wypracowania kategorii, własności tychże, związków między nimi, a także ustalenia luk;

5. pobieranie próbek w celu skonstruowania teorii, nie tworzenie próby reprezentatywnej;

6. przeprowadzanie niezależnej analizy przed przeglądem literatury.

„Teoria ugruntowana powinna umożliwiać przewidywanie, wyjaśnianie i rozumienie zachowań społecznych, których dotyczy" (Konecki 2000:28). Mają rozumieć ją nie tylko naukowcy, ale też tzw. znaczący laicy, czyli osoby będące uczestnikami badanego obszaru rzeczywistości. Według Straussa ważnym sprawdzianem dla teorii ugruntowanej jest to, czy osoby uczestniczące w badaniu są w stanie rozpoznać się w raporcie badawczym (ibidem).

Teoria ugruntowana, by spełnić wymogi przewidywania, wyjaśniania i rozumienia, musi (Konecki 2000:28-29):

- być dostosowana (*fit*); kategorie teoretyczne są wypreparowane z danych zebranych w trakcie badań, danych nie można wymuszać czy wybierać dopasowując je do istniejących kategorii (por. pkt. 2. i 6. powyżej),

- pracować (*work*); teoria wyjaśnia i pozwala interpretować badane zjawiska, umożliwia też prognozę na przyszłość,

- być istotna: (*relevant*); jest istotna dla działalności osób w badanym obszarze, ponieważ opisuje procesy czy problemy które wyłaniają się z obserwowanej rzeczywistości (por. pkt. 2. i 6.).

- być modyfikowalna (*modifable*); można ją przeformułować, jeśli była generowana przez wiele lat i staje się nieistotna lub niedopasowana

- dać się odnieść (*transcendenting*); można odnieść teorię do innych obszarów badawczych i zastosowanych tam metod badania (por. pkt. 3.)

Wymienione powyżej cechy to pojęcia należące do retoryki generowania teorii. Generowanie teorii ugruntowanej jest rozumiane jako proces, w którym zbieranie danych, budowanie i weryfikacja hipotez przeplatają się wielokrotnie. Nie są one ułożone sekwencyjne ani w inny sposób rozdzielone w czasie. Hipoteza w teorii ugruntowanej jest rozumiana jako teza lub twierdzenie, dotyczące relacji między pojęciami. Nie mierzy się siły związku, ważne jest empiryczne ugruntowanie istnienia takiej relacji. Weryfikacja hipotezy odbywa się przez porównanie określonej relacji w innym kontekście, np. w innych warunkach kulturowych (por. pkt. 3.) (Konecki 2000:30).

Do procesu generowania teorii należy teoretyczne pobieranie próbek (*theoretical sampling*). Owo pobieranie jest procesem zbierania danych w celu generowania teorii. Badacz zbiera, koduje i analizuje dane i w miarę wyłaniania się z nich teorii decyduje, gdzie i jakie dane będzie zbierać dalej (por. pkt. 1. i 5). Zbieranie danych jest więc kontrolowane przez wyłaniającą się teorię. Procedurą teoretycznego pobierania próbek jest metoda permanentnej analizy porównawczej (*constant comparative method*) (por. pkt. 1. i 3.). Polega ona na tym, że badacz porównuje różne przypadki i zjawiska celem ustalenia tego, co jest wspólne w nich, jak i w warunkach ich występowania. Porównuje ona/on pojęcia z przypadkami empirycznymi, dzięki czemu można wygenerować nowe własności teoretyczne pojęcia, a także nowe hipotezy (relacje pomiędzy pojęciami). Porównuje także pojęcia między sobą, co również może prowadzić do ich integracji w hipotezy, a więc i do tworzenia teorii (Konecki 2000:30-31).

Teoria ma umożliwiać przewidywanie, wyjaśnianie i rozumienie zachowań społecznych, których dotyczy. Praktyczne rozumienie ukończonej teorii ugruntowanej bywa jednak różne. Autorzy rozumieją ją m.in. jako: empiryczne uogólnienie; skłonność; objaśnienie procesu; związek między zmiennymi; rozumienie abstrakcyjne; opis. Istnieje obecnie wiele twierdzeń dotyczących tego, czym jest teoria ugruntowana. Sam B. Glaser w jednej z późniejszych prac (*The Grounded Theory Perspective: Conceptualization Contrasted with Description z* 2001 r.) określił ją jako teorię dotyczącą rozwiązywania głównego problemu (Charmaz 2009:171-173). Strategia MTU pozwalać ma rozpoczęcie pracy nad słabo zbadanym zagadnieniem. W założeniu ogranicza prekonceptualizacje (hipotezy, pojęcia, teorie), jest elastyczna, tzn. pozwala podążać za wyłaniającymi się w trakcie badania wskazówkami, umożliwia wprowadzanie zmian nawet w późnym etapie badania, a w tradycyjnej wersji dąży do wyprowadzania teorii średniego zasięgu z terenowych badań empirycznych (Konecki 2000:26-28, Charmaz 2009:24-25).

K. Charmaz proponuje rozumienie teorii ugruntowanej jako teorii „konstruktywistycznej", nie „obiektywistycznej". Zdaniem Charmaz rozumienie danych jako reprezentacji obiektywnych faktów, które badacz znajduje, jest charakterystyczne dla zwolenników obiektywistycznej teorii ugruntowanej. Takie podejście zakłada istnienie zewnętrznej rzeczywistości i bezstronnego badacza, który notuje fakty. Autorka twierdzi, że takie podejście do teorii ugruntowanej skutkuje brakiem krytycznego spojrzenia na przyjęte przez badacza założenia, interpretacje i interakcje. Uznaje je ona za wywodzące się z pozytywizmu, co podejmuje także Clarke. Zwolennicy konstruktywistycznej teorii ugruntowanej, do których należy K. Charmaz, zajmują stanowisko krytyczne wobec procesu badawczego i jego wyników, mając na uwadze interpretacje działań i znaczeń zarówno przez uczestników, jak i przez badacza. Istotne dla nich jest ustalenie własnych założeń i ich wpływu na badanie. Konstruktywiści każdą analizę postrzegają jako osadzoną w miejscu, czasie, kulturze i sytuacji, co przekłada się na – silnie wyrażony także u Clarke – postulat autorefleksji/samoświadomości badacza badania (Charmaz 2009:167-171). Krytyka wobec tradycyjnej MTU i specyfika podejścia Clarke również jest antypozytywistyczna, ale nie tożsama z podejściem Charmaz.

A. Kacperczyk wskazuje, że Clarke krytykuje klasyczną MTU z postmodernistycznego punktu widzenia. Chodzi o bazujące na idei dekonstrukcji Derridy oraz wiedzowładzy Foucault przekonanie, dotyczące niemożności utrzymania modernistycznej wizji „jednej prawdy" i statusu podmiotu poznającego. Tym samym w postmodernizmie uznaje się, że prawda, wiedza i wartości są społecznymi konstruktami, a rzeczywistość społeczna na charakter dyskursywny. Przy tym ten dyskurs jest heterogeniczny, wielogłosowy. W związku z tym nie można już uprawiać nauk społecznych w sposób modernistyczny. Konieczne jest rozwijanie epistemologicznej i metodologicznej samoświadomości, które Kacperczyk nazywa przekonaniem o prowizoryczności naukowych konstrukcji i konieczności ich ciągłego zabezpieczania (Kacperczyk 2007:7). Naszym zdaniem, oznacza to, że podejście Clarke, czyli analiza sytuacyjna, ma być bardziej samoświadoma niż tzw. tradycyjna MTU. Clarke mówi, że MTU w wersji klasycznej jest oporna (*recalcitrance*), co przejawia się w (ibidem):

- braku pogłębionej refleksji co do samego procesu badawczego;
- występowaniu nadmiernych uproszczeń w postaci
    - nacisku na to, co wspólne i szukaniu spójności na siłę;

- koncentrowaniu się raczej na pojedynczym niż na wielu współistniejących procesach społecznych;

- interpretowaniu zróżnicowania danych w kategoriach tzw. przypadków negatywnych;
- tendencji do szukania „czystości" w teorii ugruntowanej.

Na bazie powyższej krytyki

> Tradycyjna teoria ugruntowana zostaje przez autorkę świadomie i z rozmysłem: rozebrana na części (zdekonstruowana), uzupełniona ekologiczną metaforą teorii społecznych światów/aren, uzupełniona kartograficzną analizą elementów kluczowych alternatywnych wobec podstawowego procesu społecznego (*basic social process*) i koncentracją na gęstej złożoności szeroko zarysowanej i wewnętrznie skomplikowanej sytuacji badania (Kacperczyk 2007:7).

Clarke jednoznacznie wskazuje, że za wady (*shorcomings*) tradycyjnej MTU uznaje pozytywistyczne tendencje, niedostateczną refleksyjność, nadmierne upraszczanie zamiast uwzględniania różnorodności i brak analizy władzy (Clarke i in. 2015:12). Ta krytyka dotyczy głównie, jeśli nie wyłącznie, B. Glasera. Glaser miał pozostać wierny pozytywistycznej, funkcjonalistycznej socjologii R. K. Mertona i P. Lazarsfelda, obowiązującej na Uniwersytecie Columbia, który ukończył (Clarke i in. 2015:36-37; por. Konecki 2000:32-34). Pozytywistyczny charakter glaseriańskiej MTU Clarke z zespołem określili jako żywo realistyczny i naiwnie empirystyczny, co skontrastowano z podejściem Straussa. Strauss miał przez całe życie pozostać symbolicznym interakcjonistą, jego podejście do MTU było z czasem coraz bardziej konstruktywistyczne, a wszystko to miało odzwierciedlać jego pragmatyczna filozofię (2015:37).

Od klasycznej MTU analiza sytuacyjna różni się zatem zasadniczo w trzech wymiarach (Kacperczyk 2007:7). Po pierwsze, chodzi o podejście do danych gromadzonych przez badacza i ich status. O ile Glaser mówił, że potencjalnie wszystko może stanowić dane (*all is data*), to zdaniem Clarke dane to wyłącznie te elementy badanego wycinka rzeczywistości, które badacz jako takie spostrzega i włącza do analizy. Jednocześnie dla Clarke nie ma jakiejkolwiek zewnętrzności, skąd można by pobrać dane obiektywne – jest tylko konstrukcja i rekonstrukcja znaczeń. Glaser uważał, że pewne dane są konstruowane przez badacza, ale inne są obiektywne, czyste (*pure*)[46], badacz po prostu je znajduje. Z tą czystością Clarke bez wątpienia się nie zgadza. Drugą znaczącą różnicą jest stosunek do wielokrotnie już wspominanych czynników/aktorów/podmiotów/obiektów pozaludzkich czy nie-ludzkich

46 Zauważmy podobieństwo do postrzegania danych jako surowych (*raw*) w data science / big data i krytykę tego podejścia (por. 2.2.1.)

(*non-human actants/objects*; prezentując wyniki badań własnych będziemy mówić po prostu o aktorach). Clarke uznaje je za pełnoprawne elementy badanej rzeczywistości. Są one dla niej częścią budowanego wyjaśnienia, ponieważ oddziałują w badanej sytuacji. Rozumiemy, że Clarke przypisuje obiektom pozaludzkim sprawczość (*agency*), ale nie podmiotowość. Kacperczyk wskazuje, że w klasycznej MTU obiekty pozaludzkie nie są w ogóle brane pod uwagę, nie są włączane do analizy ani w żaden sposób dostrzegane. Po trzecie, jak wyraźnie widać w pokazanych wyżej wymogach wobec klasycznej teorii ugruntowanej[47], w tym podejściu zmierzano do ahistorycznej, akulturowej, abstrahującej od miejsca i czasu, oczyszczonej z kontekstu transcendentnej teorii pojęciowej. Analiza sytuacyjna i ugruntowane teoretyzowanie uwzględnia cechy badanej sytuacji, zanurzenie badanego wycinka rzeczywistości w lokalnych kontekstach.

Zdaniem Kathy Charmaz „(…) usytuowanie teorii ugruntowanych w ich społecznych, historycznych, lokalnych oraz interakcyjnych kontekstach wzmacnia je" (2009:232). Choć to bez wątpienia odejście od tradycyjnej MTU, to Clarke widzi rzecz inaczej. Nie chodzi jej o uwzględnienie kontekstu, bowiem „**otoczenie (*conditions*) sytuacji jest sytuacją**. [podkr. org.] Nie ma czegoś takiego jak kontekst" (Clarke 2005:71). Badacz ma przy tym wykazać się jak największą samoświadomością wobec konstruowanego przez siebie wyjaśnienia. Teoria ugruntowana ma być elastyczną strategią heurystyczną. Zmierza się w kierunku namysłu nad drogą, która doprowadziła do uzyskania końcowego efektu analiz, i to właśnie ta droga nazwana jest ugruntowanym teoretyzowaniem (Kacperczyk 2007:9). Wykorzystywana jest koncepcja „teoretyzowania" zamiast koncepcji „teorii", ponieważ najważniejszym celem analizy sytuacyjnej jest otwarcie danych i sytuacji na niekończące się (*ongoing*) negocjacje, a nie promowanie obiektywnej prawdy (Uri 2015:146). Clarke w tym kontekście mówi także – o czym wspominaliśmy w części 3.1.1. – że jednostką analizy jest dla niej nie wybrany proces czy zjawisko, a sytuacja. Sytuacja jest konstruowana za pomocą trzech rodzajów map (sytuacyjnych, społecznych światów/aren i pozycyjnych), i poprzez pracę analityczną z tymi mapami. Odróżnia to analizę sytuacyjną od tradycyjnej MTU, dążącej do ukazania ludzkiego działania jako podstawowego procesu społecznego (Clarke i in. 2015:12), co jest wynikiem rozumienia tego, czym jest sytuacja. Powtórzymy, że nie jest ona uwzględnieniem otoczenia, a jednocześnie (i nie tylko) tym, co tradycyjnie nazwalibyśmy podstawowym procesem i jego otoczeniem. Clarke pokazuje, co składa się na sytuację w formie matrycy sytuacyjnej

47 Szczególnie tego, że teoria powinna: „pracować" (*work*), czyli wyjaśniać i pozwalać interpretować badane zjawiska, umożliwiać prognozę; oraz „dać się odnieść" (*transcendenting*), czyli można odnieść teorię do innych obszarów badawczych i zastosowanych tam metod badania.

(*situational matrix*), gdzie między elementami nie ma żadnej hierarchii (Clarke 2005:73, za: Kacperczyk 2007:22). Są to:

- elementy czasoprzestrzenne;
- elementy ludzkie (*human*), indywidualne i kolektywne;
- elementy pozaludzkie (*nonhuman*);
- elementy polityczno-ekonomiczne;
- dyskursywne konstrukcje aktorów;
- elementy organizacyjne / instytucjonalne;
- główne kwestie kontestowane;
- elementy lokalne i globalne;
- elementy socjokulturowe;
- elementy symboliczne;
- dyskurs popularny i inne dyskursy;
- pozostałe elementy empiryczne.

Elementy te, tworzące „sytuację działania" (*the situation of action*), uznaje się za sprzężone i wpływające na siebie wzajemnie. Zauważmy, że wizualizacja tej matrycy sytuacyjnej – przypomina dziecięcy rysunek słońca, gdzie w centrum jako tarczę umieszczono tytuł „sytuacja działania", elementy sytuacji zaś ułożono jak promienie – umieszczona została na okładce pierwszej pracy w całości poświęconej analizie sytuacyjnej (Clarke 2005). Autorka wskazuje, że np. Strauss i Corbin we wcześniejszych tego rodzaju matrycach (*conditional matrix*) w centrum stawiali działanie (*action*), a następnie pokazywali hierarchię warunków – od m.in. interakcyjnych przez organizacyjne i dalej narodowe, graficznie ujęte jako kolejno oddalające się od centrum kręgi. Podkreśla ona, że analiza sytuacyjna zakłada coś innego. Zakłada, że elementy otoczenia sytuacji (*conditional elements*) muszą być określone w samej sytuacji, ponieważ są dla sytuacji konstytutywne. Jak powiedziano wyżej, one są sytuacją, a nie tylko otaczają ją, kształtują czy przyczyniają się do niej (Clarke 2015:96–98). Podejście Clarke jest zatem, podobnie jak Charmaz, konstruktywistyczne, ale Clarke podkreśla traktowanie sytuacji jako jednostki analizy. Tym samym Clarke odchodzi od znanego z innych nurtów jakościowych nastawienia na jednostkę ludzką jako podmiot (*individual-centered*) w kierunku nastawienia na uchwycenie pełnej

sytuacji badania (*the full situation of inquiry*) (Kacperczyk 2007:10). Celem badania dla Clarke jest „zrozumienie złożoności i heterogeniczności indywidualnych i kolektywnych sytuacji, dyskursów i interpretacji sytuacji" (Clarke 2005: xxv, za: Kacperczyk 2007:10). Jej zdaniem jest to zgodne z dwiema ideami Normana Denzina: „sytuacyjnej interpretacji" (*situating interpretation*) (Clarke 2005:xxviii) i włączeniem do interakcjonizmu myśli postmodernistycznej i poststrukturalistycznej Barthesa, Derridy, Foucaulta, Baudillarda (Clarke 2015:90).

Postulaty samoświadomości i unikania nadmiernego uproszczenia widzimy także w innych aspektach analizy sytuacyjnej. Nastawienie na samoświadomość badacza jest kolejną cechą podejścia Clarke, odróżniającą je od tradycyjnej MTU, gdzie nacisk był położony na „czystą tablicę" badacza w procesie badania – ograniczenie prekonceptualizacji, bagażu teoretycznego i własnych założeń. Clarke uważa, że jej podejście odpowiada na cechującą postmodernistyczny zwrot zmianę stanowiska badacza „z 'wszechwiedzącego analityka' *(all-knowing analyst)* do 'samoświadomego uczestnika' (*acknowledged participant*) w procesie produkcji zawsze częściowej wiedzy" (Clarke 2003:555–56). Uznaje ona, że należy dążyć nie do ograniczenia, a do odsłonięcia informacji składających się na ową „tablicę", intelektualne tło (*wallpaper*) badacza. Tło może być traktowane np. jako użyteczne lub stronnicze, ale zawsze powinno być uświadomione i odsłonięte, czy to w narracji czy w mapach (o których więcej powiemy później). W przeciwnym razie badacz nie wie, jakie są jego ukryte założenia, a zatem nie może określić jak decydują – a decydują – o postępach i kierunkach jego pracy. Związany z tym jest także postulat, nazywany przez Clarke elementem odpowiedzialności etycznej badacza, by wyartykułować jasno, co badacz uważa w swojej pracy za przemilczane (*sites of silence*). To również daje obraz tego, z jakiej pozycji wypowiada się badacz. Tym samym znane z różnych badań jakościowych zdanie: badacz jest narzędziem, u Clarke brzmiałoby: ...narzędziem jak najbardziej jawnie używanym (Kacperczyk 2007:18-19).

Co do unikania nadmiernych uproszczeń, to widać to wyraźnie w – wielokrotnie wyrażanym przez Clarke – postulacie nagłaśniania głosów grup czy pozycji słabo słyszalnych (*helping silences speak*), przemilczanych, nieobecnych w społecznym świecie, a także w konstruowanym przez badacza wyjaśnieniu (Clarke 2003:561; 569, 2005:85, 2015:108; Clarke i Star 2008:119, 124, 129). „Metody badań mają umożliwiać i zachęcać analityka do naświetlenia poprzednio marginalizowanych i nielegitymizowanych perspektyw wiedzy o życiu społecznym" (Kacperczyk 2007:10).

> Analiza sytuacyjna (SA) nadaje się w szczególności do projektów krytycznych, feministycznych, antyrasistowskich i innych, zbliżonych do nich. Dzieje się tak z wielu ważnych i intencjonalnych powodów. Po pierwsze, SA kładzie silny nacisk na wyjaśnienie złożoności i różnorodności elementów i pozycji w badanej sytuacji. Jednym z najbardziej powszechnych wzorców, zarówno w badaniach jakościowych jak i ilościowych, jest skupienie się na podobieństwach i usuwanie (*erasing*) lub pomijanie (*not attending*) niemal wszystkiego innego – w tym ludzi i stanowisk działających wbrew „temu, jak jest" (*'the way things are'*). Osoby będące przedstawicielami mniejszości, pozycje dyscyplinarne mniejszości, głosy mniejszości i inne marginalizowane sprawy po prostu się nie pojawiają. Krótko mówiąc, większość badań jest wykonywana (choć nieumyślnie) z „góry na dół", czego powodem są niezbadane założenia dotyczące hierarchii ważności. SA w bardzo wyraźnym kontraście opowiada się za artykułowaniem wszystkich elementów, wszystkich pozycji, wszystkich głosów. Oferuje narzędzia, dzięki którym badacz może pracować nie tylko „od dołu", ale także „od zewnątrz", aby uchwycić i przedstawić to, co znajduje się na marginesie sytuacji, [a co jest] szczególnie pomocne w kwestionowaniu istniejących hierarchii władzy. [SA] Nie tylko nie skupia się na dominancie, ale podejmuje próbę zdenerwowania i wyparcia milczących hierarchii (*tacit hierarchies*) [tłumaczenie własne RŻ] (Clarke i in. 2015:20–21).

Autorka zaznacza, że w ciągu czterdziestu lat rozwoju teorii ugruntowanej i jej odmian na świecie nastąpiły wielkie zmiany w sytuacji wielu grup i pozycji marginalizowanych, co znajduje odzwierciedlenie także w nauce. Wskazuje ona, na swoim przykładzie, na zmianę sytuacji kobiet w naukach społecznych. W początkach jej kariery panować miał dla nich – i osób kolorowych – chłodny klimat (*chilly climate in the classroom*), ale stopniowo sytuacja zmieniała się i obecnie mamy do czynienia z globalnym ociepleniem politycznym (*political global warming*). Clarke twierdzi, że w rozwoju teorii ugruntowanej nastąpiło przejście od „Wspaniałych Mężczyzn" (*Great Men*), czyli Glasera i Straussa, do ich uczennic i uczniów, „Obowiązkowych Córek" (*Dutiful Daughters*) – m.in. Clarke, Carrie Friese, Rachel Washburn, S. L. Star, Charmaz. Wymieniła także mężczyzn, m.in. Krzysztofa Koneckiego i Antony'ego Bryanta (Clarke 2015:84; 110).

Clarke mówi o unikaniu nadmiernych uproszczeń także w kontekście postrzegania samej analizy sytuacyjnej i innych odmian teorii ugruntowanej, oraz generalnie badań jakościowych po zwrocie postmodernistycznym/poststrukturalnym. Jej zdaniem od końca lat 60. badania i badaczy m.in. feminizmu, antyrasizmu, postkolonializmu, LGBTQ wrzuca się do jednego worka (*lump together*) pod pozornie nieograniczoną etykietą postmodernizm/poststrukturalizm. Sprzeciwia się ona także polaryzacjom, schizmom w dziedzinie badań jakościowych, uznając je za wyraz nadmiernego upraszczania, tylko pod inną nazwą (Clarke i in. 2015:46–47). W podobnym duchu mówiła o tym, że nie popiera wydzielania się dyscyplin (*discipling*), różnych wersji teorii ugruntowanej, jak glaseriańskiej i straussowskiej. Clarke przywołała słowa Straussa, który zwykł mówić, że interakcjonizm – a zdaniem autorki także teoria ugruntowana i analiza sytuacyjna – jest jak bankiet. Ludzie

przychodzą na bankiet, biorą to, co chcą, a resztę zostawiają (Clarke 2015:85). Jest to stanowisko niezupełnie spójne z przywołanym wyżej kontrastowaniem „pozytywistycznego" Glasera z „interpretatywnym" Straussem, czego w innej pracy dokonała autorka wraz z zespołem (Clarke i in. 2015:36-37).

Podsumowując, przedstawiamy sześć zalet analizy sytuacyjnej Clarke (Clarke 2005:19–31; Uri 2015), oraz uwagi krytyczne wobec niej. Pierwszą zaletą ma być uświadamianie ucieleśnienia i sytuacyjności (*acknowledging embodiment and situadness*). Oznacza to, że w projekcie badawczym badacz stara się dostrzegać swoją pozycję, wiedzę, założenia, a także dostrzegać to u uczestników badania (Uri 2015:144). Druga zaleta to ugruntowanie w sytuacji (*grounding in the situation*). Z uwagi na omawiany wyżej brak nastawienia na poszukiwanie podstawowego procesu społecznego, podczas badań stawiane są pytania eksplorujące różne wymiary sytuacji. Pozwala to zauważyć szeroką gamę elementów sytuacji i powiązań między nimi, w czym pomocne jest szkicowanie kolejnych map. Nazywa się to ugruntowaniem sytuacji w badaniu, czyli tym, o czym Clarke mówiła: „sytuacja jest zawsze czymś większym niż suma swoich części" (ibidem:144-5). Trzecia nazwana została „różnice i złożoność" (*differences and complexities*), i wynika bezpośrednio z drugiej (ibidem:145). Postulat Clarke dotyczący unikania nadmiernych uproszczeń omówiliśmy już szerzej. Zaleta czwarta dotyczy pojęć uwrażliwiających (*sensitizing concepts*), analityki i teoretyzowania. Jest to konsekwencją zarówno rozumienia pojęcia sytuacji, jak i zalet wymienionych wcześniej. To zasadniczo konstruktywistyczne podejście do badań pozwala zachować otwartość na zmianę w trakcie procesu badawczego, uniknąć nieuprawnionych generalizacji i abstrahowania od tzw. kontekstu. Jest to zdaniem Clarke niezbędne, ponieważ „nie ma sensu pisanie Wielkiej Teorii czegoś tak zmiennego [jak społeczny świat/arena]", przy czym Clarke przestrzega przed badaniem dorywczym (*haphazard*) (ibidem 145-6). Zaletą piątą ma być samo wykonywanie analizy sytuacji (*doing situation analysis*). Chodzi o pracę z trzema rodzajami map, które pozwalają na specyfikację, ponowną reprezentację i dalsze badanie (*specification, re-representation and subsequent examination*), badanie sytuacji, czyli jej elementów i relacji między nimi. Clarke podkreśla, że elementy są dla niej także tym, co tradycyjnie nazywano kontekstem. Mapowanie, czyli tworzenie map, ma jej zdaniem szereg zalet. Autorka nazywa mapowanie urządzeniem przełączającym (*shifting device*), ponieważ pozwala przełamać (*rupture*) standardową analizę danych jakościowych, i sprowokować badacza do świeżego spojrzenia. Mapy nie muszą być gotowym produktem analizy, a ćwiczeniem analitycznym, także na wczesnych etapach badania. Clarke zachęca więc do demapowania i remapowania

(*unmapping and remapping*). Mapy mają także pomagać w ujęciu elementów czasowych i przestrzennych. Ponadto mapy mają pozwalać na szybszą i łatwiejszą pracę analityczną, niż gdy mamy tekst w formie narracji (Clarke 2005:29-30). Zaletą szóstą jest zwrócenie uwagi na dyskurs(y) (*turning to discourse(s)*). Z uwagi na złożoność i heterogeniczność postmodernistycznego świata XXI w. tradycyjna analiza transkrypcji wywiadów i notatek z obserwacji ma być niewystarczająca. Do uchwycenia sytuacji potrzeba także innych rodzajów dyskursów. Stąd zaletą ma być łączenie typowego badania etnograficznego z analizą źródeł historycznych czy materiałów wizualnych, na zasadzie wspomnianej etnografii wielostanowiskowej (ibidem:30, 145-6). Analiza dyskursu jako element analizy sytuacyjnej obejmować może wszelkie formy używania języka (od mediów po przepisy czy podręczniki), obrazy (od sztuki przez ulotki czy schematy), symbole (jak logo, flagi), przedmioty codziennego użytku (m.in. kubki, T-shirty z nadrukiem, komputery) i inne środki komunikacji (np. zachowania niewerbalne, dźwięki) (Kacperczyk 2016:645). Wzorowana na Foucault koncentracja na dyskursie ma pomóc w analizie władzy, a także w nagłaśnianiu głosów i pozycji marginalizowanych. Narzędziem analitycznym są mapy pozycyjne, które mają przedstawiać pozycje artykułowane w dyskursie (Clarke i in. 2015:45; 52). Omawiane sześć zalet ma pchać teorię ugruntowaną w kierunku postmodernistycznego zwrotu (Clarke 2005:19).

Uwagi krytyczne dotyczą m.in. problemu granic sytuacji. Próba uchwycenia sytuacji, jako wszystkich, w tym marginalizowanych, elementów może być zadaniem przytłaczającym. Badacz kieruje się zatem znanym z klasycznej teorii ugruntowanej nasyceniem kategorii, tzn. decyduje, że dany element jest już wystarczająco przeanalizowany, gdy nie widzi by pojawiały się nowe wątki (Uri 2015:148). Zatem sytuacja jest czymś skonstruowanym przez badacza – Clarke w żadnym razie tego nie ukrywa. Zgadzamy się z tezą, że

> „sytuacja", która jest tu przedstawiana jako fundamentalna jednostka analizy, dla każdego podmiotu poznającego może oznaczać i zawierać coś innego. Dlatego uporczywe poszukiwania znaczenia „sytuacji" mogą nas ostatecznie co najwyżej doprowadzić do tautologicznego stwierdzenia, że **sytuacja jest tym, co bada analiza sytuacyjna, albo tym, co badacz w swoim projekcie określił jako „sytuację"** [podkr. RŻ]. (...) sytuacja, stanowiąca kontekst działania, zawsze jest konstruowana i każdorazowo powstaje na życzenie obserwatora, badacza, analityka, który ją „składa", z tego, co zaobserwuje (Kacperczyk 2007:22-3).

Zauważmy, że problem płynności granic jest charakterystyczny dla krytyki teorii społecznych światów w ogóle, i był omawiany już przez A. L. Straussa (Kacperczyk 2016:37-8, Strauss 1993:214, por. część 3.1. tej rozprawy). My nie uznajemy tego za problem, bowiem z jednej strony sądzimy, że w przypadku badanego przez nas świata DS „rzeczywiście"

granice są płynne. Koncepcja płynnych granic dobrze pracuje w wybranym wycinku rzeczywistości. Z drugiej strony uznajemy naszą odpowiedzialność, czy inaczej mówiąc arbitralność decyzji o domknięciu granic, tak w sensie decydowania o punkcie wysycenia kategorii, jak i ramach czasowych, instytucjonalnych, papierowych (w sensie ilości stron) tej rozprawy. To oznacza dla nas konstruowanie opisu społecznego świata, i kierując się postulatami Clarke o samoświadomości i autorefleksji badacza, nie udajemy, że jest inaczej.

Inny aspekt krytyki analizy sytuacyjnej, to problem marginalizowania działania (*action*) podmiotu ludzkiego. Kacperczyk uważa, że skoro sytuacja stanowi opakowanie dla wyznaczonych przez badacza granic własnych poszukiwań, czyli nie ma żadnego jednoznacznego desygnatu, to jest to opakowanie puste. Wypełnić należy je społecznymi działaniami i procesami. Dla niej „w złożoności sytuacji i współzależności jej wewnętrznych komponentów, mieści się konkluzywne wyjaśnienie i interpretacja ludzkich działań" (Kacperczyk 2007:23). Zgadzamy się z powyższym, ale widzimy marginalizację działania jako potencjalne zagrożenie dla konkretnego – w tym rzecz jasna naszego – projektu badawczego, a nie wadę koncepcji Clarke. Mówi ona bowiem jasno o „sytuacji działania" (*the situation of action*) (Clarke 2005:73) i swoim zaangażowaniu w badania, których centrum stanowi podmiot (*individual-centered materials*) (Clarke 2005:xxviii). Innowacyjność jej projektu polega na tym, że dąży do ujęcia pełnej sytuacji badania (ibidem), a nie na marginalizowaniu działania. Skłaniamy się w tej pracy ku poświęceniu większej uwagi działaniom podmiotu ludzkiego niż ujęciu pełnej sytuacji. Odzwierciedlają to cele szczegółowe rozprawy. Zresztą, jak wskazano wyżej, pełnia tego ujęcia to arbitralna decyzja badacza. Bliski jest nam koncept rekonstruowania świata życia badanych, czyli etnograficzne patrzenie na świat oczami jego uczestników. Ujęcie pełnej sytuacji wymaga tego, co Clarke nazywa spojrzeniem „od zewnątrz". Stosujemy tę strategię w znacznie mniejszym stopniu niż spojrzenie „od dołu", oraz – o ile to możliwe – spojrzenie „od środka" (por. Clarke i in. 2015:21–22).

Krytykowane są także mapy Clarke, które mają pomagać w ujęciu dynamicznego, złożonego i wielowymiarowego świata, a są zazwyczaj tylko dwuwymiarowymi, czarno-białymi, statycznymi rysunkami. Zatem takie mapy dalece nie wystarczają do tak postawionego zadania, a nierzadko powodują zamieszanie i bałagan (*confusion and clutter*), czyli nie przekazują klarownego i zrozumiałego komunikatu (Uri 2015:149). Krytyka ta jest zasadna, choć niezrozumiały może być przecież każdy sposób komunikacji naukowej (narracja, tabela, lista, wykres, schemat itd.). Z pomocą przychodzi nam DS, z metodami i

technikami wizualizacji danych ilościowych, oraz innymi wizualizacjami charakterystycznymi dla tego świata (por. rys. 2). Prezentacja graficzna danych w sposób dostarczający informacji wielowymiarowej, złożonej, jak najpełniejszej, a przy tym czytelnej, rzetelnej i atrakcyjnej wizualnie jest tematem lub elementem niezliczonej ilości podręczników i materiałów szkoleniowych, z których korzystają data scientiści i ogólnie mówiąc różni analitycy danych (Anscombe 1973; Biecek 2015b, 2016; Biecek, Baranowska, i Sobczyk 2019; Chibana 2017; Gągolewski i in. 2016; Ismay i Kim 2018; Leek 2015; Peng 2016a, 2016b; Szeliga 2017; Tufte 1983, 1990; Wickham 2010; Wickham i Grolemund 2017). Nieudolność twórcy, manipulacja mająca wywołać konkretne wrażenia na odbiorcy, duża ilość grafiki przy małej ilości informacji (*junk charts*) – np. w infografikach dziennikarskich czy wykresach naukowych – są wskazywane, krytykowane, a także wyśmiewane (Biecek 2018e; Fung 2014; Munroe 2010; Wainer 1984). Wizualizacja danych uznawana jest za bardzo ważny element w procesie analizy/eksploracji danych[48], kiedy analitycy wykonują wykresy sami dla siebie, jak również w komunikacji wyników dla interesariuszy czy zleceniodawców[49] (Leek 2015:23–33; 58–59).

Piszący te słowa przyznaje, że po pewnym czasie badań terenowych, a więc kontaktu z obowiązującym w świecie DS kanonem wizualizacji[50], a także mając doświadczenie zawodowe i badawcze w graficznym prezentowaniu danych ilościowych, stykając się po raz pierwszy z charakterystycznymi dla Clarke mapami (por. Kacperczyk 2007:12 rys. 2.; 16 rys. 4.) miał wrażenie archaizmu i amatorstwa tych rysunków[51]. Stosujemy jednak w tej rozprawie podejście do map proponowane przez A. E. Clarke, traktując je zarówno jako cenne ćwiczenie analityczne (naszą „eksplorację danych"), jak i wyjątkową na gruncie badań jakościowych prezentację konstrukcji analitycznych („komunikację wyników badań"). W pewnych przypadkach po prostu rozbudowujemy te mapy o kolejne wymiary (por. rys. 43.), tak jak czyni się to prezentując dane ilościowe w myśl „warstwowej gramatyki grafiki

48 Sztandarowym przykładem jest kwartet Anscombe'a z 1973 r., który doczekał się twórczego rozwinięcia w postaci Datasaurusa w 2016 i Datasaurusa Dozen w 2017. Podobnym przykładem jest Paradox Simpsona. Zasadniczo te przykłady wskazują, że wizualizacja danych jest niezbędna dla zrozumienia danych. Nie wystarczą ani statystyki opisowe, ani modelowanie (por. Piątkowska 2017).

49 Typowe przykłady najlepszych komunikacji wyników analiz ilościowych to: róża Nightingale (zwana wykresem polarnym lub grzebieniowym), mapa przedstawiająca losy armii napoleońskiej na różnych etapach inwazji na Rosję autorstwa Charlesa Minarda, czy wykres pudełko-wąsy Johna Tuckey'a (por. Biecek 2016).

50 Proponujemy czytelniczce / czytelnikowi samodzielne spojrzenie na możliwości wglądu (i wyglądu) w dane ilościowe przy użyciu narzędzi takich jak np. Tableau, Power BI, pakiety „ggplot2", „rCharts" i „Shiny" dla R, pakiety „plotly", „bokeh" i „dash" dla Pythona.

51 Szablony map są dostępne na https://study.sagepub.com/clarke2e/student-resources/templates, dostęp 14.07.2018 r., jako dodatek do książki *Situational Analysis: Grounded Theory After the Interpretive Turn* (Clarke, Friese, i Washburn 2017).

(*layered grammar of graphics*)" (Wickham 2010). Stosujemy też m.in. mapę myśli umożliwiającą pokazanie klasyfikowania i hierarchii (por. rys. 3.), wykres radarowy do prezentacji jakościowej konstrukcji analitycznej (rys. 38.), diagramy prezentujące proces (rys. 24., 36., także w połączeniu z mapą segmentów świata społ. por. 44.) oraz rozmaite materiały wizualne, nierzadko przygotowane przez innych autorów.

Krytyce podlega także włączenie przez Clarke analizy dyskursu do analizy sytuacyjnej, co wskazaliśmy wyżej jako zaletę szóstą (*turning to discourse(s)*). Krytycy twierdzą, że faktycznie Clarke nie daje żadnych konkretnych wskazówek do systematycznego włączenia analizy dyskursu. Przy tym nie korzysta ona z dorobku analizy dyskursu sensu stricte, jak np. metod badania wzorców składniowych wypowiedzi. Tym samym czytelnik może mieć wrażenie, że w pracy Clarke powstało „wiele hałasu o czymś (*much ado about something*)", ale ostatecznie nie wiadomo o czym, ani jak to zastosować (Martin 2006:123-24). Brak metod analizy dyskursu ma uniemożliwiać analizę zgodną z podejściem Foucault (Diaz-Bone 2013).

Krytyka wystosowana została rzecz jasna również ze strony zwolenników tradycyjnej teorii ugruntowanej. Rozumienie klasycznej teorii ugruntowanej ma być u Clarke bardzo uproszczone – co jest poważnym zarzutem przy wołaniu autorki o unikanie nadmiernych uproszczeń – i zasadza się na utożsamieniu MTU z podstawowym procesem społecznym. Vivian Martin wskazuje, że to utożsamienie jest błędne, zatem propozycja odejścia od podstawowego procesu na rzecz złożoności sytuacji jest, jej zdaniem, naprawą czegoś, co nie jest zepsute (*Clarke proceeds to fix what is not broken*). Autorka ta ma ponadto ignorować tę część klasycznej teorii ugruntowanej, która nie wyrasta z symbolicznego interakcjonizmu, tj. systematyczne, rygorystyczne, „naukowe" (*scientific if you will*) podejście Glasera charakteryzujące Uniwersytet Columbia. Ta ignorancja ma być regresem w badaniach jakościowych, powrotem do lat 60., gdy nauki społeczne krytykowały większość tego rodzaju badań jako „zupę z dowodów anegdotycznych (*a soup of anecdotal evidence*)" (Martin 2006:121–22). Akurat zarzut niesystematyczności czy braku rygoru jest, w naszej ocenie, niezasadny, ponieważ Clarke dodając nowe narzędzie analityczne (mapy) bazuje na kodowaniu i pisaniu notatek oraz not teoretycznych takim, jak w tradycyjnej teorii ugruntowanej. Poza tym mapy mają pomóc w znanym z MTU procesie jednoczesnej analizy, kodowania i pisania not wraz ze zbieraniem danych i teoretycznym dobieraniem próbek. Stymulują one badacza do myślenia i otwierają dane, co ma pomóc przezwyciężać paraliż analityczny (*analytic paralysis*) (Clarke 2005:xxi; 83-86). Chyba, że Martin mówi o nastawieniu Clarke na zauważenie w badanym wycinku rzeczywistości społecznej

sprzeczności i chaosu, a nie poszukiwanie „czystości" (czyli teoretyzowaniu zamiast teorii), z tym jednak trudno nam dyskutować, bowiem zasadniczo są to różnice ontologiczne.

W języku klasycznej teorii ugruntowanej to, że Clarke bazuje na koncepcji społecznych światów/aren nazwano wymuszaniem danych. Społeczny świat/arena jest tutaj traktowany jako silny (*potent*) kod teoretyczny, który „będzie kształtować ostateczny projekt badawczy zanim badacz zacznie badanie terenowe" (Martin 2006:122). W ogóle kwestia statusu uprzedniej wiedzy czy prekonceptualizacji badacza u Clarke jako koniecznej, uświadamianej i potencjalnie wartościowej prowadzi do pytania, czy w jej wydaniu teoria ugruntowana została ponownie ugruntowana (*re-grounded*), czy może „odgruntowana"? Zgadzamy się z Kacperczyk, która uważa, że Clarke nie dokonuje rewitalizacji czy ożywienia MTU, tylko proponuje nowe podejście, „ponowne wcielenie".

> Jest to kompletne zerwanie z tradycyjną, *pure* Grounded Theory – *pushing GT around the Postmodern Turn* [podkreślenia org.] oznacza w istocie wypchnięcie jej poza obszar jej zakorzenienia, i to już nie jest ta sama GT (Kacperczyk 2007:25).

Postmodernistyczne podejście Clarke załamuje wizję generowania teorii z danych (ibidem). Proponowany jest szereg – mówiąc językiem MTU – kodów teoretycznych: od samego pojęcia społecznych światów/aren, przez wszelkie pojęcia i procesy w takich światach, elementy sytuacji działania. Proponowane jest także zadawanie konkretnych, w żadnym razie nie ugruntowanych w danych terenowych, pytań o tak skonceptualizowany świat społeczny/arenę (Clarke 2005:115-16), i stosowanie konkretnych strategii analitycznych z użyciem map – wizualnych i konceptualnych szablonów (Clarke 2005:83-144), zatem z pewnością „prekonceptualizacja" jest silna. Podejście to jest konstruktywistyczne i jako takie je przyjmujemy. Jak napisaliśmy wyżej, **zakładamy, że data science jest społecznym światem. Tym samym dokonywany w tej pracy opis świata społecznego jest przez nas konstruowany, nie odkrywany**.

4.2. Techniki badawcze i analityczne

Zastosowaliśmy typową dla badania społecznych światów, a szczególnie dla analizy sytuacyjnej, etnografię wielostanowiskową (Clarke 2005:xxxiii; Clarke, Friese, i Washburn 2015:46; Marcus 1995). Źródłem danych jakościowych były wywiady swobodne, obserwacje uczestniczące, a także elementy autoetnografii, netnografii i etnografii opartej na współpracy (*collaborative ethnography*). Jest to zdecydowanie dominujący komponent naszych badań empirycznych. Próba dobrana została w sposób nieprobabilistyczny, celowy, zgodnie ze

strategią próbkowania teoretycznego. Materiał jakościowy był analizowany za pomocą kodowania, kilku rodzajów not i map. Wyniki prezentujemy w częściach 5.2. – 5.4. oraz całej części 6.

Do technik etnograficznych dołączyliśmy ilościową analizę (na poziomie opisu i eksploracji, nie wnioskowania statystycznego) danych zastanych dwojakiego rodzaju. Po pierwsze, wykonana została wtórna analiza wewnętrznych sondaży „środowiska" DS; sondaże te traktujemy jako akty samowiedzy badanego świata. Po drugie, przeanalizowaliśmy dane z serwisu meetup.com, popularnej platformy służącej organizacji spotkań, m.in. osób zainteresowanych data science. Wyniki, wraz z innymi informacjami ilościowymi (pozyskanymi głównie ze źródeł o charakterze aktów samowiedzy świata DS), prezentujemy przede wszystkim w części 5.1., ale pojawiają się też w innych fragmentach pracy. Celem tej części badań było przede wszystkim ujęcie pełnej sytuacji badania, zaś dodatkowo enkulturacja badacza w działaniu zbliżonym do działania podstawowego badanego świata społecznego, a w konsekwencji głębsze zrozumienie kategorii jakościowych.

W całym procesie badań i analiz kierowaliśmy się sugestiami Clarke, co do pytań badacza wobec badanych światów/aren (Kacperczyk 2016:573):

- Pytania dotyczące świata społecznego:
  - Na czym polega praca każdego świata?
  - Jakie są główne zobowiązania (*commitments*) danego świata?
  - Jak uczestnicy wyobrażają sobie, że powinni wypełniać te zobowiązania?
  - Jak świat opisuje sam siebie – prezentuje siebie – w swoich dyskursach?
  - Jak opisywane są inne światy na arenie?
  - Jakie działania były podejmowane w przeszłości i są przewidywane w przyszłości?
  - Jakie technologie są używane i uwikłane w daną sytuację?
  - Czy istnieją szczególne miejsca, gdzie organizowane są działania? Jakie one są?
  - Co jeszcze wydaje się ważne w tym społecznym świecie?
- Pytania dotyczące aren:
  - Na czym koncentruje się arena (na jakim zagadnieniu, jakiej sprawie)?
  - Jakie społeczne światy są obecne i aktywne na danej arenie?
  - Jakie społeczne światy są obecne i spodziewane (na tej arenie), a jakie nieobecne, choć można byłoby się ich spodziewać?

- Czy są jakieś światy nieobecne, jakich można byłoby oczekiwać, że się pojawią?
- Jakie są najbardziej gorące sprawy/kontestowane tematy/bieżące kontrowersje na arenach dyskursu?
- Czy w dyskursie pojawiają się jakieś zaskakujące przemilczenia? – Co jeszcze wydaje się ważne na temat tej areny?

- Możliwe kierunki bardziej pogłębionych analiz:
  - Intensywne skupienie uwagi na pracy konkretnego świata społecznego; – Intensywne skupienie uwagi na technologii, jakiej świat społeczny używa lub wytwarza oraz jak ona krąży wewnątrz oraz pomiędzy światami;
  - Intensywna koncentracja na działaniach podejmowanych przez konkretne światy w odniesieniu do konkretnych spraw;
  - Intensywne skupienie uwagi na procesie konstruowania granic pomiędzy światami przez różne światy na arenach oraz dyskursy na ten temat;
  - Intensywne skupienie uwagi na dyskursach produkowanych przez dany świat lub światy na arenie;
  - Intensywne skupienie uwagi na szerokiej arenie dyskursu (która może zawierać w sobie inne areny).

4.2.1. Badania jakościowe

Naszym najważniejszym źródłem danych jakościowych były indywidualne **wywiady swobodne ukierunkowane** (Lutyński 1974; za: Przybyłowska 1978:62–63 por. Konecki 2000:170), prowadzone wadzone w duchu wywiadu intensywnego (Charmaz 2009:39). Prowadzący taki wywiad ma przygotowaną listę potrzeb informacyjnych, w trakcie wywiadu zgłębia tematy dochodząc do zagadnień szczegółowych, pytania zadaje w sposób niestandaryzowany. W technice tej akcentuje się możliwość uzyskania od uczestnika jego interpretacji doświadczeń. Uczestnik jest traktowany jako ekspert (Charmaz 2009:41).

Pytania otwarte w naszych wywiadach dotyczyły zagadnień ogólnych i szczegółowych, odwołując się do doświadczenia Rozmówczyń / Rozmówców[52]. Wywiady przeprowadził autor rozprawy i jedyny badacz (RŻ), starając się dostosować treść i język pytań do partnerów/partnerek rozmowy. Badacz prowadził rozmowy z pozycji laika, prosząc o tłumaczenie znanych mu terminów, raczej nie używał języka świata DS jako „swojego”. W

52 Stosujemy tu wielkie litery ze względu na szacunek do osób, które uczestniczyły w naszych wywiadach. Zaznaczamy płeć danej osoby.

każdym wywiadzie stosował zapis rozmów rejestrujący (nagrywanie na dyktafon cyfrowy w smartfonie, poprzedzone świadomą zgodą Rozmówców). Zależało nam na bezpośrednim kontakcie z Rozmówcami, liczyliśmy na możliwość obserwacji w ich „normalnym" otoczeniu (co było trudne w realizacji, por. część 4.3). Wywiady miały zbliżać się do rozmowy, która „odbywa się w naturalnym kontekście potocznego życia i pracy osób objętych naszym badaniem" (Konecki 2000:179).

Można rozróżnić wywiady swobodne ze względu na cel: rozpoznawanie zjawiska albo sprawdzanie hipotez (Kvale 2004:105). W naszych badaniach znacznie ważniejszy był pierwszy z nich. Zwracamy szczególną uwagę na unikanie „naiwnego empiryzmu". Badacz, który wierzy w neutralny dostęp do obiektywnej, niezależnej od siebie rzeczywistości społecznej wykazuje się właśnie naiwnym empiryzmem. Badacz społeczny – „naiwny empirysta" – prowadząc wywiad zbiera reakcje werbalne respondenta tak, jak entomolog zbiera owady czy glacjolog próbki lodu. Ów badacz-górnik chce wydobyć wiedzę z przedmiotów tak, jak wydobywa się rudę drogocennego metalu z ziemi. Ruda czeka na odkrycie we wnętrzu, nie tknięta i nie zmieniona przez wydobywającego. Możliwa jest też inna strategia – badacz może być jak podróżnik. Taki badacz-podróżnik wędruje z tubylcami przez nieznany teren, zadaje im pytania, słucha opowieści o ich świecie. Może poszukiwać konkretnych miejsc lub wędrować swobodnie. Jego relacja z podróży jest rekonstrukcją rozbudowaną o własne interpretacje. Podróż może doprowadzić podróżnika do przemiany, skłonić go do refleksji nad uznawanymi wcześniej wartościami czy obyczajami własnego kraju (Kvale 2004).

Badacz-podróżnik uznaje, że wywiad to rozmowa, w której dane są współtworzone – rodzą się w relacji badacza z Rozmówcą. Narzędziem badawczym jest osoba prowadząca wywiad (Konecki 2000). Kvale wylicza kwalifikacje badacza prowadzącego wywiad swobodny – traktowaliśmy je w naszej pracy jako zalecenia do realizacji wywiadów (por. Kvale 2004:153-154 [ramka 8.2]). Prowadzący wywiady swobodne jest / powinien:

- być biegły w temacie wywiadu; może rozmawiać o temacie z uczestnikami, zna dobrze główne aspekty tematu, nie próbuje olśniewać swoją wiedzą.
- nadaje porządek wywiadowi; nakreśla jego ramy.
- mówi jasno i wyraźnie; zadaje krótkie pytania, nie używa języka akademickiego ani żargonu zawodowego.

- jest delikatny; pozwala kończyć wypowiedzi, dostosowuje się do rytmu myślenia i wypowiedzi badanego, toleruje milczenie, daje do zrozumienia, że badany może przedstawiać nietypowe poglądy czy silne emocje.
- jest wrażliwy; jest empatyczny, zwraca uwagę na sygnały werbalne i niewerbalne, słucha aktywnie starając się uzyskać opis emocjonalnych odcieni, dostrzega jakich tematów wywiadu nie rozwijać ze względu na towarzyszące im emocje.
- jest otwarty; słyszy jakie aspekty wywiadu są szczególne ważne dla badanego, jest otwarty na aspekty, które wprowadza badany.
- kieruje przebiegiem wywiadu, nie boi się przerwać wypowiedzi nie związanych z celem wywiadu.
- jest krytyczny, dopytuje mając na uwadze wcześniejsze wypowiedzi, jak i ich logiczną spójność.
- pamięta, co zostało powiedziane, może nawiązać do wcześniejszych części wywiadu.
- interpretuje, podczas wywiadu dostarcza interpretacji wypowiedzi badanego do której ten się ustosunkowuje.

Podane kwalifikacje nie przekładają się bezpośrednio na jakość wywiadu. Kryteria oceny tej jakości nie są wyraźne i zależą od celu i treści wywiadu. Osoby doświadczone i biegłe w prowadzeniu wywiadów mogą wyjść poza te zalecenia czy świadomie je łamać. Kvale podaje przykład badania, w którym prowadzący wywiad o sensie władzy zagrał rolę osoby autorytarnej – przerywał badanym, był wymagający i wrogi – dzięki czemu uzyskał spontaniczne reakcje badanych na władzę (2004:154-155). Pomocny dla badacza (RŻ) był także wpis na blogu Ethnography Matters (Annechino 2016), z ogólnymi wskazówkami do realizacji wywiadów jakościowych:

- wywiady z Twojego punktu widzenia [badacza] polegają głównie na słuchaniu![53]
- nie wkładaj ludziom słów w usta – mów tak mało, jak tylko się da.
- unikaj zarówno pozytywnej jak i negatywnej informacji zwrotnej.
- spodziewaj się być zaskoczony i niczego innego. Słowa do używania jako sygnał aktywnego słuchania – zamiast „mhm", „rozumiem" powiedz „ciekawe", „interesujące".[54]

53 „Jeśli mówisz więcej niż 5% czasu, to przypuszczalnie mówisz za wiele" (Babbie 2006:327).

54 „Zamiast mówić 'aha' lub kiwać głową, jakby wszystko było automatycznie jasne, badacz może stwierdzić: 'to interesujące, powiedz mi więcej na ten temat' " (Charmaz 2009:40).

- badani są ekspertami w zakresie swoich doświadczeń, Ty jesteś po to, aby słuchać i uczyć się; przekaż to stanowisko na etapie rekrutacji Rozmówców i w instrukcji przed wywiadem.
- czasami zboczenie z tematu, dygresja, mogą być najważniejsze, najbardziej informatywne, najbardziej znaczące dla wywiadu.
- rekrutowanie nieznajomych jest skuteczne – możesz np. wręczać wizytówki, wtedy ludzie mają czas na decyzję co do udziału w rozmowie.

Wywiad nie jest metodą obiektywną ani subiektywną. Jego istotą jest interakcja międzyosobowa (Kvale 2004:15-17, 75, 163). „W przeciwieństwie do sondażu, wywiad jakościowy jest interakcją między prowadzącym wywiad a respondentem" (Babbie 2006:327). Konecki (2000:179-79) przedstawił przejście od rozumienia „inter-view" jako „roz-mawiania" czyli sytuacji łączącej partnerów w relacji symetrycznej do rozumienia jako „wywiad", co w języku polskim oznacza szpiegowanie bądź instytucję szpiegowską. Przekłada się to w konsekwencji na prowadzenie wywiadów jako przesłuchania czy odpytywania. Wywiad jakościowy ma zatem formę zbliżoną do rozmowy, jednak inne zasady niż rozmowy potoczne:

> Badacz powinien wyrażać swoje zainteresowanie i chcieć wiedzieć więcej. Pytania, które są uznawane za nieprzyzwoite lub przez które można się prześlizgnąć za przyjacielską zgodą – nawet w obecności serdecznych przyjaciół – w badaniach mają kluczowe znaczenie. (…) W przeciwieństwie do normalnej konwersacji, badacz może zmieniać tok rozmowy i kierować się przeczuciami. Wywiad sięga głębiej niż zwykła rozmowa i pozwala na nowo zbadać wcześniejsze wydarzenia, poglądy i uczucia (Charmaz 2009:40).

Prowadząc wywiady swobodne RŻ starał się wykazywać empatią, dbał o wzajemne zaufanie i naturalność sytuacji. Budowanie zaufania Rozmówczyń i Rozmówców do badacza miało w naszej pracy szczególne znaczenie z uwagi na dobre wzajemne skomunikowanie osób związanych ze światem DS w Polsce. Sądzimy, że utrata zaufania u jednego Rozmówcy spowodowałaby problemy z zaufaniem u kolejnych, gdyż informacja o namawiającym do wywiadów i niewłaściwie zachowującym się doktorancie socjologii szybko obiegłaby badane środowisko. Jak wskazuje Konecki: „każde *faux pas* popełnione przez badacza może zniszczyć zaufanie budowane przez tygodnie lub miesiące" (2000:172).

To właśnie podejście do wywiadów jako rozmów oraz traktowanie Rozmówczyń i Rozmówców jako ekspertów – w zakresie świata społecznego DS i w zakresie własnego doświadczenia w tym świecie – pozwoliło nam na uzyskanie głębokiego wglądu w punkt widzenia uczestników osadzonych w badanym świecie. W trakcie indywidualnych, czasami

kilkugodzinnych wywiadów swobodnych, nie tylko stosowaliśmy pytania otwarte, dopytywanie i interpretowanie wcześniejszych wypowiedzi, ale także dosłownie prosiliśmy o wyjaśnianie konkretnych wydarzeń, technologii i innych rozmaitych elementów/procesów świata społecznego DS, z którymi zetknęliśmy się w całości naszych badań etnograficznych.

Obserwacja uczestnicząca to technika (a szerzej – rola etnografa), która służyła nam w dużej mierze jako źródło pytań badawczych do wywiadów (por. Zaród 2018:171). Czasami było odwrotnie – informację uzyskaną w wywiadzie byliśmy w stanie zinterpretować w sposób bliski punktowi widzenia uczestników badanego świata dopiero dzięki kolejnym obserwacjom uczestniczącym. Obserwacje uczestniczące były niezwykle ważne dla enkulturacji badacza (RŻ) w społecznym świecie DS (Jemielniak 2018:20). Obserwacje, w których spotykaliśmy na żywo uczestników osadzonych w świecie DS były też pomocne w nawiązywaniu kontaktów z potencjalnymi Rozmówcami.

Jak już wskazaliśmy (por. 4.1.), trudno uznać badania za etnografię, jeżeli nie stosowano obserwacji uczestniczącej. „Tym, co określa badanie etnograficzne, jest regularna i powtarzająca się obserwacja ludzi i sytuacji (…)" (Angrosino 2010:106). Obserwacja taka:

> jest prowadzona w naturalnym kontekście obserwowanych działań ludzkich i interakcji oraz podąża za naturalnym przebiegiem życia codziennego obserwowanych osób. Pozwala to badaczowi zanurzyć się w życiu potocznym, gdzie związki, korelacje oraz przyczyny zjawisk mogą być zaobserwowane bezpośrednio, tak jak się wyłaniają w trakcie obserwacji. Jakościowa obserwacja nie jest zatem ograniczona przez wcześniej ustalone kategorie pomiaru lub możliwych reakcji, ale pozwala na poszukiwanie pojęć i kategorii istotnych dla badanych osób (Konecki 2000:145-46).

Obserwacje uczestniczące autor (RŻ) przeprowadził samodzielnie. Według klasycznej typologii czterech ról badacza stosującego omawianą technikę badań jakościowych (Gold 1958), był on początkowo całkowitym uczestnikiem (nie ujawniał swojej tożsamości socjologa piszącego o DS w Polsce), później najczęściej uczestnikiem jako obserwatorem (wyjawiał tożsamość i uzyskiwał zgodę na obserwowanie, czasami był akceptowany jako członek grupy, szczególnie w sensie grupy towarzyskiej). Bez ujawnienia tożsamości RŻ nie było możliwe zapraszanie osób do udziału w wywiadach.

Według kryteriów klasyfikacji i opisu obserwacji uczestniczącej Kazimierza Doktóra (Konecki 2000:151) piszący te słowa prowadził głównie obserwację uczestniczącą ukrytą (szczególnie na początku badań, a później ukrytą przynajmniej dla znakomitej większości osób, które obserwował, a z którymi nie nawiązał rozmowy), rzadziej obserwację jawną. Ogółem w toku badań obserwacje stawały się coraz bardziej jawne, ponieważ piszący te

słowa nierzadko spotykał te same osoby (np. organizatorów meetupu) podczas kolejnych obserwacji, czasami Rozmówczynie i Rozmówców, dopytujących o przebieg badań i wyniki. W zaawansowanej fazie badań autor (RŻ) został zaproszony i wystąpił na jednym z meetupów świata DS, prezentując swoje etnograficzne szkice {obserwacja 33}. Starał się on zawsze całkowicie uczestniczyć w obserwowanym wydarzeniu, przyjmował postawę konformistyczną wobec norm badanego świata (a także wobec proponowanych działań, np. zaproszony brał udział w quizie rekrutacyjnym czy w wyjściu na piwo po warsztatach). W sensie dystansu społecznego wobec innych uczestników przyjmował pozycję równorzędną i rzadko podrzędną, podczas obserwacji częściej wykazywał bierność niż aktywność (np. raczej nie zadawał pytań prelegentom), starał się nie wyróżniać spośród uczestników zachowaniem, ubiorem, sposobem wypowiadania się, wyglądem i konfiguracją laptopa; być „kameleonem w terenie badawczym" (Geertz 1983:56).

> Pamiętaj, że jako taka, obserwacja uczestnicząca nie stanowi techniki zbierania danych, ale jest raczej rolą, jaką przyjmuje etnograf, by ułatwić sobie dostęp do potrzebnych informacji (Angrosino 2010:78).

Świadomi tego rozumienia obserwacji uczestniczącej przedstawiamy techniczne założenia realizowania tej części naszych badań. Zgodnie z zaleceniami Angrosino (2010:81-87) badacz (RŻ) starał się postrzegać działania i wzajemne interakcje ludzi w miejscu badania za pomocą wszystkich pięciu zmysłów. Podstawową formą zapisywania spostrzeżeń były notatki terenowe (*field notes*):

- zawsze opatrzone datą, godziną i miejscem obserwacji;
- z zapisem krótkiej charakterystyki osób zgromadzonych w miejscu badania (np. „jest około 15-16 osób z czego 4 kobiety, jeden mężczyzna około 45 lat pozostali 20-25")
- zawierające dosłowne fragmenty zasłyszanych wystąpień, rozmów i innych form werbalnych;
- zawierające sporządzone przez badacza opisy ludzi i nie-ludzi oraz ich działań i interakcji we wszelkich konfiguracjach – opisy staraliśmy się sporządzać w sposób jak najbardziej pozbawiony interpretacji (np. „prelegent zaczyna mówi cicho, jąka się, powtarza dwa razy zdanie, patrzy w ekran komputera" a nie „prelegent wygląda na zestresowanego");
- zawierające opis miejsca badania i jego cech fizycznych;
- zachowujące chronologię wydarzeń.

Najczęściej były to notatki terenowe w formie cyfrowej, zapisywane przez smartfon na serwerze (tzw. chmurze, np. Dropbox lub usłudze Google) w celach zabezpieczenia przed utratą danych i wygody w postaci synchronizacji plików na różnych urządzeniach. W czasie warsztatów polegających na pisaniu kodu w językach programowania badacz (RŻ) notował przeważnie na laptopie, w środowisku programistycznym (*Integrated development environment*, IDE) pomiędzy linijkami kodu, za pomocą komentarzy, czyli napisów nie interpretowanych przez komputer jako polecenia (bardziej szczegółowy opis narzędzi badacza w części 4.3., zaś pisania kodu i narzędzi używanych do tego w świecie DS w częściach 5.2. i 5.3.). Rzadziej fotografował, o ile było to możliwe zbierał artefakty fizyczne (ulotki, naklejki, firmowe gadżety itp.), zawsze zaś zbierał i wchodził w interakcje z artefaktami cyfrowymi. Były to zarówno materiały udostępniane przez organizatorów wydarzeń (np. prezentacje, grafiki, kod) jak i same strony internetowe wydarzeń, meetupów czy różnego rodzaju materiały cyfrowe, o których dowiadywał się podczas obserwacji. Tym samym obserwacje łączyły się płynnie z netnografią.

Obserwacje łączyły się także z **autoetnografią**, gdyż notatki sporządzane w przerwach lub po zrealizowaniu konkretnej obserwacji w terenie przybierały czasem autoetnograficzny charakter, tzn. dotyczyły własnych odczuć, stanów wewnętrznych, pojawiających się w trakcie obserwacji myśli oraz własnych działań i interakcji z innymi ludźmi w miejscu badania. Ponieważ „rzeczywiste praktyki wytwarzania wiedzy mieszczące się pod pojęciem ‘autoetnografii’ trudno uznać za jednolitą metodę” (Kacperczyk 2014:33), doprecyzujmy znaczenie autoetnografii w naszych badaniach. Traktujemy ją jako technikę otrzymywania danych jakościowych, w której badacz tworzy autonarracyjne notatki, bazując na umysłowym procesie autorefleksyjnym i autoanalitycznym (Kacperczyk 2014:37–38). Według Clarke po zwrocie postmodernistycznym autorefleksja badacza nie jest opcją, jest ona konieczna (Uri 2015:146).

Celem głównym stosowania autoetnografii w naszych badaniach świata społecznego DS było zwiększenie samoświadomości badacza i odsłonięcie informacji składających się na jego intelektualne tło (*wallpaper*). Zgodnie z postulowaną przez Clarke zmianą stanowiska badacza „z ‘wszechwiedzącego analityka’ *(all-knowing analyst)* do ‘samoświadomego uczestnika’ (*acknowledged participant*) w procesie produkcji zawsze częściowej wiedzy” (Clarke 2003:555–56). Staraliśmy się zatem o „jednoczesne bycie outsiderem i insiderem podczas przeprowadzania swojego badawczego projektu” (Darmach 2017:95). Jest to bliskie programowi socjologii refleksyjnej Pierre’a Bourdieu i Loïca Wacquanta, w sensie

refleksyjnego podejścia do procesu badawczego i jego uwarunkowań, mające pomagać badaczowi poznać przyczyny własnych działań badawczych i osadzić je w sytuacji społecznej i politycznej (Bielecka-Prus 2014:80–81). Tym samym stosowaliśmy przede wszystkim autoetnografię analityczną, czyli problematyzowaliśmy proces badawczy oraz relacje badacza z badanym światem, a w znacznie mniejszym stopniu autoetnografię ewokacyjną, czyli zdobywanie informacji o badanym wycinku rzeczywistości społecznej poprzez koncentrację np. na własnych emocjach, myślach, doznaniach fizycznych (ibidem:85). Zgadzamy się ze stanowiskiem, że „ostatecznie decyzja, czy tworzyć autoetnografię analityczną czy ewokatywną jest decyzją o charakterze etycznym" (Kacperczyk 2017:141). Powracamy do problemu w części 4.3.2. tej pracy.

Skoro **netnografia** jest, zgodnie z intencją twórcy terminu, badaniem etnograficznym społeczności w internecie (Kozinets 2003), to badanie świata DS nie mogłoby zostać zrealizowane bez tego elementu. Nie można bowiem uczestniczyć w tym świecie bez korzystania z rozmaitych platform, stron i narzędzi internetowych. Świadomi braku systematyczności pojęć dotyczących badań jakościowych z użyciem internetu (Jemielniak 2018:11–13) nazywamy użyte techniki netnografią w rozumieniu, że jest to „odmiana etnografii zaadaptowana do badania kultur i społeczności wirtualnych oraz do badań nad wirtualnymi aspektami życia społecznego" (Kacperczyk 2016:639). W netnografii „źródłami danych może stać się każdy dający się zauważyć i wyodrębnić element przestrzeni internetowej" (ibidem:641), jak np. dokumenty (pliki tekstowe, graficzne, audio, wideo, a w naszych badaniach także fragmenty kodu i pliki z kodem - skrypty), posty i komentarze, fragmenty rozmów i korespondencji mailowej, strony internetowe i ich elementy graficzne i sterujące (jak np. przyciski, wyszukiwarki, reklamy, banery, linki), strony/media społecznościowe, fora, blogi. Przeważnie nasze badania netnograficzne były nastawione na analizę treści, a nie analizę zdarzeń i interakcji (ibidem). Stąd w niniejszej pracy przyjęliśmy konwencję cytowania treści ze źródeł internetowych, używanych przez uczestników badanego świata (takich jak np. repozytorium kodu GitHub, forum StackOverflow, strona społecznościowa LinkedIn) w sposób identyczny jak innych źródeł literaturowych zamieszczonych w sieci.

Etnografia oparta na współpracy była przez nas wykorzystywana w niewielkim stopniu i w sposób niepełny, jednak chcemy ją w tym miejscu odnotować z uwagi na bardzo istotny wkład konsultacji z osobami biorącymi udział w naszych badaniach. Taka współpraca

pozwoliła m.in. na przygotowanie ostatecznej, kompletnie innej od pierwotnej, wersji opisu działania podstawowego badanego świata, (por. 5.2.). W założeniu Luke'a Erica Lassitera

> etnografię opartą na współpracy (*collaborative ethnography*) możemy określić ogólnie jako podejście do badań etnograficznych, które w sposób zamierzony i wyraźny ujawnia oraz podkreśla elementy współpracy na każdym etapie procesu badawczego – od konceptualizacji projektu, poprzez pracę terenową, do ważnego w tym kontekście etapu pisania (2005:16).

Chodzi jedynie o współpracę badaczy z osobami na różnych etapach badań i pisania rozprawy. Jej autor (RŻ) konsultował się w sposób nieformalny z osobami osadzonymi w badanym świecie. Najczęściej były to rozmowy po wywiadach, kilkukrotnie przesłanie zainteresowanym osobom pierwszej wersji fragmentu tekstu rozprawy, najrzadziej były to konsultacje teoretycznego pomysłu przed napisaniem właściwego fragmentu. Bez wątpienia jednak współpraca była niepełna, bo piszący te słowa pozostał w roli etnografa „mającego ostatnie słowo" i nie włączał konsultantów / konsultantek uczestniczących w badanym świecie do wielogłosowej wypowiedzi jako równorzędnych autorów / autorek swoich tekstów (por. Pietrowiak 2014:26–27).

Procedurą techniczną, która została zastosowana w procesie analizy danych jakościowych pozyskanych za pomocą wymienionych wyżej technik było **kodowanie**[55]. Generalnie jest ono wydobywaniem kategorii i ich własności z materiałów jakościowych. Nie kategoryzuje się danych jakimś „kluczem", a kategorie z tychże danych się wyprowadza (Kvale 2004:48).

> Kodowanie to główne ogniwo między zbieraniem danych a budowaniem wyłaniającej się teorii, które ma na celu wyjaśnienie tych danych. Poprzez kodowanie można zdefiniować to, co się dzieje w danych, i zacząć rozszyfrowywać znaczenie tych faktów (Charmaz 2009:63).

Kody to *summative statements*, przy czym Strauss podkreślał, że kody nie mają być tylko podsumowaniem lub nazwaniem kategorii, a powinny uwzględniać sześć elementów (Konecki 2000:48-49):

- warunki przyczynowe wystąpienia zjawiska,
- samo zjawisko,
- kontekst występowania zjawiska,

55 Słowo „kodowanie" występuje także w świecie DS w znaczeniu „pisanie kodu". Pisanie kodu oznacza czynność zapisywania serii instrukcji do wykonania przez komputer za pomocą klawiatury komputerowej, przez człowieka, w języku programowania zrozumiałym dla komputera (Nowosad 2019). Ta czynność jest w naszym ujęciu składnikiem działania podstawowego w świecie DS, o czym w części 5.2. tej rozprawy.

- warunki interweniujące (czyli to, co dodatkowo wpływa na zjawisko, np. zmienia jego nasilenie),
- działania / strategie i techniki interakcyjne (taktyki ludzi dotyczące zjawiska i konkretne zachowania realizujące te taktyki),
- konsekwencje zjawiska.

Kodowanie można podzielić na rzeczowe i teoretyczne. Pierwsze z nich służy konceptualizacji rzeczowej, drugie konceptualizacji relacji między kategoriami, czyli budowaniu hipotez (Konecki 2000:51). Konecki rozróżnia w kodowaniu rzeczowym kodowanie otwarte (I) i selektywne (II) (2000:52-53), Charmaz kodowanie wstępne (I) i skoncentrowane (II) (2009:66-81). Rozróżnienie sprowadza się do dyrektywy, by najpierw (I) kodować cały materiał jakościowy, np. wers po wersie, wydarzenie po wydarzeniu czy nawet słowo po słowie (lub zdaniami, akapitami itp.), i w sposób krytyczny, otwarty, raczej szybki i spontaniczny, zaetykietować materiał (kodowanie to prowadzi do spełnienia kryteriów dopasowania i istotności teorii ugruntowanej). W praktycznym wymiarze pracy badacz powinien więc zwrócić uwagę m.in. na język stosowany przy kodowaniu wstępnym. Kody są bowiem konstruowane, tzn. badacz aktywnie nazywa fragmenty materiału jakościowego - dobiera słowa, które jego zdaniem ujmują to, co w określonym fragmencie jest istotne. Istnieje postulat, by używać języka odzwierciedlającego działania, nie tematy. Ma to pomóc w ograniczeniu własnych prekoncepcji (Charmaz 2009:65-66).

Następnie (II), w kodowaniu selektywnym czy skoncentrowanym należy dokonać wyboru kodów analitycznie najistotniejszych (np. najbardziej interesujących badacza z uwagi na znaczenie, najczęstszych, najbardziej generalnych tzn. występujących w największej liczbie przypadków, zaś w klasycznej MTU kodów należących do kategorii centralnej) i kodować w sposób bardziej wybiórczy, kierunkowy. Kody otwarte mogą zatem na etapie kodowania selektywnego stać się samodzielnymi kategoriami, lub etykietami „klastrów" kodów (czyli inne kody grupujemy pod tą nazwą kodu np. na podstawie znaczenia lub współwystępowania kodów ze sobą); takie klastry możemy też nazwać inaczej niż istniejący kod, np. przyjąć nazwę bardziej ogólną. Z wybranych kodów otwartych możemy rezygnować, szczególnie po pierwszych cyklach kodowania otwartego, gdyż zalecane jest – i tak właśnie postępowaliśmy w naszych badaniach świata DS – by kodować jeden fragment materiału jakościowego na kilka sposobów. Tym samym każda jednostka materiału jakościowego (jak słowo/zdanie/akapit/wydarzenie) może być zakodowania więcej niż jeden raz, a przypisane jej kody mogą mieć charakter równorzędny lub hierarchiczny (zawierają się w sobie) (Babbie

2006:409). W naszych analizach nie stosowaliśmy hierarchii kodów, traktując sytuacje tego rodzaju jako wskazanie do przekodowania kodów „podrzędnych" do „nadrzędnych", włączając znaczenia stojące za kodem „podrzędnym" do własności kategorii kodu „nadrzędnego". Kody mają charakter prowizoryczny, badacz wypróbowuje różne kody na tych samych danych, odrzuca większość i stara się wyselekcjonować te pasujące najlepiej – a zazwyczaj będzie pasować więcej niż jeden kod (Clarke 2005:84).

Cykle kodowania I – II należy wykonywać wielokrotnie w trakcie realizacji projektu badań jakościowych, i tak też czynimy, o czym więcej w części 4.3. Po zakodowaniu rzeczowym jakiejkolwiek części materiałów jakościowych (III) należy wykonać notatkę (*memo*). Takie notatki, zawsze opatrzone datą, stanowią dokumentację pracy projektu badawczego i powinny zawierać uzasadnienia poczynionych podczas kodowania decyzji. Pomagają także badaczowi na obserwowanie swoich postępów w kodowaniu i utrzymanie porządku w procesie kodowania. *Memo* to zatem pisane niemalże codziennie podsumowania, dotyczące m.in. czego jako badacze się spodziewaliśmy, a co nas zaskoczyło. W notatkach zwracamy także uwagę na wzorce podobieństw lub sprzeczności, opisujemy problemy i dalsze pomysły. „Część tego, co zapiszesz podczas analizy, może znaleźć się w twoim końcowym raporcie; większość tych notatek będzie co najmniej inspiracją (…)" (Babbie 2006:411). Stosowaliśmy także – za ujęciem Straussa i Corbin (1990:197) – noty teoretyczne (o których dalej) oraz notatki kodowe. Notatki kodowe to przypisanie interpretowanego przez badaczy znaczenia do nazw kodów, co uznaje się za szczególnie ważne z uwagi na używanie w naukach społecznych ogromnej ilości pojęć mających rozmaite znaczenia potoczne (Babbie 2006:411). W używanym przez nas oprogramowaniu CAQDAS (QDA Miner, o czym więcej w 4.3.) dla notatek stanowiących dokumentację (*memo*) stosowaliśmy funkcję notatek do projektu (Project → Notes) z wygodnym przyciskiem wstawienia aktualnej daty i czasu, dla notatek kodowych zaś funkcję opisu kodu w książce kodowej (Code definition → Description), łatwą do podejrzenia i edycji w czasie kodowania materiału.

Kodowanie teoretyczne dotyczy konstruowania hipotez o relacjach pomiędzy utworzonymi kodami / kategoriami. Jest ono prowadzone na podstawie kodów selektywnych (Charmaz 2009:86); można wykonywać też kodowanie selektywne i teoretyczne równolegle (Konecki 2000:54-55) i stosowaliśmy głównie taką strategię. Przypomnijmy, że z uwagi na przyjętą ramę teoretyczną i pojęciową także sam społeczny świat/arena jest silnym (*potent*) kodem teoretycznym, który kształtuje projekt badawczy (por. Martin 2006:122). Narzędziem kodowania teoretycznego są noty teoretyczne (*theoretical memo*), czyli „zapisane w języku

teoretycznym myśli badacza o zakodowanych kategoriach i ich relacjach wobec siebie" (Konecki 2000:55-56). Noty powinny być zatytułowane i odnosić się do danego fragmentu materiału jakościowego (ibidem). Nierzadko takie myśli pojawiały się u piszącego te słowa podczas kodowania selektywnego i były natychmiast zapisywane, albo w QDA Miner za pomocą funkcji komentarza do zakodowanego fragmentu (Comment) albo w notatniku ogólnego przeznaczenia Google Keep.

Z materiałów jakościowych wyłaniają się charakterystyczne terminy używane przez uczestników badania. Można przyjąć je jako tzw. kody *in vivo* (dosł. „na żywym [osobniku]"). Terminy te kryją skondensowane znaczenie, mogą pomóc badaczom np. w odzwierciedleniu perspektywy uczestników badanego świata społecznego. Stosowanie kodów *in vivo* pomaga zachować symboliczne znaczenie pewnych wypowiedzi czy działań, odzwierciedlić założenia czy imperatywy kształtujące działanie, a w szerszym sensie uczynić analizę bliższą światu badanych. Badacz nie powinien jednak stosować terminologii uczestników tak, jak robią to oni, lecz podejść do terminów w sposób krytyczny – szukać ich znaczeń. Owe znaczenia można również przekształcić w kody innego typu (Charmaz 2009:76-79). Możliwe jest też odkrywanie hipotez *in vivo*. Badacz podczas pracy terenowej może zauważyć pewne powiązania między kategoriami na żywo, w chwili, w której zachodzą (Konecki 2000:29-30). Tego typu spostrzeżenia mogą być zapisywane np. w notatkach terenowych z obserwacji (*field notes*), w *memo*, czy jakichkolwiek szybkich notatkach nieformalnych (tekstowych lub głosowych).

A. E. Clarke uważa, że stosując analizę sytuacyjną społecznych światów/aren można i powinno się używać wywodzącej się z klasycznej MTU strategii kodowania danych jakościowych – strategii opisanej w pracach Straussa, Straussa i Glasera, Straussa i Corbin oraz Charmaz (Clarke 2005:42; 142). Tę, stosowaną w naszych analizach strategię, opisaliśmy powyżej bazując głównie na pracy Koneckiego (2000), który wprowadził MTU do polskiej socjologii. Stosowaliśmy także, na różnych etapach badań i analiz, zaproponowane przez Clarke **mapy** (Clarke 2005:86; tłum. Kacperczyk 2007:11):

- sytuacyjne (*situational maps*) – do artykułowania elementów sytuacji i sprawdzania powiązań między nimi;
- mapy społecznych światów/aren (*social worlds/arenas maps*) – jako kartografia kolektywnych zaangażowań, relacji społecznych i obszarów działania;

- mapy pozycyjne (*positional maps*) – jako strategie upraszczające do zaznaczania pozycji wyartykułowanych i niewyartykułowanych w dyskursach.

Mapy przygotowywać można zarówno dla danych jakościowych już zakodowanych, zakodowanych częściowo lub wcale. Przede wszystkim jednak badacze powinni być dobrze zapoznani z danymi

Jak wspominaliśmy przy okazji omówienia krytyki wobec podejścia A. E. Clarke, mapy mają pomóc w znanym z MTU procesie jednoczesnej analizy, kodowania i pisania not wraz ze zbieraniem danych (obserwacjami, wywiadami, netnografią itd.) i teoretycznym dobieraniem próbek. Zdaniem Clarke proces wykonywania map i same mapy stymulują badaczy do myślenia i otwierają dane jakościowe, co ma badaczom pomagać przezwyciężać paraliż analityczny (*analytic paralysis*) (Clarke 2005:xxi; 83-86).

> Funkcją map proponowanych przez Clarke jest: otwierać dane, przełamywać impasy w trakcie analizy, stymulować do analizy relacyjnej i odsłaniać obszary przemilczeń (*sites of silence*) (Kacperczyk 2007:17).

Dla piszącego te słowa mapy rzeczywiście były pomocne w powyższych kwestiach. Ponieważ wykonywanie map ma stymulować i naszym zdaniem stymuluje do myślenia m.in. o relacjach czy przemilczanych obszarach, zgodnie z zaleceniami Clarke (2005:84; 89-90) po wykonaniu danej mapy (*mapping session*) zapisywaliśmy notatkę do dokumentacji projektu (*memo*), ewentualnie w trakcie lub po wykonaniu notę teoretyczną lub nieformalną.

Dobierając próbę badawczą stosowaliśmy **teoretyczne pobieranie próbek** (*theoretical sampling*).

> Teoretyczne pobieranie próbek sprawia, że badacz cofa się lub wybiera nową ścieżkę w razie, gdy jego kategorie są niepewne, a wyłaniające się idee niepełne. Powrót do świata empirycznego i zgromadzenie większej ilości danych na temat własności danej kategorii umożliwia nasycenie tych własności danymi i napisanie większej liczby not oraz nadanie im z czasem coraz bardziej analitycznego charakteru (Charmaz 2009:125).

Clarke uznaje tę strategię za kluczową część procedury badawczej w analizie sytuacyjnej (2005:xxi). Jak wspominaliśmy w części 4.1., to pobieranie jest procesem zbierania danych w celu generowania teorii, a zbieranie jest kontrolowane przez wyłaniającą się teorię. W projekcie badawczym w sposób cykliczny / iteracyjny i równoczesny zbiera się, koduje i analizuje dane, a w miarę wyłaniania się z nich teorii decyduje, gdzie i jakie dane zbierać dalej (Konecki 2000:30-31).

Użyta strategia należy do nieprobabilistycznych metod doboru próby (Babbie 2006:204–8). Zgodnie ze strategią teoretycznego dobierania próbek próba dobierana była celowo, w pewnym stopniu metodą kuli śnieżnej (gdyż niemal wszystkich informatorów prosiliśmy o sugestię w rodzaju „z kim należy porozmawiać, gdzie się wybrać, co zrobić, żeby lepiej poznać polskie środowisko data science?" i kilka aktów badawczych wykonaliśmy z takiego polecenia). Opisany przez Babbiego (2006:207-8) problem ogólnie marginalnego statusu osób, chcących wziąć udział w badaniu jakościowym nie wystąpił w naszych badaniach. Informatorkami / informatorami były zarówno osoby z pozycji marginalizowanych czy nienagłaśnianych jak i osoby z pozycji władzy, wysokiego statusu lub dużej widoczności w badanym świecie DS. Sądzimy jednak, że do specyfiki naszych Rozmówczyń / Rozmówców (czyli osób, które udzieliły wywiadu swobodnego) należy biografia związana z karierą naukową. Wiele z tych osób ma stopień doktora, pracowało lub pracuje na uczelniach wyższych czy współpracuje z nimi, wiele jest w trakcie doktoratu lub porzuciło studia doktoranckie / nie obroniło pracy doktorskiej. Twierdzimy, że osadzone w badanym świecie osoby z „naukowym" epizodem w biografii były bardziej zainteresowane badaniami realizowanymi przez RŻ w ramach socjologicznej pracy doktorskiej o DS, niż osoby w tym świecie nie mające owych „naukowych" doświadczeń[56]. Nie jest to błędem w stosowanym przez nas podejściu, gdyż w żadnej mierze nie dążymy do uzyskania próby reprezentatywnej dla jakiejś populacji:

> Celem teoretycznego dobierania próbek jest uzyskanie danych pomocnych w wyjaśnianiu kategorii. (…) dotyczy [ono] tylko rozwoju konceptualnego i teoretycznego; **nie** chodzi tu o przedstawienie populacji lub zwiększenie możliwości statystycznego uogólnienia wyników badań (Charmaz 2009:131).

W różnych odmianach teorii ugruntowanej przyjmuje się, że momentem zakończenia etapu gromadzenia danych jest ten, w którym **kategorie są nasycone**. W potocznym języku badaczy jakościowych stosujących teorie ugruntowane mówi się o sytuacji, w której badacze z każdym nowym aktem badania odnajdują wciąż te same schematy, czyli gromadzenie nowych danych nie prowadzi ich już do nowych spostrzeżeń (Charmaz 2009:147). W artykułach przeglądowych wskazywano, że nasycenie kodu (*code saturiation*) można osiągnąć już po około dziewięciu wywiadach – badacze mają wtedy wrażenie, że słyszeli już wszystko w temacie danej kategorii – zaś nasycenie znaczenia (*meaning saturation*) osiąga się około 16 – 24 wywiadów – wtedy badacze mają wrażenie, że rozumieją daną kategorię

56 Kontakt z informatorami – naukowcami pozwolił nam na względnie wysokie „nasycenie" kategorii związanych z jedną ze ścieżek stawania się data scientistem, w aktach samowiedzy świata DS nazwanej *from science to data science* (Migdał 2016), co opisujemy w rozdziale 6.

(Aldiabat i Le Navenec 2018:248). Według Kvale często wystarcza od 10 do 15 wywiadów, a ich ilość jest uzależniona m.in. od posiadanych zasobów czasu i środków, a przede wszystkim od tego, czy udało się zrealizować założenia badawcze (Kvale 2010:87–89). Inni autorzy wskazywali, że wielkość próbki nie jest istotna w tego rodzaju badaniach, bowiem moc informacyjna (*information power*) różnych aktów badawczych, informatorów, odwiedzonych miejsc jest bardzo różna. Ponadto każdy wywiad może mieć różną długość w sensie czasu trwania, a także można przecież rozmawiać kilkukrotnie z tą samą osobą, zatem liczenie uczestników nie ma większego sensu (ibidem). Charmaz, wskazując na prace Janice Morse i Iana Dey'a, w ogóle kwestionuje pojęcie nasycenia i możliwość osiągania owego nasycenia. Naszym zdaniem, słuszna jest teza, że nasycenie jest czymś, co badacze deklarują lub zakładają, ale nie udowadniają i jak sądzimy, nie możliwe jest jednoznaczne udowodnienie nasycenia kategorii. Samo pojęcie nasycenia może być po prostu artefaktem stworzonym z połączenia tego, na czym badacze się koncentrują i ich sposobów zbierania danych (Charmaz 2009:148–49). Ponieważ niniejsza praca jest pisana ze stanowiska konstruktywistycznego, naturalnie przyjmujemy propozycję, by traktować dane jakościowe raczej jako sugerujące kategorie, a nie nasycające je w sensie klasycznej MTU (ibidem). Zgadzamy się z Charmaz co do tego, że „wytyczne zawarte w teorii ugruntowanej należy wykorzystać w celu znalezienia klucza do materiału, a nie traktować jako maszynę, która wyręczyłaby badacza w pracy" (2009:149). Clarke mówiła o nasyceniu map sytuacyjnych, jednak traktuje to w mało restrykcyjny sposób. Wskazywała, że nasycenie występuje, gdy badacz jest przekonany, iż mapa nie wymaga większych zmian, że niczego nie przeoczył i zawarł wszystkie najważniejsze elementy. Ostatecznym testem nasycenia ma być to, czy badacz utraciwszy wszystkie materiały i notatki z badań potrafiłby na podstawie tejże mapy sytuacyjnej opowiedzieć wszystkie najważniejsze historie, dotyczące owej sytuacji (Clarke 2003:570–71).

Terenowe badania jakościowe były przez nas realizowane od 21. października 2016 do 14. czerwca 2019 roku (966 dni; ponad dwa i pół roku).

4.2.2. Analizy ilościowe

Ponieważ „dobre badanie etnograficzne łączy dane z obserwacji, wywiadów oraz analiz danych zastanych" (Angrosino 2010:102), w czerwcu i lipcu 2019 r. wykonana została wtórna analiza danych ilościowych:

- pochodzących z wybranych sondaży „środowiska" DS. Autorzy ankiet udostępnili publicznie zebrane dane;

- z serwisu meetup.com, grup zlokalizowanych w Polsce i deklarujących zajmowanie się data science. Dane zostały pobrane legalnie z użyciem mechanizmu zapewnionego po stronie meetup.com (API).

Jeżeli chodzi o **dane zebrane w środowiskowych sondażach**, skorzystaliśmy z sześciu zbiorów zebranych przez cztery podmioty, wybrane przez nas jako znaczące w badanym świecie społecznym DS na podstawie własnych badań etnograficznych. Charakterystykę podmiotów realizujących te ankiety, zbiorów danych i nasze decyzje dotyczące wyboru i zawężenia tych zbiorów do dalszych analiz prezentujemy poniżej. Zasadniczo staraliśmy się korzystać z możliwie najszerszego horyzontu czasowego i zawężać zbiory danych do respondentów mieszkających w Polsce.

Najstarszym źródłem danych ilościowych jest formularz zgłoszeniowy – wielokrotnie przez nas wymienianej w różnych punktach niniejszej pracy – pierwszej w Polsce konferencji użytkowników języka programowania R o nazwie **WhyR? z 2017 r**. Organizatorzy udostępnili zbiór danych w serwisie GitHub[57], uczestnicy konferencji zostali zaproszeni do konkursu wizualizacji tych danych. Do naszych celów zawęziliśmy zbiór (n = 246) do respondentów deklarujących kraj zamieszkania – Polskę, zatem analizowana próba n = 202. Zbiór zawiera 18 zmiennych. Zbiór ten, tak jak wszystkie z analizowanych zbiorów z ankiet środowiskowych, udostępniono w pliku csv[58]. Był to jedyny polskojęzyczny zbiór z analizowanych przez nas zbiorów danych. W roku 2018 organizatorzy WhyR? nie udostępnili danych zgłoszeniowych, zaś analizę prowadziliśmy kilka miesięcy przed konferencją w roku 2019 (odbywała się 26. – 29. września).

Analizowaliśmy także dane z formularza zgłoszeniowego drugiej w Polsce konferencji **PyData 2018**. Organizacja PyData skupia się wokół zastosowań języka Python (ale także m.in. R i Julia) w data science i także jest w naszej pracy wielokrotnie wspominania. Zbiór danych również udostępniono na GitHubie[59]. Zawiera on 11 zmiennych i był najmniej dla nas informatywnym spośród analizowanych zbiorów. Po zawężeniu zbioru (n = 435) jw.

57 https://raw.githubusercontent.com/WhyR2017/konkursy/master/dane_z_formularza_rejestracyjnego.csv, dostęp 19.07.2019 r.

58 Rozszerzenie .csv oznacza *Comma Separated Value* (dosł. wartości rozdzielane przecinkami) i jest typowym, międzyplatformowym formatem przechowywania danych w tzw. plikach płaskich, czyli dwuwymiarowych – zorganizowanych w wiersze (rekordy / jednostki analizy) i kolumny (pola / zmienne) por. http://www.creativyst.com/Doc/Articles/CSV/CSV01.htm, dostęp 20.08.2019.

59 https://raw.githubusercontent.com/stared/random_data_explorations/master/201811_pydatawaw2018/pdwc2018_anonym.csv, dostęp 19.07.2019 r.

analizowana próba to n = 284. Nie skorzystaliśmy z danych[60] z pierwszej polskiej edycji wydarzenia z uwagi na brak interesujących dla nas zmiennych demograficznych. Edycja warszawskiej PyData 2019 odbywała się dopiero w grudniu tegoż roku, zatem dane nie były dostępne.

Innym źródłem danych były ankiety **Kaggle ML & DS Survey z 2017 i 2018 r**. Kaggle to będąca częścią Google LLC platforma internetowa, w której ludzie z wolnej stopy mogą brać udział w konkursach na najlepsze modele predykcyjne i opisywanie zestawów danych, przesyłanych przez firmy i użytkowników oraz wygrywać nagrody (Kaggle 2018). Jak później pokażemy, platforma Kaggle jest popularna w badanym świecie społecznym i służy np. kreowaniu własnego wizerunku. Jednak jeżeli czyjaś aktywność dotycząca modelowania danych sprowadza się wyłącznie do tzw. „grania w Kaggle" (kod *in vivo*), to taka osoba raczej nie będzie uznawana za uczestnika świata DS, zaś z pewnością nie za data scientista. Analizowane zbiory danych ML & DS Survey także były przedmiotem owych konkursów, można je pobrać po zalogowaniu się na konto użytkownika platformy[61]. Zbiór[62] z 2017 r. to odpowiedzi z pierwszego w historii sondażu użytkowników Kaggle. Wzięło w nim udział 16 716 osób z całego świata, my analizowaliśmy próbę n = 130 deklarujących zamieszkanie w Polsce. Liczba zmiennych to 288. Zbiór z roku 2018[63] zawiera odpowiedzi 23 860 ochotników ujęte w 395 zmiennych, analizowana przez nas „polska" próba to n = 252.

Źródła te były dla nas znaczące także w sensie jakościowym, a proponowane ujęcie działania podstawowego (por. część 5.2. tej pracy) wzorowane jest na definicji z pierwszego sondażu Kaggle (2017) „The State of Data Science and Machine Learning". W czasie wykonywania naszych analiz dane z sondażu 2018 były najnowszymi z dostępnych.

Do najnowszych źródeł danych należały badania **Stack Overflow Annual Developer Survey 2018 i 2019**. Stack Overflow (SO) jest to forum internetowe przeznaczone raczej do pomocy technicznej, dotyczącej różnych języków programowania do wszelkich zastosowań, nie tylko DS. SO jest częścią portalu Stack Exchange, w skład którego wchodzi m.in. forum datascience.stackexchange.com, przeznaczone raczej do pomocy merytorycznej w dziedzinie DS. Posługiwanie się SO należy do podstawowych umiejętności uczestników świata DS i

60 https://github.com/stared/random_data_explorations/blob/master/201711_pydatawaw/pdwc2017_anonym.csv, dostęp 19.07.2019 r.

61 Piszący te słowa (RŻ) utworzył własne konto na Kaggle 6. lutego 2017 w ramach badań etnograficznych, nie uczestniczył w konkursach.

62 https://www.kaggle.com/kaggle/kaggle-survey-2017, dostęp 19.07.2019 r.

63 https://www.kaggle.com/kaggle/kaggle-survey-2018, dostęp 19.07.2019 r.

generalnie osób piszących kod. W ankiecie z 2018 r. wzięło udział 98 855 ochotników zrekrutowanych spośród użytkowników forum SO, odpowiedzi ujęto w 129 zmiennych. Respondenci z Polski, deklarujący jednocześnie przynależność do kategorii programistów (*DevType*) o nazwie „Data scientist or machine learning specialist" stanowili analizowaną przez nas grupę n = 121. Dane organizatorzy upublicznili przez Dysk Google[64]. W badaniu z roku 2019[65] uczestniczyło 88 883 ochotników, odpowiedzi zapisano za pomocą 85 zmiennych. Analizowana przez nas próba n = 111 została odfiltrowana w ten sam, co w starszym zbiorze z SO sposób.

Przeanalizowaliśmy także **dane pozyskane z serwisu meetup.com**, gdzie tworzone są grupy osób o podobnych zainteresowaniach, spotykające się na żywo. Meetup to bardzo popularna forma organizacji spotkań osób zainteresowanych data science, nie tylko w Polsce. W dniu 4. czerwca 2019 za pomocą API serwisu pobraliśmy dane o wszystkich 86 ówczesnych grupach meetupowych zlokalizowanych w Polsce. Informacje o tych grupach dostępne są publicznie w internecie na stronach grup, jednak możliwość pobrania ich masowo poprzez API dostępna jest jedynie dla użytkowników posiadających konto[66] w serwisie i wymagała wygenerowania unikalnego tokenu autoryzującego operację[67]. Dane pobraliśmy w plikach JSON[68], a następnie wykonaliśmy parsowanie plików JSON na tabele płaską, zawierającą 86 jednostek analizy (czyli grup meetupowych) opisanych za pomocą 31 zmiennych.

Pobranie danych przez API wykonaliśmy w języku Python, zaś wszystkie inne czynności związane z pracą na danych z meetup.com oraz z wyżej opisanych sondaży środowiskowych wykonaliśmy w języku R. Kod umożliwiający powtórzenie całości naszych analiz, wraz ze wszystkimi plikami wejściowymi i wyjściowymi upubliczniliśmy na koncie GitHub Remigiusza Żulickiego[69], w duchu powtarzalnych badań ilościowych (*reproducible*

---

64 https://drive.google.com/uc?export=download&id=1_9On2-nsBQIw3JiY43sWbrF8EjrqrR4U, dostęp 19.07.2019 r.

65 https://drive.google.com/file/d/1QOmVDpd8hcVYqqUXDXf68UMDWQZP0wQV/, dostęp 19.07.2019 r.

66 RŻ założył swoje konto na meetup.com 26. września 2016 r. na potrzeby prowadzonych badań etnograficznych, wielokrotnie uczestniczył w spotkaniach grup w różnych miastach, w momencie pobierania danych przez API był zapisany jednocześnie do ośmiu grup zajmujących się DS.

67 https://secure.meetup.com/meetup_api, dostęp 19.07.2019 r.

68 *JavaScript Object Notation*, ustrukturyzowany format wymiany danych, może być odczytywany np. przez przeglądarki internetowe czy inne aplikacje.

69 Analiza meetupów: https://github.com/zremek/meetup-harvesting,
sondaży: https://github.com/zremek/survey-polish-data-science.
Badacz (RŻ) założył konto na GitHubie 3. czerwca 2018 r. na potrzeby badań etnograficznych, później używał go także do innych projektów analitycznych (por. Żulicki i Żytomirski 2020).

science) (Leek 2016; Peng 2011a). Podobne (i znacznie bardziej zaawansowane) projekty dotyczące przetwarzania, analizy i modelowania danych zastanych za pomocą języków R i Python są zbliżone do tego, co uznajemy za działanie podstawowe badanego świata DS. Takie projekty to typowa dla uczestników badanego świata interakcja z aktorami nie-ludzkimi (np. dane, kod, środowisko programistyczne i inne narzędzia cyfrowe) i były dla RŻ ważnym elementem enkulturacji w świecie społecznym DS i bez wątpienia pomogły w zrozumieniu specyfiki działania podstawowego badanego świata.

Użyliśmy **technik analitycznych o charakterze opisowym i eksploracyjnym** (Leek i Peng 2015). Opis (*descriptive data analysis*) oznacza tutaj podsumowania wykonywane na konkretnych zbiorach danych, zaś eksploracja (*exploratory data analysis*) to bazujące na opisie poszukiwanie zależności: trendów, korelacji czy związków pomiędzy zmiennymi w celu wygenerowania pomysłów i hipotez (ibidem). Nasze analizy nie mają charakteru wnioskowania statystycznego (*inferential statistics / inferential data analysis*), zatem nie dążymy do uogólnienia zależności, zaobserwowanych w konkretnych zbiorach danych, na populacje (ibidem). W naszej ocenie takie postępowanie byłoby nieuprawnione, bowiem analizujemy dane zastane z ankiet zrealizowanych na próbach niereprezentatywnych, dobranych w sposób nieprobabilistyczny z niedookreślonych populacji (Babbie 2006:498–503). Tym samym nie proponujemy testów istotności statystycznej[70]. Użyte przez nas techniki analityczne to głównie podsumowania częstości względnych i bezwzględnych w podziale na grupy (tradycyjnie prezentowane w tabelach krzyżowych, u nas przeważnie w postaci wykresów słupkowych), statystyki opisowe, np. miary tendencji centralnej, miary symetrii rozkładu, współczynniki korelacji, także w podziale na grupy (co przedstawiamy m.in. w postaci wykresów pudełkowych lub punktowych).

4.3. Realizacja badań

Łącznie autor niniejszej pracy wykonał 26 wywiadów swobodnych ukierunkowanych (w sumie 30 godzin i 44 minuty nagrań), 46 aktów obserwacji uczestniczącej (zarówno online jak i offline, łącznie około 295 godzin obserwacji) oraz sporządził 12 not autoetnograficznych. Badań zrealizowanych z użyciem netnografii i etnografii opartej na współpracy ani analiz ilościowych danych zastanych nie ujmujemy w ten sumaryczny sposób

70 Używanie tego rodzaju testów bazujących na p-value jako kluczowego kryterium weryfikacji hipotez jest w ostatnich latach krytykowane także przy próbach reprezentatywnych i badaniach eksperymentalnych (por. Amrhein, Greenland, i McShane 2019).

z uwagi na ich inny charakter, choć terminy realizacji analiz ilościowych zostały utrwalone (w repozytorium GitHub). Spis zrealizowanych badań znajduje się w Aneksie.

Pierwszym aktem badawczym był wywiad, zrealizowany przez autora (RŻ) 21. października 2016 r. z osobą poznaną na gruncie prywatnym, a zajmującą się zawodowo analizą danych. Trwał on 28 minut, badacz zrozumiał swój brak przygotowania i po tym doświadczeniu postanowił skupić się najpierw na obserwacjach uczestniczących oraz netnografii. Miało to na celu przygotowanie listy potrzeb informacyjnych do kolejnych wywiadów, szeroko pojętą enkulturację badacza, poznanie języka charakterystycznego dla badanego wycinka rzeczywistości społecznej (na tym etapie nie posługiwaliśmy się jeszcze ramą teoretyczną społecznych światów), a także poznanie osób, z którymi potencjalnie będzie można przeprowadzić wywiady oraz generalnie dopracowanie koncepcji pracy doktorskiej[71]. Wywiad drugi zrealizowaliśmy w dwadzieścia miesięcy po pierwszym, 21. czerwca 2018 r. Trwał on 47 minut i był zdecydowanie bardziej informatywny, zaś przygotowana lista potrzeb informacyjnych pozytywnie przeszła test w terenie i posłużyła jako baza do dalszych rozmów. Od tego czasu wywiady był realizowane z naprzemienną intensywnością – kilka wywiadów w miesiącu / dwóch i podobnej długości przerwa na ich kodowanie oraz analizę – aż do 30. maja 2019, kiedy to wykonano ostatni, dwudziesty szósty wywiad (167 minut[72], najdłuższy).

Lista potrzeb informacyjnych do wywiadów zmieniała się w trakcie realizacji badań z uwagi na zgłębianie różnych elementów lub procesów badanego świata społecznego. Jednak niemal we wszystkich rozmowach badacz (RŻ) nawiązywał interakcję za pomocą następującej sekwencji pytań „rozgrzewkowych", nierzadko uzupełnianych przez badacza o bieżące prośby o wyjaśnianie słów Rozmówców bądź zweryfikowanie podawanych przez badacza interpretacji:

- Czy Ty jesteś data scientist?
- Co to dokładnie znaczy „data scientist"?
- Czym dokładnie Ty się w tym wszystkim [o czym Rozmówcy opowiadali powyżej] zajmujesz?
- Czy możesz opowiedzieć o jakimś ostatnim projekcie od początku do końca? Tak, żebym zrozumiał od czego się zaczęło, co było później i na czym się skończyło?

71 Później także w sensie formalnym, na potrzeby otwarcia przewodu doktorskiego w dniu 9.04.2018 r.

72 Średnia długość wywiadu to prawie 71 minut, mediana równa jest 56 min. Długość wywiadów przeważnie rosła wraz z ich kolejnością: dla wywiadów 1 – 9 średnia długość to 50,5 minuty, dla 10 – 18 to 57 minut, dla wywiadów 19 – 26 wynosi 106 minut.

Po narracji zazwyczaj nawiązywała się rozmowa o różnych aspektach projektów, o których opowiadano badaczowi i czasami w ten sposób udawało się zaspokoić niemal wszystkie potrzeby informacyjne. Jeżeli Rozmówcy nie rozwijali narracji, badacz pytał o ich wykształcenie i karierę zawodową (bardzo skuteczne okazało się proste pytanie „skąd się wzięłaś / wziąłeś w data science?") lub o narzędzia pracy. W zależności od etapu realizacji badań i od wypowiedzi Rozmówców były poruszane rozmaite tematy. Cenne poznawczo – głównie z uwagi na temat wartości świata społecznego – okazały się pytania stosowane na etapie wychodzenia z wywiadu, jak: „czy lubisz swoją pracę / co Ci się w niej podoba?", „jakie masz plany na przyszłość?", „jaka będzie przyszłość data science?". Wywiad kończyła przeważnie wspomniana już prośba o sugestię w rodzaju „z kim należy porozmawiać, gdzie się wybrać, co zrobić, żeby lepiej poznać polskie środowisko data science?". Nierzadko Rozmówcy polecali kontakt z osobami, z którymi badacz zrealizował już wywiad, co było dla nas dowodem trafnego zidentyfikowania osób silnie osadzonych i rozpoznawalnych w badanym świecie[73]. Wątek polecania kolejnych informatorów czasami rozwijał się w kierunku dopytywania badacza przez Rozmówców o metodologię doboru próby oraz uogólniania wyników badań jakościowych i udzielania badaczowi porad do realizacji badań. Ogólnie pytania badacz (RŻ) zadawał zgodnie ze wskazówkami Charmaz (2009:44-51).

Instrukcja przed wywiadem była inspirowana wpisem na blogu Ethnography Matters (Annechino 2016):

> Mogę zadawać pytania, które wydadzą Ci się głupie lub oczywiste. Jest to normalną częścią wywiadu, dlatego, że to Ty jesteś ekspertem, Ty najlepiej znasz swoje doświadczenia. Chcę spróbować zrozumieć, jak Ty widzisz te wszystkie sprawy, Twoimi własnymi słowami.

Wszyscy Rozmówcy / Rozmówczynie wyraziły świadomą zgodę na nagrywanie wywiadów. Piętnaście wywiadów zrealizowano twarzą w twarz, pozostałe w sposób zapośredniczony przez komunikatory głosowe lub wideo (Google Hangouts: 1 wywiad, telefonicznie: 4, Skype: 6). Po realizacji zaplanowanych potrzeb informacyjnych i zakończeniu ostatniej odpowiedzi lub narracji Rozmówców badacz (RŻ) mówił zawsze: „to wszystko, co na dzisiaj przygotowałem, czy chciałabyś / chciałbyś coś dodać, do czegoś wrócić, czy o czymś jeszcze powinniśmy porozmawiać?". Po odpowiedzi przeczącej albo po dokończeniu rozmowy badacz uprzedzał o wyłączeniu dyktafonu i kończył nagranie, przechodząc do podziękowań i prośby o możliwość późniejszych konsultacji. Rozmówcy

73 Jednak strategia dobierania Rozmówców najbardziej rozpoznawalnych i posiadających wysoki status w społecznym świecie była w późniejszych etapach badań niekoniecznie zgodna ze strategią teoretycznego doboru próbki, stąd badacz nie realizował jej przez cały czas.

nierzadko wyrażali zainteresowanie wynikami naszych badań, wszyscy zgodzili się na późniejsze konsultacje. Zdarzało się, że jeszcze w tym momencie badacz prosił o możliwość zadania dodatkowego pytania do wywiadu, ponieważ wpadł na jakiś pomysł lub spojrzał w swoje notatki – ku jego zdziwieniu za każdym razem akceptowano te sytuacje. Badacz początkowo w małym stopniu korzystał z notatek z listą potrzeb informacyjnych (wyłącznie w formie papierowej), starając się podczas wywiadów poświęcić całą uwagę na aktywne słuchanie Rozmówców. Niemniej w toku badań zauważył, że korzystanie z notatek pomaga, a nie przeszkadza w aktywnym słuchaniu, bowiem zwalnia umysł od obowiązku przypominania sobie listy potrzeb informacyjnych w celu zadawania kolejnych pytań. Wówczas wystarczy aktywnie słuchać i dopytywać / proponować interpretacje słów Rozmówców, a po wyczerpaniu wątku przełączyć uwagę na notatki i rozpocząć kolejny wątek. Po wywiadach badacz zawsze starał się kontynuować rozmowę w swobodnej formie (o ile Rozmówcy mieli na to czas; jeżeli nie, badacz przesyłał ponowne podziękowania), a także mówić możliwie wiele o sobie, by zredukować asymetrię sytuacji wywiadu. Zdarzyło się raz na tym etapie wrócić do wątków z wywiadu i nagrać kolejnych siedem minut rozmowy.

Wszystkie wywiady były nagrywane na dyktafon cyfrowy w smartfonie (z włączonym trybem samolotowym). Jeden wykonano w języku angielskim[74], resztę po polsku[75]. Transkrypcje wywiadów początkowo realizował autor rozprawy, później zostały one zlecone i zrealizowane dzięki finansowaniu ze środków Katedry Socjologii Kultury, Instytut Socjologii, Wydział Ekonomiczno-Socjologiczny Uniwersytetu Łódzkiego. W trzech przypadkach poproszono badacza o autoryzację transkrypcji. Pierwsza osoba, która poprosiła o to przed rozpoczęciem nagrywania, zrezygnowała z prośby po zakończeniu wywiadu wyjaśniając, że z pewnością nie ujawniła żadnych tajemnic firmy swojej ani firm klientów. Dwie kolejne osoby prosiły o autoryzacje po zakończeniu wywiadów. Druga osoba także wskazywała na obawę przed ujawnieniem tajemnicy firmowej, ostatecznie zadecydowała o pozostawieniu całości transkrypcji. Trzecia osoba wskazała na wątek osobisty[76] i informacje umożliwiające zidentyfikowanie jej, zdecydowała o wykreśleniu około 270 słów, czyli 3,5% transkrypcji

74 Wywiad z osobą pochodzącą z Europy Zachodniej, pracującą od wielu lat w Polsce; angielski jest dla niej drugim językiem.

75 Jeden z Rozmówców sugerował nam przeprowadzanie wszystkich wywiadów po angielsku, wskazując na możliwość wykonywania dość dobrej jakości automatycznych transkrypcji z tego języka używając oprogramowania bazującego na modelach uczenia maszynowego. Z sugestii nie skorzystaliśmy, oceniając możliwość posługiwania się językiem ojczystym dla wszystkich parterów rozmowy, a zatem rozumienie subtelności w odcieniach znaczeniowych rozmaitych wyrażeń, jako dalece ważniejszą dla celów naszego badania.

76 Czujemy się w obowiązku odnotować, że w tym usuniętym fragmencie mowa była o doświadczaniu dyskryminacji ze względu na płeć, studiując na państwowej uczelni wyższej w Polsce.

(całość transkrypcji to 7 869 słów). W jednym przypadku RŻ poprosił o zgodę na wykorzystanie w badaniach fragmentu nagranej wypowiedzi zaczynającej się od słów „oficjalnie do wywiadu tego nie powiem", osoba (wcześniej nie zgłaszająca uwag) po zapoznaniu się z transkrypcją takiej zgody udzieliła. Transkrypcje sporządzono dosłowne (*verbatim*), usunięto z nich wszystkie imiona i nazwiska, nazwy firm (Rozmówców i ich klientów, pozostawiono w innych przypadkach np. rozmowy o narzędziach / wydarzeniach wspieranych przez konkretne podmioty), a także nazwy miejscowości i nazwy własne, mogące posłużyć do identyfikacji osób (np. nazwy meetupów, gdy wywiad prowadzono z osobą organizującą go). Pozostawiono nazwy własne technologii, gdyż bez tego nie sposób byłoby mówić o narzędziach pracy, a także nazwy firm, kiedy mowa o nich w sensie nie dotyczącym Rozmówców (np. Facebook wspiera rozwój pakietu PyTorch).

Wszystkim Rozmówczyniom / Rozmówcom zagwarantowaliśmy poufność, rozumianą jako sytuację badawczą, w której „badacz może zidentyfikować autora wypowiedzi, ale jednoznacznie obiecuje tej informacji nie ujawniać" (Babbie 2006:518). Wprost deklarowaliśmy przed każdym wywiadem, że rozmowy służą wyłącznie celom naukowym (przygotowaniu socjologicznej pracy doktorskiej i ewentualnych artykułów w czasopismach naukowych), a autor (RŻ) nigdzie nie opublikuje nazwiska, imienia, nazwy miejsca pracy ani innych nazw pozwalających na identyfikację Rozmówców, oraz, że nie będzie publikowana cała transkrypcja wywiadu ani żaden fragment nagranego pliku audio. We własnych zasobach cyfrowych (plikach audio z wywiadami, transkrypcjach, różnego rodzaju notatkach) nie zapisywaliśmy danych identyfikujących osoby. Każdy wywiad i notatki do niego nazywaliśmy wymyślonym identyfikatorem, kojarzącym się nam z daną rozmową, np. dla Rozmówcy, który na początku wywiadu powiedział, iż formalnie jego stanowisko pracy nie nazywa się data scientist badacz używał nazwy „formalnie_nie_ds". Zwracaliśmy szczególną uwagę na poufność wywiadów przy realizowaniu transkrypcji przez osoby trzecie. Osoby wykonujące transkrypcje w swoich umowach cywilnoprawnych miały zapis o poufności badań i zakazie prezentowania, publikowania, udostępniania i kopiowania otrzymanych plików audio i sporządzanych transkrypcji.

Przy większości wywiadów badacz sporządzał dwie notatki: jedną poświęconą sytuacji wywiadu, drugą obserwacjom w trakcie wywiadu. Pierwsza miała na celu rozwijanie rekrutacji na wywiady, interakcji z Rozmówcami i techniki prowadzenia wywiadu. Druga służyła odnotowaniu aspektów niewerbalnych rozmowy, zapisaniu pomysłów w formie not teoretycznych, rozwijaniu dyspozycji do wywiadu. Nierzadko te dwa cele badacz mieszał w

jednej notatce, sporządzanej w pośpiechu w formie nagrania na dyktafon w swoim smartfonie. Była to wygodna forma uchwycenia myśli „na gorąco", np. wsiadając do autobusu w obcym mieście, a jednocześnie komfortowa dla badacz w tym sensie, że w jego odczuciu dla postronnych takie notowanie wyglądało jak prowadzenie rozmowy telefonicznej. Notatki głosowe badacz często stosował także podczas obserwacji uczestniczących w terenie offline. Ze wszystkich sporządził transkrypcje.

W sensie metodologicznym dobieraliśmy próbkę zgodnie ze strategią doboru teoretycznego teorii ugruntowanej (por. 4.2.1.), w sensie organizacyjnym rekrutowaliśmy do wywiadów osoby zarówno poznane osobiście podczas obserwacji terenowych, osoby nieznajome, które w trakcie badań rozpoznaliśmy jako osadzone w badanym świecie lub ważne z uwagi na teoretyczny dobór próbki, oraz osoby z polecenia innych Rozmówczyń / Rozmówców. Część wywiadów była zatem umawiana twarzą w twarz, inne w sposób zapośredniczony. Portal LinkedIn (o którym więcej w części 5.3.4.) służył nam często do umawiania się na rozmowy z osobami poznanymi osobiście lub poleconymi, a czasami do nawiązywania kontaktu po raz pierwszy. Rzadziej była to korespondencja emailowa. RŻ nie korzystał z Facebooka od 2017 r. i nie używał tego medium do rekrutacji informatorów, choć już pod koniec badań wskazywano w wywiadach, że jest ono przydatne[77]. Badacz nie zdecydował się jednak na ponowne założenie konta w tym celu[78], uznając, że stosowane kanały i strategia rekrutacji jest odpowiednia z uwagi na przyjętą metodologię badań. Niemniej bez wątpienia wśród naszych Rozmówczyń / Rozmówców zdecydowanie dominują osoby biorące udział w wydarzeniach świata społecznego (jak meetupy i konferencje), jak zatem sądzimy – i jak wskazywano nam w trakcie badań – są to osoby bardziej „towarzyskie", o wyższej sprawności werbalnej i interpersonalnej, niż uczestnicy badanego świata nie biorący udziału w takich spotkaniach[79]. W jednym przypadku wywiad z osobą poleconą odbył się prawie natychmiast po wywiadzie z osobą polecającą – badacz był w miejscu pracy Rozmówców i po pierwszej rozmowie został przedstawiony innemu członkowi zespołu, który z rekomendacji kolegi od razu wyraził chęć udzielenia wywiadu.

---

77 „(...) jeżeli byś chciał dotrzeć [do osób nie biorących udziału w konferencjach i meetupach], to nie obejdziesz się bez social mediów. (...) No bo powiem ci że ja na Facebooku no do teraz nie mam zdjęcia (...) gdzieś tam też było dla mnie ważne, powiedzmy trzymanie się swojej prywatności, no ale potem się okazało że coś za coś.(...) bez tego yyy po prostu nie, nic się nie zdziała jeśli chcesz mieć kontakt z ludźmi pracującymi z nowymi technologiami, nie?" {wywiad 21}.

78 Później jednak założył – i skasował – ponownie konto w celu pobrania danych z publicznej grupy facebookowej Data Science PL (por. 4.3.3.), ostatecznie nieudanego.

79 Później przedstawimy kategorię nerda, postrzeganego m.in. jako osoba mało „towarzyska", zainteresowana raczej abstrakcjami lub technologiami niż ludźmi, a będącą punktem odniesienia dla świata DS.

W kilkunastu przypadkach wywiad nie doszedł do skutku. Część osób, z którymi badacz nawiązał kontakt po raz pierwszy np. za pośrednictwem LinkedIn nie odpowiedziała na pierwszą wiadomość. Inni, także osoby poznane osobiście i z polecenia, przestawali odpisywać po kilku wiadomościach, zazwyczaj kiedy przychodziło do ustalenia terminu rozmowy. Jednoznaczne odmowy udzielenia wywiadu zdarzały się rzadko. W jednym przypadku taka odmowa okazała się testowaniem badacza (przez osobę silnie osadzoną w badanym świecie, mającą w nim wysoki status i rozpoznawalność), ostatecznie rozmowa doszła do skutku.

Wywiady planowaliśmy realizować w miejscach, które byłyby „naturalnym kontekstem życia i pracy" Rozmówczyń / Rozmówców, co miało przyczyniać się do ich deformalizacji i zbliżenia wywiadu do „naturalnej sytuacji rozmowy", oraz możliwości poczynienia dodatkowych obserwacji (Konecki 2000:179). Niestety w większości przypadków (20 z 26) nie udało się tego wykonać. Poza wyżej wspomnianymi dwoma wywiadami (w dużej firmie) w miejscu pracy zrealizowano jeszcze trzy (jeden także w dużej firmie, dwa na uczelniach), a jeden podczas konferencji. Pozostałe wywiady[80] przeprowadzono w kawiarniach (9) albo zdalnie (11). Sądzimy, że z uwagi na silne akcentowanie tajemnicy firmowej w badanym świecie nasi Rozmówcy nie chcieli lub nie mogli zapraszać badacza do miejsca pracy. Nawet kiedy tak było, prowadzący wywiady (RŻ) był zapraszany do pomieszczeń konferencyjnych lub pokoi do rozmów rekrutacyjnych, nie miał szansy zobaczyć zespołu data science przy stanowiskach pracy. Nie dotyczyło to uczelni wyższych, gdzie badacz rozmawiał z pracownikami naukowymi.

Z powodu tajemnicy nie udało się także zrealizować obserwacji w miejscu pracy zespołu data science. Po wstępnej zgodzie poznanego w trakcie badań informatora – szefa działu DS – informator ten powiadomił nas, że kierownictwo firmy nie zgodziło się na wizytę, ponieważ „mają teraz super tajny projekt i nie mogą przyjmować nikogo z zewnątrz". Badacz nie ponawiał próśb w przekonaniu, że nie wypada mu ponownie wychodzić z inicjatywą, szczególnie, że ów informator w rozmowach nieformalnych kilkukrotnie pomagał nam podczas badań.

Obserwacje były podstawową formą naszych badań terenowych od października 2016 do czerwca 2018. W tym czasie autor (RŻ) zrealizował większą ich część (około 3/5 – 176,5 z

80 Podsumowując, z 26 wywiadów 11 zrealizowaliśmy zdalnie, a 15 twarzą w twarz. Z owych 15 rozmów 3 miały miejsce w dużych firmach, 2 na uczelniach (zatem łącznie 5 w miejscu pracy), 1 wywiad podczas konferencji świata DS, pozostałe 9 wywiadów w kawiarniach lub innych lokalach gastronomicznych.

295 godzin). Miejsca obserwacji starał się wybierać zgodnie ze strategią teoretycznego doboru próbki. Zdarzało się, że wybrane miejsca / wydarzenia okazywały się być inne, niż badacz się spodziewał, i że nie można mówić w ich przypadku o obserwowaniu świata DS. Niemniej takie obserwacje okazały się również informatywne, pozwalając badaczowi lepiej zrozumieć:

- miejsca przecięć badanego świata z innymi światami oraz praktyki komunikacyjne (np. legitymizujące / autoprezentacyjne / marketingowe) osób reprezentujących badanych świat skierowane do osób z innych światów (np. do biznesu);
- praktyki „podszywania się" innych grup związanych z analizą danych pod świat DS;
- dyskurs o data science / big data / uczeniu maszynowym / tzw. sztucznej inteligencji w innych światach czy grupach bez jawnego udziału uczestników świata społecznego DS.

Ponad połowę czasu obserwacji uczestniczących badacz spędził online (około 149,5 godziny z 295), choć ilość aktów obserwacji polegających na uczestnictwie online była mniejsza niż tych offline (łącznie 46 wszystkich aktów, z czego 16 online). Czas spędzony przed komputerem, głównie na uczestnictwie w kursach MOOC i samodzielnych próbach pisania kodu w R (a w mniejszym stopniu w Pythonie) był przeżyciem bliskim doświadczeniom – szczególnie początkujących – uczestników świata lub osób chcących wejść do tego świata, zatem pomocnym w zrozumieniu wielu jego aspektów.

O ile narzędzia stosowane przy realizacji wywiadów – dyktafon cyfrowy w smartfonie i i notatnik papierowy z dyspozycjami – sprawdzały się przez cały okres realizacji badań, to niełatwo było znaleźć skuteczne narzędzia do notowania. Skuteczne oznacza dla nas: wygodne (pozwalające łatwo i szybko notować, zacząć / przerwać / wznowić / zakończyć notatkę jak najmniej zakłócając badaczowi uczestnictwo w obserwowanym miejscu), komfortowe (takie, że notowanie wygląd „normalnie" w trakcie obserwacji, nie wzbudza zainteresowania ludzi w miejscu obserwacji, badacz jak najmniej obawia się rozpoznania jako ktoś, kto prowadzi obserwację uczestniczącą), efektywne na dalszych etapach badań (generujące jak najmniej dodatkowych czynności podczas przeglądania, kodowania, sporządzania map, analizowania kodów / materiału i pisania rozprawy). Notatniki papierowe były dość komfortowe i wygodne podczas konferencji – szczególnie, że organizatorzy często rozdawali logowane notatniki i część osób z nich korzystała – ale już nie podczas meetupów („nienormalne" zachowanie badacza, a także niewygoda – meetupy odbywały się często w

ciemnych pubach), zaś bezużyteczne podczas kursów i warsztatów offline i online (podczas korzystania z laptopa). Przy kodowaniu notatki papierowe także nie sprawdziły się. Po krótkiej i zniechęcającej próbie ich przepisywania badacz zdecydował się na skanowanie do plików .jpg i kodowanie ich jak obrazów. Było to jednak nieefektywne, pracochłonne, pliki graficzne zajmują znacznie więcej przestrzeni dyskowej niż tekstowe, co utrudniało zarówno cyfrową archiwizację badań, jak i pracę w narzędziu CAQDAS – QDA Miner.

Znacznie bardziej skuteczne były dla nas notatki cyfrowe. W przypadku obserwacji polegających na pisaniu kodu badacz, jak wspominano wyżej, notował głównie w IDE za pomocą komentarzy między linijkami kodu w języku programowania, a do kodowania jakościowego eksportował całe skrypty jako pliki .txt, co pozwoliło kodować jakościowo także kawałki kodu skryptów. W przypadku innych obserwacji, zarówno online i offline, stosowano notatki w tzw. chmurach. Na różnych etapach badacz używał plików .txt synchronizowanych przez Dropbox, aplikacji Notes by Firefox oraz Google Keep. Żadne z rozwiązań nie było idealne, niemniej z uwagi na łatwość późniejszego importowania plików do QDA Miner notatki z obserwacji (*field notes*) wykonywaliśmy głównie w plikach .txt na Dropboxie (tam także zapisane były pliki programu QDA Miner i sam plik .odt z rozprawą doktorską), a inne notatki w Google Keep z uwagi na możliwość nadawania etykiet i łatwość przeglądania, wyszukiwania, dołączania adresów url i zdjęć. Wszystkie te narzędzia umożliwiały notowanie zarówno z komputera jak i smartfona, sortowanie, filtrowanie, przeglądanie i uzupełnianie notatek, ściąganie ich i importowanie do QDA Miner. Notowanie na smartfonie, jak również bieżące odwiedzanie czy ściąganie różnych zasobów internetowych, fotografowanie, sporządzanie notatek głosowych, zapraszanie właśnie poznanych osób na LinkedIn było dla badacza pomocne w pracy offline[81]. Badacz zawsze też starał się korzystać z narzędzi danego wydarzenia, np. na konferencji korzystał ze specjalnej aplikacji przygotowanej przez organizatorów na urządzenia mobilne, pozwalającej m.in. na przeglądanie i wyszukiwanie wystąpień oraz własne notatki do każdego wystąpienia.

Poświęcamy tyle uwagi narzędziom badań jakościowych, ponieważ po pierwsze, zależy nam na jawności naszych działań, po drugie, zgadzamy się z teza, że wsparcie narzędziowe pomaga zająć się problemem długotrwałości w badaniach etnograficznych. Oczywiście długi czas przebywania w terenie jest niezbędny dla zrozumienia badanego wycinka rzeczywistości społecznej: „w istocie, badania etnograficzne generują ‘samorodki’ (*nuggets*) użytecznych

81 Także w prozaicznych czynnościach, jak np. sprawdzanie rozkładów jazdy transportu publicznego i kupowanie biletów, orientacja w terenie dzięki cyfrowym mapom, wpisywanie terminów do kalendarza, smartfon był niezastąpionym narzędziem pracy terenowej etnografa.

informacji w nieprzewidywalnych odstępach czasu" (Sommerville i in. 1993:171). Chodzi nam o długotrwałość w sensie żmudnej i pracochłonnej walki z własnymi narzędziami – do notowania, kodowania i analizy danych jakościowych oraz pisania rozprawy.

Kodowanie jakościowe, z założenia długotrwałe, pracochłonne i iteracyjne było także walką z własnymi narzędziami (o czym więcej w 4.3.3.), ostatecznie badacz posłużył się oprogramowaniem QDA Miner Lite v2.0.5 firmy Provalis Research oraz w małym stopniu jego wersją płatną (QDA Miner 4.1.37). Zakodowano wszystkie transkrypcje dwudziestu sześciu wywiadów wraz z poczynionymi do nich notatkami, wszystkie notatki autoetnograficzne, około 90% materiałów z obserwacji uczestniczących oraz wybrane materiały netnograficzne. W przypadku netnografii przeważnie korzystaliśmy ze źródeł tekstowych (np. artykuły, książki, strony internetowe, blogi, wpisy w mediach społecznościowych) i przyjęliśmy konwencję cytowania tych źródeł jako literatury, umieszczając wszystkie w menadżerze bibliografii Mendeley[82] i tam nadawaliśmy etykiety, sporządzaliśmy notatki, dodawaliśmy pliki tekstowe czy zrzuty ekranu – tak samo, jak postępowaliśmy z literaturą naukową. Wyjątkowo potraktowaliśmy dwie wczesne, amerykańskie książki zawierające wywiady z osobami zajmującymi się data science (Brooks 2014; Gutierrez 2014), oraz transkrypcje wybranych odcinków pierwszego polskiego podkastu „Data Science po polsku"[83] (także zawierającego wywiady) kodując je w QDA Miner. W przypadku obserwacji pominęliśmy część zebranych materiałów z uwagi na ich małą informatywność (od pewnego etapu badań), a dużą pracochłonność. Mowa o plikach z prezentacjami/slajdami, które prelegenci – uczestnicy świata DS niemal zawsze udostępniają publicznie (oczywiście obok kodu w językach programowania), a które szczególnie na początku badań bardzo chętnie i w dużej ilości pobieraliśmy, ciesząc się z bogactwa materiału jakościowego. W QDA Miner nie istnieje jednak możliwość załadowania plików zawierających jednocześnie tekst i grafikę, jak .pdf, zatem badacz był zmuszony najpierw konwertować prezentacje do plików graficznych i tekstowych, a następnie importować do QDA Miner.

Materiały kodowaliśmy systematycznie od lipca 2018, wykonując kilka iteracji kodowania rzeczowego (otwartego i selektywnego), przeplatanego kodowaniem

82 Mendeley to należący do wydawnictwa Elsevier darmowy, dostępny na licencji zamkniętej program do zarządzania bibliografią naukową, udostępniania tekstów naukowych i współpracy zdalnej, oraz serwis internetowy dla naukowców: https://www.mendeley.com/, dostęp 19.03.2019 r.

83 Prowadzonego przez Szymona Drejewicza od lipca 2017 do marca 2018, https://www.youtube.com/channel/UC7DGutbNdAjEFbOV-Al8aRA/featured, dostęp 05.09.2019 r.

teoretycznym. Przykładowo, po pierwszym kodowaniu selektywnym w dniu 23.10.2018 sporządzono *memo* o następującej treści: w pierwszej iteracji kodowania otwartego wygenerowano 248 kodów. W wyniku pierwszej iteracji kodowania selektywnego usunięto 129 kodów, z czego 45 uznano za nieprzydatne lub zapisano jako uwagi do kodów, a 84 połączono z innymi kodami (częściowo z nowo powstałymi). Pozostawiono 119 kodów, dla każdego zapisując znaczenie (tzn. notatkę kodową – opis kodu w książce kodowej), a w kilku przypadkach zmieniając nazwę. W wyniku łączenia kodów powstało 18 nowych kodów, więc łącznie pierwsza iteracja kodowania selektywnego dała wynik 137 kodów.

Głównie w kodowaniu rzeczowym pomocna była funkcja przeszukiwania materiału według słów kluczowych, czasami z użyciem operatorów logicznych. Piszący te słowa chciałby podać datę zakończenia kodowania, jednak z uwagi na charakter tego projektu badawczego nie jest to możliwe, a co najmniej do momentu ukończenia pierwszej wersji niniejszej rozprawy kody najpewniej ulegną jeszcze zmianom.

W analizie danych jakościowych – zarówno w kodowaniu selektywnym, kodowaniu teoretycznym, sporządzaniu tzw. map czy innych schematów, jak i w pisaniu rozprawy – koncentrowaliśmy się przede wszystkim na znaczeniach tworzonych kodów / kategorii, oraz występujących między nimi podobieństwach i różnicach. Ważne były także podobieństwa i różnice pomiędzy przypadkami, tzn. wywiadami czy obserwacjami. Korzystaliśmy z szeregu możliwości programu QDA Miner: obliczania częstości występowania kodów[84] (zarówno w całości zakodowanego materiału jak i w pojedynczych wywiadach / notatkach), sprawdzania reprezentacji danego kodu tzn. sprawdzenia, w jakich przypadkach on występuje lub nie, przeglądania lub eksportowania do pliku wszystkich opatrzonych wybranymi kodami fragmentów materiału – czynność tę Konecki (2000:54-55) nazywał tworzeniem kart kodowych – oraz sprawdzania współwystępowania kodów. To współwystępowanie było obliczane za pomocą indeksu Jaccarda (por. Tomanek i Bryda 2015) i wizualizowane w dość pomocnej dla badacza postaci dendrogramu. Bardzo cenne analitycznie było dla nas także zwyczajne przeglądanie materiału jakościowego i nadanych kodów, pomagało to także zauważać sekwencje kodów nie współwystępujących (tzn. nie nachodzących na siebie).

W porównaniu z badaniami jakościowymi wykonana przez nas ilościowa analiza danych zastanych była krótka i bezproblemowa. Wrócimy do tego zagadnienia w części 4.3.3., na razie zaznaczmy jedynie, że wygoda korzystania z narzędzi open source do analiz

84 Ta funkcja była pomocna także przy kodowaniu rzeczowym, bowiem pozwalała sprawdzić postęp pracy – jaki odsetek konkretnego tekstu jest już zakodowany.

ilościowych – języków R, Python i edytorów RStudio oraz Jupyter Notebook – była dla badacza bez porównania większa niż w przypadku używanych przez niego narzędzi CAQDAS. Być może to kwestia spędzonego czasu oraz faktu, że wykonano proste analizy ilościowe, niemniej były one dla autora (RŻ) przyjemną odmianą.

4.3.1. Usytuowanie badacza

„Zacznę, à la Strauss, od odrobiny mojej intelektualnej autobiografii w sensie tego, jak stałam się tym, kim jestem teraz i jak znalazłam się w obecnym miejscu" (Clarke 2015:85).

Piszący te słowa (RŻ) przedstawia, jak postrzega obecnie własne usytuowanie w formie notatki autoetnograficznej, napisanej specjalnie w tym celu:

> Mam ochotę napisać - "jaki jest mój background", bo tak mówi się w świecie DS i chyba dość mocno nasiąknąłem tym językiem. Zatem mój background jest taki, że jako dzieciak miałem trochę zacięcia do matematyki i w szkole średniej poszedłem do klasy matematyczno-informatycznej. Nie interesowały mnie komputery, tylko zagadki. Jednak pod koniec szkoły (2005 - 2006) zacząłem się coraz mocniej interesować polityką, filozofią, literaturą, i doszedłem do wniosku, że chciałbym zajmować się sondażami wyborczymi oraz politycznymi diagnozami i będę studiował socjologię. Podczas studiów byłem muzykiem-amatorem, początkowo miałem bardzo dobre oceny, później byłem niespecjalnie dobrym studentem. Naukowo wciągnąłem się zdecydowanie w socjologię jakościową, symboliczny interakcjonizm, co wydaje mi się w Łodzi szczególnie popularne. Pracę magisterską osadzoną w MTU napisałem o łódzkiej scenie muzycznej i pożegnałem się z uczelnią. Gdzieś w 2013, pracując w Obserwatorium Rynku Pracy dla Edukacji w ŁCDNiKP dostałem zadanie opisania perspektyw zatrudnienia dla osób kończących technikum informatyczne. W sieci szybko spotkałem się z "najbardziej seksownym zawodem świata XXI wieku" oraz mnóstwem tekstów o Big Data - to określenie było wówczas na fali. Nie bardzo zadowolony z ówczesnej pracy zainteresowałem się tematem pod kątem swojej przyszłej kariery, jednak szybko zrozumiałem, że nie chodzi o robienie tabel przestawnych w Excelu, ani nawet formularzy w HTMLu (co wówczas uważałem za technologię kosmiczną). W 2014 zacząłem studia doktoranckie z innym tematem pod opieką prof. Kazimierza Kowalewicza (z którym pracowałem także przy magisterce), niemniej big data wciąż chodziło mi po głowie. Po lekturze książki - Cukier, Kenneth i Victor Mayer-Schönberger. 2014. Big Data. Rewolucja, która zmieni nasze myślenie, pracę i życie. Warszawa: MT Biznes Ltd. - i akceptacji opiekuna naukowego w 2015 zacząłem intensywne poszukiwanie literatury. Około rok trwało określenie właściwego tematu pracy doktorskiej. Początkowo szukałem tematu jak najbardziej wąskiego, konkretnego, chciałem także wykorzystać przede wszystkim metody text miningowe. Jednak gdy zorientowałem się, że nie ma obszernej, etnograficznej pracy o data science, pod koniec 2016 postawiłem na szeroko zakrojoną pracę eksploracyjną z wykorzystaniem dobrze mi znanych i rozumianych metod oraz ramy teoretycznej. Chyba nie potrafiłbym napisać innej pracy doktorskiej, zatem moje "jakościowe" wykształcenie z czasów magisterskich ma silny wpływ na kształt tej rozprawy. Widzę jeszcze trzy obszary mojego usytuowania, które ją kształtują. Po pierwsze - i o tym powiem jeszcze nie raz - na początku badań trochę bardziej chciałem sam zostać data scientistem niż pisać pracę socjologiczną o data science. Nazywam to "byciem wannabesem". Po drugie, uważam swoje dawne zbliżenie z matematyką za zaletę w badaniu tego obszaru. Choć nauczyłem się bardzo niewiele programowania, tyle samo analizy danych i jeszcze mniej uczenia maszynowego, to nie bałem się tych tematów (a nawet niekiedy zajmowanie się nimi było przyjemne). Na podstawie obserwacji potocznych sądzę, że wśród zdecydowanej większości moich koleżanek i kolegów humanistów występuje w różnych konfiguracjach

> strach, niechęć i brak zainteresowania podobnymi tematami (nawet analizą danych ankietowych w SPSSie; nawet zapoznaniem się z tabelami / wykresami sporządzonymi przez kogoś innego). Po trzecie, osobiście ważna jest dla mnie prywatność. Jestem introwertykiem i nigdy nie miałem ochoty na silną obecność w mediach społecznościowych, niemniej bez wątpienia lektury dotyczące zagrożeń prywatności i innych konsekwencji społecznych związanych ze zbieraniem i analizą danych o ludziach sprawiły, że zmieniłem swoje zwyczaje i narzędzia cyfrowe. W rodzinie i wśród znajomych jestem znany jako zachęcający do skasowania konta na Facebooku, używania innych niż Google wyszukiwarek, sprawdzania uprawnień instalowanych aplikacji, używania szyfrowanych email, stosowania adblocków, dbania o bezpieczeństwo haseł i pinów. {autoetnografia 12}

Opisywaną wyżej chęć wejścia do świata data science w sensie marzeń o podjęciu pracy zarobkowej w tym obszarze badacz zdiagnozował u siebie pod koniec 2018 roku – wtedy, gdy już osłabła:

> Na moje wyniki badań może mieć wpływ moje umocowanie w świecie społ, jako początkującego i - teraz już mniej ale wcześniej bardzo - wannabesa. Świat początkującego jest inny niż doświadczonego, także muszę się postarać żeby nie opisywać swojego świata data science a w ogóle świat oczami uczestników. A wannabes to jeszcze bardziej niebezpieczny dla sensowności wyników, bo idealizuje i wzoruje się, dąży do bycia tym więc nie ma w tym krytyki lub marginalna {autoetnografia 8}

Niecałe dwa miesiące później badacz był już w stanie zrekonstruować swoje doświadczenia w dłuższej perspektywie zauważając, że przez jakiś czas był uwiedziony nie tylko wyobrażoną atrakcyjnością pracy w data science, ale także obietnicami epistemologicznymi obecnymi w nietechnicznym dyskursie o tzw. big data:

> Przeszedłem drogę zmiany poglądów co do data science.
> Początkowo ~2014 zwracałem uwagę tylko na termin big data, czytałem głównie po polsku rzeczy biznesowe. Zafascynowałem się tematem po lekturze "Big Data. Rewolucja..." i będąc na drugim roku [studiów doktoranckich] stałem się entuzjastą, wydawało mi się że to jest nowy, bardzo dobry sposób zdobywania wiedzy o świecie, a to zawsze interesowało mnie najbardziej. Jednocześnie zainteresowałem się mocno statystyką dzięki zajęciom na drugim roku, dowiedziałem się też że można pisać kod w SPSSie i uznałem, że to jest ekstra. Nabrałem z czasem spojrzenia krytycznego w sensie możliwości zdobywania wiedzy dzięki big data, i to widać w moim artykule z 2017 r. Wtedy czyli 2016 już zacząłem dużo czytać o data science i poszedłem na studia podyplomowe z analizy / data minig, zacząłem się sam uczyć R. Praca w BK [Biurze Karier UŁ] pozwalała mi na testowanie pomysłów i umiejętności na danych z badań absolwentów, starałem się dać raportom nową jakość a jednocześnie usprawnić swoją pracę, pisząc wszystko w syntaxie [języku programowania w IBM SPSS], do tego dołożyłem VBA dla zadań w wordzie, excelu. Zawód data scientista zafascynował mnie, oczarował, stałem się wannabesem, uczyłem się R nie z myślą o doktoracie, a mojej karierze zawodowej, chciałem zostać data scientistem. Kilka razy radziłem się żony co do rzucenia studiów doktoranckich i zmiany pracy. Miałem też wielokrotnie poczucie stresu, bo tyle jeszcze nauki przede mną w drodze do pracy w data science, przecież nie mam wykształcenia matematycznego, nie umiem programować. To czasami było frustrujące. W czasie badań, właściwie obserwacji czułem się czasem w podwójnej roli - badacza, ale i wannabesa, chciałem dowiedzieć się i nauczyć rzeczy po prostu dla siebie. Rzadko, ale dochodziła też znana mi z BK perspektywa administracji uniwersytetu, kiedy mowa była o studiach, programach studiów, tematach prac mgr utd. Dopiero w połowie 2018, kiedy ruszyłem z wywiadami, zacząłem chyba tracić albo słabło to

> moje wannabe. Dziś jest już bardzo słabe (...) Czasami rozmówcy pytają mnie, czy chcę się zająć data science, zawsze mówię że nie, ale dodaję że umiem zrobić to i owo w R i pracuję z danymi ankietowymi. Mówię im, że np. badania marketingowe mnie interesują. (...) {autoetnografia 9}

Wyżej badacz (RŻ) stwierdził, że uznaje swoje „swoje dawne zbliżenie z matematyką za zaletę w badaniu tego obszaru [DS]" {autoetnografia 12}, bowiem dzięki temu nie bał się podejmować prób pisania kodu, analizy i modelowania danych. Niemniej z wczesnych notatek autobiograficznych wynika, że czasami odczuwał on silny dyskomfort związany z niezrozumieniem rozmaitych szczegółów matematycznych i informatycznych działań podejmowanych przez uczestników badanego świata:

> Zaczynam pisanie tej refleksji z powodu jakiegoś stresu czy spięcia wywołanego nauką nowych rzeczy w R. Poczułem, że kiedy tylko usiadłem do nauki wspomnianego rozdziału, poczułem napięcie wywołane tym, ze nie rozumiem dokładnie jak działa funkcja (…) boję się czegoś nie rozumieć – tak z powodów technicznych / matematycznych, jak i nieznajomości angielskiego. Boję się nie rozumieć czegoś, co w moim przekonaniu jest rzeczą absolutnie podstawową, czymś co każdy data scientist ma w małym palcu. (...) {autoetnografia 3}

Piszący te słowa zdaje sobie sprawę, że jego usytuowanie jako osoby nie bojącej się matematyki i jednocześnie humanisty (socjologa jakościowego) prowadziło do tego, że poświęcił on szczególnie dużo uwagi technologiom badanego świata i starał się udowodnić swoje – nietypowe dla humanistów – rozeznanie w tej materii. Przy tym badacz jest pewien, że rozeznanie w stosowanych technologiach jest niezbędne do jakiegokolwiek zrozumienia świata społecznego DS, a właściwie również każdego innego świata społecznego. Zauważa on jedynie, że było to dla niego szczególnie ważne. Początkowo szukał nawet zewnętrznego potwierdzenia swojego rozeznania matematycznego i informatycznego np. dokładnie pamięta swoje zadowolenie z ukończenia pierwszego kursu używania języka R {obserwacja 3}, ze zdania korporacyjnego testu umiejętności numerycznych będącego jednym z elementów rekrutacji na warsztaty {obserwacja 7}, zadowolenie z ukończenia studiów podyplomowych „Analiza danych i data mining" na Wydziale Matematyki i Informatyki UŁ.

Opisane usytuowanie badacza przyczyniło się do głębszego zrozumienia dwóch aspektów badanego świata. Chodzi o zrozumienie perspektywy osób chcących wejść do świata DS (tzw. wannabesów) lub osób początkujących w DS oraz rolę tego, co nazywane bywa zdolnościami matematycznymi lub znajomością matematyki w praktykach legitymizacji i zaświadczania o autentyczności wśród uczestników świata DS.

Utrudnieniem w zrozumieniu sytuacji kobiet w badanym świecie społecznym było to, że badacz jest mężczyzną i wszystkie wywiady oraz obserwacje prowadził osobiście. Bez

wątpienia płeć jest czynnikiem wpływającym na interakcje w czasie badań jakościowych (Charmaz 2009:42–43; Konecki 2000:173–74). Pomimo rozpoznania sytuacji kobiet w DS jako areny tego świata społecznego początkowo nie potrafił on rozmawiać o tym z kobietami osadzonymi w świecie DS. Badacz nie był w stanie przyjąć perspektywy kobiet jako grupy mniejszościowej i w pewnych aspektach defaworyzowanej – może dlatego, że sam poza doświadczeniem rzadkiego i łagodnego wyśmiewania jego niskiego wzrostu i wątłej postury zawsze należał do grupy o uprzywilejowanej pozycji, tj. białych, heteroseksualnych mężczyzn z klasy średniej. Udało się badaczowi uzyskać głębsze narracje od Rozmówczyń/Rozmówców dopiero po konsultacjach z bardziej doświadczonymi socjolożkami i gruntownym przebudowaniu pytań dotyczących sytuacji kobiet. Dobrze otwierało rozmowę z kobietami pytanie „jak Ci się pracuje jako kobiecie?", zaś z mężczyznami „czy data science to jest męski zawód?" (pamiętając, że w badanym świecie nie zawsze mowa o działalności zarobkowej). Zgodnie z postulatami A. E. Clarke, co do nagłaśniania w badaniach światów społecznych marginalizowanych w dyskursie pozycji, badacz starał się opracować tę arenę jak najbardziej dokładnie, wychodząc poza męską perspektywę oraz – mamy nadzieję – unikając łatwej i pozornie „szlachetniej" roli obrońcy uciśnionych. Pozostał czujny i krytyczny wobec tego zadania, kwestionując swoje motywacje:

> Obawiam się że poprzez "nagłaśnianie dyskryminowanych kobiet w ds" próbuję uspokoić swoje sumienie co do tego, jak źle, seksistowsko i przedmiotowo zdarza mi się traktować kobiety. (…) {autoetnografia 10}

Możliwe, że w powyższych poszukiwaniach autora motywowało także to, że w 2017 r. urodziło się jego pierwsze dziecko – córka.

Innym aspektem wpływającym na wyniki naszych badań jest znacznie większa styczność badacza z językiem R niż z Python. Te dwa służące do wykonywania działania podstawowego i najbardziej popularne w badanym świecie języki programowania skryptowego wyznaczają dwa przenikające się subświaty społecznego świata DS w Polsce, a spór o przewadze jednego narzędzia nad drugim opisujemy jako arenę. Decyzja o koncentracji na jednej tylko technologii – czyli R – nie była przez nas podjęta świadomie. Już na etapie wczesnych poszukiwań w 2015 r. autor (RŻ) natrafił na dwuczęściowy, polskojęzyczny kurs „Pogromcy Danych" (Biecek 2015b, 2015a), który później ukończył w ramach badań {obserwacja 3}. W tym czasie język R był w sensie ilościowym bardziej popularny niż Python, około 2018 r. rozpoczęła się trwająca przewaga popularności Pythona (Nunns 2017; Piatetsky 2017b, 2018c; Strong 2018), na świecie i w Polsce. Ponad 4 z 5

badanych w Polsce firm AI deklarowało, w odpowiedzi na pytanie wielokrotnego wyboru, używanie Pythona, zaś około 2/5 używało GNU R (Borowiecki i Mieczkowski 2019:36–37). Badacz po zrozumieniu, że zajmuje się przede wszystkim subświatem DS posługującym się R, a także że społeczność użytkowników R jest znacznie szersza i nierzadko rozłączna ze światem DS (np. R powszechnie używają naukowcy akademiccy rozmaitych dziedzin obliczeniowych, a także analitycy danych z wielu dziedzin niekoniecznie uznawani za część świata DS), starał się zwrócić w kierunku subświata użytkowników Pythona {obserwacja 30, 31, 39, 43}. Ten aspekt usytuowania badacza przyczynił się do lepszego zrozumienia aren technologicznych (czy narzędziowych) w świecie DS.

4.3.2. Kwestie etyczne w badaniach własnych

Badania etnograficzne, a szczególnie obserwacja uczestnicząca ukryta jest, z etycznego punktu widzenia, problematyczną techniką badawczą (Angrosino 2010:121–25; Babbie 2006:336; Kacperczyk 2016:668–72; Li 2008). W odniesieniu do tej techniki podnoszone są przede wszystkim kwestie braku świadomej zgody osób obserwowanych na udział w badaniach oraz posługiwania się fałszywą tożsamością przez wykonujących obserwacje badaczy.

W odniesieniu do problemu świadomej zgody, to zdecydowaliśmy się w żadnym z aktów obserwacji nie ogłaszać na forum faktu prowadzenia badań, a zatem nie pytaliśmy o taką zgodę ani w obserwacjach offline (np. na meetupie), ani online (np. biorąc udział w kursie). Badacz obserwował jednak wyłącznie to, co ma charakter ogólnodostępny. Oznacza to, że nie zatrudniał się on, nie wstępował do żadnej formalnej ani nieformalnej organizacji, nie prowadził także obserwacji na spotkaniach prywatnych. Nie wszystkie obserwacje realizowano w miejscach publicznych (np. nie możemy uznać za publiczne warsztatów online z ograniczoną liczbą miejsc lub konferencji z udziałem płatnym), jednak bez wątpienia miały one charakter ogólnodostępny (nie były miejscami prywatnymi ani z dostępem ograniczonym wyłącznie dla członków organizacji), a udział w nich był niezbędny do pozyskania danych o badanym wycinku rzeczywistości społecznej:

> Metody badań niejawnych pozostają w sprzeczności z zasadami świadomej zgody i mogą naruszać prywatność badanych. Do obserwacji uczestniczącej i nieuczestniczącej w sferze niepublicznej bez wiedzy badanych (…) należy uciekać się wyłącznie wtedy, gdy inne metody nie wystarczają do pozyskania podstawowych danych (pkt. 17. Walne Zgromadzenie Delegatów Polskiego Towarzystwa Socjologicznego 2012).

By chronić prywatność osób obserwowanych publikujemy jedynie fragmenty z notatek terenowych badacza, materiały wizualne pozyskane z publicznych zasobów internetu lub świadomie przekazane badaczowi, ewentualnie wytworzone przez badacza (jak zdjęcia, ale nie zawierające wizerunku ani informacji identyfikujących osoby). Badacz (RŻ) nie korzystał w trakcie obserwacji z rejestrowania nagrań audio ani wideo. Brak świadomej zgody w naszych obserwacjach uczestniczących nie zaszkodził osobom obserwowanym.

Autor (RŻ) nie posługiwał się fałszywą tożsamością, jednak w początkowej fazie badań najczęściej prezentował najbardziej komfortowy dla siebie wycinek prawdy, zatajając fakt prowadzenia badań. Przez około 18 pierwszych miesięcy rzadko ujawniał on fakt prowadzenia badań socjologicznych społecznego świata DS do swojej pracy doktorskiej. Najczęściej pozostawał przy tożsamości socjologa ankietowego, pracującego w administracji Uniwersytetu Łódzkiego (w Biurze Karier), głównie z wykorzystaniem MS Excel i IBM SPSS, a zainteresowanego możliwościami analizy danych w R i generalnie zaawansowaną analizą danych ilościowych oraz uczeniem maszynowym. Tożsamość ta, zgodna z prawdą, była dla badacza stosunkowo wygodna i nie wzbudzająca wielu pytań wśród osób poznawanych w trakcie obserwacji. Na „korzyść" badacza przemawiał też fakt, że wizualnie raczej nie odróżniał się od zawsze licznych grup początkujących data scientistów czy tzw. wannabesów (osób dla których badany świat jest grupą odniesienia normatywnego; generalnie zainteresowanych DS, chcących wejść do świata) – młodo wyglądający, nie wyróżniający się strojem mężczyzna z plecakiem na laptop to najbardziej typowy widok na wydarzeniach badanego świata. Później badacz zdecydował przedstawiać się bez ogródek jako socjolog piszący o data science w Polsce, co, jak wspominaliśmy, miało niewątpliwy walor praktyczny, gdyż umożliwiało rekrutowanie w terenie osób do wywiadów swobodnych. Poza praktyką, badacz odczuwał narastający dyskomfort z powodu posługiwania się półprawdą, a z drugiej strony stawał się coraz bardziej pewnym siebie badaczem – mającym orientację w języku badanego świata, wiedzącym co i po co robi, potrafiącym podzielić się własną hipotezą. Obawy o zbyt wczesne „bycie zdemaskowanym" był naszym zdaniem słuszne z metodologicznego punktu widzenia. Badacz ujawniając swoją tożsamość wobec uczestników badanego świata DS był zawsze przepytywany mniej lub bardziej szczegółowo o swoją metodologię i wyniki. Zatem gdyby w takiej pierwszej interakcji dowiódł swojej niekompetencji, miałby zapewne kłopoty z zaufaniem, rekrutacją i co najmniej z poważnym traktowaniem w (co najmniej) kręgach towarzyskich lub zawodowych danego uczestnika

świata społecznego. Ostatecznym „ujawnieniem się" był dla autora występ na meetupie świata DS z prezentacją swoich roboczych wyników {obserwacja 33}.

Brak zgody na udział w naszych badaniach zachodził także w analizach danych zastanych. Nie informowaliśmy o badaniach ani podmiotów udostępniających te dane, ani podmiotów będących jednostkami analizy (respondenci albo grupy meetupowe; zresztą w przypadku respondentów nie mogliśmy zidentyfikować personaliów), bowiem korzystaliśmy wyłącznie z danych dostępnych legalnie i dostępnych publicznie – poza danymi platformy Kaggle, gdzie zbiory mogą pobrać wyłącznie użytkownicy zalogowani na swoje konto. W przypadku danych z meetup.com same dane o grupach są dostępne publicznie w internecie, ale skorzystaliśmy z metody pobierania (API) udostępnianej jedynie użytkownikom posiadającym konto na portalu. Wątpliwości etyczne mógłby budzić fakt obejścia mechanizmu wyszukiwania grup przez badacza – portal domyślnie pozwala na wyszukanie wszystkich grup zlokalizowanych w promieniu do 100 mil od lokalizacji podanej na koncie użytkownika[85]. W inny miejscu można jednak sprawdzić liczbę grup dla całej Polski. Znając tę liczbę, autor (RŻ) wykonał najpierw zapytanie dla swojej prawdziwej lokalizacji (Łódź), później zmieniał lokalizację ręcznie[86] tak, by wyszukiwanie o promieniu 100 mil pokryło cały obszar Polski, po czym zmienił ją z powrotem na Łódź. Pobrane pliki badacz połączył, odfiltrował grupy spoza Polski i usunął zduplikowane rekordy. Naszym zdaniem, to postępowanie nie szkodzi osobom badanym ani interesowi grup meetupowych. Miało ono na celu pobranie publicznie dostępnych danych, zatem nie zachodzi to, co Jemielniak (2019:160) nazwał dostępnością techniczną, a nie legalną.

Ponieważ zgadzamy się ze stanowiskiem, że „ostatecznie decyzja, czy tworzyć autoetnografię analityczną czy ewokatywną jest decyzją o charakterze etycznym" (Kacperczyk 2017:141) i opowiadamy się raczej za typem analitycznym, musimy wyjaśnić nasze stanowisko etyczne szerzej. Kierujemy się w naszej pracy tym, co Kacperczyk nazwała etyką scjentystyczną, zakładającą, że

> badacz ma prawo użyć instrumentalnie każdego źródła danych, by zgłębić analizowany temat i wydrzeć mu jego tajemnice. (…) Analiza jest zawsze cięciem i kawałkowaniem, enzymatycznym procesem toczącym się na „danych", którym od początku do końca zarządza badacz (ibidem).

85 Można zmienić ustawienie na globalne, ale nie działa wtedy filtr kraju i ilość pobranych rekordów jest ograniczona do 500. Badacz zastosował to zapytanie, jednak nie pozwoliło ono pobrać danych o wszystkich 86 grupach z zapisanym tematem „data science" w Polsce.

86 Kolejno zmiany na: Poznań, Wrocław, Rzeszów, Olsztyn, Szczecin, Biała Podlaska, Koszalin, Bielsko-Biała.

Stąd w tej pracy także wąskie zastosowanie etnografii opartej na współpracy – badacz konsultował się z osobami uczestniczącymi w społecznym świecie DS, ale samodzielnie nadawał ostateczny kształt konsultowanym fragmentom. Jest to z jednej strony zgodne z przyjętym przez nas stanowiskiem konstruktywistycznym – zakładamy, że dokonywany w tej pracy opis świata społecznego jest **konstruowany** przez badacza, a nie **odkrywany –** z drugiej strony zgodne z podejściem etycznym badanego przez nas świata społecznego DS.

4.3.3. Problemy w postępowaniu badawczym

> Generalnie postęp w nauce jest efektem uczciwości i otwartości; kłamstwa i próba obrony własnego „ja" jedynie go opóźniają. Najlepiej służy się społeczności naukowej – i nauce jako takiej – ujawniając prawdę o wszystkich pułapkach i problemach, które napotkało się na konkretnej drodze naukowej analizy. Być może pozwoli to uniknąć ich innym (Babbie 2006:523).
> (…) ujawnienie problemów w postępowaniu badawczym spotyka się z tak jednoznacznie pozytywnym odbiorem recenzentów [socjologicznych pracy doktorskich] (Kretek-Kamińska 2016:288).

Jak wspomniano badacz planował wykonanie ilościowej analizy danych tekstowych – postów, komentarzy i ewentualnie reakcji z publicznej, polskojęzycznej grupy facebookowej Data Science PL. Nie doszło to do skutku z powodu braku zgody firmy Facebook. Oczywiście odnalezienie w sieci skryptów do pobrania takich danych w języku Python nie stanowiło kłopotu. Z uwagi na jasną instrukcję badacz wybrał rozwiązanie Maxa Wolfa[87]. Legalne pobranie publicznie dostępnych danych miało być wykonane za pomocą technologii Graph API[88] i wymagać identyfikatora grupy (dla Data Science PL to 971898396201529), oraz identyfikatora i hasła aplikacji Facebooka uzyskującej dostęp do danych (*App ID* oraz *App Secret*). Badacz, podając swoje dane osobowe, utworzył prywatne konto na Facebooku, a następnie konto deweloperskie oraz „pustą" aplikację, służącą jedynie skorzystaniu z API. Dzięki temu posiadał on *App ID* oraz *App Secret*, przygotował zatem i uruchomił skrypty, uzyskując powtarzalny błąd. Inaczej niż przed 2018 rokiem, owa „pusta" aplikacja wymagała autoryzacji[89] do pobrania danych publicznych. Badacz wypełnił zatem odpowiednie formularze z prośbą o autoryzację, opisując swoją aplikację jako służącą jedynie do pobrania danych z publicznej grupy do celów naukowych i podając swoją afiliację wraz z pełnymi

87 https://github.com/minimaxir/facebook-page-post-scraper, dostęp 05.03.2019 r.

88 Jednym z najsłynniejszych przypadków zastosowania Graph API jest pobranie danych przekazanych później firmie Cambridge Analytica (por. 2.2.3.4); oficjalne materiały techniczne Graph API dostępne są na https://developers.facebook.com/docs/graph-api, dostęp 19.02.2019 r.

89 Komunikat brzmiał: *(#10) To use 'Groups API', your use of this endpoint must be reviewed and approved by Facebook. To submit this 'Groups API' feature for review please read our documentation on reviewable features:* https://developers.facebook.com/docs/apps/review, dostęp 12.05.2019 r.

danymi kontaktowymi Uniwersytetu Łódzkiego. Otrzymał odpowiedź odmowną (w ocenie badacza formularz był sprawdzany automatycznie) i zrezygnował z analiz ilościowych grupy Data Science PL, ograniczając się do netnografii.

Badaczowi odmówiono także dostępu do danych źródłowych sondażu ponad 260 zlokalizowanych w Polsce firm zajmujących się tzw. sztuczną inteligencją autorstwa Fundacji digitalpoland (Borowiecki i Mieczkowski 2019), powołując się na „zgody i RODO”[90], zatem korzystaliśmy tylko z informacji opublikowanych w raporcie Fundacji (ibidem).

Pod koniec badań problemem dla badacza stało się utrzymanie aktywnego słuchania podczas wywiadów. Zupełnie nowe dane pojawiały się rzadko, w konsekwencji czego badacz spostrzegł u siebie nieprofesjonalne zachowania w interakcjach z Rozmówcami:

> po wywiadzie 25.
> kończę już swoje badania. zauważyłem że podczas wywiadów cieszę się, kiedy rozmówczyni/rozmówca mówi coś zgodnego z moimi spostrzeżeniami i daję temu wyraz uśmiechem, tonem głosu, entuzjastycznym potakiwaniem. Na tym etapie rzadko dowiaduję się czegoś zupełnie nowego i niespodziewanego, zatem podczas wywiadów albo cieszę się, albo jestem wręcz znudzony, kiedy słyszę to samo po raz dwudziesty któryś. To są trudne momenty, gdyż staram się traktować z jak największym szacunkiem uczestników, mam nadzieję że nikt nie odczuł mojego spadku uwagi. Kończę zatem z poczuciem, że wiele kategorii zostało nasyconych. Problemem jest kodowanie w qda miner, mam poczucie że te kody są za mało abstrakcyjne, za mało ogólne, za blisko transkrypcji. Bardziej wartościowe, abstrakcyjne kategorie mam w notach teoretycznych i innych, luźniejszych pomysłach, w memo. Chcę to złożyć w całość (…) {autoetnografia 11}

Takie odczucia badacza mogą wskazywać na to, że zebrał w terenie wystarczający w odniesieniu do celów niniejszej pracy materiał jakościowy. Jeżeli kolejne badaczki i badacze terenowi mogą skorzystać na tym doświadczeniu, to niech przyświeca im myśl, że znacznie lepiej pracuje się w terenie będąc autentycznie zaciekawionym drugim człowiekiem i badanym wycinkiem rzeczywistości społecznej, niż udając takie zaciekawienie wobec kogokolwiek.

---

90 Odpowiedź przesłana badaczowi:
W dniu 2019-05-31 o 13:03, Piotr Mieczkowski <piotr.mieczkowski@digitalpoland.org> pisze:
> Dzień dobry,
>>
Nie jest niestety to możliwe z racji zgód i RODO. Podmioty tylko dlatego tak ochoczo wypełniły nasz kwestionariusz, gdyż zagwarantowaliśmy im publiczną anonimowość i brak wglądu strony trzeciej do źródłowych danych ☹ Firmy podawały konkretne pola jak przychody i to nie są informacje publicznie a przez wiele firm traktowane jako "wrażliwe".
>>>
Ukłony,
> Piotr

Innego rodzaju problemem badacza było doświadczanie tzw. paraliżu analitycznego, czyli sytuacji, kiedy nie wiedział on, jaką kolejną czynność ma wykonać i jakie decyzje podjąć. Szczególnie dotyczyło to kodowania jakościowego, które było nie tylko żmudne, długotrwałe i dość niewygodne (o czym za chwilę), ale przede wszystkim obciążające kognitywnie. Ilość decyzji interpretacyjnych, szczególnie w początkowym stadium kodowania, kiedy co chwila powstaje potrzeba tworzenia nowego kodu, była dla badacza (RŻ) męcząca i momentami paraliżująca. W przezwyciężeniu tego problemu pomagały zwyczajne przerwy w kodowaniu i naprzemienna koncentracja pracy na realizacji badań i kodowaniu, sporządzanie map (por. Clarke 2005:xxi; 83-86) oraz konsultacje lub dyskusje z rozmaitymi osobami, także nie mającymi nic wspólnego z data science ani naukami społecznymi.

Niewygoda kodowania wynikała z prozaicznych problemów dotyczących oprogramowania CAQDAS. Pierwszą niedogodnością był sam wybór narzędzia. Badacz nie miał dostępu do licencji na najpopularniejsze komercyjne pakiety (Atlas.it, NVIVO), zaś polecane kilka lat temu rozwiązania open source (Niedbalski 2013) okazały się bezużyteczne[91]. Badacz zdecydował się zatem na bezpłatny, dostępny na licencji zamkniętej program QDA Miner Lite. Cechą, na której nam zależało była możliwość kodowania plików testowych i graficznych, a ten program firmy Provalis Research dawał tę możliwość. Praca z aplikacją była dla badacza niewygodna – głównie z powodu konieczności bezustannego klikania, spowodowanej niemal zupełnym brakiem skrótów klawiaturowych do obsługi programu[92]. Ciągłe klikanie połączone z innymi kłopotami, brakiem przystępnych materiałów z rozwiązaniami, a co najgorsze brakiem zapisu swoich poczynań krok po kroku połączonym z niemożnością pewnego odtworzenia wcześniejszego stanu pracy było dla piszącego te słowa wręcz zniechęcające, gdy porównał doświadczenie braku komfortu i bezpieczeństwa pracy w QDA Miner z pracą na danych ilościowych w języku R (a nawet w IBM SPSS lub arkuszach kalkulacyjnych, jak MS Excel czy Google Sheets). Dodatkową niedogodnością było to, że QDA Miner Lite z niewyjaśnionego dla badacza powodu nie pojawia się w widoku *Task Switcher* systemu Windows 7, zatem pracując w innej aplikacji badacz nie mógł przełączyć się do QDA Miner Lite za pomocą skrótu Alt + Tab, do którego jest od wielu lat

91 Najbardziej rozbudowany i obiecujący program Weft QDA nie jest wspierany ani rozwijany od 2009 roku. Twórca Weft QDA Alex Fenton poinformował, że program zawiera błędy mogące spowodować utratę danych lub brak możliwości uruchomienia zapisanych plików i nie poleca swojego programu do dużych projektów, jak badania do pracy doktorskiej (por. http://www.pressure.to/qda/, dostęp 29.04.2019 r.).

92 Użytkownik ma jedynie możliwość dodania skrótów klawiaturowych do używania konkretnych kodów, tj. można za pomocą kursora zaznaczyć wybrany tekst lub obszar grafiki i wciskając skrót zakodować ten fragment.

przyzwyczajony. Program jest dostępny tylko na system Windows, co także przeszkadzało badaczowi pracującemu czasem na Ubuntu 18.04. Po kilku miesiącach pracy badacz uzyskał legalną kopię pełnej wersji programu QDA Miner od kolegi wcześniej używającego tej aplikacji (Burski 2014). Wersja pełna oferuje o wiele większe możliwości, np. obliczenie współwystępowania kodów z użyciem indeksu Jaccarda i wizualizację klastrów kodów za pomocą dendrogramu. Niestety poważnym problemem okazał się inny w wersji Lite a wersji pełnej system zapisu znaków w aplikacji. W pokaźnej porcji zakodowanego materiału zaimportowanej z wersji Lite do wersji pełnej polskie znaki diakrytyczne były odczytywane nieprawidłowo w materiale, zaś prawidłowo w kodach. Badacz otrzymał informację z pomocy technicznej firmy Provalis Research, że system zapisu znaków różni się pomiędzy wersjami i problem jest niemożliwy do rozwiązania. Uznając poprawianie błędnie odczytanych znaków za zbędne, ostatecznie badacz kodował materiał jakościowy w wersji Lite, zaś z wersji pełnej korzystał jedynie doraźnie, na potrzeby użycia bardziej zaawansowanych funkcji analitycznych, importując specjalnie przygotowaną kopię pliku. Doświadczył on również kłopotów z zapisaniem wyników swojej pracy oraz nieoczekiwanym wyłączanie się programu QDA Miner Lite. Kontakt z pomocą techniczną po raz kolejny nie był pomocny, zdaniem pracowników plik o wielkości naszego (niecałe 2 GB) powinien być obsługiwany prawidłowo. Podsumowując, nie polecamy nikomu pracy z oprogramowaniem QDA Miner. Gdyby badacz zaczynał kodowanie jakościowe teraz, prawdopodobnie ograniczyłby się do danych tekstowych i próbowałby skorzystać z któregoś pakietu dla języka R (Huang 2018; Peters 2019; Rinker 2020).

Rozdział 5. Elementy społecznego świata data science

W rozdziale prezentujemy elementy społecznego świata DS w Polsce, posługując się podstawowymi pojęciami przyjętej ramy teoretycznej (por. część 3.1. i rys. 3). Arenom i procesom zachodzącym w społecznym świecie poświęcamy osobne części. Poniżej koncentrujemy się na trzech elementach. Po pierwsze, na opisie działania podstawowego i przykładach wyników pracy data scientistów (część 5.2.), by pokazać jak najbardziej dogłębnie, co właściwie robią uczestnicy badanego świata społecznego. Jest to realizacja pierwszego z celów szczegółowych niniejszej rozprawy (por. część 1.2). Po drugie, na technologiach tego świata (5.3.), które pomagają zrozumieć jak odbywa się działanie podstawowe oraz są elementem ściśle związanym z procesami wyznaczania granic społecznego świata, procesami legitymizacji i zaświadczania o autentyczności, i procesami segmentacji świata społecznego (w tym profesjonalizacji jako jednej z konsekwencji segmentacji). W ten sposób – wraz z ujętą w kolejnym rozdziale częścią 6.1. – realizujemy drugi z celów szczegółowych rozprawy. Po trzecie, rekonstruujemy i opisujemy wartości społecznego świata DS (5.4.). Traktujemy wartości jako punkty odniesienia w ocenianiu sposobu wykonania działania podstawowego i działań wspomagających oraz element nierozerwalnie spleciony z tożsamością uczestników świata. Tym samym wartości są powiązane z wymienionymi wyżej procesami (wyznaczania granic, legitymizacji i zaświadczania o autentyczności, segmentacji). Cześć 5.4. służy realizacji trzeciego celu szczegółowego rozprawy. Części poświęcone elementom społecznego świata DS w Polsce poprzedzamy szkicem o charakterze ilościowym (5.1.), który pomaga zrozumieć pełny obraz sytuacji badania – pojąć skalę świata, rozmieszczenie geograficzne, jego demografię, charakter ekonomiczny firm i rynku pracy DS, rynku edukacyjnego oraz zmiany następujące w czasie. Cześć ta bazuje na źródłach zastanych, poddanych wtórnej analizie, w tym aktach samowiedzy badanego świata społecznego. Części 5.2.-5.4. to głównie wyniki badań własnych.

5.1. Data science w Polsce – ujęcie ilościowe

5.1.1. Szkic geograficzny

Bazując na badaniu przeprowadzonym przez Fundację digitalpoland w październiku 2018 r. na grupie ponad 160 firm[93] zajmujących się sztuczną inteligencją[94] (Borowiecki i Mieczkowski 2019), informacjach o studiach i badaniach naukowych (Borowiecki i Mieczkowski 2019:52; Czarnowski i in. 2018; Gontarz 2019:34–35; Ministerstwo Cyfryzacji RP 2018b:111–13), aktach samowiedzy – wynikach sondaży wewnątrz świata DS (Migdał 2017; Prokulski 2017) oraz własnych analizach dotyczących wydarzeń w polskim świecie społecznym DS stwierdzamy, że **badany świat koncentruje się w największych polskich ośrodkach miejskich, przy dominacji Warszawy**[95].

Zaledwie 15% firm ulokowanych jest poza sześcioma największymi ośrodkami miejskimi, z których każdy liczy więcej niż pół miliona mieszkańców. W Warszawie ma swoje biura 43% badanych podmiotów, 11% w Trójmieście (Gdańsku, Gdyni lub Sopocie), 9% w Krakowie, po 8% w Poznaniu i Wrocławiu, 6% w aglomeracji katowickiej (Borowiecki i Mieczkowski 2019:14). Autorzy zauważają, że choć Łódź jest trzecim co do wielkości polskim miastem, to nie ma wyraźnej reprezentacji w ilości firm zajmujących się AI. Firmy w Łodzi zaklasyfikowano do 15% podmiotów ulokowanych poza sześcioma wyróżnionymi ośrodkami miejskimi (ibidem).

Studiowanie kierunków DS jest możliwe m.in. na dziesięciu najlepszych – według Rankingu Szkół Wyższych magazynu „Perspektywy" – polskich uczelniach (Gontarz 2019:34-35). Dodajemy do nich inne, znane nam studia nie wymienione w raporcie Andrzeja Gontarza, podając źródło informacji. W podziale na lokalizację są to studia:

- W Warszawie:
    - Politechnika Warszawska:

93 „Ponad 160 małych i dużych firm wzięło udział w naszym sondażu. Nasze badania wskazują, że w Polsce w obszarze AI działa ponad 260 firm." (Borowiecki i Mieczkowski 2019:10). Z materiałów dołączonych do raportu wynika, że są to 264 podmioty.

94 Co zdefiniowano jako „firmy rozwijające lub oferujące usługi lub produkty AI (*companies which develop or offer AI services or products*)" (Borowiecki i Mieczkowski 2019:14).

95 We wszystkich analizowanych przez nas wymiarach w Warszawie zlokalizowana jest co najmniej 1/4 (wymiar: ilość grup meetupowych) działalności w Polsce. W innych wymiarach (ilość członków grup meetupowych, ilość kierunków studiów, ilość firm) w Warszawie zlokalizowane jest co najmniej 2/5 działalności (por. rys. 4.). Szczegóły prezentujemy dalej.

        - studia stacjonarne I stopnia na kierunku „Inżynieria i analiza danych”, Wydział Matematyki i Nauk Informacyjnych;
        - studia podyplomowe „Data Science – algorytmy, narzędzia i aplikacje klasy Big Data”;
        - studia podyplomowe „Big Data – przetwarzanie i analiza dużych zbiorów danych”;
        - studia I stopnia „Inżynieria i analiza danych” (Wydział Matematyki i Nauk Informacyjnych PW 2019).
    - Uniwersytet Warszawski:
        - studia II stopnia „Informatyka i Ekonometria, specjalność Data Science”, Wydział Nauk Ekonomicznych;
        - studia podyplomowe „Data Science w zastosowaniach biznesowych – warsztaty z wykorzystaniem programu R”.
    - Szkoła Główna Handlowa:
        - studia II stopnia „Advanced Analytics – Big Data” prowadzone w języku angielskim;
        - studia podyplomowe „Inżynieria danych – Big Data”;
        - studia I stopnia „Metody ilościowe w ekonomii i systemy informacyjne” (Szkoła Główna Handlowa 2018).
    - Szkoła Główna Gospodarstwa Wiejskiego:
        - specjalność „Big Data Analytics” na studiach II stopnia kierunku „Informatyka i ekonometria”.
    - Polsko-Japońska Akademia Technik Komputerowych:
        - studia podyplomowe „Big Data – Inżynieria dużych zbiorów danych” .
- W Poznaniu:
    - Uniwersytet im. Adama Mickiewicza:
        - studia II stopnia „Analiza i przetwarzanie danych”, Wydział Matematyki i Informatyki;
        - studia podyplomowe „Przetwarzanie danych – Big Data” na tym samym wydziale.
    - Politechnika Poznańska:
        - specjalność „Technologie Przetwarzania Danych” na studiach II stopnia kierunku „Informatyka”;

        - studia podyplomowe „Hurtownie danych i analiza danych”
- W Krakowie:
    - Uniwersytet Jagielloński:
        - ścieżka edukacyjna „Data Science” na Wydziale Fizyki, Astronomii i Informatyki Stosowanej;
        - studia podyplomowe „Analiza Biznesowa” w Instytucie Informatyki i Matematyki Komputerowej.
    - Akademia Górniczo-Hutnicza w Krakowie:
        - studia podyplomowe „Analiza danych – Data Science” na Wydziale Informatyki, Elektroniki i Telekomunikacji;
        - studia podyplomowe „Metody statystycznej analizy danych”, Wydział Zarządzania, Samodzielna Pracownia Zastosowań Matematyki w Ekonomii.
- W Łodzi:
    - Uniwersytet Łódzki:
        - studia I stopnia „Analiza danych” (Wydział Matematyki i Informatyki UŁ 2018);
        - studia podyplomowe „Analiza danych i data mining”[96] (Wydział Matematyki i Informatyki UŁ 2016);
    - Politechnika Łódzka:
        - specjalność „Inżynieria Oprogramowania i Analiza Danych” na studiach I stopnia kierunku „Informatyka” (Wydział Fizyki Technicznej i Matematyki Stosowanej PŁ 2018b);
        - specjalność „Matematyczne metody analizy danych biznesowych” na studiach I stopnia kierunku „Matematyka” (Wydział Fizyki Technicznej i Matematyki Stosowanej PŁ 2018a);
- W Gliwicach:
    - Politechnika Śląska:
        - specjalizacja „Data Science” w ramach studiów II stopnia na Wydziale Automatyki, Elektroniki i Informatyki.
- We Wrocławiu:
    - Politechnika Wrocławska:

96 Autor niniejszej rozprawy ukończył te studia w roku akademickim 2016/2017.

        - studia II stopnia „Data Science (Danologia)", Wydział Informatyki i Zarządzania.
- W Białymstoku:
    - Politechnika Białostocka:
        - studia podyplomowe „Data Science" (Wydział Informatyki Politechniki Białostockiej 2019).
- W Gdańsku:
    - Politechnika Gdańska:
        - studia podyplomowe „Inżynieria danych – data science" (Politechnika Gdańska 2019).

> Według danych Ministerstwa nauki i Szkolnictwa Wyższego dotyczących rekrutacji na studia na rok akademicki 2018/2019 (w uczelniach nadzorowanych przez ministerstwo) najwięcej kandydatów na jedno miejsce przypadało na inżynierię farmaceutyczną (18,3) oraz inżynierię i analizę danych (17,8). Rok wcześniej **inżynieria i analiza danych była na pierwszym miejscu – 54,6 zgłoszeń na jedno miejsce – a dwa lata temu taki kierunek studiów w ogóle nie pojawił się w raporcie ministerstwa** [podkr. RŻ]. Nie świadczy to o spadku zainteresowania studiami w tym obszarze. Spadek liczby chętnych na jedno miejsce odzwierciedla raczej rozwój oferty studiów związanych z danymi w całej Polsce (Gontarz 2019:31).

Podsumowując, z 27 wyróżnionych wyżej kierunków studiów w Warszawie zlokalizowanych jest 11 (ponad 2/5, co jest wynikiem zbliżonym do przytoczonych 43% obecnych w Polsce firm zajmujących się AI, a mających biura w tym właśnie mieście). Po cztery kierunki studiów znajdują się w Poznaniu, Krakowie i Łodzi, zaś po jednym w Gliwicach, Wrocławiu, Białymstoku i Gdańsku. Można się kształcić w dziedzinie DS – przynajmniej częściowo – poza uczelniami. Znane są nam dwa polskie kursy dla początkujących (Alekseichenko 2017; Kodołamacz.pl i Centrum Zarządzania Innowacjami i Transferem Technologii Politechniki Warszawskiej 2017). Fundacja digitalpoland wymienia osiem działających w Polsce firm, które oferują kursy/szkolenia (*AI Trainings*) (Borowiecki i Mieczkowski 2019:59). Polacy kształcą się także na międzynarodowych kursach MOOC. Podkreślmy dodatkowo, że w świecie społecznym DS, nie tylko w Polsce, uczestniczą – czy węziej rzecz ujmując: pracują – osoby z różnym wykształceniem (por. Burtch 2018:20). Przedstawiamy powyższe informacje, by zilustrować, gdzie w sensie geograficznym koncentruje się w Polsce dyskurs dotyczący badanego świata społecznego.

W działalności naukowej związanej ze sztuczną inteligencją dokument Ministerstwa Cyfryzacji RP wymienia jedynie ośrodki warszawskie i krakowskie. Łącznie w Warszawie

wyróżniono 33 zespoły[97], a w Krakowie 13 (Ministerstwo Cyfryzacji RP 2018b:111–17). Istnieje Polskie Porozumienie na Rzecz Rozwoju Sztucznej Inteligencji (PP-RAI), złożone z pięciu polskich organizacji naukowych: Polskiego Stowarzyszenia Sztucznej Inteligencji, Polskiego Towarzystwa Sieci Neuronowych, Polskiej Grupy Systemów Uczących się, Polskiego Oddziału IEEE SMC, Polskiego Oddziału IEEE Computational Intelligence Society (Czarnowski i in. 2018:29). W Polsce działa jeszcze pięć innych tego rodzaju organizacji nie będących częścią PP-RAI. Ogółem z owych 10 organizacji naukowych cztery mają prezesów pracujących w Warszawie, po dwie we Wrocławiu i Poznaniu, a po jednej w Krakowie, Częstochowie, Gdyni i Toruniu. Osób pełniących funkcje prezesek/prezesów jest 12 przy 10 organizacjach, ponieważ Polska Grupa Systemów Uczących się ma trzech prezesów (z Warszawy, Wrocławia i Poznania) (Borowiecki i Mieczkowski 2019:52). Na Uniwersytecie Łódzkim w styczniu 2019 powstało Centrum Badań nad Sztuczną Inteligencją i Cyberkomunikacją. Jednostka na ma celu badanie wpływu sztucznej inteligencji na zarządzanie i organizacje, zarządzanie komunikacją i relacjami w mediach społecznościowych oraz na platformach cyfrowych, czy zarządzanie przedsiębiorstwem cyfrowym i wirtualnym zespołem (Polska Agencja Prasowa 2019). Studia doktoranckie, reklamowane terminem DS, to studia prowadzone wspólnie przez Uniwersytet Warszawski i Politechnikę Warszawską w dziedzinach matematyki i informatyki (Biecek 2017b), oraz z zakresu obliczeniowej ekonomii i psychologii (*Quantitative Psychology and Economics*) na UW (Uniwersytet Warszawski 2019).

Wydarzenia publiczne w polskim świecie społecznym DS dzielimy na konferencje, hackatony i meetupy. Są to wydarzenia prawie zawsze ogólnodostępne, a poza konferencjami także bezpłatne. Mamy na myśli spotkania twarzą w twarz, bo jeśli chodzi o spotkania online to zdecydowanie najważniejszym otwartym forum jest grupa facebookowa Data Science PL założona przez Piotra Migdała[98]. Fundacja digitalpoland wymieniła 11 najważniejszych polskich konferencji w obszarach AI, ML, DS i big data (Borowiecki i Mieczkowski 2019:53-54) w sposób zgodny z naszymi obserwacjami, zaznaczając, że większość z nich odbywa się w Warszawie. Przyjrzyjmy się największej w Polsce, międzynarodowej konferencji PyData Warsaw. Dotychczas odbyły się dwie edycje, w 2017 i 2018 roku, jak sama nazwa wskazuje – w Warszawie, a dokładnie w Centrum Nauki Kopernik. W roku 2017 uczestnicy z Warszawy

97 Są to jednostki na różnych poziomach struktury uczelni: instytuty, katedry, zakłady. Zauważmy, że powtarzają się jednostki z różnych poziomów, np. wymieniono oddzielnie Instytut Ekonometrii w Kolegium Analiz Ekonomicznych SGH i oddzielnie Zakład Wspomagania i Analiz Decyzji, będący częścią tego Instytutu Ekonometrii (Ministerstwo Cyfryzacji RP 2018:115).

98 Podajemy kilka informacji ilościowych o grupie w części 5.1.4.

stanowili prawie 44% ogółu (181 z 415 osób), zaś w 2018 było to około 37,5% (163 z 435) (Borowiecki i Mieczkowski 2019:57; Migdał 2017).

Co do hackathonów, nie dysponujemy niestety możliwością rzetelnego ujęcia ilościowego tego zjawiska w Polsce. Powiedzmy jednak, czym one są. Termin jest połączeniem słów „hack" i „maraton" i generalnie oznacza wydarzenie, na którym kilkuosobowe zespoły pracując np. przez 24 godziny proponują rozwiązanie postawionego przez organizatorów zagadnienia w postaci prototypu, przeważnie aplikacji internetowej/komputerowej/mobilnej. Prototypy są prezentowane przed jury. Organizatorami mogą być firmy lub podmioty państwowe, często są oferowane nagrody pieniężne, rzeczowe lub wsparcie inwestorów w skomercjalizowaniu zwycięskiego prototypu. Udział w hackathonie traktowany jest także jako okazja do rozwijania swoich umiejętności, budowania marki osobistej na rynku pracy i nawiązywania kontaktów (*networking*). Hackathony nie dotyczą wyłącznie DS, przeważnie są to imprezy ogólnie programistyczne. W dziedzinie DS w Polsce organizowane były hackatony m.in. przez platformę Kaggle[99] (Kaggle 2018), przez Główny Urząd Statystyczny (Główny Urząd Statystyczny 2018), Bank Zachodni WBK (Bank Zachodni WBK 2018) czy EMEA Coders League[100] (ChallengeRocket.com 2018).

Nazwa meetup, (odmieniana zgodnie z regułami języka polskiego, np. „meetupy", „meetupowe") pochodzi od nazwy serwisu internetowego meetup.com, i dosłownie oznacza spotkanie. Serwis ten pozwala na tworzenie grup osób o podobnych zainteresowaniach, które spotykają się na żywo. W Polsce meetupem nazywa się taką grupę, np. „słyszałem, że powstał nowy meetup Sztuczna Inteligencja Łódź", mówi się tak również o spotkaniu jako pojedynczym wydarzeniu, np. „idę dzisiaj na meetup". Zdanie „w Krakowie jest ostatnio dużo meetupów" oznacza jednak dużo spotkań, bowiem podstawową działalnością grup jest spotykanie się. Z naszych obserwacji wynika, że w badanym świecie za bardzo aktywny uważa się meetup odbywający się raz w miesiącu. Częstszych spotkań nie zaobserwowaliśmy. Są meetupy spotykające się raz na 2-3 miesiące, zdarzają się kilkumiesięczne przerwy w spotkaniach. Każdy meetup ma co najmniej jednego organizatora. Są to osoby zaangażowane komercyjnie lub naukowo w DS, które organizują spotkania oddolnie, zdecydowanie rzadziej na zlecenie. Serwis meetup.com w Polsce pobiera od każdej grupy miesięczne opłaty w wysokości 14,99 USD lub 9,99 USD, gdy opłata wykonywana jest za pół roku z góry[101].

99 Kaggle jest popularną, międzynarodową platformą internetową oferującą konkursy polegające przeważnie na uzyskaniu modelu uczenia maszynowego jak najlepiej dopasowanego do zadanego zbioru danych.

100 Jest to seria międzynarodowych konkursów technologicznych.

101 Istnieją także plany Meetup Pro, które mogą mieć inne opłaty.

Opłaty dokonuje prywatnie organizator/organizatorzy, część meetupów jest sponsorowana lub organizowana przez firmy, część jest dofinansowana przez większe organizacje. Te organizacje to:

- PyData: międzynarodowy program edukacyjny dotyczący pracy z danymi, założony przez amerykańską organizację non-profit NumFOCUS (por. Travis Oliphant i środowisko Anaconda, część 2.1.). Określa się on jako forum wymiany pomysłów i wzajemnej nauki dla międzynarodowej społeczności twórców i użytkowników narzędzi do analizy danych. Społeczności PyData stosują w DS języki Python, Julia i R, lecz nie są tylko do tych ograniczone. Organizacja obejmuje łącznie 140 meetupów w 54 krajach z sumą ponad 126 tys. uczestników. W Polsce istnieją grupy w Łodzi, Warszawie, Gdańsku i Wrocławiu. Adres https://www.meetup.com/pl-PL/pro/pydata/, dostęp 03.06.2019 r.
- R User Groups: międzynarodowa społeczność użytkowników języka R sponsorowana przez R Consortium, wspierające i współpracujące z R Fundation. R Fundation to organizacja non-profit dbająca o prace nad rozwojem języka R, tzw. R Project. Organizacja obejmuje łącznie 72 meetupy w 32 krajach z sumą ponad 46 tys. uczestników. W Polsce istnieje grupa „Spotkania Entuzjastów R-Warsaw RUG Meetup & R-Ladies Warsaw". Adres https://www.meetup.com/pl-PL/pro/r-user-groups/, dostęp 03.06.2019 r.
- Data Community: polska społeczność IT działająca w pięciu miastach, łącznie prawie 3,5 członków. Jak sami określają, uczestnicy dzielą się wiedzą w zakresie przechowywania, transformowania i przetwarzania danych poprzez bazy danych, usługi z nimi związane oraz aplikacje z nimi współpracujące. Społeczność organizuje lub współorganizuje wydarzenia związane z językiem SQL. Wcześniej pod nazwą Polish SQL Server User Group, zorientowana na rozwiązania firmy Microsoft, ale w żaden sposób z tą firmą nie związana. Tylko 3 z 5 grup mają jako temat spotkań wpisaną nazwę „data science". Adres https://www.meetup.com/pro/data-community/, dostęp 03.06.2019 r.
- About the Talend User Groups: międzynarodowa społeczność użytkowników rozwiązań do integracji danych firmy Talend. Obejmuje 41 grup w 19 krajach, łącznie nieco ponad 6 tys. uczestników. Adres https://www.meetup.com/pro/talend-user-group/, dostęp 03.06.2019 r.

- GDG: międzynarodowa społeczność Google Developer Groups, jak sami piszą, dotycząca wszystkiego od systemu operacyjnego Android, przez Chrome, Drive, platformę Google Cloud, do API produktów jak Cast API, Maps API, YouTube API. 993 grupy w 139 krajach przy sumie ponad 667 tys. uczestników. Adres https://www.meetup.com/pro/gdg/, dostęp 03.06.2019 r.
- Qlik Global Meetups: międzynarodowa społeczność użytkowników oprogramowania Qlick Sense, platformy analitycznej tzw. Business Intelligence. Obejmuje 50 grup w 23 krajach, razem około 6,3 tys. członków. Adres https://www.meetup.com/pro/qlik/, dostęp 03.06.2019 r.

Z naszych badań wynika, że w badanym świecie jednoznacznie z DS kojarzona jest tylko PyData. R User Groups to społeczność nastawiona na język R, także kojarzona z DS, ale również z zastosowaniami R w wąskich dziedzinach badań naukowych, jak hydrologia lub badania genetyczne. Data Community i Talend, choć różniące się znacząco, dotyczą narzędzi i problemów składowania danych, a nie analizy czy modelowania, zatem zagadnień z obszaru inżynierii danych (*data engineering*), przecinającego się ale nie tożsamego z DS. GDG to szeroka społeczność programistyczna. Qlick jest komercyjnym oprogramowaniem analitycznym, z którego korzysta się przede wszystkim za pomocą interfejsu graficznego (*graphic user interface*, GUI), czyli klikając na wybrane elementy. Tym samym nie jest to narzędzie DS, ponieważ przede wszystkim za takie uznawane są te wolne/otwarte (*open source*) i obsługiwane za pomocą interfejsu tekstowego – pisania kodu. Z 86 polskich meetupów, gdzie zapisano temat „data science", do sześciu omówionych organizacji należy 11 grup[102]. Meetupy są postrzegane jako okazja do nawiązywania kontaktów, co nazywane jest – nie tylko w DS – networkingiem, nauki od praktyków, wymiany doświadczeń, w przypadku osób występujących do budowania marki osobistej, w przypadku wspierających grupy firm/uczelni/organizacji do działań marketingowych i rekrutacji.

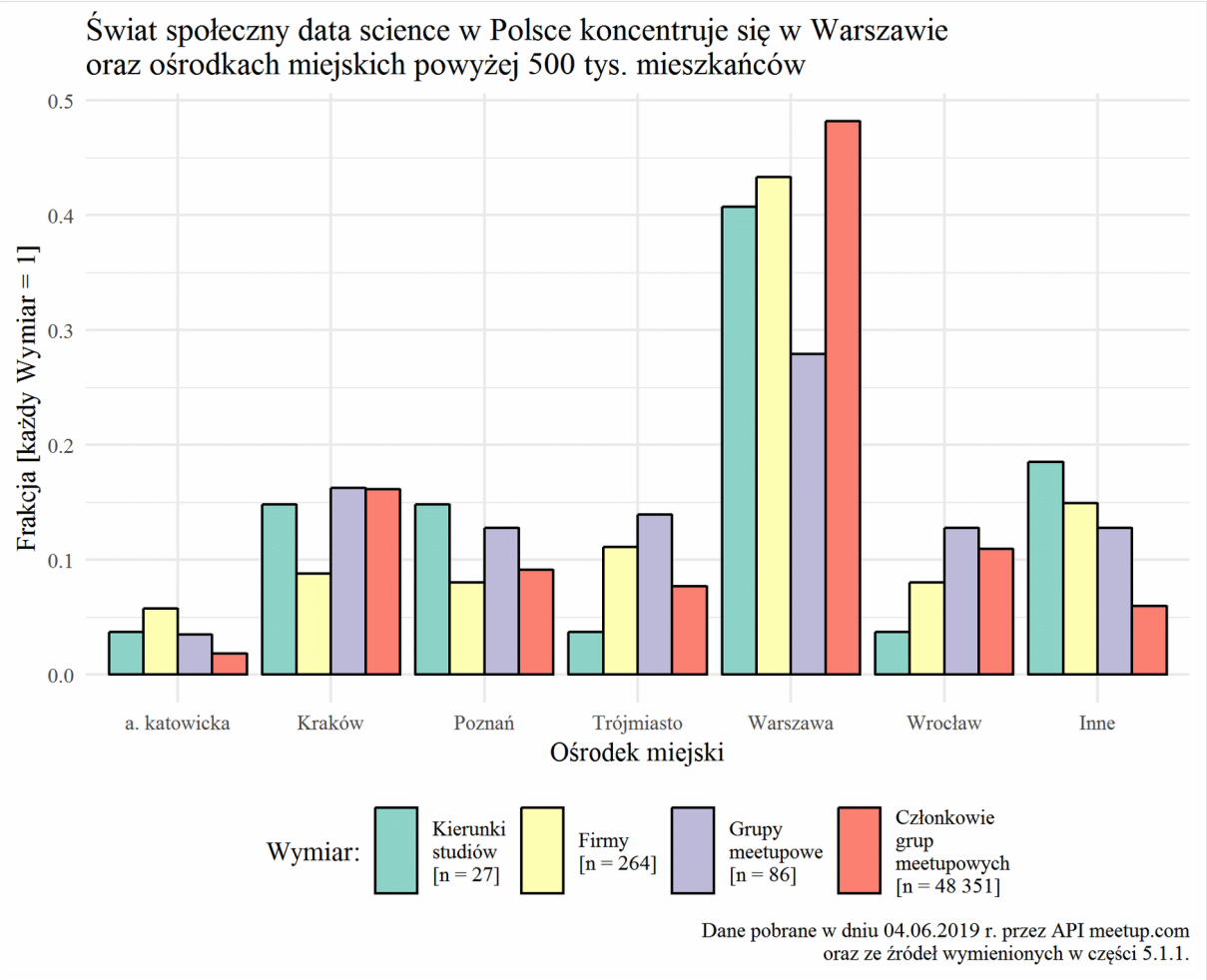
> Tradycyjne podejście do rozpoczęcia kariery jako data scientist może być często niewystarczające. Niezwykle **ważne okazuje się nawiązywanie nieformalnych kontaktów – zarówno online jak i offline – z osobami o podobnych zainteresowaniach, które już pracują lub planują rozpoczęcie kariery w obszarze data science. Najlepszym sposobem jest udział w meetupach, warsztatach i seminariach czy hackatonach** [podkr. RŻ] – mówi dr Dominik Batorski socjolog z interdyscyplinarnego centrum Modelowania Matematycznego i komputerowego na Uniwersytecie Warszawskim, współzałożyciel firmy Sotrender, organizator

102 O ile nie wskazano inaczej, wszelkie wyniki ilościowe dotyczące polskich meetupów data science przedstawiamy na podstawie analiz własnych. Sposób pozyskania danych źródłowych z API meetup.com omówiliśmy w części 4.2. Kod umożliwiający powtórzenie całości analiz wraz ze wszystkimi plikami wejściowymi i wyjściowymi na dostępny jest publicznie na koncie GitHub Remigiusza Żulickiego, https://github.com/zremek/meetup-harvesting.

> data science Warsaw Meetup. (...) Data Science Warsaw to warszawska społeczność skupiająca data scientist. to organizacja non-profit działająca na rzecz upowszechniania informacji związanych z data science. Uczestnicy spotkań rozmawiają o narzędziach, metodach i technologiach stosowanych do pozyskiwania, przekształcania, badania, analizowania i wizualizacji danych. Dyskutują o możliwościach prognozowania, rozwoju produktów związanych z danymi i perspektywach ich biznesowego wykorzystania. (…) Dlaczego warto brać udział w tych spotkaniach? Trzeba mieć na uwadze fakt, że to, co dzieje się na uczelniach, ma ograniczony charakter. Uczelnie nie dysponują dużymi zbiorami danych. (...) Czym wyróżniają się meetupy? **Meetupy to chyba najlepszy sposób, żeby spotkać bezpośrednio i posłuchać praktyków, zobaczyć, jak robi się konkretne rzeczy. To także doskonała okazja, żeby wymienić się doświadczeniami.** [podkr. RŻ] Ma to tym większe znaczenie, że na kursach online, ale także na uczelniach wyższych wymiar praktyczny nauczania jest ograniczony. Zresztą, niecała wiedza zdobywana na zajęciach akademickich czy szkoleniach później się przydaje (Gontarz 2019:18–19).

Kontynuując wątek geograficznej koncentracji badanego świata społecznego w największych ośrodkach miejskich przy dominacji Warszawy, prezentujemy omawiane informacje dotyczące firm, studiów i meetupów. Zauważmy, że we wszystkich analizowanych wymiarach w Warszawie zlokalizowana jest co najmniej 1/4 (wymiar: ilość grup meetupowych) działalności w Polsce. W innych wymiarach (ilość członków grup meetupowych, ilość kierunków studiów, ilość firm) w Warszawie zlokalizowane jest co najmniej 2/5 działalności[103]. Wśród pozostałych ośrodków miejskich nie ma jednoznacznych różnic w udziale w działalności w czterech analizowanych wymiarach, co przedstawiamy na rysunku 4.

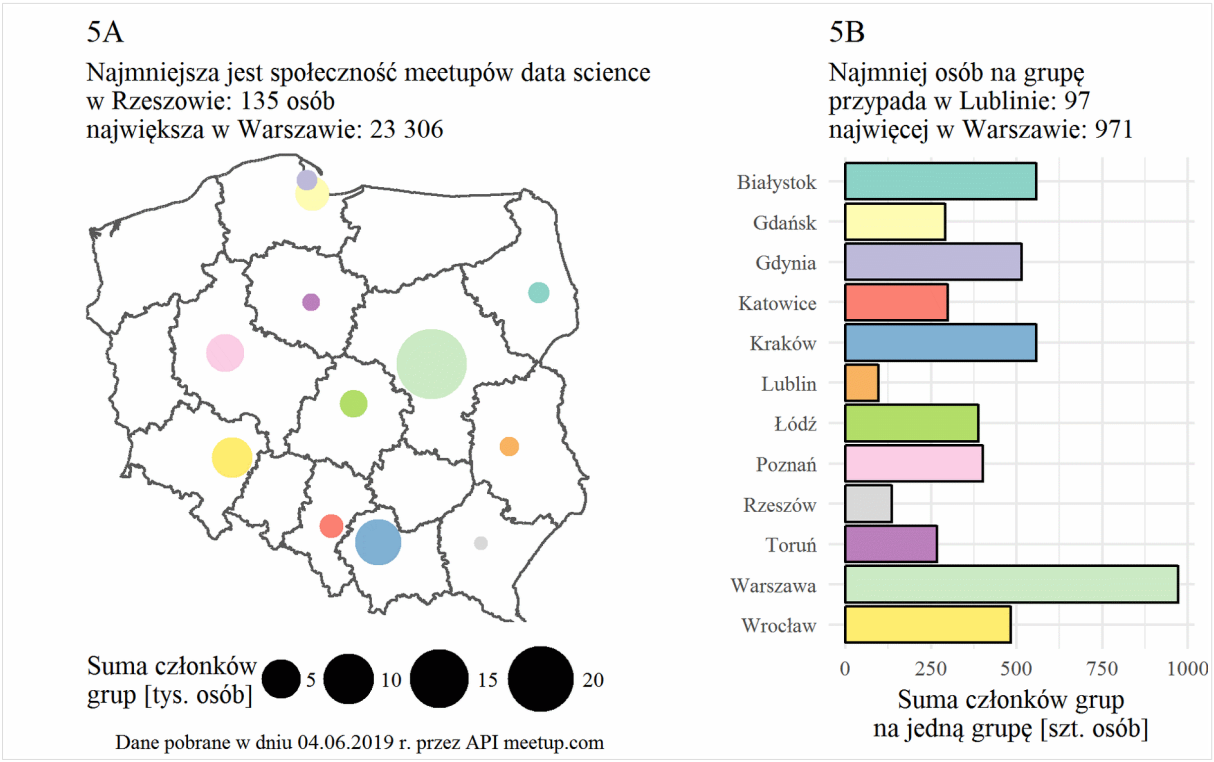
103 W przypadku firm i kierunków studiów nie porównujemy ilości osób (pracowników, studentów), a tylko podmioty. W przypadku firm Fundacja digitalpoland zbierała informacje o wielkości firmy, ale w raporcie „Map of the Polish AI" nie pokazano tej zmiennej w podziale na ośrodki miejskie. Fundacja odmówiła nam dostępu do kompletu danych źródłowych powołując się na RODO i zgody udzielone przez podmioty uczestniczące w sondażu Fundacji. Na rysunku 4. podajemy liczbę 264 firm, zaznaczając, że jest to jedyna znana ilość firm AI w Polsce ogółem, wynikająca z materiałów dołączonych do raportu Fundacji. Faktycznie w sondażu udział wzięło „ponad 160 firm" (por. Borowiecki i Mieczkowski 2019:10). W przypadku kierunków studiów odstąpiliśmy od pracochłonnego, a mającego znikomą wartość dla celów pracy, zadania wystosowywania próśb o podanie liczby studentów do każdej z uczelni lub odpowiedniej jej jednostki administracyjnej.



Rysunek 4: Lokalizacja działalności świata społecznego data science w Polsce – firmy, studia, meetupy

Źródło: opracowanie własne za pomocą GNU R

Powyżej (rys. 4.) zgrupowano lokalizację studiów i meetupów do sześciu ośrodków miejskich w sposób zaproponowany przez autorów badania firm (Borowiecki i Mieczkowski 2019), by umożliwić to zestawienie. Dysponujemy bardziej szczegółowymi informacjami, szczególnie jeżeli chodzi o meetupy. Na rysunku 5. prezentujemy sumy ilości członków grup meetupowych we wszystkich polskich miastach, w jakich grupy istnieją (część 5A), oraz liczbę członków na jedną grupę, także w podziale na miasta (5B). Najmniejsza jest społeczność meetupowa w Rzeszowie, przy 135 osobach w jednej grupie, największa zaś w Warszawie – 23 306 członków w 24 grupach (por. 5A). W pięciu województwach: zachodniopomorskim, warmińsko-mazurskim, lubuskim, opolskim i świętokrzyskim nie ma żadnej grupy meetupowej o temacie „data science". Najmniej osób na jedną grupę przypada w Lublinie (około 97 osób, przy sumie 387 członków i czterech grupach w tym mieście). Najwięcej – dziesięciokrotnie więcej: 971 osób na grupę – przypada w Warszawie (por. 5B).

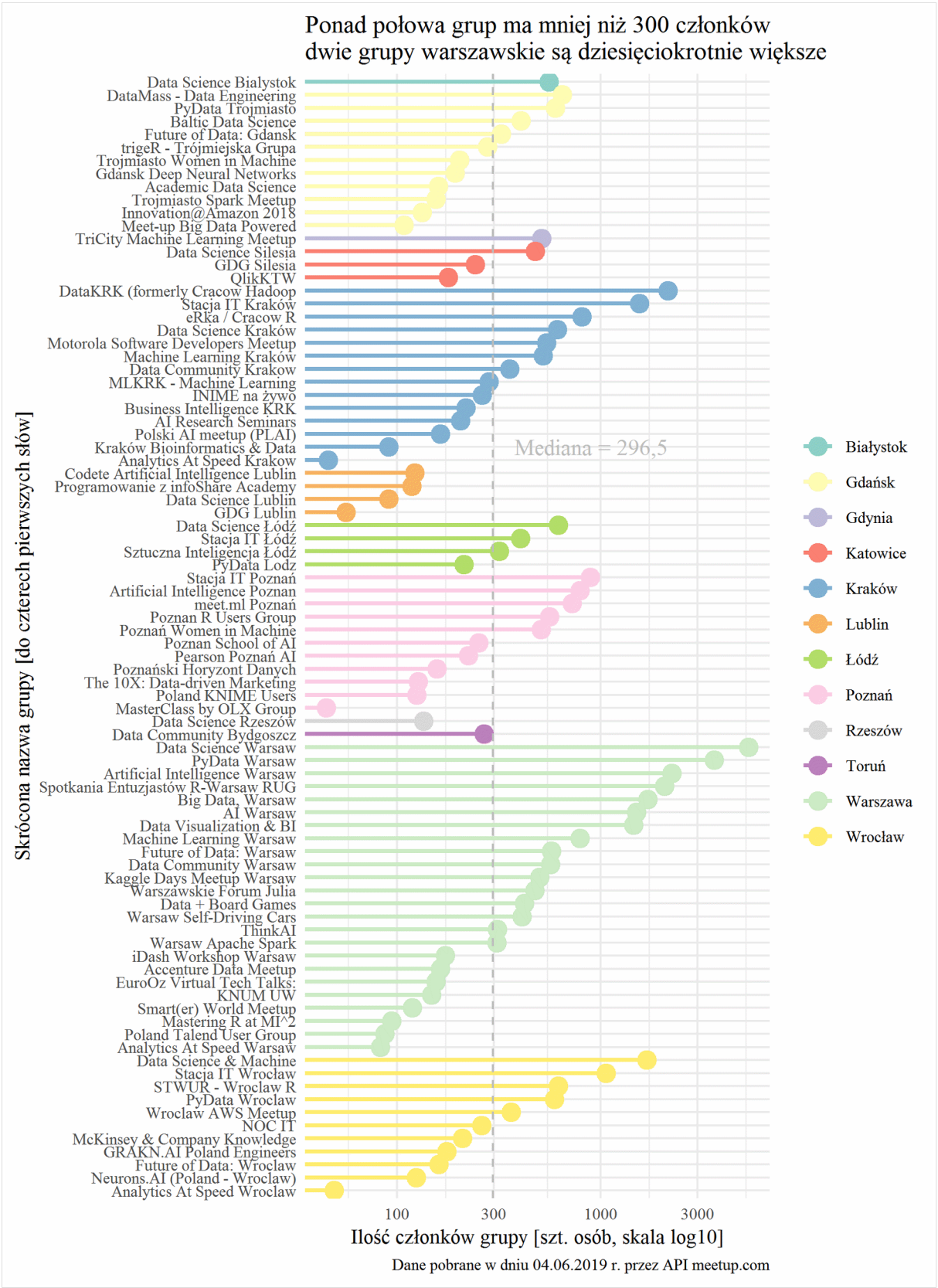


Rysunek 5: Suma członków grup meetupowych data science i ilość członków na grupę w podziale na dwanaście polskich miast

Źródło: opracowanie własne za pomocą GNU R

Liczebność grup meetupowych jest zróżnicowana w kraju i w obrębie miast. Ogółem minimalna liczebność to 45 osób, maksymalna zaś 5 339. Średnia liczebność wynosi około 562 osoby, mediana jest równa 296,5. Oznacza to, że liczebność grup ma rozkład niesymetryczny, prawoskośny – wiele jest grup nielicznych, a mało bardzo dużych. Prawoskośność tego rozkładu wskazuje wysoka, dodatnia wartość współczynnika skośności A = 3,6. Podobny rozkład występuje w miastach z największą liczbą grup[104]. Liczebność wszystkich 86 analizowanych grup prezentujemy poniżej (rys. 6.). Suma członków grup w mieście jest silnie, dodatnio skorelowana z wielkością populacji miasta, współczynnik korelacji rang Spearmana r_s = 0,91. Przyjęliśmy stan populacji na 31.12.2018 r. (System Wspomagania Analiz i Decyzji GUS 2019).

104 Warszawa: 24 grupy, A = 2,1; Kraków 14 grup, A = 1,7; po 11 grup: Gdańsk, A = 0,95; Poznań, A = 0,4; Wrocław A = 1,5.



Rysunek 6: Ilość członków 86 grup meetupowych w podziale na dwanaście polskich miast

Źródło: opracowanie własne za pomocą GNU R

Prezentowane dane o grupach meetupowych mają cztery cechy, które trzeba wziąć pod uwagę interpretując wyniki analizy. Po pierwsze, liczba członków grup jest zdecydowanie niższa niż liczba osób, biorących udział w spotkaniach[105]. Po drugie, te same osoby są zapisane do wielu grup o tej samej lub zbliżonej tematyce[106]. Liczbę członków należy traktować jako wyraz zainteresowania grupą, nie można wnioskować na jej podstawie o wielkości świata społecznego. Po trzecie, fakt, że meetup istnieje, nie oznacza, że jest aktywny, tzn. organizuje spotkania. Temu zagadnieniu przyjrzymy się bliżej omawiając ramy czasowe i zmiany w czasie badanego świata społecznego. Po czwarte, wybraliśmy prostą technicznie metodę wyszukania wszystkich meetupów z zapisanym tematem „data science". Nie wszystkie te meetupy są postrzegane jako część świata społecznego DS w Polsce, co sygnalizowaliśmy wcześniej, mówiąc o sześciu organizacjach je dofinansowujących.

5.1.2. Szkic demograficzny

Demografię badanego świata próbujemy przedstawić w wymiarach struktury płci, wieku i wykształcenia uczestników. **Ogółem występuje zdecydowana przewaga mężczyzn[107], dominacja osób w wieku 25-34 lat[108], posiadających wykształcenie magisterskie[109] i wykształcenie w kierunkach informatycznych[110]**. Jest tak również dla skrzyżowania tych cech. Mężczyźni w wieku 25-34 lat z wykształceniem magisterskim w kierunku informatycznym stanowią nieco ponad 13% respondentów[111] i są najczęściej występującą grupą[112].

---

105 Wśród analizowanych 86 grup, w dniu pobrania przez nas danych 17 miało zaplanowane kolejne spotkanie. Maksymalna ilość kliknięć „RSVP", oznaczających potwierdzenie uczestnictwa w spotkaniu przez członków, wynosi 45% ilości wszystkich członków grupy (meetup Pearson Poznań AI & Tech, 224 uczestników, 101 potwierdzeń). W przypadku najliczniejszych grup odsetek potwierdzeń jest znacznie niższy, np. dla PyData Warsaw – 7% (261 potwierdzeń).

106 Piszący te słowa (RŻ) jest członkiem ośmiu z analizowanych grup meetupowych.

107 Dla każdego z analizowanych źródeł danych mężczyźni stanowią co najmniej 65% wszystkich odpowiedzi, licząc całość łącznie z brakami/odmowami odpowiedzi na pytanie o płeć (por. rys. 7.).

108 Grupa wiekowa 25-34 lata to co najmniej 25% respondentów dla źródeł zawierających informację o tej zmiennej, j.w. (por. rys. 8.).

109 Wykształcenie magisterskie deklarowało co najmniej 40% badanych, j.w. (por. rys. 9.)

110 Ukończenie kierunku informatycznego wskazywało co najmniej 28% uczestników badań, j.w. (por. rys. 10.)

111 Przy połączeniu źródeł danych ankietowych Kaggle (2017 i 2018) oraz Stack Overflow (2018 i 2019) oraz usunięciu braków i odmów odpowiedzi, n = 493 osoby. Wymienione źródła danych omawiamy dalej.

112 Druga najczęściej występująca grupą różni się od pierwszej tylko kierunkiem wykształcenia – matematycznym lub statystycznym i stanowi niecałe 8% z 493 respondentów zdefiniowanych j.w.. Trzecia to mężczyźni w wieku 18-24 lata z wykształceniem informatycznym na poziomie licencjata lub inżyniera (około 7%), czwarta to mężczyźni w wieku 35-44 lata z wykształceniem informatyczny na poziomie magistra (około 4,5%).

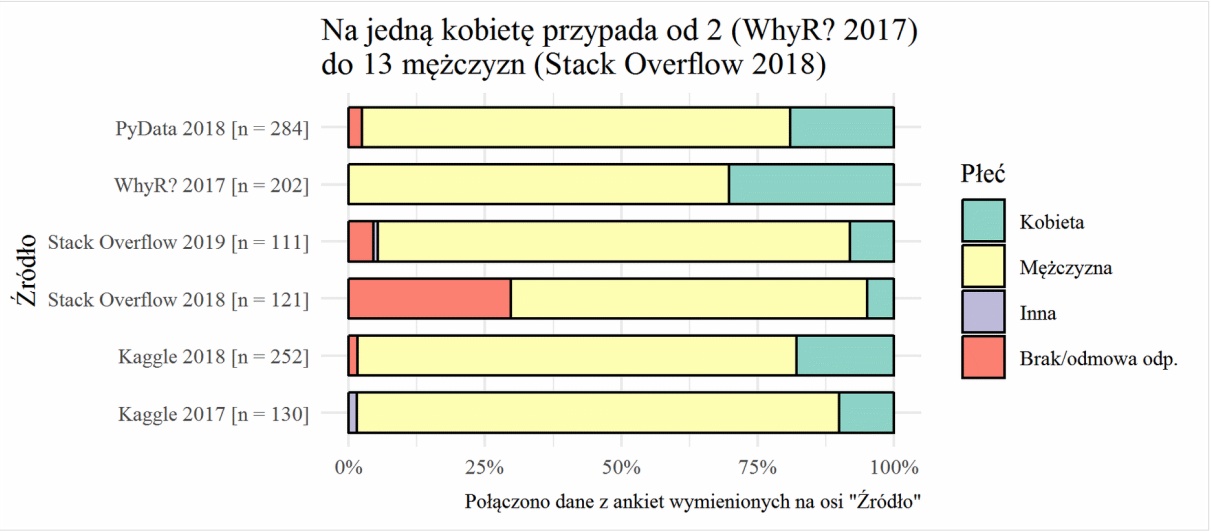
Korzystamy przede wszystkim z analiz własnych. Wykonaliśmy je na publicznie dostępnych danych, pochodzących z międzynarodowych ankiet przeprowadzonych wśród użytkowników Stack Overflow (w 2018 i 2019 r.) i Kaggle (w roku 2017 i 2018) oraz z formularzy rejestracyjnych odbywających się w Polsce konferencji WhyR? w 2017 i PyData w 2018. Charakterystykę podmiotów realizujących te ankiety, zbiorów danych i nasze decyzje dotyczące wyboru i zawężenia tych zbiorów do dalszych analiz opisaliśmy w części 4.2.2.

Choć analizujemy tylko odpowiedzi respondentów, którzy mieszkali w Polsce, to żaden z tych zbiorów nie może być traktowany jako próba reprezentatywna, odzwierciedlająca populację osób zajmujących się DS w Polsce. Pomijamy kwestię rozmytych granic świata społecznego DS, ponieważ pojęcia „populacji" i „społecznych światów" pochodzą z różnych paradygmatów socjologii. W przypadku czterech ankiet międzynarodowych są to próby złożone z ochotników korzystających z serwisów internetowych podmiotów realizujących badania. W przypadku dwóch ankiet polskich są to próby złożone z osób, które zarejestrowały się na ww. konferencje. Znając te cechy uważamy, że analiza danych z ww. źródeł pozwala na wgląd w demografię badanego świata. W każdym z analizowanych wymiarów wyniki z różnych źródeł są niesprzeczne, choć widać specyfikę każdego z nich.

W świecie społecznym DS zdecydowanie najwięcej jest mężczyzn. Spośród analizowanych źródeł najniższą proporcję około dwóch mężczyzn na jedną kobietę ma konferencja WhyR? 2017. Niemniej oznacza to, że mężczyźni stanowili prawie 70% ankietowanych przy nieco ponad 30% kobiet (organizatorzy nie uwzględnili w ankiecie innych kategorii płci ani możliwości ominięcia lub odmowy odpowiedzi). Najwyższa proporcja 13/1 (około 65% mężczyzn do 5% kobiet) wystąpiła w ankiecie Stack Overflow 2018. Zauważmy prawie 30% braków/odmów odpowiedzi w tym źródle. Kategoria płci „Inna" została uwzględniona we wszystkich analizowanych ankietach międzynarodowych[113], zaś w polskich – nie. W formularzu PyData 2018 istniała możliwość ominięcia odpowiedzi na pytanie o płeć. Odsetek wskazań „Inne" nie przekroczył 2% dla żadnego źródła. W ankiecie Stack Overflow 2019 taką kategorię wskazała jedna osoba, w Kaggle 2017 dwie (por. rys. 7.). Zauważmy, że odsetek kobiet jest wyższy dla późniejszego względem wcześniejszego badania międzynarodowego, w ramach tego samego realizatora[114].

113 Przykładowo w badaniu Kaggle 2017 kategorię nazwano „[osoba] niebinarna, genderqueer lub gender-niekonformistyczna" *(Non-binary, genderqueer, or gender non-conforming).*

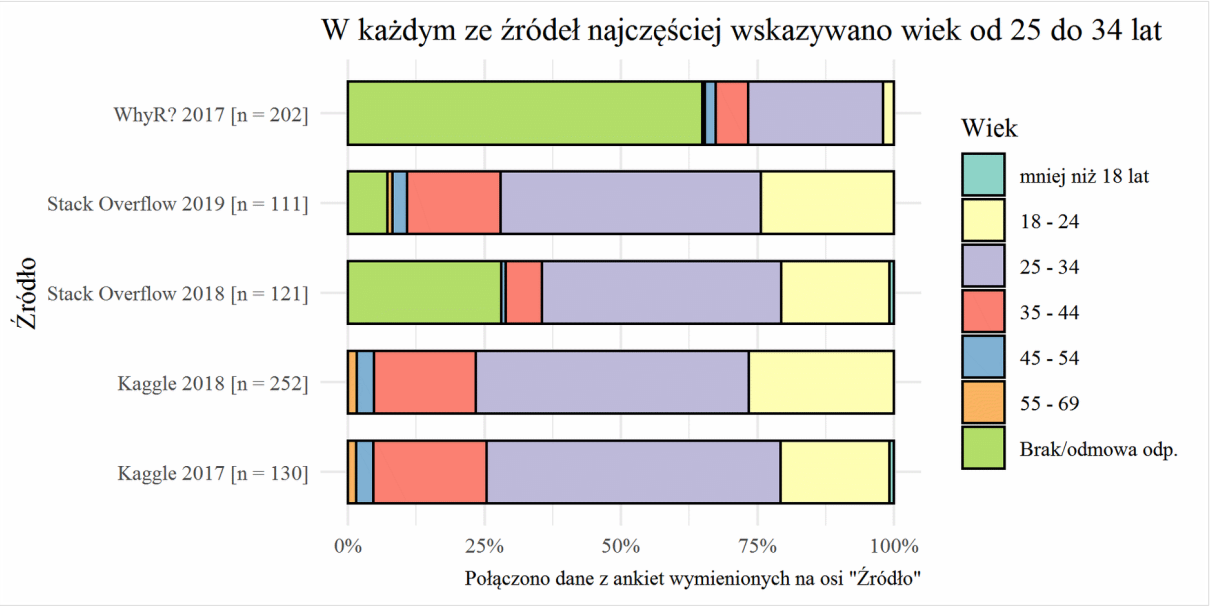
114 Dla Kaggle 2018 jest to wzrost o około 8 pp. względem 2017 (odpowiednio 18% i 10%), dla Stack Overflow 2019 o 3 pp. względem roku 2018 (odpowiednio 8% i 5%).



Rysunek 7: Rozkład płci uczestników według ankiet związanych z polskim światem społecznym data science

Źródło: opracowanie własne za pomocą GNU R

W każdym ze źródeł najczęściej wskazywano wiek od 25 do 34 lat (rys. 8.). Nie możemy podać bardziej szczegółowych informacji, ponieważ takie, obejmujące dziewięć lat przedziały wiekowe przyjęto w ankiecie Stack Overflow 2018. Węższe przedziały z ankiety Kaggle 2018 i wartości liczbowe z pozostałych – poza PyData – ankiet pogrupowaliśmy, by umożliwić zestawienie. W ankiecie PyData nie zbierano informacji o wieku. Wiek 25 – 34 lat podała około ½ respondentów Kaggle 2017 i 2018 oraz Stack Overflow 2019. W Stack Overflow było to ponad 2/5 badanych. W ankiecie WhyR? 2017 tę kategorię wskazała ¼ badanych, ale należy podkreślić prawie 65% braków/odmów odpowiedzi na pytanie o wiek. Dla czterech źródeł drugą najczęściej wskazywaną kategorią był przedział wiekowy 18 – 24 lata. W Kaggle 2018 i Stack Overflow 2019 wybrało ją około ¼ respondentów, dla Kaggle 2017 i Stack Overflow 2018 około 1/5. Dla WhyR? drugim najczęściej wskazywanym był przedział od 35 do 44 lat (około 6% odpowiedzi), zaś 18 – 24 był trzecim pod względem częstości przedziałem (2% wskazań). Wiek poniżej 18 lat deklarowano jedynie w ankietach Kaggle 2017 i Stack Overflow 2018. Górny przedział wiekowy 55 – 69 jest domknięty, ponieważ w żadnym ze źródeł respondenci nie wskazali wieku wyższego niż 69 lat. Osoby w tym górnym przedziale stanowiły maksymalnie niewiele ponad 1,5% (obie ankiety Kaggle). W ankiecie Stack Overflow nie było osób w tym przedziale wiekowym (por. rys. 8.).



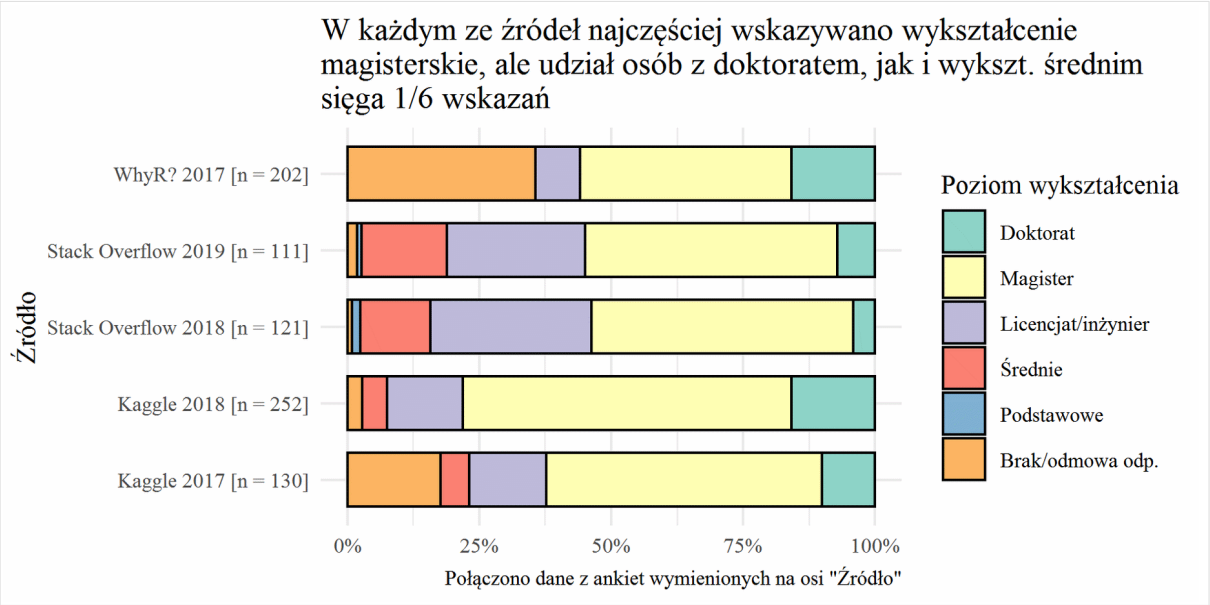
Rysunek 8: Rozkład wieku uczestników według ankiet związanych z polskim światem społecznym data science

Źródło: opracowanie własne za pomocą GNU R

W każdym ze źródeł najczęściej podawano wykształcenie magisterskie. Najrzadziej tę kategorię wskazywano w WhyR? 2017 (około 2/5 odpowiedzi), ale zwracamy uwagę na najwyższy udział braków/odmów odpowiedzi dla tego źródła, który tłumaczymy specyfiką zadanego w ankiecie pytania[115]. Osoby z wykształceniem magisterskim stanowiły około 62% respondentów Kaggle 2018 i około ½ w innych źródłach. Udział osób z wykształceniem pierwszego stopnia (licencjat lub inżynier) to od około 8% (WhyR? 2017) do 31% (Stack Overflow 2018). Dla każdego źródła jest to druga najczęściej wskazywana kategoria, przy czym dla obu ankiet Stack Overflow udział jest o ponad 10 pp. wyższy niż dla obu ankiet Kaggle. Udział osób posiadających doktorat sięga 1/6 dla WhyR? 2017 i Kaggle 2018. Podobny jest udział osób z wykształceniem średnim dla obu ankiet Stack Overflow. Wykształcenie średnie może oznaczać osobę w trakcie studiów; osobę, która studiowała, ale nie uzyskała dyplomu; osobę, która zakończyła formalną edukację na poziomie szkoły średniej. Zauważmy, że odsetek osób z doktoratem jest – podobnie jak w przypadku odsetka kobiet (por. rys. 7.) – wyższy dla późniejszego względem wcześniejszego badania międzynarodowego, w ramach tego samego realizatora[116].

115 Pytanie, na podstawie którego prezentujemy poziom wykształcenia dla WhyR? 2017 brzmiało „Stopień naukowy do umieszczenia na certyfikacie uczestnictwa w konferencji (o ile dotyczy)". Zatem poziom wykształcenia podały tylko osoby, które chciały otrzymać potwierdzenie uczestnictwa w konferencji.

116 Dla Kaggle 2018 jest to wzrost o około 6 pp. względem 2017 (odpowiednio 16% i 10%), dla Stack Overflow 2019 o 3 pp. względem roku 2018 (odpowiednio 7% i 4%).



Rysunek 9: Poziom wykształcenia uczestników według ankiet związanych z polskim światem społecznym data science

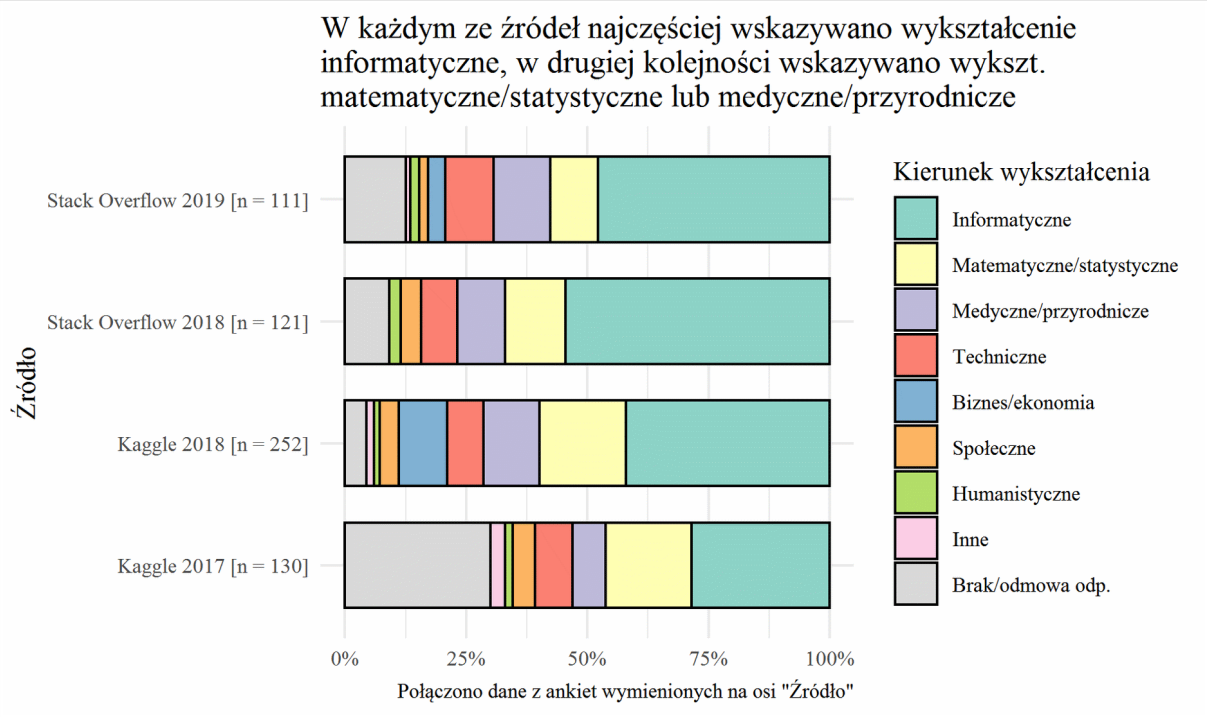
Źródło: opracowanie własne za pomocą GNU R

Ankiety międzynarodowe zawierają także informację o kierunku wykształcenia badanych. Kierunki określano na każdym poziomie wykształcenia, w tym przy odmowie udzielenia odpowiedzi na pytanie o poziom wykształcenia (niemniej dotyczy to łącznie tylko trzech z 614 osób[117]). W każdym ze źródeł międzynarodowych najczęściej wskazywano wykształcenie informatyczne, co prezentujemy na rysunku 10. Najniższy odsetek tego kierunku wykształcenia występuje w Kaggle 2017 (nieco ponad 28%), najwyższy zaś w Stack Overflow 2018 (około 54%). Drugim najczęściej wymienianym było, w przypadku trzech źródeł, wykształcenie matematyczne lub statystyczne (około 12% dla Stack Overflow 2018 i po około 18% dla obu ankiet Kaggle). Drugi najczęściej wskazywany kierunek był inny dla Stack Overflow 2019 – tam wykształcenie medyczne lub przyrodnicze wybierano nieco częściej (około 12%) niż matematyczne/statystyczne (10% odpowiedzi). Nauki medyczne/przyrodnicze wskazywano nieznacznie częściej niż techniczne. Jedynie w przypadku Kaggle pierwszy z wymienionych obszarów wybrało 7%, a drugi niecałe 8% badanych. Autorzy ankiet uwzględnili możliwość podania wykształcenia biznesowego/ekonomicznego[118] w badaniach późniejszych, tzn. 2018 dla Kaggle i 2019 dla Stack Overflow. Ten kierunek wykształcenia wyróżnia się dziesięcioprocentowym udziałem

117 Jednej osoby w Kaggle 2018 i dwóch w Stack Overflow 2019.

118 Jako przykłady wskazano marketing, finanse, rachunkowość, ekonomię.

wskazań w badaniu Kaggle 2018[119]. Wykształcenie społeczne wskazywano rzadziej, było to od 2% odpowiedzi dla Stack Overflow 2019 do około 4% w pozostałych źródłach. Obszar humanistyczny nie przekraczał 2,5% wskazań (Stack Overflow 2018). Odsetek odpowiedzi „Inne" był najwyższy w ankiecie Kaggle 2017 (około 3%), w tym źródle najczęściej odnotowano odmowę/brak odpowiedzi (około 30%).



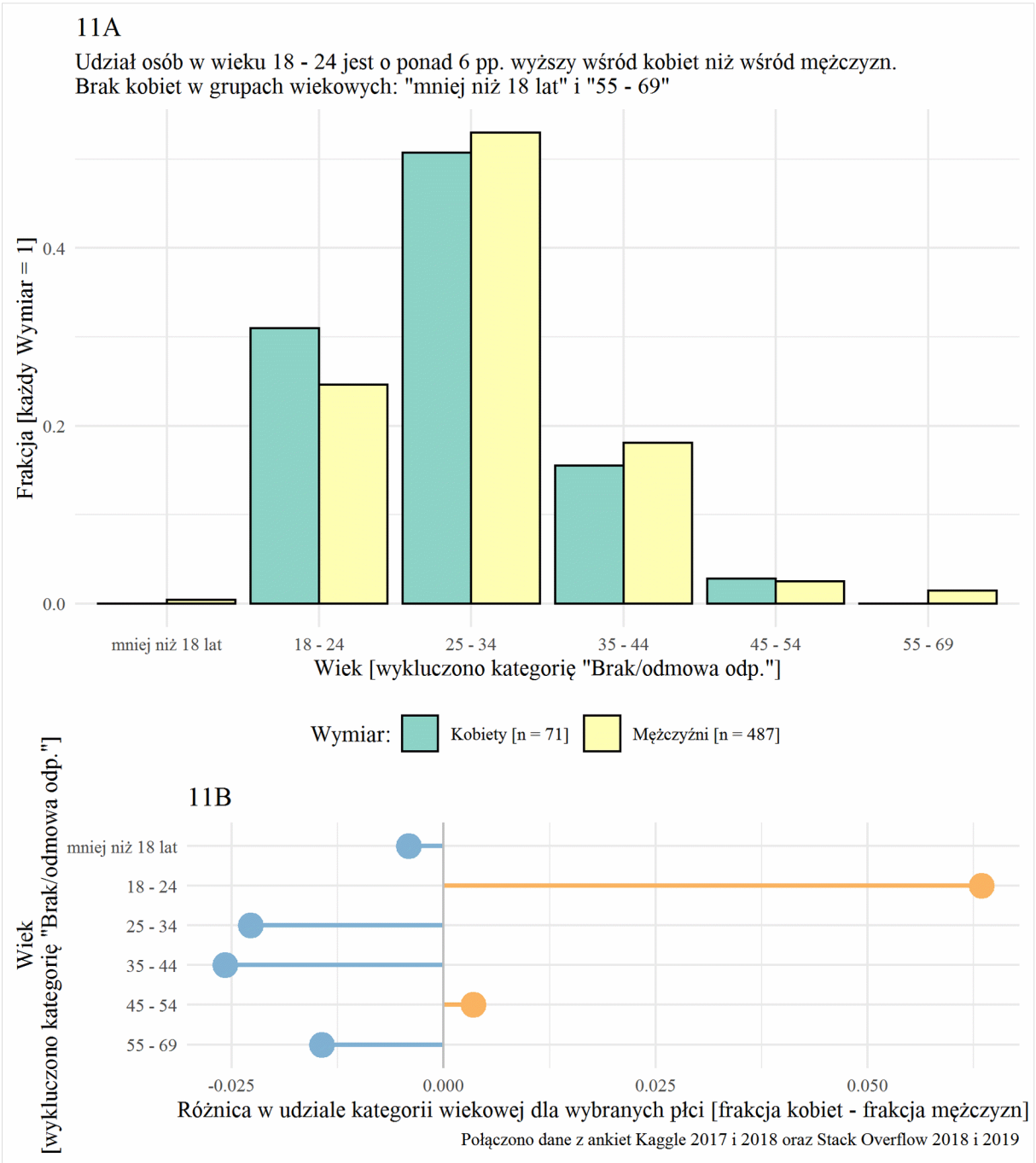
Rysunek 10: Kierunek wykształcenia uczestników według ankiet związanych z polskim światem społecznym data science

Źródło: opracowanie własne za pomocą GNU R

Wśród przekrojów powyżej analizowanych zmiennych demograficznych chcemy przedstawić różnice w wybranych grupach. Na rysunku 11. (11A i 11B) prezentujemy, że po połączeniu czterech analizowanych źródeł danych widoczna jest różnica ponad sześciu punktów procentowych w udziale osób w wieku 18-24 lat dla kobiet i mężczyzn. Na rysunkach prezentujemy nie procenty, a frakcje, ponieważ kobiet było tylko 71, jednak w opisie dla ułatwienia posługujemy się procentami. Łącznie wśród kobiet ta grupa wiekowa stanowi 31% respondentek, zaś wśród mężczyzn jest to niecałe 25%. Różnica wynosi ponad 6 pp., co oznacza , że kobiety wskazywały wiek 18-24 lata około 1,24 razy częściej niż mężczyźni. Nie odnotowano respondentek w wieku niższym niż 18 lat. Udział kobiet był

119 Z naszych badań jakościowych wynika, że w polskim świecie społecznym data science osoby, które ukończyły studia w obszarze ekonometrii lub metod ilościowych w ekonomii mogą identyfikować się zarówno jako ekonomiści, jak i statystycy.

minimalnie (o 0,35 pp.) wyższy w – stosunkowo nielicznym – przedziale 45-54 lat, zaś po około 2,5 pp. niższy w przedziałach 25-34 i 35-44. Podkreślmy, że nie było żadnej respondentki w najwyższej grupie wiekowej 55-69 lat.

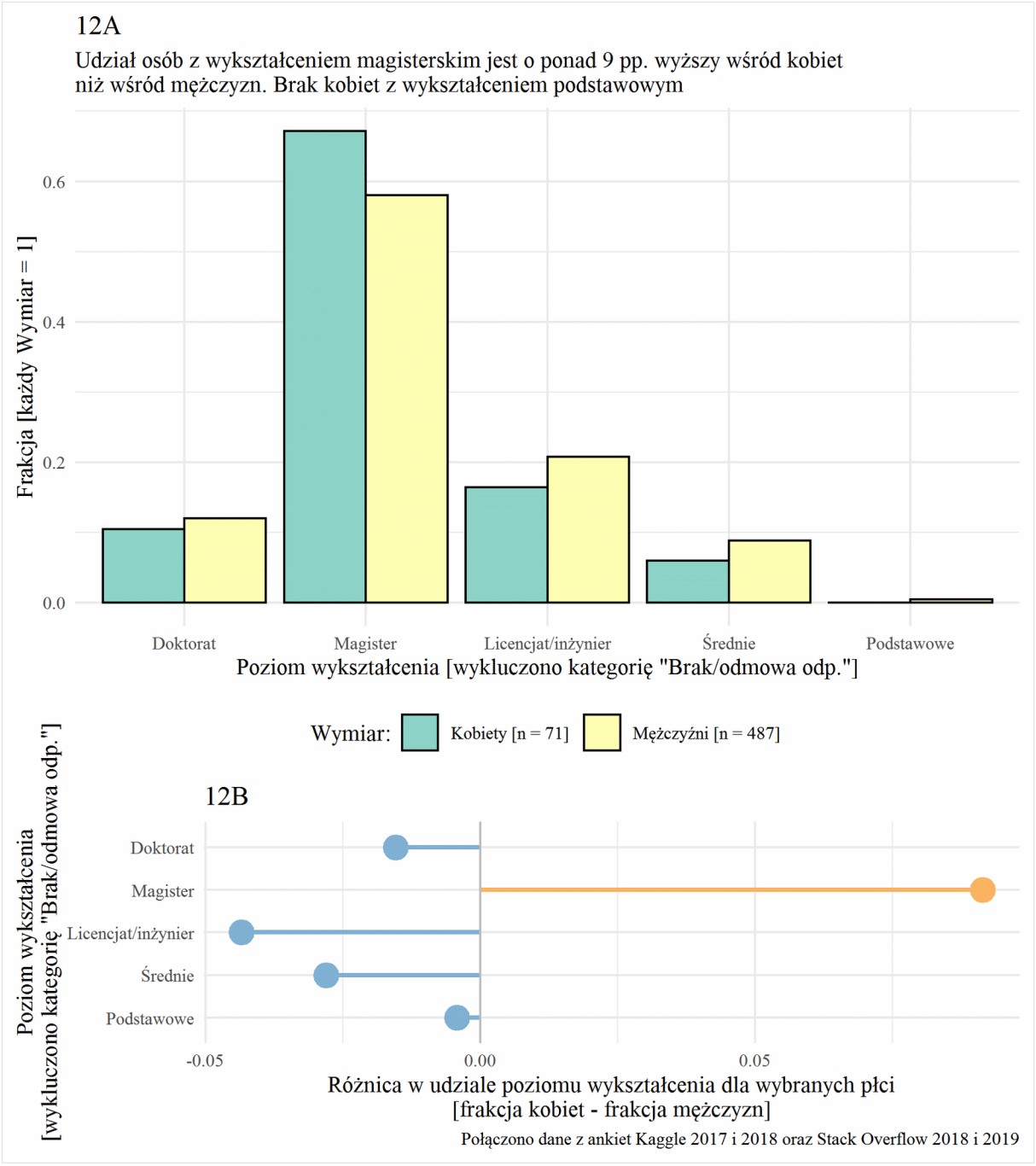


Rysunek 11: Różnica w udziale kategorii wiekowej uczestników dla wybranych płci (mężczyzn i kobiet)

Źródło: opracowanie własne za pomocą GNU R

Na rysunku 12. (12A i 12B) przedstawiamy poziom wykształcenia w podziale na płeć badanych. Odsetek osób z wykształceniem magisterskim był o ponad 9 pp. wyższy w gronie kobiet niż mężczyzn (odpowiednio 67% i 58%), zatem kobiety wskazywały tę kategorię

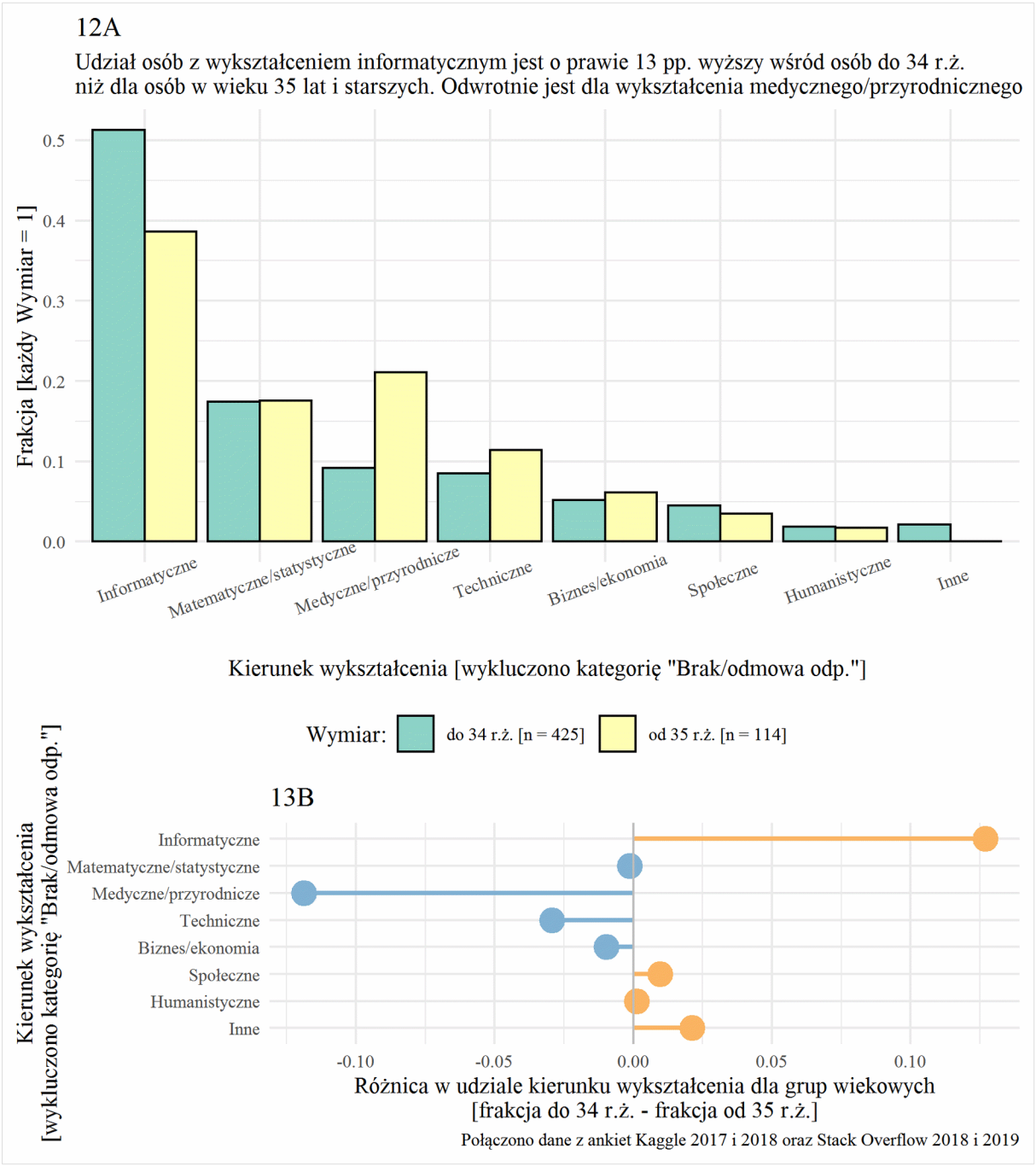
około 1,16 razy częściej. Inne poziomy wykształcenia występowały częściej u mężczyzn niż u kobiet, szczególnie wykształcenie na poziomie licencjata lub inżyniera – tę kategorię wybrało ogółem prawie 21% mężczyzn przy nieco ponad 16% kobiet (różnica ponad 4 pp.; mężczyźni wybierali ją około 1,31 razy częściej niż kobiety). W żadnym źródle danych nie było kobiet deklarujących wykształcenie podstawowe.



Rysunek 12: Różnica w udziale poziomu wykształcenia uczestników dla wybranych płci (mężczyzn i kobiet)

Źródło: opracowanie własne za pomocą GNU R

Na rysunku 13. prezentujemy różnice dotyczące kierunku wykształcenia respondentów w podziale na grupy wiekowe. Zredukowaliśmy kategorie wiekowe do dwóch – do 34 roku życia i od 35 wzwyż. Kierunki informatyczne wskazywano około 1,3 razy częściej w grupie młodszej (około 51% badanych) niż w starszej (niecałe 39%), zatem różnica wynosi prawie 13 pp. Przeciwna i większa różnica wystąpiła we wskazaniach kierunków medycznych lub przyrodniczych. W gronie starszym kategorię tę wybrało 21%, a w młodszym 9%, zatem osoby od 35 r.ż. wskazały wykształcenie medyczne/przyrodnicze ponad dwa razy częściej niż te do 34 lat. Grupa starsza około 1,3 razy częściej niż młodsza wskazała także wykształcenie techniczne. Udział kierunków matematycznych i statystycznych jest zbliżony dla obu grup wiekowych (po około 17,5%). W starszej grupie wiekowej nie było wskazań na kategorię „Inne".



Rysunek 13: Różnica w udziale kierunku wykształcenia uczestników dla grup wiekowych (do 34 r.ż. i od 35 r.ż.)

Źródło: opracowanie własne za pomocą GNU R

5.1.3. Szkic ekonomiczny

Przyjrzyjmy się ogólnym informacjom dotyczącym charakterystyki ekonomicznej podmiotów zajmujących się AI, oraz rynkowi pracy dla data scientistów w Polsce. W pierwszym z wyróżnionych obszarów bazujemy na raporcie Fundacji digitalpoland (Borowiecki i Mieczkowski 2019), skupiając się na ośmiu wnioskach:

1 Około **1/3 badanych podmiotów uzyskuje dochód głównie –** w co najmniej 70% **– ze zleceń zagranicznych** (2019:6, 20).

2 Ponad **20% z badanych firm nie uzyskała dotychczas dochodu z produktów i usług bazujących na AI**. (2019:19).

3 Około **2/3 badanych firm finansuje swoje działanie całkowicie samodzielnie** (2019:15).

4 Ponad **1/3 firm jest przynajmniej częściowo finansowana przez kapitał zagraniczny**. Należą do tej grupy także podmioty wymienione w punkcie 3. jako finansujące swoje działania całkowicie samodzielne (2019:15).

5 Firmy **najczęściej świadczą usługi w zakresie „analityki, big data i business intelligence" (43%)**, **najrzadziej zaś w obszarach „wojsko i obronność" (3%)** (2019:23-24).

6 Zadania, jakie wykonywane są w ramach oferowanych usług to **najczęściej przetwarzanie i rozpoznawanie obrazów (62%)** (2019:27-28).

7 Pomimo tego, że wielkość badanych firm zajmujących się AI jest zróżnicowana, to **zespoły techniczne zazwyczaj są nie większe niż pięć osób (58%)** (2019:32).

8 Około **77% firm współpracuje z uczelniami wyższymi**. **Prawie 2/5 firm opublikowało artykuł w czasopiśmie naukowym** (2019:44-46).

W odniesieniu do wniosku pierwszego, z raportu wynika, że prawie ¾ spośród badanych około 160 firm zajmujących się sztuczną inteligencją w Polsce (por. Borowiecki i Mieczkowski 2019:10, 14) pracuje dla klientów zagranicznych, tzn. realizuje dla nich jakiekolwiek zlecenia. Około 1/3 badanych podmiotów uzyskuje dochód głównie – w co najmniej 70% – ze zleceń zagranicznych. Od 30% do 70% dochodu od klientów zagranicznych wskazało około 13% podmiotów, zaś do 30% dochodu z tego źródła podało 27% firm. Do tej grupy nie należą firmy, uzyskujące dochody wyłącznie od klientów polskich (także 27%). Zdaniem autorów, popyt na rozwiązania bazujące na AI w Polsce jest ograniczony przez niechęć do ryzyka związanego z wdrażaniem nowych i nieprzetestowanych technologii w największych przedsiębiorstwach państwowych oraz przez ogólną liczbę dużych firm w Polsce – nie wyjaśniono dokładnie o co chodzi, ale sądzimy, że jest to sugestia, że niewiele jest w Polsce dużych firm, które mogą być zainteresowane dużymi wdrożeniami AI. Poza tym, budżety dużych firm w Polsce są niższe niż budżety firm

w Europie Zachodniej i innych krajach rozwiniętych. Autorzy wspierają swoje tezy informacją, że 35% badanych firm ma „biuro lub innego rodzaju stałą reprezentację" poza granicami Polski. Ma to oznaczać, że polski rynek nie jest gotowy ani otwarty na używanie AI w biznesie, i stanowi, ich zdaniem, jasny sygnał dla rządu RP i największych spółek państwowych o konieczności stymulacji popytu na rozwiązania bazujące na AI w Polsce. To chyba apel do nich o rozpisywanie przetargów i składanie zamówień. Autorzy wskazali także na opinię Katarzyny Ludki, pracującej jako AI Director w Ringer Alex Springer Polska. Ludka uważa, że prezentowane wyżej wyniki świadczą o relatywnej niedojrzałości polskiego rynku dla produktów i usług AI. Polskie firmy, a właściwie osoby na stanowiskach kierowniczych jakoby nie rozumieją korzyści płynących z wdrażania rozwiązań opartych na AI. W podobnym tonie wypowiadają się inni zaproszeni eksperci. Autorzy podkreślają potrzebę edukowania polskich kierowników w zakresie szans i zagrożeń związanych z wdrażaniem AI, oraz potrzebę ogólnego zwiększania świadomości w zakresie AI. Wskazano, że omawiana nieświadomość oraz niska jakość danych[120] są w opinii badanych firm większymi utrudnieniami we wdrażaniu AI niż ograniczenia kosztowe (Borowiecki i Mieczkowski 2019:6, 15, 20-23, 25).

Co do wniosku drugiego – 1/5 badanych firm nie uzyskała dotychczas dochodu z produktów i usług bazujących na AI – to autorzy wskazują na wczesne stadium rozwoju wielu podmiotów[121]. Około połowa badanych firm zaczęła zajmować się AI nie wcześniej niż w ciągu ostatnich dwóch lat, tzn. przełomu 2016 i 2017 r. Nieco ponad 1/3 podmiotów wskazała, że co najmniej 70% ich dochodów pochodzi z produktów i usług bazujących na AI, i te autorzy określają jako „AI-only companies". Pozostałe nieco ponad 2/5 firm uzyskuje część dochodów z tego źródła. Wskazano, że wiele firm oferuje szersze niż AI portfolio, a część dołączyła AI jako kolejny oferowany produkt/usługę. Podkreślmy, że model biznesowy w niemal połowie badanych podmiotów polega wyłącznie na usługach wykwalifikowanej pracy zleconej przez inne firmy, co określono jako oferowanie „siły umysłowej" (*brain power*). Te firmy nie oferują zatem żadnych produktów, a usługą jest, ogólnie rzecz ujmując, wynajmowanie pracowników (Borowiecki i Mieczkowski 2019:19, 31).

120 W naszych badaniach jakościowych pojawiają się w tym kontekście, jako kody in vivo, pojęcie kultury danych. Dotyczy ona zarówno aspektów technologicznych jak i biznesowych korzystania z danych w organizacji. Zagadnienie rozwijamy w części 6.2.

121 Jak wskażemy w części 6.1.2. powoływanie się na wczesne stadium rozwoju jest jedną z praktyk legitymizacji świata społecznego data science.

W trzecim wniosku podkreśliliśmy, że 2/3 badanych firm finansuje swoje działanie całkowicie samodzielnie. Wśród innych źródeł finansowania w pytaniu wielokrotnego wybory wskazywano na subsydia i granty (34%), tzw. venture capital (VC)[122] (23%), fundusze tzw. aniołów biznesu (10%), inkubatory akademickie i akcje (po 4%) oraz różnego rodzaju kredyty (3%). Wskazano jednak, że 21% firm określających się jako samodzielnie finansujące się deklaruje jednocześnie otrzymanie subsydium lub grantu (Borowiecki i Mieczkowski 2019:14-15). Autorzy i zaproszeni eksperci twierdzą, że badane firmy szukają i korzystają z inwestorów zagranicznych, ponieważ VC w Polsce to raczej inwestycje w mniej ryzykowne niż AI technologie – sprawdzone, adresowane dla konkretnego sektora lub rozwiązujące konkretny problem. Do rozwiązań bazujących na AI ma polskich inwestorów zniechęcać też ryzyko dotyczące długiego czasu oczekiwania na komercjalizację produktu/usługi. Obecnie polskie inwestycje VC w omawianej dziedzinie „zwykle nie przekraczają 3 mln PLN" i jak rozumiemy są uznawane za niewielkie (Borowiecki i Mieczkowski 2019:16). Ekspert Wojciech Walniczek uważa, że zaskakująco duży jest odsetek firm AI deklarujących samofinansowanie się. Jego zdaniem oznacza to, że firmy są zakładane przez osoby, które prowadzą już inny, dochodowy biznes lub są to działy międzynarodowych korporacji o dużym zapleczu finansowym. Walniczek uznaje za niski udział finansowania z VC, aniołów biznesu i innych źródeł. Ma to oznaczać, że polskie firmy AI są niedojrzałe, są we wczesnej, „garażowej" fazie rozwoju (Borowiecki i Mieczkowski 2019:18). Inni eksperci – Piotr Pietrzak z firmy IBM Poland i wielokrotnie wymieniany w naszej pracy socjolog Dominik Batorski – zwracają uwagę na rolę finansowania prac badawczo-rozwojowych dotyczących AI z grantów Narodowego Centrum Badań i Rozwoju (NCBiR). Batorski podkreśla, że nie ma w NCBiR programów przeznaczonych konkretnie dla sztucznej inteligencji czy uczenia maszynowego, ani dla konkretnych dziedzin zastosowania tychże[123]. Dodajmy, że Ministerstwo Nauki i Szkolnictwa Wyższego w dniu 29. maja 2019 ogłosiło nabór do programu doktoratów wdrożeniowych, gdzie część II dotyczy

122 Finansowanie za pomocą venture capital jest odnoszone do małych i średnich innowacyjnych firm niepublicznych we wczesnych fazach ich rozwoju. Podkreślane jest wysokie ryzyko związane z inwestycją i możliwość osiągnięcia wysokiej stopy zwrotu przez inwestora. Inwestor kupuje akcie lub udziały „młodego" przedsiębiorstwa, stając się współwłaścicielem spółki. Okres zaangażowania inwestora jest z góry określony. Zysk inwestora ma polegać na odsprzedaży akcji/udziałów po określonym czasie (Lulek 2013).

123 Z naszej (RŻ) korespondencji z NCBiR wynika, że do czerwca 2019 nic się nie zmieniło. Zapytaliśmy mailowo „(…) czy istnieją lub są planowane programy NCBiR dedykowane technologiom uczenia maszynowego/sztucznej inteligencji/data science?", na co otrzymaliśmy odpowiedź: „(…) przesyłam Panu informacje na temat konkursu Szybka Ścieżka. Jest to horyzontalny konkurs, w którym nie ma wymogów co do profilu badań B+R jeśli chodzi o dziedzinę dlatego jeżeli Państwa pomysł jest innowacją co najmniej w skali kraju, koszt prac B+R wyniesie minimum 1 mln złotych do tego większość tych kosztów będzie ponoszona poza województwem mazowieckim - proszę zapoznać się z poniższą dokumentacją i aplikować (…)".

„wykorzystania sztucznej inteligencji w procesach technologicznych lub społecznych" (Muenchen 2019), rozpoczęło ono także badanie (we współpracy z Fundacją digitalpoland) wśród 50 tys. polskich naukowców na temat rozwoju sztucznej inteligencji w sektorze nauki (Kaczorowska-Spychalska 2019). Pietrzak uważa, że NCBiR jest instytucją wiodącą, jeśli chodzi o pozarynkowe finansowanie polskich firm AI. Twierdzi, że jeżeli „zwiększy się apetyt na ryzyko inwestorów VC", to w roku 2019 będzie widoczny wzrost konkurencji pomiędzy różnego rodzaju inwestorami, i coraz więcej firm AI będzie uzyskiwać dodatkowy kapitał. Batorski zaznacza, że nie ma wątpliwości, iż rozwój produktów/usług opartych na AI opiera się mocno na pracach badawczo-rozwojowych. Dlatego dla rozwoju firm ważne jest zarówno finansowanie pozarynkowe i inwestycje, jak i współpraca firm z uczelniami lub naukowcami (Borowiecki i Mieczkowski 2019:16-17).

Z powyżej omówionym wnioskiem łączy się kolejny. Z badania Fundacji digitalpoland wynika, że 1/3 firm jest przynajmniej częściowo finansowana przez kapitał zagraniczny. Należą do tej grupy także podmioty wymienione wyżej jako finansujące swoje działania całkowicie samodzielne. Autorzy raportu wskazują, że z jednej strony polskie firmy zajmujące się AI są otwarte na kapitał zagraniczny, z drugiej zaś strony międzynarodowe korporacje lokują w Polsce swoje działy AI i centra badawczo-rozwojowe. Twierdzą także, że obcokrajowcy zakładają w Polsce startupy, by oferować usługi bazujące na AI klientom ze swoich rodzinnych krajów, co ma zapewniać tym startupom większy margines finansowy m.in. ze względu na tańszą siłę roboczą w Polsce niż w owych krajach (Borowiecki i Mieczkowski 2019:15).

Firmy zajmujące się AI w Polsce dostarczają usług/produktów dla różnego rodzaju zastosowań. Autorzy badania, w pytaniu wielokrotnego wyboru, wyróżnili tutaj 24 kategorie (Borowiecki i Mieczkowski 2019:23-24):

- analityka, big data i business intelligence (kategoria najczęściej wskazywana – 43% podmiotów realizowało zlecenia w tym zakresie zastosowań);
- sprzedaż, marketing lub reklama (37%);
- technologie dla finansów lub ubezpieczeń (*fintech, insurtech*) (28%);
- internet rzeczy (*internet of things*, IoT) i przemysł 4.0 (27%);

- zrobotyzowana automatyzacja procesów (*robotic process automation*, RPO), usługi dla klientów biznesowych (*business-to-business*, B2B)[124] (25%);
- handel detaliczny, handel elektroniczny (*e-commerce*) (23%);
- obsługa klienta, chatboty (22%);
- technologie dla medycyny (*MedTech*), biotechnologia (21%);
- bezpieczeństwo, cyberbezpieczeństwo, wykrywanie oszustw (16%);
- paliwa lub energia (15%);
- inteligentne miasto, środowisko, usługi publiczne (15%);
- transport, logistyka, motoryzacja (15%);
- media społecznościowe (14%);
- edukacja (13%);
- elektronika lub robotyka (12%);
- telekomunikacja (10%);
- produkty konsumenckie (*consumer goods*) (8%);
- technologie dla zastosowań prawnych (*LegalTech*) (8%);
- inne (8%);
- turystyka, czas wolny (7%);
- zatrudnienie, zasoby ludzkie (*HR Tech*) (7%);
- rozrywka, gry, muzyka, media, sport (6%);
- technologie w rolnictwie i żywności (*FoodTech*) (5%);
- wojsko i obronność (3%).

Autorzy wskazują ponadto, że rozkład wymienionych kategorii zastosowań AI jest różny w zależności od stażu działania firm, oraz klientów dla których głównie pracują. Porównano zatem firmy, które zaczęły zajmować się AI do dwóch lat wstecz – co oznacza 2017 r. – i więcej niż dwa lata wstecz. Przypomnimy, że takie grupy mają równą liczbę badanych

124 Nie jest dla nas jasne, czy to RPO jest oferowane dla klientów biznesowych – prawdopodobnie tak, ale kierując się konwencją z innych kategorii oznaczałoby to RPO lub B2B, co wydaje nam się pozbawione sensu, bo za szerokie i nierozłączne z innymi kategoriami. Z naszych badań wynika, że niemal wszystkie z wymienionych kategorii to produkty/usługi B2B.

podmiotów. W tym przekroju wskazano, że stare firmy częściej niż nowe zajmowały się, jak napisano, „tradycyjnymi" zastosowaniami czyli analityką, big data i BI oraz sprzedażą, marketingiem i reklamą. Wśród firm starych kategorię pierwszą wskazało 50% podmiotów, wśród nowych 36% („stare" prawie 1,4 razy częściej), kategorię drugą odpowiednio 46% i 29% firm („stare" prawie 1,6 razy częściej). W przekroju dotyczącym klientów porównano dwie grupy. W jednej znalazły się firmy uzyskujące dochód wyłącznie od polskich klientów i te, gdzie dochód zagraniczny stanowi do 30% całości. W drugiej zgrupowano firmy z wyższym dochodem zagranicznym. Firmy pracujące dla klientów zagranicznych częściej niż te działające głównie w Polsce potwierdzały zastosowania w handlu detalicznym i elektronicznym (odpowiednio 35% do 25%; 1,4 razy częściej). Przeciwnie było dla zastosowań w dziedzinie paliw i energii – kategorię wskazało 22% firm działające lokalnie przy 12% działających głównie za granicą („lokalne" ponad 1,8 razy częściej). Ekspert Tomasz Wesołowski z firmy Edward.ai twierdzi, że powyższe wyniki oznaczają gotowość działów marketingu i sprzedaży do wprowadzania technologii opartych na AI. W następnej kolejności Wesołowski przewiduje gotowość w dziedzinach obsługi klienta i RPO. W jego wypowiedzi zaakcentowane są dwa, omawiane już wcześniej wątki – niskiej świadomości klientów co do możliwości zastosowania AI oraz niechęci do ponoszenia ryzyka związanego z wdrażaniem technologii słabo sprawdzonych, z niedookreśloną wartością biznesową (Borowiecki i Mieczkowski 2019:25-27).

Zadania, jakie wykonywane są przez algorytmy uczenia maszynowego, w ramach oferowanych produktów/usług podzielono na 15 kategorii w pytaniu wielokrotnego wyboru (Borowiecki i Mieczkowski 2019:27-28):

- przetwarzanie i rozpoznawanie obrazów (kategorię wskazało 62% badanych firm);
- eksploracja danych (55%);
- systemy rekomendacyjne (52%);
- przetwarzanie języka naturalnego (43%);
- systemy eksperckie (30%);
- ocena ryzyka, wykrywanie oszustw (27%);
- optyczne rozpoznawanie znaków (*optical character recognition*, OCR) (25%);
- uczenie maszynowe dla robotów (22%);
- rozpoznawanie mowy (17%);

- przetwarzanie dźwięków (16%);
- tłumaczenie maszynowe (15%);
- logika rozmyta (*fuzzy logic*) (13%);
- inne (12%);
- sztuka generowana za pomocą AI (8%);
- algorytmy grające w gry (6%).

W przekroju na firmy działające w AI od 2017 r. i starsze wskazano, że firmy nowe rzadziej niż stare zajmowały się eksploracją danych (odpowiednio 44% względem 66%, „nowe" około 1,5 razy rzadziej) i oceną ryzyka/wykrywaniem oszustw (21% do 30%, „nowe" ponad 1,5 razy rzadziej). Autorzy tłumaczą to tym, że takie zadania są z powodzeniem rozwiązywane przez – jak kolejny raz napisano – „tradycyjne" techniki uczenia maszynowego. Nowe firmy mają wykazywać „naturalną tendencję" do stosowania nowszych technik, przede wszystkim w postaci głębokich sieci neuronowych (*deep learning*). Deep learning jest stosowany w zadaniach rozpoznawania i przetwarzania obrazów, rozpoznawania mowy, przetwarzania dźwięku i przetwarzania języka naturalnego. W tych obszarach mają „szukać swoich szans" nowe firmy (Borowiecki i Mieczkowski 2019:28).

Siódmy wniosek dotyczy tego, że zespoły techniczne zajmujące się AI zazwyczaj są nie większe niż pięć osób (co deklarowało 58% badanych firm), choć wielkość firm jest zróżnicowana. Łączna ilość pracowników badanych firmy to 1-3 osoby dla 15% podmiotów, 4-10 osób dla ¼, 11-20 oraz 21-50 wskazało po 19% firm i powyżej 50 dla 22%. Przy tym zespoły 1-2 osobowe ma 27% firm, na 3-5 osób w zespole wskazało 31% podmiotów (zatem łącznie 58% firm ma zespół nie większy niż 5 osób, jak powiedziano wyżej). Od 6 do 10 osób w zespole miało 22% badanych podmiotów, 11-20 – 8%, a na zespół powyżej 20 osób wskazywało około 15% firm (Borowiecki i Mieczkowski 2019:31-33). Autorzy uznali, że wielkość zespołów jest związana z zadaniami, które omawialiśmy wyżej. Podzielono firmy na zespoły „większe" (mające sześć i więcej osób w zespole AI) oraz „mniejsze" (te, gdzie zespół stanowi 1-5 osób). Firmy z większymi zespołami częściej niż te z mniejszymi wskazywały na osiem zadań, wśród których ponad dwukrotna różnica częstości wskazań wystąpiła dla uczenia maszynowego dla robotów (34% do 14%; „większe" 2,4 razy częściej) oraz przetwarzania dźwięku (25% do 10%; „większe" 2,5 razy częściej) (Borowiecki i Mieczkowski 2019:29-30).

W ostatnim wyróżnionym przez nas wniosku mowa o współpracy firm i akademii. Około 77% firm współpracuje z uczelniami wyższymi. Odpowiedzi na pytanie wielokrotnego wyboru wskazują, że około 1/3 firm współpracuje z zespołem badaczy akademickich, a po 43% wskazało na zatrudnianie badaczy lub indywidualną współpracę z badaczem. Prawie połowa badanych podmiotów ma w zespole technicznym osobę, która posiada doktorat. Prawie 2/5 firm opublikowało artykuł w czasopiśmie naukowym, a 8% ma ich więcej niż 10. Widoczna jest zdecydowana różnica w skali współpracy firm z akademią w podziale na wielkość zespołu technicznego AI. Wśród firm z większym zespołem (6 i więcej osób) tylko 12% wskazało brak współpracy, wśród tych z zespołem 1-5 osobowym było to 37% („większe" współpracowały z akademią ponad trzy razy częściej). Ekspert Michał Staśkiewicz, CEO w firmie Alphamoon.ai, stwierdził, że „w pewnym sensie nigdy wcześniej nie istniała dziedzina, która w takim stopniu [jak AI] łączy biznes i naukę" (Borowiecki i Mieczkowski 2019:44-46). To samo powiedział Batorski: „obecnie nie istnieje żadna inna branża tak bardzo powiązana z sektorem naukowym, jak AI" (Borowiecki i Mieczkowski 2019:50). Współpraca firm z akademią polega zazwyczaj na włączaniu naukowców lub studentów w tworzenie/rozwijanie konkretnych rozwiązań (48% wskazań), organizowaniu staży lub zajęć dla studentów (31%), dostarczaniu danych lub ich wymianie (18%), prowadzenie zajęć kursów lub studiów (13%), innych (12%), współpracy ze studenckimi kołami naukowymi (11%), a najrzadziej na organizacji konkursów lub grantów dla studentów (6%). Podkreślono fakt, że działania firm nastawione na studentów mają pozytywny wpływ na proces rekrutacyjny. Firmy oczekują od akademii dostarczenia pracowników – zarówno absolwentów jak i doświadczonych badaczy. Nazwano to relacją jednostronną, bowiem współpraca nastawiona jest jakoby głównie na rozwój firm, nie uczelni. Przywołano przykład, wspomnianej przez nas w innym kontekście, najważniejszej międzynarodowej konferencji naukowej w dziedzinie uczenia maszynowego NeurIPS 2018 (por. część 2.2.3.5. tej pracy). Spośród tysięcy opublikowanych w jej ramach artykułów, tylko cztery były autorstwa osób z afiliacją polskiej uczelni, a wśród nich zaledwie jeden powstały we współpracy z polską firmą AI (Borowiecki i Mieczkowski 2019:46-48).

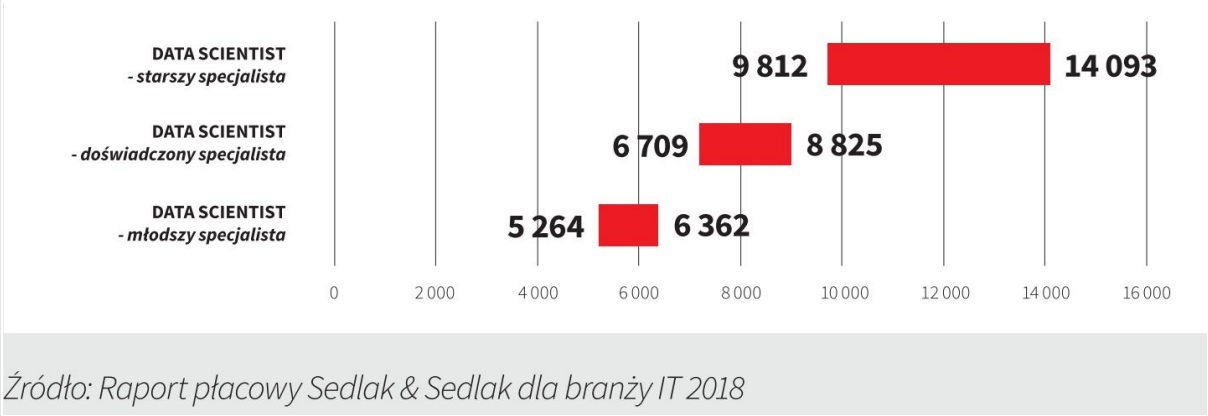
Co do rynku pracy dla data scientistów w Polsce, to przypominamy o ogromnym entuzjazmie wobec tego zajęcia (por. część 1.1.4). W Polsce także pojawiło się wiele entuzjastycznych tekstów, np. data scientist jako zawód przyszłości (Kodołamacz.pl i Centrum Zarządzania Innowacjami i Transferem Technologii Politechniki Warszawskiej 2017; Nurczyk i Ramza 2017) czy specjaliści od big data „na wagę złota" (Błaszczak 2016).

Światowe Forum Ekonomiczne wskazuje na zawód „Data Analysts and Scientists" w latach 2018-2022 jako rozwijający się (*emerging*) we wszystkich branżach (*industry profile*), i we wszystkich regionach geograficznych (Centre for the New Economy and Society 2018:43–65; 68–125). **Ministerstwo Cyfryzacji RP twierdzi, że w Unii Europejskiej brakuje 400 tys. pracowników na „rynku danych" i luka ta będzie się powiększać** (Borowik i in. 2018).

Istnieje jedna praca dotycząca polskiego rynku pracy dla, ogólnie rzecz ujmując, zawodów związanych z danymi cyfrowymi – nie tylko DS – przygotowana przez portal Enterprise Software Review (Gontarz 2019). Opracowanie to autor rozprawy otrzymał uczestnicząc w konferencji Big Data Technology Summit {obserwacja 37}. Nie zawiera ono żadnych wyników własnych analiz i niewiele ilościowych informacji z innych źródeł (o czym poniżej), niemniej stanowi dla nas wartościowe źródło, głównie jako materiał etnograficzny. „Wstęp" pokazuje entuzjazm wobec rosnącego popytu na pracę w tych zawodach. **Podkreślono wysokie zarobki, elitarność umiejętności i kompetencji umysłowych oraz zagrożenie polskiego rynku wysysaniem pracowników za granicę**:

> Oddaję w Państwa ręce pionierskie dzieło (...) – pierwszy w Polsce raport poświęcony rynkowi pracy dla ekspertów od danych, począwszy od obszaru sztucznej inteligencji, Data Science i Machine Learning, poprzez Big Data, aż do osoby, które zajmują się zarządzaniem danymi, ich jakością, rolą w organizacji, Data Master Management czy Data Governance. To ta część rynku pracy z obszaru technologii, która rozwija się niezwykle dynamicznie. **Zapotrzebowanie na ekspertów w wybranych specjalizacjach jest ogromne, a deficyt potrzebnych specjalistów stale się pogłębia. Pensje osób o zaledwie kilkuletnim doświadczeniu sięgają pułapów dyrektorskich (często ku frustracji tych ostatnich). Wszystko przez klasyczne prawa ekonomii, gdy popyt znacznie przewyższa podaż** [podkr. RŻ]. Ekspertów jest mało, bo po pierwsze wymaga to określonych predyspozycji umysłowych, po drugie jest to wciąż dziedzina młoda i możliwości kształcenia są tu ograniczone, a po trzecie specjaliści są angażowani przez firmy zagraniczne, które, nawet jeśli fizycznie nie wyciągają potencjalnych kandydatów z polski, to oferują im znakomite warunki pracy zdalnej (Gontarz 2019:4).

W tekście podano przedziały miesięcznych zarobków brutto data scientistów w Polsce, w podziale na poziom doświadczenia, bazując na danych z raportu płacowego dla branży IT firmy Sedlak & Sedlak. Nie wyjaśniono dokładnie, co oznaczają przedziały. Chyba należy rozumieć je jako minima i maksima, co odpowiada potocznemu rozumieniu tzw. widełek wynagrodzenia. Te minima i maksima rosną, a widełki poszerzają się wraz ze wzrostem poziomu doświadczenia, co przedstawiamy na rysunku 14.



Rysunek 14: Przedział miesięcznych zarobków brutto data scientistów w Polsce w podziale na poziom doświadczenia za Sedlak & Sedlak 2018

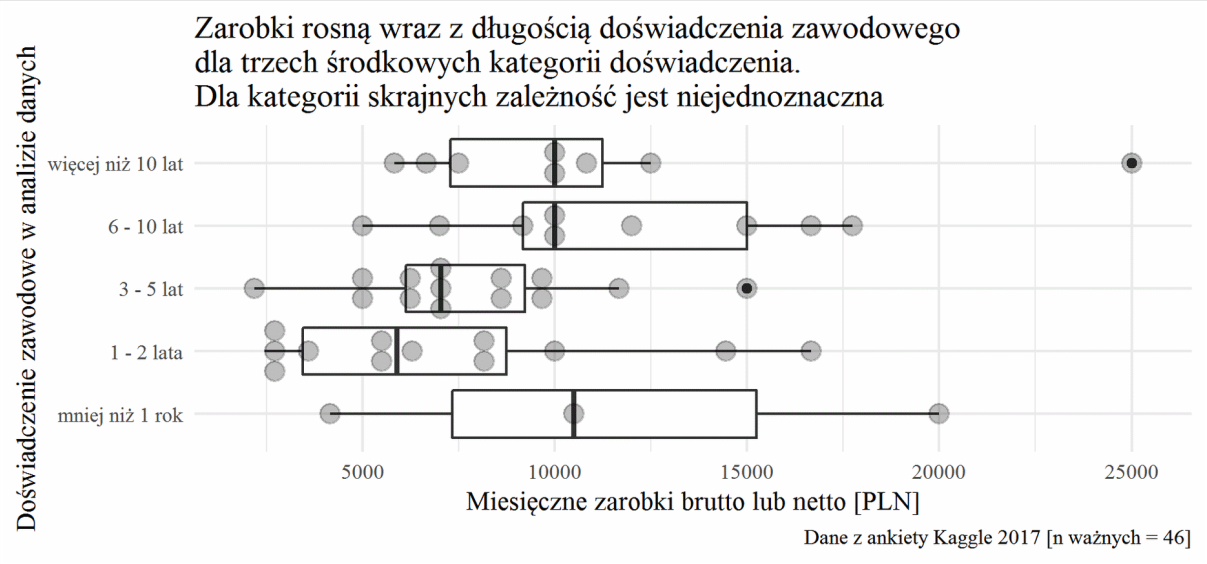
Źródło: (Wykres 1., Gontarz 2019:25)

O zarobki i poziom doświadczenia pytano także w międzynarodowych ankietach Kaggle 2017 i 2018 oraz Stack Overflow 2018 i 2019. Wyniki przedstawimy poniżej na oddzielnych rysunkach, bowiem sposób zbierania danych różni się znacząco między tymi badaniami.

W ankiecie Kaggle z 2017 r. poproszono respondentów o wpisanie dokładnej kwoty. Nie wskazano, czy chodzi o zarobki brutto czy netto. Pytanie o doświadczenie miało pięć kategorii – przedziałów różnej szerokości (por. rys. 15.). Zależność wyraźnie widoczna na rys. 14., czyli minima i maksima oraz długość przedziałów zarobków rosnąca wraz z długością doświadczenia zawodowego, tutaj zachodzi dla trzech środkowych kategorii doświadczenia zawodowego: 1-2 lata, 3-5 i 6-10. Ogółem na rys. 15. widać, że minima i maksima są odpowiednio znacznie niższe i wyższe niż na rys. 14[125]. W ankiecie Kaggle 2017 dla najkrótszego doświadczenia, poniżej jednego roku, statystyki (minimum, Q1, mediana, Q3, maksimum) są wyższe niż dla doświadczenia 1-2 lata. Może oznaczać to, że zarobki oferowane osobom rozpoczynającym pracę w DS wzrosły w czasie. Niemniej to tylko spekulacja, szczególnie, że w omawianej ankiecie dla doświadczenia „mniej niż 1 rok" zarobki podały jedynie trzy osoby (na rys. 15. każdą odpowiedź zaznaczono półprzezroczystym punktem). W przypadku kategorii dotyczącej najdłuższego doświadczenia zawodowego powyżej 10 lat część statystyk (Q1, Q3) jest niższa niż dla kategorii 6-10 lat. Minimum i maksimum są jednak wyższe, a zatem najwyższe w opisywanym przekroju.

125 Z raportu Sedlak & Sedlak wynika, że najniższe wynagrodzenie dla młodszego specjalisty to 5 264 zł brutto. W Kaggle 2017 dla doświadczenia 3-5 lat zanotowano zarobki od 2 200 zł (najpewniej netto). Maksimum w Sedlak & Sedlak to 14 093 zł. W Kaggle 2017 wskazywano wyższe zarobki miesięczne w każdej kategorii doświadczenia zawodowego (od ponad 16,5 tys. zł dla 1-2 lat doświadczenia do 25 tys. dla doświadczenia powyżej 10 lat).

Osoby z najdłuższym doświadczeniem rozpoczynały karierę, gdy nie używano terminu data scientist. Jeżeli obecnie nie posługują się tym terminem, ich praca nich może być postrzegana jako mniej prestiżowa, zatem nieco niżej opłacana niż tych, którzy zbudowali swoją karierę jako data scientist. Podkreślmy, że swoje wynagrodzenie w omawianej ankiecie Kaggle podała mniejszość respondentów, a porównywane grupy są niewielkie[126].



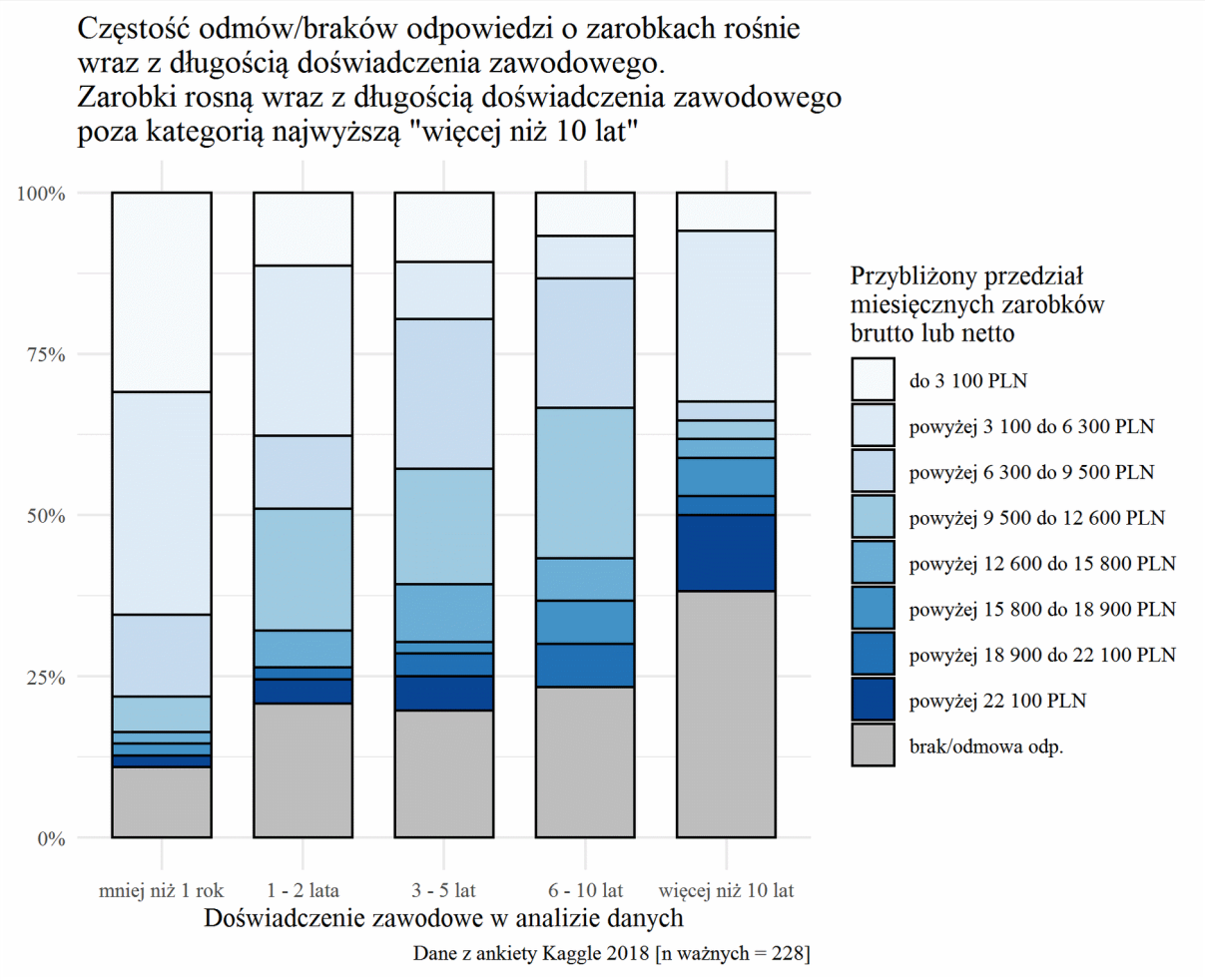
Rysunek 15: Miesięczne zarobki data scientistów w Polsce w podziale na ilość lat doświadczenia za Kaggle 2017

Źródło: opracowanie własne za pomocą GNU R

W ankiecie Kaggle 2018 respondenci wybierali przedział zarobków. Także nie ma jednoznacznej informacji, czy podano kwoty brutto/netto. Taki sposób zbierania danych w ankiecie ma tę zaletę, że pozwala nam w czytelny sposób wskazać udział braków/odmów odpowiedzi, jeśli chodzi o podanie swoich zarobków. Poniżej pokażemy zależność dotyczącą właśnie tych odmów/braków. Przeliczyliśmy zapisane w danych źródłowych roczne kwoty zarobków w USD na kwoty miesięczne w PLN. Te przedziały na rysunku 16. są przybliżone do setek. Widać wyraźnie, że – tak jak w poprzednim źródle Kaggle 2017 – omawiana wcześniej zależność wzrostu zarobków wraz z doświadczeniem zawodowym nie zachodzi jednoznacznie dla najwyższej kategorii doświadczenia. W kategorii „powyżej 10 lat" widoczny jest znacznie większy niż w kategoriach 3-5 oraz 6-10 lat udział wskazań na zarobki powyżej 3 100 do 6 300 PLN. Niemniej w tej samej kategorii udział wskazań na zarobki powyżej 22 100 PLN jest najwyższy ze wszystkich kategorii doświadczenia

126 Zarobki podało około 36% badanych (47 ze 130 osób). Ostatecznie uwzględniono 46 osób, ponieważ jedna nie podała waluty zarobków. Rys. 15. powstał dla trzech osób w grupie „mniej niż 1 rok", 12 osób w „1-2 lata", 14 „3 – 5 lat", 9 „6 – 10 lat", 8 osób „więcej niż 10 lat". Każdą osobę reprezentuje półprzezroczysty, szary punkt.

zawodowego. Podkreślmy także, że częstość odmów/braków odpowiedzi rośnie wraz z długością doświadczenia zawodowego (por. rys. 16.). Odsetek odmów/braków nie różni się jedynie w porównaniu kategorii 1-2 lata i 3 – 5 lat (odpowiednio 21% i 20%).

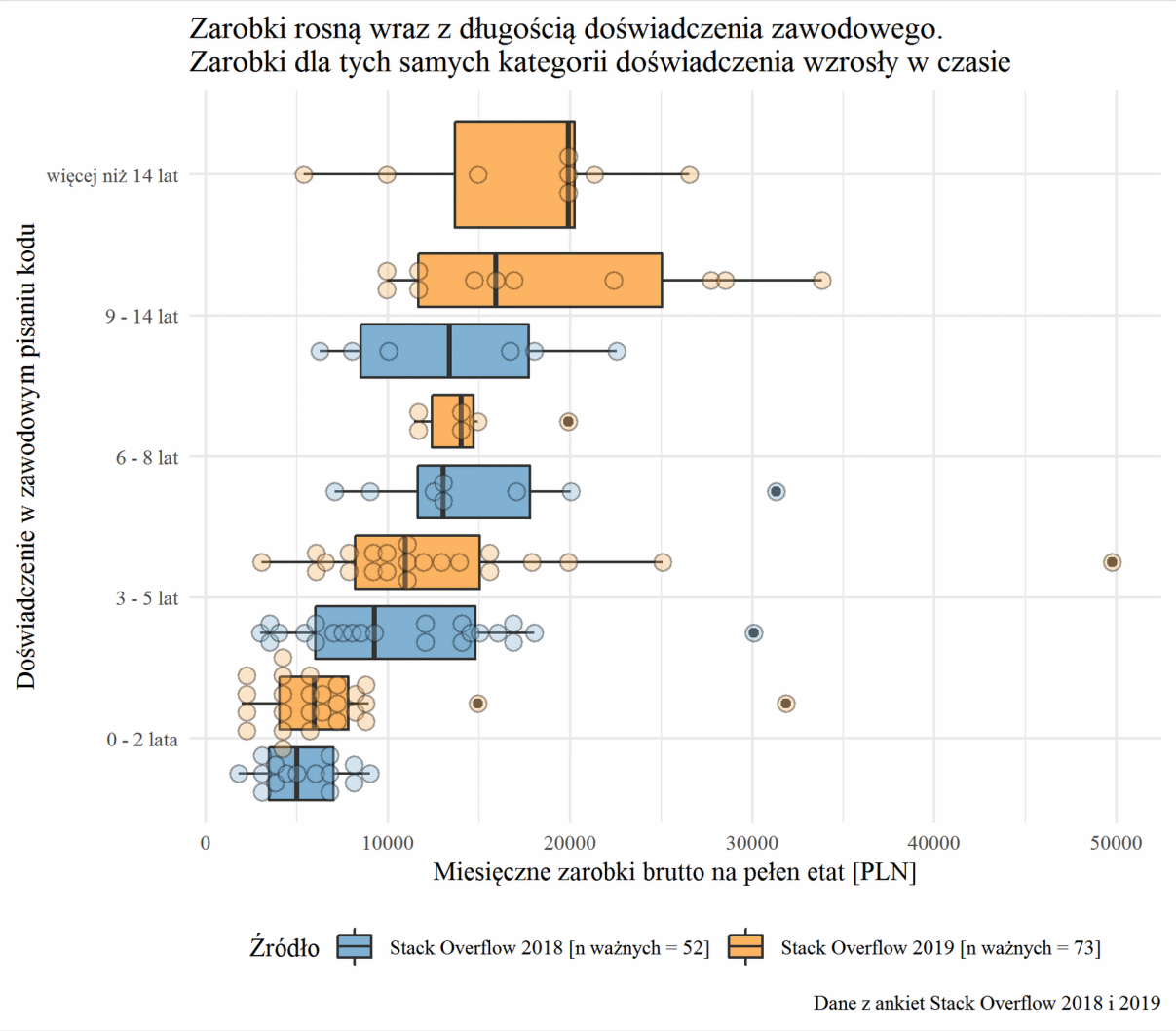


Rysunek 16: Przedział miesięcznych zarobków data scientistów w Polsce w podziale na ilość lat doświadczenia za Kaggle 2018

Źródło: opracowanie własne za pomocą GNU R

W ankietach Stack Overflow 2018 i 2019 respondentów poproszono o wpisywanie dokładnej kwoty zarobków brutto i przeliczono to na zarobki na pełen etat. Kategorie dotyczące doświadczenia zawodowego można było ujednolicić, zatem dane źródłowe są spójne. Ponieważ mamy do czynienia z tym samym realizatorem ankiety, to porównujemy wyniki dwóch ankiet zwracając uwagę także na zmianę w czasie (rys. 17.). Zarobki dla tych

samych kategorii doświadczenia zawodowego[127] wzrosły w czasie, tzn. porównując rok 2018 i 2019.



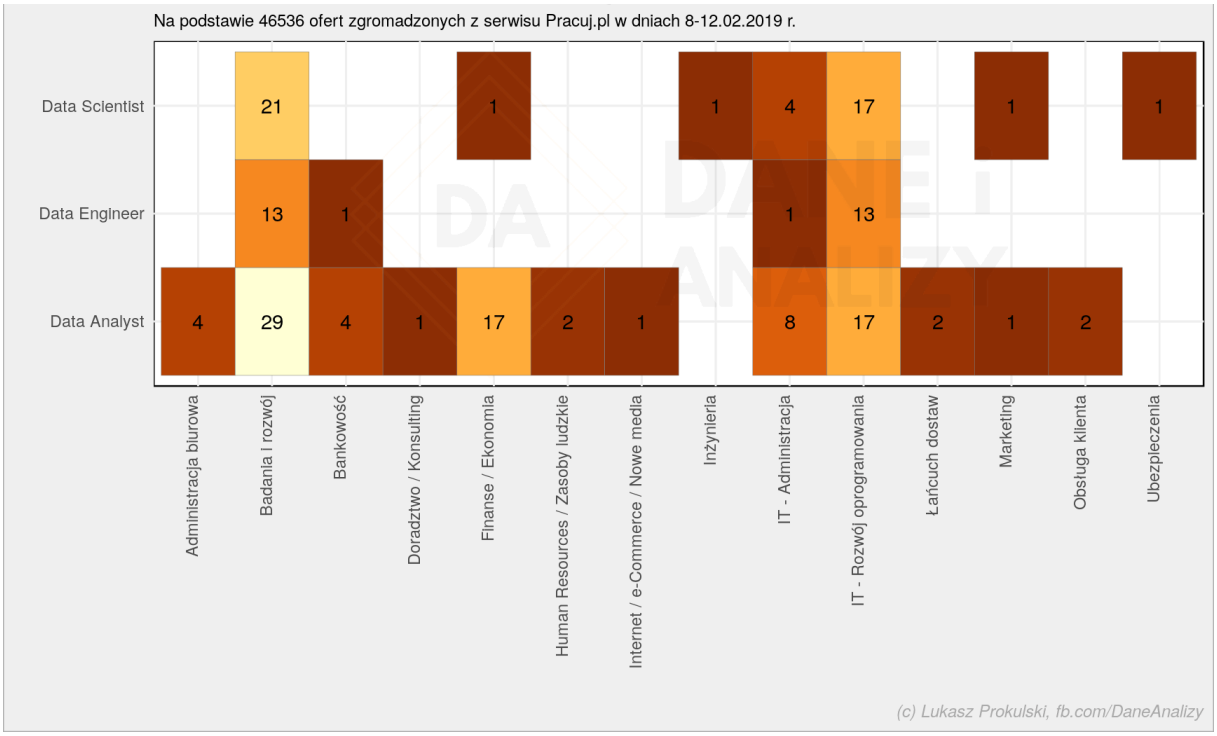
Rysunek 17: Miesięczne zarobki brutto data scientistów w Polsce w podziale na ilość lat doświadczenia za Stack Overflow 2018 i 2019

Źródło: opracowanie własne za pomocą GNU R

Zauważmy na rys. 17. wysokie wartości odstające zarobków rzędu 30 do 50 tys. złotych brutto, podczas gdy w raporcie Sedlak & Sedlak (rys. 14.) wskazywano pułap około 14 tys. jako górną granicę tzw. widełek wynagrodzenia dla najbardziej doświadczonych data scientistów. Jak wynika z badań własnych oraz np. omawianego wyżej tekstu o polskim rynku pracy związanej z danymi (Gontarz 2019:4), te najwyższe wynagrodzenia mogą uzyskiwać osoby pracujące zdalnie, z Polski, w ramach kontraktów dla USA lub innych krajów.

127 Doświadczenie w ankietach Stack Overflow zdefinowano jako „lata profesjonalnego pisania kodu (*For how many years have you coded professionally (as a part of your work)?*)". Inaczej było w Kaggle: „lata pisania kodu w celu analizy danych (*How long have you been writing code to analyze data?*)". Jak wskażemy w części 5.2. ta druga definicja jest bliższa definicji działania podstawowego społecznego świata data science w Polsce.

Łukasz Prokulski analizując na swoim blogu oferty pracy z portalu Pracuj.pl zauważył, że oferty pracy dotyczące przetwarzania i analizy danych pojawiają się na tym portalu najczęściej w tzw. branży „Badania i rozwój", a następnie w „IT – rozwój oprogramowania". Prokulski podzielił stanowiska na trzy grupy, sprowadzając nazwy stanowisk do, jak napisał, „podstawowych form": **data scientist, data engineer i data analyst** (por. rys. 17.). **Te stanowiska to łącznie niecałe 0,35% (162 oferty) spośród 46 536 ofert pracy na terenie Polski** zebranych przez Prokulskiego z Pracuj.pl w dniach 8-12.02.2019 r. Najrzadziej występowały stanowiska data engineer (28 ofert, około 17% ze 162 ofert pracy dotyczących przetwarzania i analizy danych, zaledwie 0,06% wszystkich ofert), więcej było ofert dla **data scientist (46 ofert, prawie 28,5% z przetwarzania i analizy danych, około 0,1% wszystkich)**, najczęściej oferowano stanowiska analityków danych (88 ofert, ponad 54% z przetwarzania i analizy danych, około 0,19% wszystkich). Stanowiska analityków danych (*data analyst*) występowały równie często w kolejnej kategorii „Finanse / Ekonomia", co w wymienionej wyżej branży rozwoju oprogramowania. Łącznie omawiane stanowiska pojawiły się w 14 różnych branżach (Prokulski 2019), co prezentujemy na rysunku17.



Rysunek 18: Liczba ofert pracy z nazwami stanowisk: data scientist, data engineer, data analyst w podziale na branże na podstawie Pracuj.pl za Łukaszem Prokulskim

Źródło: (Prokulski 2019)

Prokulski zaznaczył, że Pracuj.pl jest serwisem ogólnobranżowym. Oferty pracy dotyczące konkretnych branż mogą pojawiać się na innych, wyspecjalizowanych serwisach. Prokulski mówi o odpływie ofert branży IT z Pracuj.pl (2019). Sądzimy, że uzyskana z Pracuj.pl liczba ofert pracy dotyczącej przetwarzania i analizy danych jest około dwukrotnie zaniżona[128], ale rozkład ilości ofert dla grup stanowisk data analyst, data scientist i data engineer oceniamy jako oddający postrzeganie tych stanowisk w badanym świecie społecznym[129].

Ta i zbliżona, wcześniejsza analiza polskich ofert pracy związanych z przetwarzaniem i analizą danych biorą pod uwagę także znajdujące się w treści ofert oczekiwania co do umiejętności pracownika i stosowane technologie (Prokulski 2019; Więcko i in. 2016). Technologiom poświęcono miejsce także w badaniach ankietowych DS, jak analizowanych Kaggle i Stack Overflow oraz innych (Borowiecki i Mieczkowski 2019; Piatetsky 2017b, 2018d, 2018c), co pokażemy w części 5.3. tej pracy.

5.1.4. Zmiana w czasie

Choć w Polsce **już w latach 90. XX w. stosowano komercyjnie to, co obecnie bywa nazywane sztuczną inteligencją, to data science jako nazwa pojawia się od 2014 r., zaś rozkwit zainteresowania data science widoczny jest od roku 2017**.

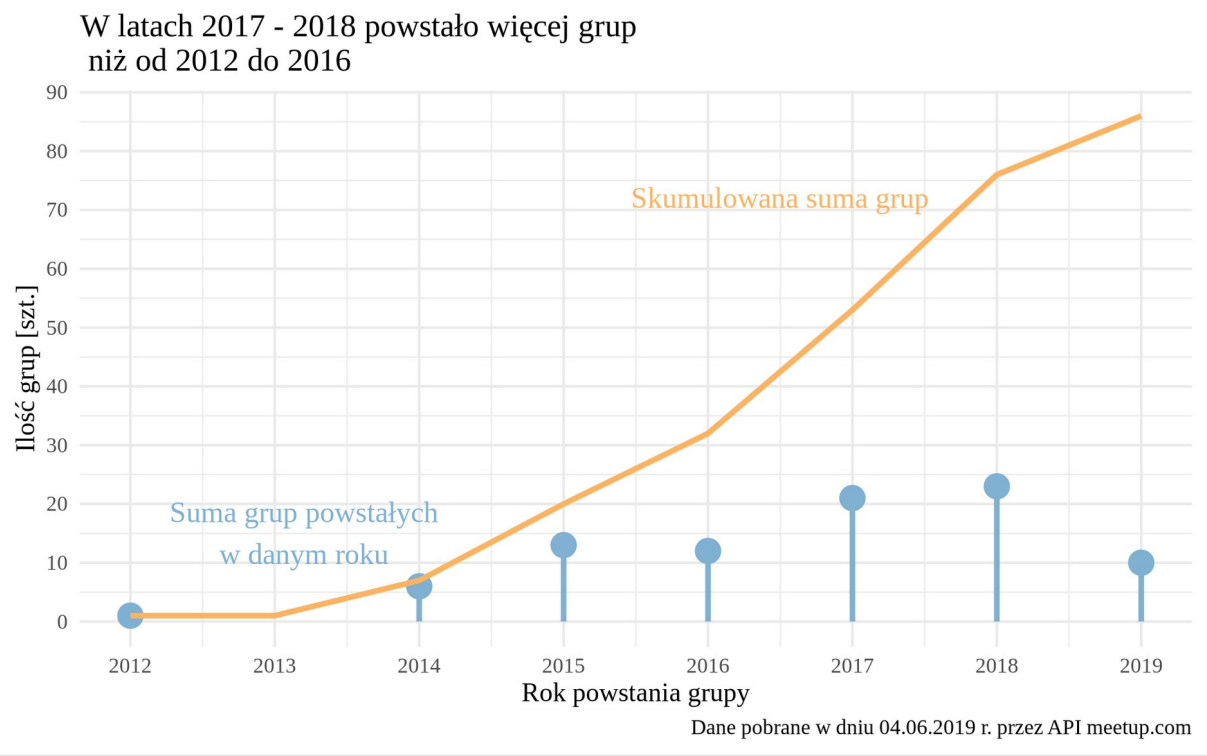
W raporcie Fundacji digitalpoland wskazano, że 3% badanych podmiotów zaczęło stosować AI w latach 90., zaś połowa w latach 2017-2018, przy czym badanie zrealizowano w październiku 2018 (Borowiecki i Mieczkowski 2019:13). Z własnych analiz grup meetupowych wynika, że najstarszą z działających obecnie grup o tematyce DS jest DataKRK (formerly Cracow Hadoop User Group) z Krakowa, założona w styczniu 2012 r. W roku 2013 nie powstały żadne do dziś istniejące grupy, zaś w 2014 założono ich sześć[130]. W kolejnych

128 Ofert publikowane są np. na stronach www firm i udostępniane na grupie facebookowej Data Science PL, na LinkedIn, na portalach nofluffjobs.com, pl.indeed.com, careerjet.pl, glassdoor.com, upwork.com (portal służy raczej do oferowania i przyjmowania zleceń, nie zatrudnienia). Działania związane z oferowaniem pracy to także to, co dzieje się w kontaktach twarzą w twarz: targi pracy w formule stoisk odbywały się np. na konferencji Data Science Summit w 2019, a przedstawiciele firm mówią bądź prezentują wizualnie (na slajdach, roll-upach) co najmniej informację o fakcie poszukiwania kandydatów do pracy podczas niemalże każdej konferencji czy na meetupach.

129 Ogólnie data analyst postrzegany jest jako praca najmniej prestiżowa, najłatwiejsza i najmniej dochodowa z tych trzech. Data scientist jest stanowiskiem uznawanym za bardziej wymagające, bardziej prestiżowym i wyżej opłacanym niż analityk, ale nazwa ta jest używana w odniesieniu do różnych zakresów obowiązków, wobec czego występuje duża zmienność wymagań, prestiżu i zarobków. Data engineer cechuje się porównywalnym prestiżem i zarobkami co data scientist, ale nazwa ta określa węższy wycinek zakresu obowiązków niż data scientist, więc zmienność będzie niższa niż dla data scientist. Rozwijamy ten problem w dalszej części rozprawy (6.1.3).

130 Były to grupy (por. rys. 20.):

latach ilość powstających grup rosła, poza rokiem 2016 (por. rys. 19.). W 2017 i 2018 powstało odpowiednio 21 i 23 grupy. Do dnia pobrania danych w 2019 r. powstało 10 grup, co może wskazywać na podobną liczbę pojawiających się grup w trzecim roku z rzędu. Podkreślmy, że analizujemy dane dotyczące grup istniejących na dzień pobrania danych. Nie możemy na tej podstawie powiedzieć nic o grupach, które działały wcześniej, ale rozwiązano je przed 4. czerwca 2019.



Rysunek 19: Powstawanie grup meetupowych data science w latach 2012 - 2019 (do czerwca)

Źródło: opracowanie własne za pomocą GNU R

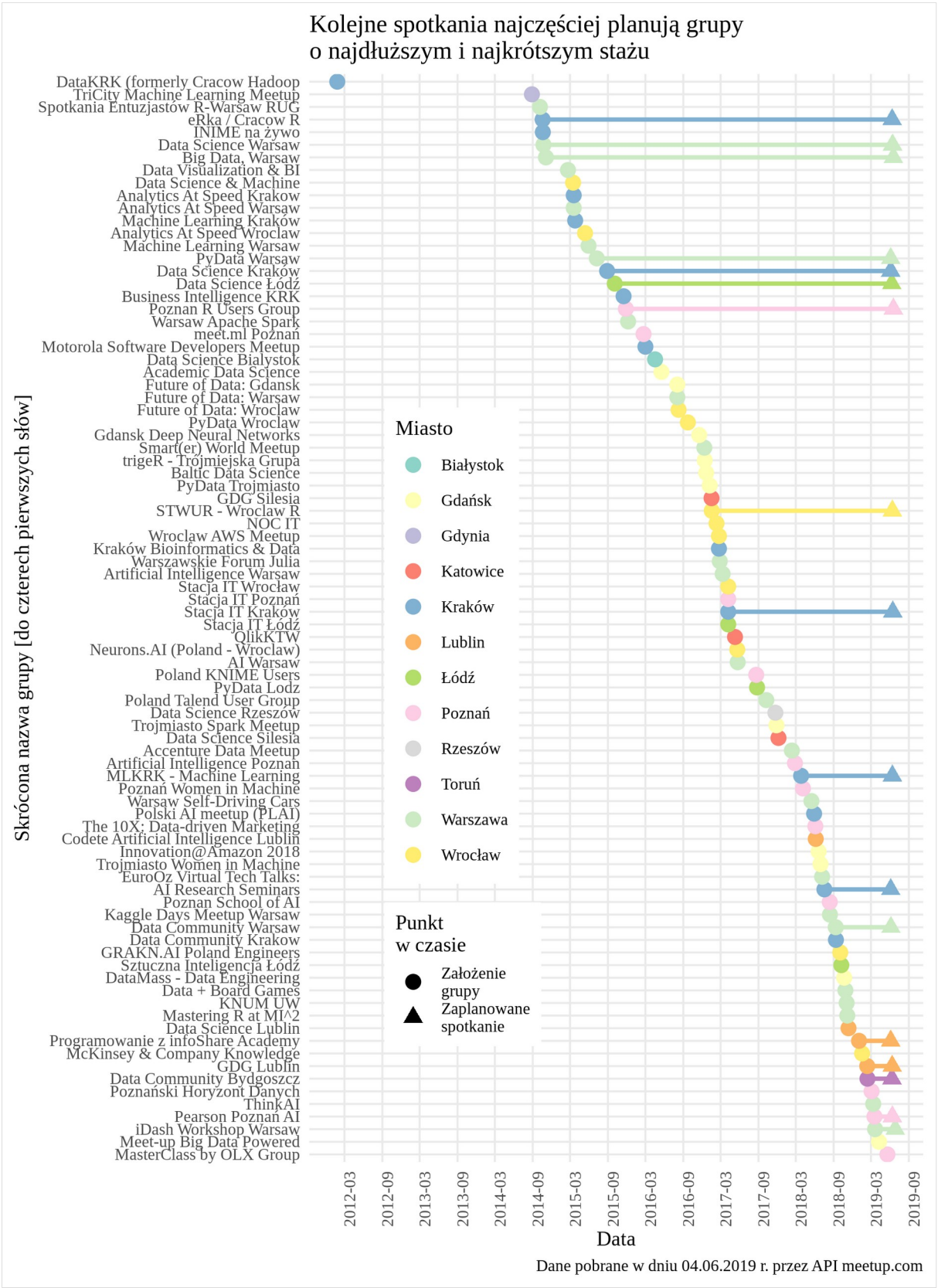
Od 19. lipca 2015 r. funkcjonuje najważniejsza platforma komunikacji polskiego świata społecznego data science, czyli grupa facebookowa Data Science PL[131]. Określamy ją z pełną odpowiedzialnością jako najważniejszą, ponieważ nie ma innych polskich platform komunikacji osób zajmujących się DS ani na Facebooku, ani na innych portalach społecznościowych czy w jakichkolwiek mediach. Piotr Migdał planował założenie grupy facebookowej dla grona swoich znajomych, zajmujących się DS, którego liczebność oceniał wówczas na około 50 osób. Na dzień 25. czerwca 2019 r. grupa ta liczyła 9 047 osób, z czego

=> kont. s. 221 • Data Science Warsaw; Big Data, Warsaw; Spotkania Entuzjastów R-Warsaw RUG Meetup & R-Ladies Warsaw w Warszawie;
- Rka / Cracow R Users Group Kraków; INIME na żywo Kraków w Krakowie;
- TriCity Machine Learning Meetup Gdynia w Gdyni.

131 Informacje podane w tym akapicie pochodzą z nieformalnych rozmów autora (RŻ) z założycielem i administratorem grupy, Piotrem Migdałem.

w dniach 29.05. - 25.06. łącznie aktywnych (co najmniej jeden post, komentarz, reakcja) członków było 8 089 (około 90% wszystkich). Facebook umożliwia administratorowi grupy śledzenie ilości jej członków od lipca 2018 r., zatem nie możemy pokazać całej historii zmian tej ilości w czasie. Wiemy jedynie, że w lipcu 2018 było to około 5 tys. osób. Do 25.06.2019 ilość ta wzrosła do wspomnianych już ponad 9 tys. osób, zatem o ponad 80%.

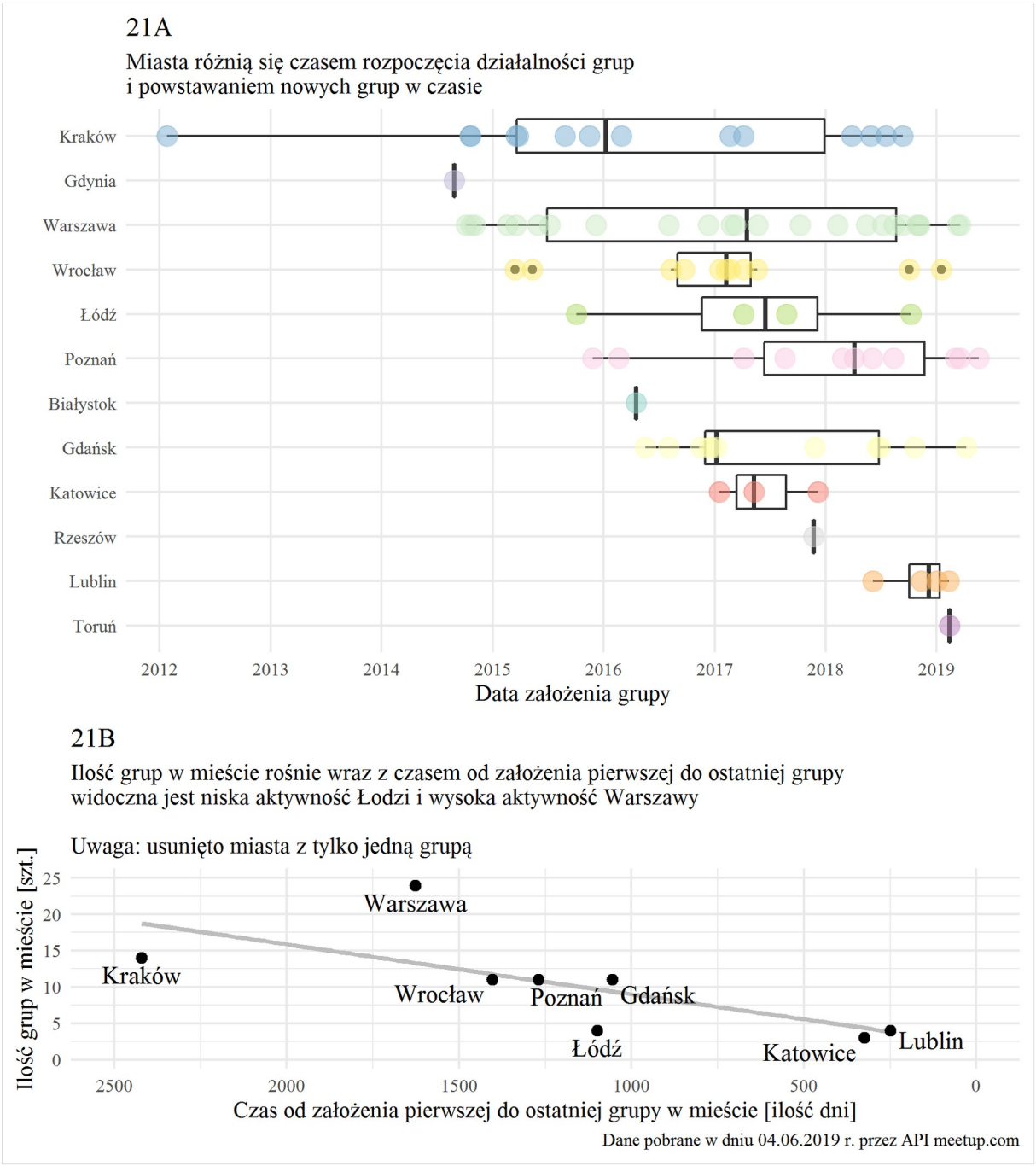
Jak wskazaliśmy wcześniej fakt, że meetup istnieje, nie oznacza, że jest aktywny, tzn. organizuje spotkania. Spośród 86 analizowanych grup 16 miało zaplanowane spotkanie. Kolejne spotkania planują najczęściej grupy najstarsze – w sensie stażu – i najmłodsze (rys. 20.). Te najstarsze mają ustabilizowaną częstotliwość spotkań i sposób ich organizacji, zaś najmłodsze cechuje początkowy zapał organizatorów. Co do samego powstawania grup, to widać, że od drugiej połowy 2014 r. nie ma wyraźnych przerw w zakładaniu nowych meetupów. Zwracamy uwagę na sekwencje pojawiania się grup zakładanych przez danego organizatora (co na rysunku 20. możemy odczytać z nazwy grupy). Przykładowo grupy Analytics At Speed założono w Warszawie i Krakowie (2015-03-19), a następnie we Wrocławiu (2015-05-13). Występują też sekwencje w danym mieście (miasto zaznaczono kolorem), np. w Lublinie do połowy 2018 r. nie było żadnego meetupu o tematyce DS. Pierwsza powstała grupa Codete Artificial Intelligence Lublin (2018-06-06), a później w ciągu czterech miesięcy założono trzy: Data Science Lublin (2018-11-12), Programowanie z infoShare Academy | Lublin (2019-01-02), GDG Lublin (2019-02-11) (por. rys. 20.).



Rysunek 20: Termin powstania 86 grup meetupowych data science oraz termin zaplanowanego spotkania w podziale na dwanaście polskich miast

Źródło: opracowanie własne za pomocą GNU R

Z rysunku 20. wynika także, że miejsce powstania meetupu może mieć związek z czasem powstania. Na rysunku 21. (część A) pokazujemy, że o ile do 2015 powstawały grupy w Krakowie, Gdyni, Warszawie, Wrocławiu, Łodzi i Poznaniu, to w Katowicach, Rzeszowie, Lublinie czy Toruniu nie było żadnych grup do 2017 r. W miastach różny jest też rozkład i częstotliwość powstawania grup w czasie. Porównajmy przypadki skrajne. W Łodzi od założenia pierwszej do ostatniej z analizowanych grup minęło 1 100 dni. W tym czasie powstały cztery grupy (jedna grupa na 275 dni). W Lublinie cztery grupy powstały w ciągu 250 dni (jedna na 62 dni, w stosunku do Łodzi to ponad cztery razy częściej). Rzecz jasna ilość grup w mieście rośnie wraz z czasem od powstania pierwszej do ostatniej grupy (część B rys. 21.), przy czym widać różnice w częstotliwości powstawania nowych grup, co rozumiemy jako jeden z wymiarów aktywności lokalnego świata DS. Ta częstotliwość jest zbliżona w Gdańsku, Katowicach, Poznaniu i Wrocławiu, niższa od nich w Łodzi i nieco niższa w Krakowie (ale wpływ ma grupa założona w 2012 r.), wyższa w Warszawie i nieco wyższa w Lublinie (por. rys. 21B).



Rysunek 21: Powstawanie 86 grup meetupowych data science w podziale na dwanaście polskich miast i w zależność od czasu

Źródło: opracowanie własne za pomocą GNU R

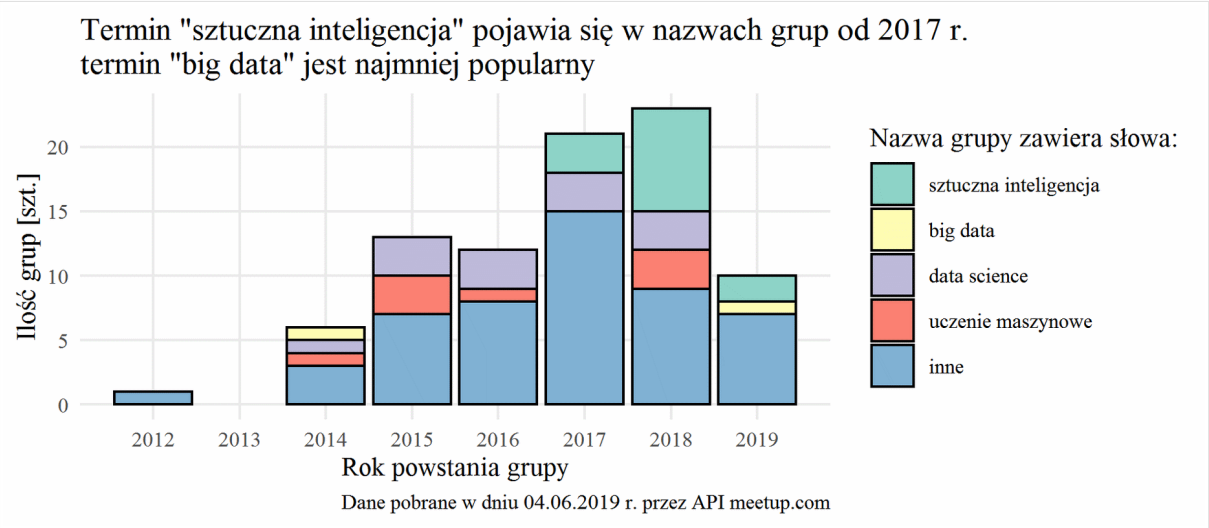
Kończąc próbę przedstawienia badanego świata w czasie, prezentujemy wybrane terminy w nazwach analizowanych grup (rys. 22.; por. rys. 19.). Koncentrujemy się na czterech kluczowych terminach (por. część 1.1. tej pracy). Rozpoczęcie stosowania terminu „sztuczna inteligencja”[132] w nazwach meetupów zaczęło się w 2017 r. i było kontynuowane

132 Dokładny sposób ekstrakcji terminów z nazw grup prezentuje kod w R. Sądzimy, że jest on czytelny także dla czytelniczek/czytelników nie programujących (por. Aneks).

przez kolejne lata. Podkreślamy zatem, że **rok 2017 jest momentem rozpoczęcia trwającego wciąż rozkwitu zainteresowania data science w Polsce**. Termin AI jest w świecie społecznym DS postrzegany jako termin nietechniczny, a co za tym idzie nieprecyzyjny. Stosowany jest jako hasło marketingowe, mające zaciekawiać i wzbudzać emocje. Terminu AI używają także przez osoby niebędące częścią świata społecznego DS, a przede wszystkim używany jest w komunikacji skierowanej do osób spoza badanego świata[133]. Używanie terminu AI w nazwach grup meetupowych może być elementem strategii rekrutowania członków grup wśród osób na bardzo wczesnym stadium zainteresowania problematyką przetwarzania i analizy danych. Zbliżoną funkcję pełnił wcześniej termin „big data", jednak w nazwach obecnie istniejących grup występuje on najrzadziej z wybranych terminów (dwie na 86 nazw). Obecnie termin big data stracił swą niedawną siłę oddziaływania marketingowego i jest używany przede wszystkim jako termin techniczny, dotyczący metod i technologii składowania i przetwarzania zbiorów danych (por. 1.1.1., 5.3.2. i 5.3.3.), nie dotyczy analizy danych ani uczenia maszynowego. Termin big data w sensie technicznym należy do szeroko pojętej informatyki, ewentualnie świata przenikającego się z DS – inżynierów danych (*data engineer,* por. rys. 18.)[134]. Brak jest wyraźnych trendów jeżeli chodzi o posługiwanie się określeniami „data science" i „uczenie maszynowe" w nazwach meetupów (rys. 22.).

133 Zagadnienie to analizujemy w części 6.2.1. dotyczącej aren.

134 Świat inżynierii danych traktujemy jako świat lub subświat powstały na przecięciu działu informatyki zajmującej się bazami danych i infrastrukturą do ich przechowywania oraz przetwarzania, który obecnie staje się odrębny od nieco bardziej dojrzałego świata data science. W wypracowanym przez A. Straussa języku procesów światów społecznych przedstawimy zagadnienie bliżej w części 6.1.



Rysunek 22: Wybrane terminy w nazwach 86 grup meetupowych data science według roku powstania grupy

Źródło: opracowanie własne za pomocą GNU R

5.2. Sposoby definiowania działania podstawowego w data science

> Serce świata społecznego stanowi *podstawowe działanie* (*primary activity*). Wokół niego koncentruje się cała energia uczestników. Stanowi ono także kryterium wyodrębnienia społecznego świata jako pewnej całości, bo jego uczestnikami są aktorzy podejmujący takie działanie. (...) Oczywiście podstawowe działanie nie wyczerpuje sumy działań podejmowanych w społecznym świecie. Odnajdziemy tu zawsze działania wspomagające (*auxiliary activities*), towarzyszące temu podstawowemu, obecne na każdym poziomie socjologicznej obserwacji, zarówno w perspektywie mikro-, jak i w makroskali. (Kacperczyk 2016:34-35)

Co właściwie robią ludzie zajmujący się DS? Jak można ująć ich działanie podstawowe? **Działania podstawowego badanego świata nie można ująć jako „analiza danych”**. Analizą danych – jako czynnością, nie w sensie działania podstawowego – zajmuje się co najmniej kilkanaście środowisk, które nie są tożsame z DS (Granville 2014). Dotyczy to także różnego rodzaju naukowców – akademików, którzy przecież też analizują dane (Azam 2014).

Nie można zatem ująć działania podstawowego w ten sposób – jako analiza danych – ponieważ nie stanowi to kryterium wyodrębnienia całości społecznej. Z drugiej strony jest to ujęcie zbyt wąskie, bo świat DS po części zajmuje się analizą danych (por. rys. 24.), ale nie tylko tym. Ponadto istnieje nazwa stanowiska pracy „analityk danych” (*data analyst*) (por. rys. 18. i część 5.1.3.), a ludzie wykonujący pracę analityka, traktowaną często jako coś, co

robiło się za pomocą arkuszy kalkulacyjnych i zapytań do baz danych w języku SQL już 30 lat temu, niekoniecznie postrzegani są jako uczestnicy świata społecznego DS.

5.2.1. Działanie podstawowe data science

Działanie podstawowe społecznego świata data science w Polsce określamy jako „pisanie kodu do przetwarzania, analizy i modelowania danych”. Przyjrzyjmy się naszemu określeniu dokładniej.

Pisanie kodu oznacza czynność zapisywania serii instrukcji do wykonania przez komputer za pomocą klawiatury komputerowej. Podkreślmy, że chodzi o własnoręczne pisanie instrukcji w formie tekstu, tzw. linijek kodu, przez człowieka w języku programowania[135] zrozumiałym dla komputera (Nowosad 2019). Celowo nie używamy tutaj słowa „programowanie”, po pierwsze, żeby zwrócić uwagę na **fizyczną czynność[136] pisania kodu**, a po drugie, by zaznaczyć, że **data science jest światem odrębnym od świata programowania**, w sensie inżynierii oprogramowania (*software engineering*) i tworzenia oprogramowania (*software development*). Między tymi światami występują przecięcia, ale te światy nie są tożsame, ani nie pozostają w relacji hierarchicznej – DS nie jest podzbiorem, podgrupą czy jakkolwiek ujętą częścią, zawierającą się w całości wewnątrz świata programowania. Rozróżnienie na kodowanie i programowanie występuje też w żargonie osób potrafiących programować, o czym dalej. Najpierw trzeba wyjaśnić to, co sygnalizujemy w poprzednim zdaniu – że programowanie oznacza zarówno umiejętność (osoba potrafi programować w sensie znajomości składni języka programowania i koncepcji programowania[137]), oznacza także proces umysłowy (osoba konceptualnie opracowuje, co ma zostać wykonane, tłumaczy wymyśloną czynność na język programowania), oznacza samą czynność pisania (tutaj słowa programowanie i kodowanie oznaczają to samo), oraz oznacza programowanie jako tworzenie oprogramowania (czyli np. aplikacji, stron internetowych –

135 Jak wspomniano najpopularniejsze języki programowania w data science to Python i R. Poświęcamy im osobna część w rozdziale 6. tej rozprawy

136 Jest to zatem element namacalny (*palpable*) (Strauss 1978:121). Tę namacalność, czyli pisanie kodu jako czynność fizyczną na klawiaturze komputera przy patrzeniu w ekran podkreśla się w komunikatach skierowanych do osób uczących się programowania. Porównuje się pisanie kodu do gry na instrumencie muzycznym, mówiąc o konieczności regularnego ćwiczenia i stopniowego nabierania wprawy w początkowo niewykonalnych czynnościach. „Musisz **napisać na klawiaturze** (*type*) [podkr. org.] każdy z tych przykładów, własnymi rękoma. Jeśli kopiujesz i wklejasz [przykłady z książki], możesz równie dobrze wcale tego nie robić. Te przykłady mają ćwiczyć Twoje ręce, mózg (*brain*) i umysł (*mind*) w czytaniu, pisaniu i dostrzeganiu (*see*) kodu. Jeśli kopiujesz i wklejasz, oszukujesz siebie (...)” (Shaw 2014:2–3).

137 Na podstawie naszego niewielkiego doświadczenia z językiem R wymieńmy np.: pojęcia obiektów prostych (wektorów, macierze) i złożonych (listy, ramki danych), indeksowanie, zmienne i przypisywanie zmiennych, typy danych, operatory logiczne, funkcje, pakiety, pętle, instrukcje warunkowe (Biecek 2015a; Nowosad 2019; Wickham i Grolemund 2017).

tym nie zajmuje się świat DS). **Zatem uczestnicy świata społecznego DS potrafią programować, wykonują proces umysłowy przetłumaczenia ludzkiego pomysłu na język programowania, ale programując – piszą kod, a nie tworzą oprogramowanie**.

Zapisany w języku programowania zbiór instrukcji wykonujący określone zadania – w badanym świecie społecznym zadania te dotyczą danych – określa się jako skrypt.

> *Badacz: (...) ciężko mi jest zrozumieć taką różnicę, niektórzy mówią, że to jest jakaś duża różnica, a niektórzy że nie, różnica pomiędzy pomiędzy data science a programowaniem, tak w ogóle, nie? Czy mógłbyś mi to jakoś przybliżyć? Bo też trochę dla ludzi z zewnątrz to się wydaje że to jest to samo, ale chyba, no nie wiem.*
> Rozmówczyni/Rozmówca: No, no, no, no. Znaczy no, na pewno nie jest to to samo. Znaczy jeżeli mówimy o programowaniu to też jest bardzo obszerny temat.
> *B: No tak!*
> R: Więc na pewno data scientiści programują, aczkolwiek też to no można, można to próbować rozgraniczyć w ten sposób, że jeżeli ktoś pisze skrypt który coś robi, no powiedzmy bierze jakieś dane, je przetwarza w jakiś sposób i wypluwa nie wiem, tabelkę na koniec, no to, można powiedzieć że to nie jest, chociaż to też jest taki **żargon** trochę, że to nie jest programowanie tylko **kodowanie**, albo **pisanie skryptów**, podczas gdy programowaniem można by nazwać, faktycznie właśnie jakąś aplikację albo program, który byłby przygotowany do robienia czegoś **ciągle**, albo [krótka pauza] No też mi ciężko jest to uchwycić, bo no ja, ja na pewno nie jestem programistą, tak? To na pewno nie. Yyy nie znam się zbyt dobrze na właśnie zasadach programowania, jakiegoś tam programowania obiektowego chociażby, jakby, no jakbym miał stworzyć od zera jakąś aplikację, czy jakiś program, do użytku przez ludzi, to raczej by to nie zabanglało [kolokwialnie: zadziałało]. Eee no, no bo właśnie data scientiści bardziej się zajmują danymi, tak? Bardziej się zajmują wysnuciem jakichś wniosków na bazie na przykład jakiegoś modelu statystycznego, a niekoniecznie nad tym na przykład jak zrobić żeby dane automatycznie nie wiem, przepływały z jednego serwera na drugi i żeby potem one się nie wiem, żeby właśnie ten wykres był w internecie online cały czas, no to to jest kwestia już właśnie programisty, inżyniera, inżyniera danych, no to też różne role można pod to podciągać. No ale tak, teoretycznie w takim naj najbardziej powszechnym rozumieniu no to chyba, data science zajmuje się właśnie, no tą węższą częścią wyciągania wniosków z danych, przygotowywaniem modeli, a niekoniecznie właśnie, tworzenie programu który będzie ten model nie wiem, obsługiwał. Jakoś tak, coś takiego. {wywiad 9}[138]

Podkreślmy, że **DS to przede wszystkim zajmowanie się danymi, nie tworzenie oprogramowania czy automatyzacja operacji**. Kontynuując wątek pisania skryptów – przykładowo prosty zbiór instrukcji wykonujący operacje na danych w języku R wygląda następująco (por. Biecek2015a):

138 Cytaty z badań własnych wyróżniamy formatowaniem jak inne cytaty oraz szarą, pionową linią po lewej stronie tekstu. W cytowanych fragmentach wywiadów wypowiedzi autora rozprawy oznaczamy kursywą i skrótem B (badacz – RŻ), a wypowiedzi uczestników zwykłym krojem pisma i skrótem R (Rozmówczyni/Rozmówca). Wykonano transkrypcje dosłowne, tak samo dosłowne są cytaty. **Pogrubieniem** zaznaczamy słowa wypowiadane z naciskiem, w nawiasach kwadratowych elementy niewerbalne [np. pauza, śmiech lub R pokazuje zdjęcie na telefonie] lub wyjaśnienia. Podkreśleniem wyróżniamy wybrane przez nas fragmenty. Numer wywiadu/obserwacji zapisujemy w nawiasach klamrowych, np. {wywiad 15}. Spis wywiadów i obserwacji znajduje się w Aneksie.

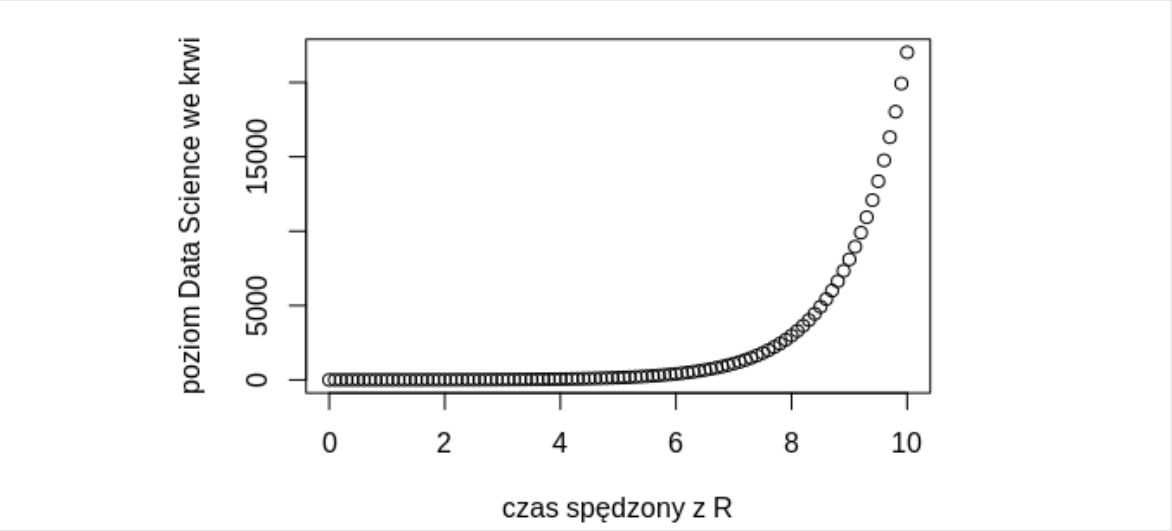
```
sekwencja <- seq(0, 10, 0.1)
poziom <- exp(sekwencja)
plot(x = sekwencja, y = poziom, xlab = "czas spędzony z R",
  ylab = "poziom Data Science we krwi")
```

Powyższe instrukcje to kolejno utworzenie sekwencji liczb funkcją „seq()” i przypisanie jej do zmiennej „sekwencja”, dalej obliczenie eksponenty funkcją „exp()” dla każdej liczby w „sekwencji” i przypisanie wyniku do zmiennej „poziom”, na koniec narysowanie wykresu funkcją „plot()”, dla której podajemy wartości argumentów – dwie przypisane we wcześniejszych liniach zmienne i dwie wartości tekstowe. Ten zbiór instrukcji zapisany do pliku „wykres_ds_we_krwi.R” to właśnie skrypt.

Po uruchomieniu skryptu instrukcją:

```
source("wykres_ds_we_krwi.R")
```

Otrzymamy wynik w postaci wykresu (rys. 23.):



Rysunek 23: Poziom data science we krwi według czasu spędzonego z GNU R za Przemysławem Bieckiem

Źródło: kurs „Pogromcy Danych” (Biecek 2015a)

Na wykresie widać efekt działania skryptu. Pozycję każdego narysowanego punktu określają liczby zapisane w zmiennych „sekwencja” (oś x) i „poziom” (oś y), które podano jako wartości argumentów „x” i „y” funkcji „plot()”, a osie opisane są wartościami argumentów „xlab” i „ylab” tej funkcji. Nim odniesiemy się do treści powyższego wykresu, chcemy pokazać, czym jest skrypt wykonujący operacje na danych. Choć jest on w podanym przykładzie bardzo prosty, a dane fikcyjne, to wykonywana operacja – wizualizacja danych na wykresie – jest jednym z typowych zadań data scientistów. Skryptu nie uznaje się za

oprogramowanie (*software*). Posłużmy się przejaskrawionym porównaniem powyższego skryptu ze znanym w naukach społecznych oprogramowaniem analitycznym SPSS. Ten skrypt można uruchomić w środowisku R lub poprzez wiersz poleceń[139], a następnie zobaczyć wyniki jego działania, tj. wykres. Oprogramowanie SPSS można także uruchomić przez wiersz poleceń, ale typową strategią użytkownika SPSSa jest raczej dwukrotne kliknięcie na odpowiednią ikonę na pulpicie. Po uruchomieniu aplikacji użytkownik ma możliwość klikania w menu, ewentualne pisania kodu wewnątrz aplikacji, i wykonania szeregu działań – od załadowania danych, przez operacje na nich, do zapisania plików wynikowych w wybranym formacie. Omawiany skrypt nie daje możliwości ani załadowania danych, ani wyboru operacji, ani zapisu wyniku, a ponad wszystko nie daje możliwości klikania czegokolwiek, bo nie ma interfejsu graficznego. Nie jest to „program do użytku przez ludzi" {wywiad 9}. Ten skrypt i skrypty w ogóle nie są przeznaczone do interakcji z użytkownikiem nie potrafiącym programować[140]. Osoba potrafiąca programować może ręcznie zmienić kod zapisany w skrypcie i uzyskać inny wynik. W powyższym przykładzie może być to zmiana wartości argumentów funkcji „seq()", co spowoduje wykonanie tych samych operacji na innych danych. Skrypt po modyfikacji, lub tylko jego element, można zatem wykorzystać ponownie do wykonania podobnej operacji. Skrypt nie zmodyfikowany umożliwia prześledzenie i sprawdzenie operacji wykonywanych na konkretnym zestawie danych, oraz zapewnia (lub powinien zapewniać[141]) powtarzalność wyniku. Powtarzalność to niezwykle ważna cecha działania skryptu, o czym powiemy później.

Omówiliśmy, co oznacza w działaniu podstawowym świata społecznego DS pisanie kodu. Kod ten służy **do przetwarzania, analizy i modelowania danych**. Uczestnicy świata społecznego DS mogą pisać kod nie wykonujący operacji na danych w ramach działań wspomagających, ale nie jako działanie podstawowe. **Działanie podstawowe jest nakierowane wyłącznie na dane**, zawsze są one w formie cyfrowej i zawsze są traktowane ilościowo. To nakierowanie wyłącznie na dane wyraźnie odzwierciedla sama nazwa DS. W Polsce mało znane i marginalnie używane są nazwy business science czy decision science, choć DS może wspomagać podejmowanie decyzji w organizacjach czy szeroko pojęty biznes. Zaznaczmy, że **nasze ujęcie działania podstawowego data science obejmuje tak**

139 Zwany też konsolą, terminalem czy powłoką (*shell*), o czy więcej w części poświęconej technologiom

140 Chyba, że skrypt jest jednym z wielu elementów gotowej aplikacji, niemniej użytkownik końcowy nie ma z takimi skryptami bezpośredniej interakcji, ewentualnie interakcję zapośredniczoną przez interfejs graficzny.

141 Powinien w sensie, że jest to traktowane jako wartość – zaleta pisania skryptów, ale zdarza się inaczej, zarówno z powodu błędów w pisaniu kodu, jak i zmieniającego się środowiska programowania. Szczególnie dotyczy to nowych wersji rozwiązań otwartych (*open source*), czyli w data science np. pakietów dla języków Python i R, oraz samych języków.

działalność komercyjną, jak akademicką i hobbystyczną, co wyjaśnimy dalej. Przyjrzyjmy się, jak jeden z Rozmówców opisał, co robią ludzie zajmujący się DS.

> R: Gdybym miał podać swoją definicję, to powiedziałbym, że ludzie zajmujący się data science zdobywają dane, poznają dane, przetwarzają dane, wydobywają z danych użyteczne wzorce, wreszcie komunikują dane.
> W ramach powyższych zadań wydzieliłbym
> - zdobywanie danych
> * parsowanie publicznych danych
> * łączenie danych z różnych repozytoriów
> * czyszczenie danych
> - poznawanie danych
> * ekstrakcja cech
> * wejście w domenę
> * eksploracja analityczna danych
> - przetwarzanie danych
> * inżynieria cech
> * czyszczenie i transformacja
> * redukcja wymiarowości
> - wydobywanie wzorców
> * klasyfikacja
> * regresja
> * klastrowanie
> * ...
> - komunikowanie danych
> * wizualizacja
> * business intelligence
> * inteligentne raportowanie {konsultacja email z R w sprawie działania podstawowego – {wywiad 23}

Rozmówca posłużył się terminami przetwarzania, analizy i modelowania danych nieco inaczej niż proponujemy, ale zauważmy, że te terminy ani w języku polskim, ani w angielskim nie są jednoznacznie dookreślone, i uczestnicy badanego świata doprecyzowują je przez podawanie przykładów. Przybliżmy, co oznaczają dla nas operacje przetwarzania, analizy i modelowania danych. Wzorujemy się na ujęciu Wickhama i Grolemunda (2017), adaptując także zaproponowany przez nich schemat ujęcia tych operacji jako procesu, w ramach projektu DS (por. rys. 24.).

Przetwarzanie danych (*data wrangling*) **jest pierwszym etapem całego procesu. Składa się na nie kolejno zaimportowanie danych, ich porządkowanie i wykonanie transformacji danych**. Dane mogą być zaimportowane z różnych źródeł, ale najbardziej typowym scenariuszem w zastosowaniach komercyjnych jest import z bazy danych. Takie dane mają już zaprojektowany sposób dostępu i pewną strukturę[142], zatem jest to scenariusz

142 Szczególnie w przypadku danych w bazach relacyjnych z dostępem przez SQL. Istnieją także bazy nierelacyjne, tzw. data lake (dosłownie jezioro danych), gdzie dane są słabiej ustrukturyzowane

uznawany za relatywnie prostszy. Może też istnieć potrzeba pozyskania danych, np. ze źródeł publicznie dostępnych (dane rządowe, naukowe, do celów szkoleniowych czy demonstracyjnych np. z pakietów R/Python, dane z konkursów np. Kaggle), ze źródeł wewnętrznych (np. firmowa poczta email czy repozytorium plików w formatach biurowych) czy zewnętrznych (np. media społecznościowe, strony www, co składa się na tzw. *web scraping*, ewentualnie zakup danych ale to nie jest element działania podstawowego), co uznawane jest za scenariusz trudniejszy. Te czynności w powyższym cytacie nasz Rozmówca nazwał parsowaniem danych[143]. Dodajmy, że w zastosowaniach naukowych czy szkoleniowych dane mogą być fikcyjne, jak w pokazanym wcześniej prostym skrypcie. Podstawową czynnością jest **import** danych do środowiska programowania, czyli jak mawiają uczestnicy badanego świata: pobranie, wciągnięcie, wrzucenie, załadowanie danych do wykonywania na nich dalszych czynności. Pobranie z bazy danych czy za pomocą metod *web scrapingu* przeważnie odbywa się też w danym środowisku programowania. Dalej wykonywane jest **porządkowanie** (*tidy*) danych, na które ogólnie rzecz ujmując może składać się np. łączenie danych z różnych źródeł i ujednolicanie sposobu ich zapisu, definiowanie sposobu zapisu danych[144], definiowanie co jest jednostką obserwacji, a co zmienną. Mówimy dla uproszczenia o danych w formie tabelarycznej, ale przypomnijmy, że opisywany proces, co do zasady odnosi się też do innych typów danych, np. fotografii, dźwięku, czy tekstów. W Polsce używane jest określenie **czyszczenia danych**, stosuje je także Rozmówca w powyższym cytacie. Czyszczenie i w ogóle czynności przetwarzania danych są uznawane za najbardziej czasochłonne, a także żmudne i nudne części procesu DS. Zasadniczo chodzi o doprowadzenie danych do stanu, który będzie umożliwiał przejście do iteracyjnego procesu transformowania, wizualizacji i modelowania danych[145]. **Transformacje danych** to, np. tworzenie podzbiorów danych za pomocą filtrowania, wybierania zmiennych czy próbkowania. Może być wykonywane grupowanie i podsumowanie danych, np. z poziomu jednej obserwacji na sekundę do średniej tych obserwacji na minutę. Transformacje

143 To słowo pochodzi z informatyki (*computer science*), jest używane w różnych kontekstach w znaczeniu: przeanalizowania, zrozumienia treści i wykonania czynności. Mówi się np. że w Pythonie czy R instrukcje w danym języku po wpisaniu w konsolę są parsowane i od razu widoczny jest wynik działania instrukcji (języki te działają interaktywnie, nie wymagają kompilowania jak np. język Java). Mówi się także o parsowaniu w sensie zaimportowania ciągów znakowych typu „20190605" jako daty 5. czerwca 2019 r. Spotkaliśmy się także z określeniami typu „nie sparsowałem tego, co powiedziałeś" co oznacza niezrozumienie słów drugiej osoby.

144 Jak w przypadku wspominanego wyżej tzw. parsowania ciągu znaków na datę, może być to parsowanie np. na zapis czasu, cechy ilościowej czy kategorialnej

145 Podkreśla się, że w konkursach (np. Kaggle), czy w nauce na kursach lub studiach, uczestnikom udostępniane są dane już uporządkowane/przeczyszczone. Zatem choć trzeba je zaimportować do środowiska programistycznego, to do rozpoczęcia transformacji/wizualizacji/modelowania (z naciskiem na modelowanie) nie potrzeba lub potrzeba bardzo niewiele pracy uznawanej za czasochłonną, żmudną i nudną.

danych to także tworzenie nowych zmiennych. Na przykładzie szkoleniowym (Wickham i Grolemund 2017) dla danych o lotach samolotów pasażerskich może być to utworzenie nowej zmiennej „prędkość lotu" w wyniku podzielenia zmiennej „dystans" przez „czas w powietrzu". Kolejna nowa zmienna może wyrażać tę prędkość w m/s zamiast km/h, inna może być logarytmem tej prędkości. Tworzenie nowych zmiennych, selekcja zmiennych czy bardziej zaawansowane matematycznie metody np. analizy składowych głównych (*principal component analysis,* PCA) są generalnie metodami na tzw. redukcję wymiaru danych. Chodzi o zmniejszenie ilości zmiennych na rzecz jakości tych zmiennych[146], przed modelowaniem. Za transformacje lub czyszczenie uznaje się także znane naukowcom zadania identyfikacji i zaadresowania braków danych czy wartości błędnych, nietypowych i odstających (*outliers*). Mogą być także stosowane podsumowania danych w formie np. statystyk opisowych. Opisywane czynności są częścią zarówno etapu przetwarzania jak i analizy danych, na którą składają się transformacje i wizualizacja danych.

Kończąc wątek przetwarzania zauważmy, że polskie słowo „przetwarzanie" nie oddaje emocjonalnego nacechowania określenia *data wrangling* u Wickhama i Grolemunda – *wrangle* dosłownie oznacza kłótnię, spór. Autorzy ci wskazują, że doprowadzenie danych do formy nadającej się do pracy często sprawia na osobie piszącej kod wrażenie walki (Wickham i Grolemund 2017). Używane w Polsce słowo „czyszczenie" akcentuje bardziej żmudny i nudny charakter tych czynności. Mówi się w tym kontekście także o preprocessingu danych (*data preprocessing*), nie spotkaliśmy się z polskim tłumaczeniem terminu. Podkreślmy, że niektóre operacje w fazie przetwarzania danych mogą być wykonywane przez osoby, które nie są uznawane za zajmujące się DS. Chodzi przede wszystkim o zadania ręcznego etykietowania danych, np. fotografii, na potrzeby późniejszego użycia metod nadzorowanego uczenia maszynowego, zwanego także uczeniem z nauczycielem (*supervised learning*). Thomas wraz z zespołem (2018:5-6) twierdzą, że choć osoby uważające się za rozwijające algorytmy (*algorithm developers*) lub oprogramowanie (*software developers*) – autorzy nie używają określenia data scientist – czasami wykonują czynność przetwarzania danych, to te czynności kojarzą się im z praktykantami i osobami zatrudnionymi zewnętrznie do powtarzalnych prac (*crowdsourced workers*) tzw. klikaczami (*clickworkers*)[147] (Crawford i in. 2018:33–34). W badanym przez nas świecie ma to częściowo zastosowanie, jeżeli chodzi o

146 Jest to np. usunięcie zmiennych silnie ze sobą skorelowanych. W naukach społecznych stosowane jest np. wyliczenie indeksu przez sumowanie kilku zmiennych, których wartości reprezentują odpowiedzi na pytanie oparte na skali Likerta – to także redukcja wymiaru danych.

147 Przykład pozyskania danych dzięki „klikaczom" za pomocą serwisu Amazona „Mechanical Turk" przedstawiliśmy w części 2.2.3.4., omawiając tzw. aferę Cambridge Analytica

etykietowanie danych. Nie możemy się zgodzić z tą tezą, jeżeli mowa o wielu innych czynnościach, które są częścią przetwarzania danych, a zapisuje się je w kodzie. Także etykietowanie danych może być postrzegane jako wartościowa część pracy data scientista, pozwalająca lepiej poznać zbiór danych i uzyskać lepszy model {obserwacja 40 i 46}, powstają narzędzia mające wspomagać data scientistów w etykietowaniu (Ratner i in. 2017). Niemniej cały etap przygotowania danych jest raczej postrzegany jako nudny i odtwórczy, choć niezbędny, zaś modelowanie jako etap najciekawszy, najbardziej satysfakcjonujący, jako „to, po co tam jesteś [jako data scientist]":

> *B: Ok. [pauza] Dotychczas ludzie mi tak mówią różnie że, znaczy wiem że, zasadniczo, eee w każdym projekcie bardzo ważne jest przygotowanie danych, a modelowanie jest później, iii ludzie mówią mi różnie że no niektórzy się zajmują obiema rzeczami, niektórzy nie, jak to jest w twoim przypadku?*
> R: Ja się zajmuje obiema rzeczami. I właściwie nie kojarzę żeby ktoś się zajmował tylko jedną z tych rzeczy a druga nie dlatego że, to jest trochę **brudna robota** z tym przygotowywaniem danych i nie wyobrażam sobie żeby ktoś miał taką rolę że robi tylko brudną robotę i nie ma na koniec, tego fajnego. No i w drugą stronę tak samo. Może tak się dzieje w jakiś wielkich korporacjach, może w innych strukturach tak jest, u nas ludzie raczej robią obie rzeczy i wszyscy których znam tak robią.
> *B: Ale dlaczego nazywasz to brudną robotą?*
> R: No bo to jest coś powtarzalnego maksymalnie, tu nie ma tej sztuki, nie ma tej nauki, nie ma tego rozkminiania, tylko bardziej coś w stylu powtarzalnych czynności które doprowadzają do tego żeby dane były czyste. Czasem to, czasem w tym jest sztuka, ale zasadniczo nie ma tak do końca, zazwyczaj to jest jednak przygotowanie żeby to się poprawnie ładowało, żeby się dobrze zapisywało do plików, czyli to już jest taka deweloperska robota, taka co właśnie od tego chcieliśmy odejść jako data scientiści mam wrażenie. Od tego co ciągle trzeba robić podobne sekwencje, czynności, nie? Chociaż to też nie jest tak na 100%, no to też jest w miarę ciekawe, to przygotowanie tam [niezrozumiałe] Po prostu w pewnym momencie to się kończy i wtedy sobie dopiero przypominasz po co tam jesteś [z uśmiechem]
> *B: Mhm, mhm, czyli rozumiem że ciekawa część projektu, to jest robienie modelu i, a właściwie takie dokręcanie go, tak, żeby co raz lepiej działał?*
> R: Tak mi się zdaje, noo
> *B: Oczywiście dla ciebie, oczywiście dla ciebie.*
> R: Tak, tak. Znaczy to nie jest tak zero jedynkowo, ale faktycznie wydaje mi się, że wtedy rozpoczyna się taka zabawa pierwsza, i to tak jeszcze że nie wiesz do końca czemu ale działa i to jest takie w sumie jakby najśmieszniejsze w tym wszystkim [z uśmiechem] [pauza] albo właśnie wiesz czemu, bo i działa, a przygotowanie danych no to właściwie nie masz z tego efektu jeszcze, dlatego to nie jest takie satysfakcjonujące. {wywiad 10}

Zwracamy uwagę na wyraźne odróżnienie pracy data scientistów od programistów – programiści, a właściwie ich „deweloperska robota" to „żeby się poprawnie ładowało", „coś powtarzalnego maksymalnie, tu nie ma tej sztuki, nie ma tej nauki, nie ma tego rozkminiania" {wywiad 10}. Twierdzimy przy tym, że czynności przetwarzania danych za pomocą kodu, także tzw. czyszczenie, są w badanym świecie postrzegane jako „brudna robota" w podwójnym sensie. Po pierwsze, jako praca żmudna, nieciekawa, nieprzyjemna, łatwa i

rutynowa. Po drugie, jako prawdziwa, ciężka, trudna i ważna dla całego projektu – pojawia się określenie „get your hands dirty" (dosłownie: ubrudź sobie ręce), np. na warsztatach czy kursach, kiedy prowadzący chcą podkreślić, że uczestnicy zrobią coś naprawdę, nie na szkolnym, przygotowanym specjalnie na te potrzeby zbiorze danych. Problem rozwijamy w części 5.4.

Drugim etapem procesu jest analiza i modelowanie danych. Składa się na nią **iteracyjny proces transformowania, wizualizacji (te dwie części nazywamy analizą) i modelowania danych**. O ile transformacje mogą inicjować proces, a modelowanie kończyć go, to **te czynności nie mają jednoznacznie określonej kolejności**. Najbardziej typowym będzie proces od transformacji, przez wizualizację, gdzie wzrokowo ocenia się wyniki transformacji w odniesieniu do relacji między zmiennymi, do modelowania, kiedy następuje próba ilościowego podsumowania tejże relacji. Wyniki miar modelu mogą prowadzić do pomysłów na kolejne transformacje w celu poprawy wybranych miar, i tak dalej, ale powtórzmy, że nie ma jednoznacznej kolejności tych czynności. Iteracje pomiędzy transformacjami a wizualizacją danych nazywamy analizą, zaś Hadley i Grolemund (2017) używają określenia eksploracja danych. Stosują oni to określenie, ponieważ ich książka dotyczy właśnie tego rodzaju analiz, ale wskazują, że istnieją także analizy konfirmacyjne i wnioskowanie formalne. Upraszczając, konfirmacja i wnioskowanie formalne są to znane w nauce, statystyczne metody testowania hipotez. Podejście eksploracyjne jest wykorzystywane w DS częściej, ale z pewnością nie wyłącznie. Dla danych tabelarycznych może być to podsumowywanie danych w różnych przekrojach z użyciem wykresów lub tabel, za pomocą rozmaitych statystyk. Hadley i Grolemund (2017) mówią w tym kontekście o dokonywaniu odkryć w danych za pomocą stawiania pytań i uzyskiwania odpowiedzi dwojakiego rodzaju. Po pierwsze, jaka jest wariancja wewnątrz każdej ze zmiennych? Po drugie, jaka jest kowariancja pomiędzy zmiennymi? Autorzy podkreślają, że stawianie kolejnych i kolejnych pytań na bazie wcześniejszych odkryć jest kluczowe dla etapu analizy danych (ibidem). Cały czas mówimy o danych w formie tabelarycznej, pokażmy więc przykład analizy dla innego rodzaju danych. W przypadku obrazów modelowanych za pomocą konwolucyjnej sieci neuronowej (*convolutional neural network,* CNNs), stosuje się metody pomnażania danych (*data augumentation*). Ogólnie rzecz biorąc wynikiem pomnażanie danych obrazowych jest większa liczba bardziej zróżnicowanych obrazów, co ma zapobiegać zjawisku nadmiernego dopasowania, tzw. przeuczenia[148] (*overfitting*) modelu. Pomnażanie obrazów polega na

148 Przeuczenie oznacza sytuację, w której model „nauczył się" nie tylko sygnału, ale i szumu z danych treningowych, w związku z czym nie jest zdolny do poprawnej generalizacji zależności między => c.d. s. 238

przekształcaniu obrazu oryginalnego na różne sposoby, np. przycinaniu, obracaniu, odbijaniu lustrzanym, przesuwaniu, dodaniu szumu, zamianie kolorów, dodaniu zakłóceń (mających symulować np. opady śniegu), zmianie perspektywy. Możliwe są kombinacje tych zmian. Obrazy zmienione dołączane są do oryginalnego zbioru danych, zatem zbiór wynikowy jest większy i bardziej różnorodny niż zbiór oryginalny (por. Jung 2019). Pomnażanie jest rodzajem transformacji. Wizualizacja polega nie na sporządzaniu wykresów podsumowujących obrazy, a na wyświetlaniu obrazu oryginalnego w porównaniu do tych przekształconych (ibidem). Wizualizacje na etapie analizy są tworzone na potrzeby osób wykonujących tę analizę lub ich najbliższych współpracowników i przełożonych. Wizualizacja końcowa, służąca komunikowaniu wyników całego procesu jest przygotowywana z myślą o innym odbiorcy, także spoza świata DS, o czym powiemy nieco niżej.

Czym jest **modelowanie,** w sensie modeli uczenia maszynowego, mówiliśmy już krótko we wstępie do tej pracy (por. 1.1.3.), a pojęcia modelu i modelowania pojawiały się wielokrotnie w rozdziale 2. Wickham i Grolemund (2017) wskazują, że

> **Modele są narzędziem do wyodrębniania wzorów z danych** [podkr. RŻ]. (...) W końcowej części tej książki dowiesz się, jak działają modele i pakiet „modeler". Zostawiamy modelowanie na koniec, ponieważ zrozumienie, czym są i jak działają modele jest łatwiejsze, gdy masz opanowane narzędzia do przetwarzania danych i programowania (Wickham i Grolemund 2017).

W języku nadzorowanego uczenia maszynowego, tj. takiego, w którym znane są wartości zmiennej zależnej, można określić modelowanie jako znajdowanie funkcji łączących zmienne niezależne z zależną. Znaną w nauce metodą tego rodzaju modelowania jest regresja, liniowa czy logistyczna. „Regresja jest babcią modeli nadzorowanego uczenia maszynowego – naukowcy pracowali nad nią już na początku XIX w." (Foreman 2017:229). W regresji liniowej człowiek zakłada występowanie liniowej zależności między zmiennymi, nie zakłada przecież z góry wartości współczynników kierunkowych ani wyrazu wolnego – są one ustalane na podstawie danych za pomocą metody najmniejszych kwadratów. Innego rodzaju modele uczenia maszynowego – czy ich podgrupa bazująca na sieciach neuronowych,

---

=> kont. s. 237 zmiennymi. Larose podaje przykład modelu przeuczonego, w którym zadanie polegało na klasyfikacji osoby do przedziału zarobków, a jako jedną ze zmiennych niezależnych włączono do modelu imię osoby. W zbiorze treningowym istniał przypadkowy związek konkretnego imienia z jednym przedziałem zarobków. Generalizowanie tej zależności spowoduje błędy na danych innych niż treningowych. Model jest przeuczony w tym sensie, że jest przesadnie dopasowany do danych treningowych, traktuje zależność przypadkową – szum – jako zależność znaczącą – sygnał (Larose 2005:91–93). Krótko mówi się o przeuczeniu, że „model nauczył się danych na pamięć".

nazywana uczeniem głębokim (*deep learning*) – nie wymagają założenia o charakterze zależności. Model zatem „uczy się przez doświadczenie", dzięki czemu może wykonywać zadanie klasyfikacji czy predykcji nie na podstawie zadanych z góry reguł, a na podstawie wzorów wyodrębnionych ze zbioru danych.

> To rozwiązanie [deep learning] pozwala komputerom uczyć się na doświadczeniu i rozumieć świat w kategoriach hierarchii pojęć, gdzie każda koncepcja jest zdefiniowana przez związek z koncepcją prostszą. Dzięki gromadzeniu wiedzy z doświadczenia, podejście to pozwala uniknąć formalnego formułowania przez człowieka całej wiedzy potrzebnej komputerowi (Goodfellow i in. 2016: 1).

Jak wskazywaliśmy, mówi się o podziale zbioru danych na treningowy (*training set*), testowy (*test*) i kontrolny (*validation*) (Szeliga 2017:111–12). Pierwszy zbiór służy do iteracyjnego przygotowania, zwanego „uczeniem" lub „trenowaniem" modelu, drugi do dostrojenia parametrów, a za pomocą trzeciego, wcześniej nie używanego zbioru danych sprawdzana jest skuteczność działania modelu. Zobaczmy, jak wyjaśnia zagadnienie modelowania osoba osadzona w badanym świecie społecznym – później opiszemy to jako majsterkowanie:

> *B: (...). Yyy co to, co to w ogóle jest modelowanie? Bo wszyscy mi mówią o modelowaniu, a ja jeszcze nikogo nie zapytałem w ten sposób* [R śmieje się] *No bo ja już o tym dużo razy słyszałem, wiem co to są najbliżsi sąsiedzi, drzewa, sieci neuronowe, spoko. Ale muszę o to kogoś w końcu zapytać, co to w ogóle jest modelowanie?*
> R: Takie lepienie z plasteliny, układanek [śmiech]
> *B: OK.[z uśmiechem] Brzmi ciekawie.*
> R: No weźmy sobie na przykład sieć neuronową.
> *B: No.*
> R: Sieć może mieć różną budowę. Może mieć jeden neuron, może mieć dwa neurony, może mieć dwieście neuronów. Może mieć jedną warstwę, może mieć pięć warstwy, może mieć dwieście warstw różnych. Do tego jeszcze dochodzą inne parametry między, w międzyczasie. Więc modelowanie polega na tym, żeby ten model był jak najbardziej optymalny, i to się robi naaa podstawie, ym nauczenia modelu, bo uczysz model na jednym zbiorze danych i on tam sobie wykminia swoje reguły, jakieś tam wzory matematyczne. I na przykład może mieć precyzję na tą 99% tak, ale musisz go później przetestować na zbiorze walidacyjnym, czyli na takim którego w ogóle nie widział. I dopiero wtedy patrzysz jaka tak naprawdę jest ta precyzja, jak jak dobrze się nauczył. Bo może tak być że on się po prostu wykuł na pamięć. I jak zobaczy nowe jakieś dane to na koniec sobie nie potrafi z nimi poradzić w ogóle. Więc modelowanie polega na tym żeby znaleźć ten balans, żeby błąd na jednym zbiorze i na drugim zbiorze był jak najbardziej zbliżony do siebie i jak najmniejszy, oczywiście.
> *B: Mhm, no dobra, a czy modelowanie to jest najważniejsza część w data science?*
> R: Najważniejsze to jest przygotowanie danych, bo jak masz dobry model, ale masz chujowe dane to ci nic nie wyjdzie tak naprawdę. Także wyczyszczenie danych i dobre przygotowanie danych to jest kluczowe. {wywiad 8}

Zwracamy uwagę na zakończenie powyższego fragmentu. Rozmówczyni podkreśliła, że najważniejszą częścią projektu DS jest faza pierwsza, tzn. przygotowanie danych. W myśl

powiedzenia *garbage in – garbage out* (śmieci na wejściu to śmieci na wyjściu) żaden model nie zrekompensuje braku przygotowania danych. To samo mówił Dominik Batorski:

> (...) większość kursów kładzie nacisk na przegląd modeli i metod machine learningowych. Tymczasem w praktyce dużo większe znaczenie ma tzw. feature engineering, przygotowanie danych, preprocesing, zapewnianie jakości danych. To, czy zastosuje się takie lub inne metody – lasy, regresje czy sieci – ma o wiele mniejsze znaczenie. A feature engineeringu nie można się nauczyć na kursach. Dlatego z punktu widzenia zatrudniających kluczowe znaczenie ma rozróżnienie na ludzi z wiedzą teoretyczną i praktyków (Gontarz 2019:19).

W powyższym fragmencie Batorski mówi, że na modelowanie kładzie się nacisk ucząc zagadnień związanych z DS, zaś w praktyce komercyjnej ważniejsze jest przygotowanie oraz analiza danych. Szczególnie analiza danych, z powodu swojej czasochłonności, uznawana jest za wąskie gardło (*bootleneck*) w procesie DS (Staniak i Biecek 2019). Powstaje szereg pakietów, które mają przyspieszać realizację analizy danych. Od roku 2018 cieszą się one coraz większą popularnością (Staniak i Biecek 2019).

Widzimy, że **modelowanie jest uznawane za najprzyjemniejszy, najciekawszy, najbardziej emocjonujący i najtrudniejszy element projektów DS**. To model jest jakoby „inteligentny", zatem wokół możliwości modeli narosła popkulturowa, eksploatowana przez marketing otoczka tzw. sztucznej inteligencji. To modele wdrożone do większych systemów informatycznych czy jakichkolwiek urządzeń automatyzują zadania w czasie rzeczywistym. To dzięki modelom roboty jakoby zastąpią ludzi i zabiorą nam pracę. Takie postrzeganie modelowania jest charakterystyczne dla laików i dla początkujących uczestników badanego świata. Jest ono także charakterystyczne dla komunikatów zaadresowanych do laików (np. marketingu usług/produktów zawierających tzw. AI) i do początkujących (np. marketingu studiów/kursów DS). Jednak **do kanonu wiedzy uczestników z doświadczeniem należy to, że modelowanie zajmuje od 10% do 40% czasu[149] w projekcie DS**, bowiem jak wskazaliśmy powyżej – przygotowanie i analiza danych są zarazem najbardziej czasochłonne, jak i są niezbędne dla sensowności modelowania:

> R: (…) to wcale nie jest tak że data scientist to spija śmietankę i cały czas sobie buduje te modele i w ogóle to ma tak świetnie. (...) tak jak w tym projekcie związanym z wykrywaniem wyłudzeń, po prostu [pauza] **musieliśmy siedzieć** i **sprawdzać cecha po cesze**. (…). To jest żadna przyjemność i spora część projektów taka jest. Jak projekt data science jest źle

149 Naszym zdaniem popularne jest przekonanie o podziale czasu 80/20 (przygotowanie danych/modelowanie), nawiązującym do tzw. zasady Pareto (por. Gulipalli 2019). Badany świat społeczny chętnie powołuje się na tę zasadę w także w innych kontekstach. Co do podziału czasu pracy, to badania ankietowe data scientistów w USA pokazują, że 53% czasu zajmuje respondentom „zbieranie, etykietowanie, czyszczenie i organizowanie danych", zaś „modelowanie danych" 19% (CrowdFlower 2017:5). Przy tym 60% respondentów wskazało, że najmniej lubi zajmować się „czyszczeniem i organizowaniem danych", a 78% najbardziej lubi zajmować się „modelowaniem danych" (ibidem).

> poprowadzony to wtedy te rzeczy **przygotowawcze** zżerają prawie cały czas projektu i to co jest najfajniejsze o czym się mówi [pauza] **modelowanie**, właśnie ta **zabawa,** z dużymi komputerami, zabawa z GPU, wiele procesorów, to na to zostaje 10% czasu . (...) wydaje mi się że to jest efekt tego, że ludzie od wielu wielu lat myślą, że machine learning i sieci neuronowe to są takie magiczne skrzynki. Że niektórzy umieją te skrzynki obsługiwać, że do tych skrzynek, mmm, **wrzuca** się dane [pauza] i z tych skrzynek wychodzi usprawnienie biznesu. [pauza] A to jest **kompletna bzdura**, ludzie myślą że to się dzieje **samo**. Że żeby być data scientistem wystarczy umieć kręcić śrubkami w tych magicznych skrzyneczkach i to załatwia całość. I dlatego się wydaje że to jest takie super świetne i że te modele będzie się budować cały czas. To jest właściwie największy **mit** i dlatego ludzie chcą sztuczną inteligencję, dlatego ludzie uważają, że w projektach data science tak cały czas robi się takie super ekstra rzeczy. {wywiad 5}

Jak widać modelowanie jest „tym co najfajniejsze, o czym się mówi", ale mitem jest robienie „ekstra rzeczy" przez cały czas – jest to zdecydowana mniejszość czasu w projekcie {wywiad 5}. Co oznacza „zabawa, z dużymi komputerami, zabawa z GPU, wiele procesorów" {wywiad 5} przybliżymy w części 5.3.3. rozprawy, poświęconej tym technologiom badanego świata. Podejmiemy także wątek magii, „magicznych skrzyneczek" {wywiad 5} opisując modelowanie jako arenę, a model jako obiekt graniczny (część 6.2.1. rozprawy).

To, że doświadczeni uczestnicy wiedzą, że modelowanie to niewielka część projektów i że uznają etap przygotowania danych za kluczowy nie oznacza, że modelowanie nie jest jednym z najbardziej lubianych elementów pracy – także dla osób z dużym doświadczeniem. Rozmówca najpierw żartuje, a następnie mówi o swojej radości z działania modelu:

> *B: OK. A co jest tak w ogóle w pana pracy najfajniejsze?*
> R: Obiad.
> B i R: [śmiech]
> R: Nie no serio [ze śmiechem]
> *B: To tak jak u mnie [ze śmiechem]*
> R: Najpierw jest ta walka z głodem do obiadu a potem walka z sennością po obiedzie [z uśmiechem]
> *B: Ehe [ze śmiechem] [pauza]*
> R: Nie wiem. Każdy ma pewnie eee swoje co mu się najbardziej podoba [pauza]
> *B: Mhm?*
> R: Ja odczuwam niesamowitą radość kiedy model jest w stanie robić coś rozsądnego. Tak, czyli na przykład bierzemy sobie właśnie jakieś aaa, jakieś zagadnienie rozpoznawania czegoś na obrazkach, zaczynamy uczyć ten model i on zaczyna **rzeczywiście** wykazywać takie po wynikach widzimy że on coś rozpoznaje potem jeszcze robimy parę różnych takich tricków żeby zobaczyć co powoduje że on rozpoznaje dany obiekt i patrzymy i mówimy no kurcze to ja bym zrobił tak samo. Nagle się człowiek orientuje, że yyy, że to **coś** to już jeeest takie coś bardzo **mądrego** [powoli] co udało mu się stworzyć. I ja jeszcze mam ten background eee informatyczny i ja wiem na przykład że narzędziami informatycznymi bym tego nie zrobił. {wywiad 2}

Dodajmy, że Wickham i Grolemund (2017) całość etapu drugiego, na który składają się analiza i modelowanie danych nazywają rozumieniem danych (por. rys. 24.). Rozumienie danych nie oznacza jednak znajomości domeny, z której dane pochodzą. Przy tym **nie znajdujemy w badanym świecie tak ostro krytykowanych poglądów w rodzaju „dane mówią same za siebie"** (por. 2.2.1.), w projektach komercyjnych **odnajdujemy za to przekonanie o konieczności współpracy data scientista z ekspertem domenowym** (Alekseichenko 2019a). Jednocześnie – pozostając przy projektach komercyjnych – data scientist nie wie, jak rozwiązać problem, z którym przychodzi do niego klient. Nie jest znawcą domeny (choć znajomość domeny jest traktowana jako zaleta, por. 6.1.1.), jest potrafiącym programować ekspertem do danych. Problem jest rozwiązywany (a przynajmniej dokonywana jest próba rozwiązania) z użyciem modelu, który wyodrębnia wzorce z danych. Rolą data scientistów jest przygotowanie środowiska do modelowania:

> R: (…) Programista to jest taka osoba, która dostaje informacje, słuchaj musisz zrobić to to to i to, i on sobie siedzi i wymyśla aha, muszę to zrobić tak, muszę to zrobić siak, gdzieś tam doczyta no i rozwiąże problem. My rozwiązujemy problemy inaczej. My mamy postawione zadanie, na przykład zwiększyć sprzedaż, albo zmniejszyć ilość wypadków, albo zwiększyć ilość prawidłowych, aaa, znajdowań na przykład jakiegoś obiektu na zdjęciach. I my nie wiemy jak to zrobić.
> *B: OK.*
> R: Nasza robota polega na tym, że tworzymy takie [pauza] **pole**, taką przestrzeń, w której **modelujemy** różne zjawiska i to modelowanie to jest tak naprawdę wyciąganie wniosków i my nie robimy tego wyciągania wniosków, one się **same** wyciągają, my tylko jesteśmy w stanie zbadać czy one są poprawne, czy są logiczne, czy **czasami** mamy takie, taki sposób zaglądania do modelu gdzie stwierdzamy, że OK, robi coś rozsądnego, a czasami nawet tego nie mamy i tylko po efektach działania modelu stwierdzamy czy jest dobrze czy jest źle. Nasza praca polega na [pauza] **pomaganiu** modelowi żeby **dobrze** działał. Aczkolwiek te różnica między nami a normalnymi programistami czy w ogóle normalnymi inżynierami jest taka, że my w zasadzie nie wiemy jak rozwiązać problem. My tylko tworzymy takie środowisko, w którym problem, powiem, **sam** się rozwiązuje. Tak dobrze, jak tylko można. {wywiad 2}

W powyższym fragmencie widoczne jest także kontrastowanie pracy data scientistów i programistów, o czym powiemy więcej w części 5.4. dotyczącej wartości badanego świata. Poświęcamy tutaj modelowaniu nieco więcej miejsca, by dokładniej zilustrować ten wyróżniający się element w działaniu podstawowym badanego świata. Powtórzmy, że rozwijamy wątki widoczne już w powyższych wypowiedziach Rozmówczyń/Rozmówców w części 6.1. rozprawy.

Wickham i Grolemund (2017) jako **ostatni element procesu DS podają komunikowanie** uzyskanych wyników. Twierdzimy, że **te elementy etapu komunikowania, które są realizowane w postaci kodu są traktowane jako część działania podstawowego**

badanego świata, zaś inne jako działania wspomagające. W kodzie realizowane są wizualizacje danych – od wykresów na potrzeby osób analizujących dane (o czym mówiliśmy wcześniej), do wizualizacji wyników także dla osób spoza badanego świata. Są to zarówno wykresy obrazujące zależności między zmiennymi (jak te prezentowane w części 5.1. naszej rozprawy), jak i wizualizacje objaśniające działanie modeli uczenia maszynowego (Biecek 2018d, 2018b). Jak mówiliśmy w części 4.1., przy okazji omawiania map A. E. Clarke, świat społeczny DS uważa za wartościowe tworzenie wizualizacji danych w sposób dostarczający wielowymiarowej, złożonej, jak najpełniejszej informacji. To wszystko w sposób czytelny, rzetelny i atrakcyjny wizualnie, z uwzględnianiem tego, jak ludzkie oko i mózg przetwarzają obraz (por. Biecek 2016). Jako działanie wspomagające traktowane będzie komunikowanie wyników w formie np. spotkania z klientem (dla projektów komercyjnych), opublikowania artykułu (dla projektów naukowych), zamieszczenia wpisu na blogu (dla realizacji hobbystycznych). Granica jest jednak płynna i rozmyta, bowiem w Pythonie i R istnieją narzędzia do publikowania wyników w postaci np. slajdów, notatników, plików .pdf czy .html, publikacji w serwisie GitHub (Allaire i in. 2019; Perez i Granger 2015; Pérez i Granger 2007; Xie i in. 2018), a nawet prowadzenia strony internetowej, bloga i pisania dokumentacji lub książek (Xie 2016; Xie, Thomas, i Hill 2019). Komunikacja między uczestnikami badanego świata odbywa się także poprzez sam kod. Zarówno Python jak i R bywa określany, niekoniecznie przez uczestników badanego świata, jako *lingua franca* DS (ActiveState 2018; Barry 2016; Nunns 2017; Thieme 2018).

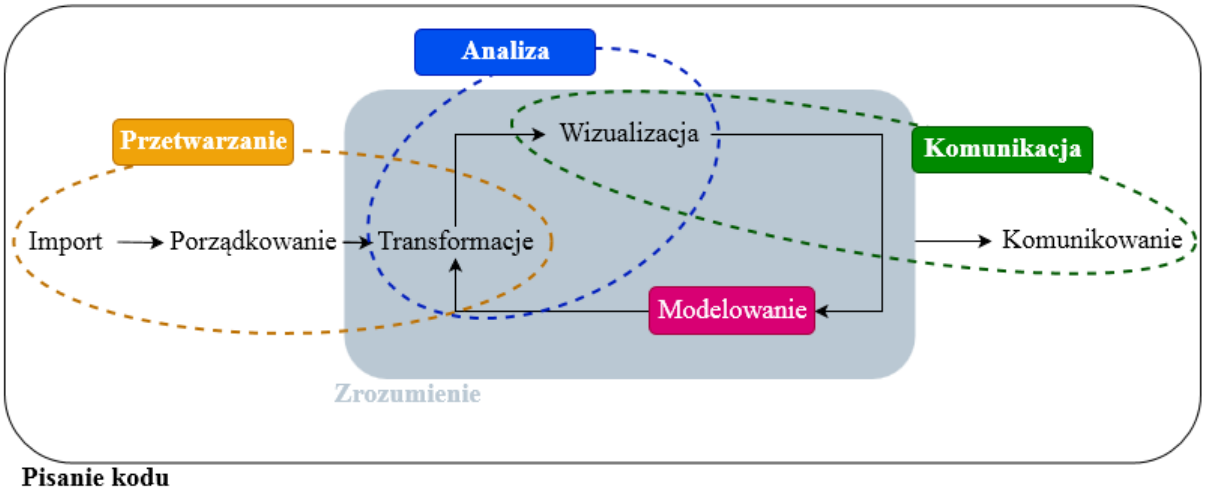
Istnieje kilka innych sposobów – innych niż ten autorstwa Wickhama i Grolemunda (2017), na których podejściu bazujemy – sposobów określenia procesu, który realizuje projekt DS. Właściwie nazwa data science nie jest tutaj zasadna, chodzi jednak bez wątpienia o projekty bazujące na analizie i modelowaniu danych cyfrowych. Wciąż popularnym jest powstały pod koniec lat 90. CRISP-DM (*cross-industry standard process for data mining* – międzysektorowy standard procesu data mining) (Azevedo i Santos 2008; Biecek 2019). Poza elementami omówionymi wyżej (przygotowaniem, analizą, modelowaniem danych i komunikacją wyników), w różnych podejściach wymieniane są następujące części: zrozumienie lub dookreślenie problemu, co równoznaczne jest ze zdefiniowaniem celu projektu; testowanie lub ewaluacja uzyskanych wyników; wdrożenie wyników (ibidem). Nie ujęliśmy, a dokładniej – nie wyraziliśmy wprost elementu testowania czy ewaluacji wyników, ponieważ w naszym rozumieniu taki element oceny jest integralną częścią każdego z omawianych elementów procesu. To na podstawie takich ocen podejmowane są decyzje o

podejmowaniu kolejnych iteracji albo o przejściu do kolejnego etapu. Nie ujęliśmy elementów: określania celu projektu ani wdrożenia wyników, ponieważ niekoniecznie są one postrzegane jako to, czym zajmuje się badany świat. Bez wątpienia ani definiowanie celów, ani wdrożenie wyników nie są składowymi działania podstawowego.

Podsumowując nasze ujęcie, **w ramach działania podstawowego świata społecznego DS w Polsce realizowany jest proces**, na który składają się:

1. w fazie pierwszej kolejno operacje importu, porządkowania i transformowania danych, co określamy **przetwarzaniem danych**;
2. w fazie drugiej iteracyjnie i niekoniecznie następujące po kolei operacje transformowania i wizualizacji danych na potrzeby analityczne, co określamy **analizą danych oraz modelowanie danych**.
3. W fazie trzeciej **komunikowanie wyników** uzyskanych w fazach 1. i 2., **może być częściowo uznawane za działanie podstawowe, częściowo za działania wspomagające**.

Prezentujemy ten proces na rysunku 24.



Rysunek 24: Operacje wykonywane na danych z użyciem programowania jako proces

Źródło: opracowanie własne na podstawie badań etnograficznych oraz tzw. R4DS (Wickham i Grolemund 2017) za pomocą draw.io

Całość procesu jest zapisywana w kodzie – ramka „pisanie kodu”[150] otacza wszystkie elementy procesu (rys. 24). Stąd nasze ujęcie działania podstawowego badanego świata jako **pisania kodu do przetwarzania, analizy i modelowania danych**.

5.2.2. Data science – komercyjne, akademickie, hobbystyczne – zapisany w kodzie proces operacji na danych

Twierdzimy, że **w polskim świecie społecznym DS osoba może zostać uznana za uczestnika tego świata przez innych uczestników tylko, jeżeli wykaże się doświadczeniem w pisaniu kodu do przetwarzania, analizy i modelowania danych, bez względu, czy jest to doświadczenie komercyjne, akademickie czy hobbystyczne**. Chodzi o pisanie kodu w ramach wszystkich trzech wyróżnionych obszarów: przetwarzania i analizy i modelowania, w jednym procesie, by osiągnąć postawiony cel. To określane jest jako przeprowadzenie projektu od początku do końca, od a do z:

> *B: Mhm, ok, ale to nie zgadzasz się z tym, że tylko ci którzy pracują komercyjnie są eee data scientist?*
> R: Eee nie, zdecydowanie się nie zgadzam, ze względu na to że ludzie którzy pracują komercyjnie bardzo często **korzystają** z dorobku osób które pracują naukowo, bądź też są to osoby które eee łączą pracę eee i w biznesie i nauce, to jest bardzo częsty przypadek naaa, na przykład na uniwersytetach technicznych czy ekonomicznych i to w ogóle, myślę że bardzo poszerza perspektywę. Eee także, ja w ogóle nie, nie lubię takich sztywnych etykietek, może inaczej wtedy kiedy one są konieczne, tak, bo jak definicja jest ostra to jak najbardziej, natomiast tutaj wydaje mi się że to niczego nie wnosi, eee ograniczanie eee właśnie tylko do jednego obszaru, aaa zaczęłam mówić o tym że, że korzystają z dorobku, bo bardzo częstooo eee żeby zbudować jakiś model, eee który chcesz ymm potem zaaplikować do jakichś swoich **własnych** zastosowań, albo chcesz stworzyć yyy, nie wiem jakiś start-up, czy czy czy jakieś rozwiązanie techniczne to musisz korzystać najpierw z jakichś danych wejściowych nie? Żeby żeby móc to skonceptualizować, czy czy przetestować, czy przetrenować jakiś model, a potem przetestować go na jakichś no nowych danych. No i bardzo często to jest tak, że to są dane po prostu zebrane przez naukowców w jakimś jednym ośrodku, czy przez konsorcjum, eee ośrodków tak, dla przykładu ludzie którzy pracują z danymi językowymi, z przekładaniem języka naturalnego, bardzo często opierają się na korpusach stworzonych przez językoznawców na uczelni, nie. No i przykłady można by mnożyć, nie? (...) Także wydaje mi się że to się bardzo yyy bardzo przenika i wzajemnie stymuluje ten, ten rozwój do dalszych działań.
> *B: Mhm, mhm, ok. Jeszcze przy okazji spytam, a czy sądzisz że [krótka pauza] też ktoś kto tak **hobbystycznie** by się zajmował analizą danych, to też jest wtedy data science, czy czy czy data scientist?*
> R: Mmm, hobbystycznie? No wydaje mi się że zależy co przez to rozumiemy, boo to czy ktoś siedzi w domu i robi, e projekt badawczy na własnym komputerze, czyy, czy w jakimś, wiesz, ekskluzywnym biurowcu przy firmowym sprzęcie, to nie definiuje tego w jaki sposób pracuje z konkretnym problemem ii jakie udaje mu się osiągnąć rezultaty, więc jeżeli ktoś robi hobbystycznie ale jest w stanie poprowadzić cały projekt od a do z, w takim rozumieniu że uczy się metod analizy danych, formułuje sobie problem na podstawie danych do których ma dostęp,

150 Oryginalnie użyto słowa „programowanie” (*programming*) (Wickham i Grolemund 2017), ale pozostajemy przy określeniu „pisanie kodu”, by zaznaczyć odrębność od programowania w sensie budowy oprogramowania (*software development*)

> jest w stanie efektywnie tworzyć na przykład jakieś modele predykcyjne, no to, no to jak najbardziej. Natomiast yyy bardzo często hobbystyczne zajmowanie się tym, eee wiąże się po prostu z, eee takim etapem bardzo wstępnym, nie? Robieniem sobie jakichś początkowych kursów eee analizy danych tak,? Czy nie wiem, początkowym etapem nauki programowania i tak dalej. To jak najbardziej wszystko jest w porządku, no [westchnienie] czy, czy [śmiech] nie wiem no, nie chcę nikomu odbierać prawa do identyfikacji tak, w tym obszarze. No wydaje mi się że jak już jest osiągnięty jakiś poziom kompetencji właśnie, eee przejścia przez wszystkie etapy realizacji takiego projektu, zrozumienia tego co się robi, no to wtedy, wtedy jak najbardziej [ze śmiechem] można się określić jako data scientist tak, hobbysta. Czemu nie? Czemu nie? {wywiad 21}

Jeżeli osoba nie przeprowadziła przynajmniej jednego projektu w całości, zgodnie z wyżej omówionym sensem, ale wykonywała części działania podstawowego, może być ona uznana za uczestnika jednego ze światów przecinających się z DS. Przykładowo dla osób koncentrujących się na przetwarzaniu i przechowywaniu danych będzie to świat inżynierii danych (*data engineering*), dla koncentrujących się na analizie – świat analityki danych, dla koncentrujących się na modelowaniu świat nauki akademickiej (chodzi przede wszystkim o rozwijanie metod uczenia maszynowego). Istnieje także grupa tzw. wannabesów[151], skoncentrowanych na uczeniu się DS, a szczególnie modelowania – to o nich mówi Rozmówczyni: „bardzo często hobbystyczne zajmowanie się tym, eee wiąże się po prostu z, eee takim etapem bardzo wstępnym, nie?" {wywiad 21}. W grupie wannabesów można wyróżnić nierozłączne grupy studentów, tzw. graczy w Kaggle lub kursantów MOOC, których ogólnie traktujemy jako aspirujących do uczestnictwa w badanym świecie.

Przy tym **nie każdy uczestnik świata społecznego DS uznaje się sam lub jest uznawany przez innych uczestników świata za data scientista**. Dotychczas używaliśmy tych określeń zamiennie, należy jednak to doprecyzować. Z naszych badań wynika, że **określeniem data scientist są bez wątpliwości nazywane osoby realizujące komercyjnie opisane wyżej działanie podstawowe**. Tym samym „data scientist" to w Polsce raczej nazwa stanowiska pracy lub zakresu obowiązków zawodowych niż nazwa określająca każdego uczestnika świata społecznego DS. Stwierdzenie „zajmuję się data science, ale nie jestem data scientistką/scientistem" jest całkowicie dopuszczalne. Dzieje się tak również w przypadku osób, które realizują działanie podstawowe komercyjnie. Nazwy stanowisk pracy takich osób mogą być różne od „data scientist", osoby te mogą także z różnych powodów unikać określania siebie jako data scientist. Określenie to może być postrzegane jako tytuł nic nieznaczący, przereklamowany lub zdezawuowany, o czym więcej powiemy w rozdziale 6.

151 Wyjaśnimy to określenie szerzej w rozdziale 6.

Działanie podstawowe może być realizowane przez tę samą osobę w ramach działalności komercyjnej, akademickiej, hobbystycznej równolegle lub jednocześnie, możliwe są także wszystkie inne połączenia tych sfer działalności w ramach działania jednej osoby. Jak wspominaliśmy, koncentracja działania osoby na jednym z elementów działania podstawowego – przetwarzaniu, analizie lub modelowaniu danych – może wskazywać na jej przynależność do świata/subświata przecinającego się ze światem DS (np. inżynieria danych, analityka danych, nauka akademicka, wannabesi). Pogłębimy to zagadnienie w części 6.1., gdzie opisujemy procesy wyznaczania granic badanego świata.

Mówimy także o DS hobbystycznym, podkreślmy jednak, że w toku badań **nie spotkaliśmy się z osobą zajmującą się DS wyłącznie hobbystycznie**. Istnieje w międzynarodowym środowisku DS nie tłumaczone pojęcie *citizen data scientist*, które oznacza jednocześnie hobbystę i podejmowanie działań w ważnej sprawie obywatelskiej, niemniej w Polsce nie spotkaliśmy osoby definiującej się w ten sposób. Oczywiście brak dowodu nie jest dowodem braku, zatem mogliśmy po prostu nie dotrzeć do takich ludzi. Twierdzimy jednak, że działanie podstawowe świata DS jako hobby podejmowane jest raczej równolegle bądź zamiennie z tym samym działaniem jako praca akademicka lub komercyjna. Osoby, które w czasie wolnym próbują swoich sił w różnych elementach działania podstawowego świata DS są w Polsce raczej określanie jako wannabe czy w ogóle nie mające wiele wspólnego z badanym światem – szczególnie, kiedy nie biorą udziału w dyskursie badanego świata.

5.2.3. (Re)Konstrukcja sposobów definiowania działania podstawowego

Nasze ujęcie działania podstawowego wzorowane jest na definicji z pierwszego sondażu Kaggle (2017) „The State of Data Science and Machine Learning". Napisano tam, że „data scientist to ktoś, kto używa kodu do analizy danych (*we define 'data scientist' as someone who uses code to analyze data*)" (ibidem). Podstawą tego ujęcia jest całość badań własnych, od końca 2016 do połowy 2019 roku (por. Aneks), przy znaczącej roli konsultacji z Rozmówczyniami/Rozmówcami.

W proponowanym tutaj ujęciu działanie podstawowe nie jest etnograficzną rekonstrukcją definicji DS ani proponowaną przez nas z pozycji naukowego autorytetu definicją tej dziedziny. Nasze ujęcie nie odnosi się do wszystkiego, co określa się jako DS, ani nawet tego, co z punktu widzenia uczestników badanego świata może być w DS

najważniejsze. Jest to zatem nasza konstrukcja, bazująca na badaniach terenowych i konsultacjach z uczestnikami osadzonymi w polskim świecie DS. **Twierdzimy jednak, że nasze ujęcie działania podstawowego jest trafne w tym sensie, że osoba, która nie realizuje takiego działania nie może być uznana za uczestnika badanego świata**. Nie zachodzi odwrotność powyższego zdania na zasadzie: „każda osoba, która realizuje … będzie uznana za uczestnika", wyłącznie „może być uznana ...". Granice świata społecznego są bez wątpienia zamazane, niechlujne, rozmyte (Clarke 1991), a problem tożsamości uczestnika złożony, o czym więcej w części 6.1.

Rozmaite definicje DS (Doran 2018) koncentrują się na trzech różnych aspektach:

- interdyscyplinarności DS, co w odniesieniu do osoby data scientista przedstawiane jest jako zestaw wiedzy i umiejętności tej osoby oraz zestaw technologii, którymi się ona posługuje;
- utylitarnego celu prowadzenia projektów DS, co określane jest np. jako „użyteczne wyniki", „dostarczanie wartości biznesowej", „wydobywanie wiedzy", „rozwiązywanie praktycznych problemów", „dokonywanie zmiany w organizacji", „wsparcie decyzji biznesowych";
- danych cyfrowych, co jest uznawane w badanym świecie za tak oczywiste, że nie podlega refleksji – różne może być ujęcie co do ilości danych (upraszczając DS jest wtedy, kiedy danych jest dużo) lub struktury danych (także upraszczając DS to przynajmniej częściowo posługiwanie się danymi nieustrukturyzowanymi, tzn. obrazem, tekstem, dźwiękiem).

Nasze ujęcie działania podstawowego odnosi się do umiejętności ludzi i niezbędności danych cyfrowych, a pomija cel projektów, który odniesiemy do wartości społecznego świata w części 5.4. Do konsultacji z Rozmówczyniami/Rozmówcami przystąpiliśmy jednak z następującą propozycją ujęcia działania podstawowego: „**rozwiązywanie problemów** za pomocą danych, kodu i metod analitycznych". Pytanie otwierające konsultacje[152] oceniamy z perspektywy czasu jako nieudane, ponieważ wskazuje na chęć uzyskania esencjalistycznego ujęcia, co mogło być odebrane jako wskazanie najważniejszego elementu DS. Niemniej

152 Dokładna treść wiadomości to: Cześć, piszę z prośbą o konsultacje do mojej pracy socjologicznej o data science. Chcę odpowiedzieć na pytanie: "co ludzie robią w data science?". Chodzi zarówno o ds komercyjne jak i akademickie i hobbystyczne. Propozycja: "rozwiązują problemy za pomocą danych, kodu i metod analitycznych." Oczywiście ludzie robią też dużo innych rzeczy, ale czy dla Ciebie to ujmuje esencję data science?

doprowadziło nas ono do cennych wniosków. Część Rozmówczyń/Rozmówców wskazywała więc na cel analiz, ale tłumacząc, że nie zawsze chodzi o rozwiązywanie problemów:

> R: Rozszerzył bym jakoś to "rozwiązują problemy". Bo często chodzi o odkrycie czegoś o czym nikt nie miał pojęcia, więc nawet nie było to postrzegane jako problem / zagadnienie. {konsultacja na LinkedIn z R w sprawie działania podstawowego – wywiad 16}

Wskazywano nam także, że określenie „rozwiązywanie problemów" jest biznesową kliszą, którą bezrefleksyjnie powtarzamy:

> R: jak chodzi Ci o jedno data science to "rozwiązują problemy za pomocą danych, kodu i metod analitycznych." ujdzie, ale wg mnie nie jest idealne. (Też: same metody analitycznie na pewno nie. Jako dodatek czasem.) O ile ciężko streścić coś to sugerowałbym na szybko: "Stosowanie programowania i statystyki do przetwarzania danych oraz tworzenia modeli predykcyjnych." czy coś w tym stylu. Co jest kluczowe: dane są to na pierwszym miejscu. Programowanie i statystyka są narzędzeniem, i koniecznym. (Bez tego to business analytics / klasyczny big consulting.)
> *B: Wychodzę z tym "rozwiązywaniem problemów", ponieważ tłumaczono mi że DS to nie jest analiza dla samej analizy, sztuka dla sztuki - że chodzi o dostarczenie rozwiązania, które ma jakąś wartość. Czy to ma sens według Ciebie?*
> R: Niby tak, ale nudzi mnie jak to powtarzają. W biznesie to można prowiedzieć o każdym zadaniu - też grafik nie rysuje kresek do galerii, albo po zachęcą klientów/inwestorów lub ułatwią użytkowanie produktu. "Writes code to analyze data" chyba najlepsze i najogólniejsze. {konsultacja email z R w sprawie działania podstawowego – wywiad 26}

W toku konsultacji wpadliśmy na pomysł stosowania porównania data scientista i jego działania do kucharza i gotowania. Uświadomiliśmy bowiem sobie, że Rozmówczynie/Rozmówcy w przeciwieństwie do nas nie myślą kategoriami działania podstawowego z teorii światów społecznych (sic!), i powinniśmy zadbać o bycie zrozumianymi. Porównanie uznajemy za informatywne, bo naszym zdaniem kucharz[153] może być np. gotującym w domu amatorem, szefem kuchni w pięciogwiazdkowym hotelu lub pracować w przydrożnym barze, ale każdy kucharz gotuje. Uznaliśmy to za wystarczające dla potrzeb konsultacji ujęcie działania podstawowego. W tym kontekście Rozmówczynie/Rozmówcy akceptowali definicję z Kaggle 2017, ewentualnie mówiąc że jest za szeroka lub nie odnosi się co celu:

> *B: Chcę odpowiedzieć na pytanie: "co ludzie robią w data science?". Chodzi zarówno o ds komercyjne jak i akademickie i hobbystyczne. Propozycja: "rozwiązują problemy za pomocą danych, kodu i metod analitycznych." Oczywiście ludzie robią też dużo innych rzeczy :) ale czy dla Ciebie to ujmuje esencję data science? Liczę na krytykę i z góry dziękuję za wszelkie komentarze. Pozdrowienia! PS a może w ogóle to pytanie jest bez sensu i należy podejść do problemu zupełnie inaczej? (...)*
> R: Czyli co ludzie robią w data science w sensie jak można streścić ich pracę? Czy chodzi o to

153 Badając socjologicznie ten wycinek rzeczywistości społecznej należałoby mówić raczej o społecznym świecie gotowania, ale pomijamy to celowo.

> dlaczego ludzie zabierają się za data science?
> *B: Streścić. Tak jak kucharze gotują*
> R: aś wymyślił zadanie ;D nie mam pojęcia jak to streścić ;D żeby jeszcze pasowało do biznesu, akademi i hobbistów
> *B: to może to zadanie jest bez sensu? ;)*
> R: To co zaproponowałeś jest dosyć odpowiednie
> *B: Hmm, coraz bardziej wątpię :) A jak oceniasz definicję ds z ankiet Kaggle "writes code to analyse data"?*
> R: No to wtedy jest takie 1 do 1 z tym że kucharz gotuje. Ale brakuje mi elementu że kucharz gotuje żeby smakowało {konsultacja na LinkedIn z R w sprawie działania podstawowego – wywiad 15}

Powyżej widać, że obstawaliśmy jeszcze w „streszczaniu", ale ta krótka rozmowa była dla nas przełomowa. Koncentracja na celu czy konkretnie rozwiązywaniu problemów byłaby z naszej strony nieprawidłowym ujęciem działania podstawowego także z uwagi na to, że pokazuje nie „jak jest", a „jak powinno być" zdaniem części uczestników badanego świata, a także światów przecinających się z nim, np. szeroko pojętego biznesu:

> R: (...) Ja mam dużą łatwość komunikacji z klientem biznesowym, a jednak część chłopaków, yyyyy [cmoknięcie] no nie ma takiego myślenia biznesowego, że na przykład rozmawiam z kolegą co teraz robi w pracy, z tej poprzedniej pracy, a on a robię tu taki model tam, jaki to był model? Sieci społecznych. Ooo ciekawe! A ja się zawsze zastanawiałam **po co firmom ten model sieci społecznych**. No bo tu można takie i takie rzeczy wyliczyć. No dobra ale co dalej z tym robić, po co to komu? [śmiech] Po co komu te rzeczy? I on mi powiedział, że a w sumie nigdy o tym nie myślałem. [śmiech] Nie, nie mają takiego przełożenia na biznes, na rzeczy które potencjalny odbiorca może z takim rozwiązaniem robić, może chcieć robić. {wywiad 18}

Zauważmy, że **w różnych źródłach uznaje się rozwiązywanie problemów za to, czym zajmują się dziedziny, na przecięciu których powstało DS: informatyka** (*computer science*) (Gallopoulos, Houstis, i Rice 1994; Lanthier 2011; de St. Germain 2008) i **matematyka** (Schoenfeld 1992:339; Skura i Lisicki 2018). W badanym świecie wskazuje się, że DS jest po prostu innym sposobem rozwiązywania problemów niż informatyka (Kuncewicz 2019a), a także, że dookreślenie rozwiązywanego problemu jest pierwszym i najważniejszym etapem projektu DS (Alekseichenko 2019b). Hasło „walki z rzeczywistymi problemami" (*fight with real-world big data problems*) może zachęcać do udziału w konkursie DS, w warsztacie, w rekrutacji (OLX 2018). Słynny w badanym świecie zespół Google Brain przedstawiał, jak można wykorzystać uczenie maszynowe i TensorFlow do rozwiązywania trudnych problemów (*solve challenging problems*), odwołując się do dokumentu „Wielkich Wyzwań Inżynierii na XXI wiek" z 2008 r. (*the National Academy of Engineering's Grand Engineering Challenges for the 21st Century*). Mowa o rozwiązaniach dla służby zdrowia, robotyki i narzędzi naukowych odkryć (Dean 2019). Już pod koniec lat

70., w pracy socjologicznej dotyczącej świata komputerów (*computer world*) – świata w rozumieniu światów społecznych A.L.Straussa – wskazano, że

> Kluczową charakterystyką świata komputerów jest jego koncentracja (*inherent focus*) na uogólnionym rozwiązywaniu problemów (*generalized problem solving*) jako aktywności centralnej (*central activity*) [tłum. własne RŻ] (Kling i Gerson 1978:27).

Badający rosyjskich data scientistów Lowrie stwierdził, że z ich perspektywy za prawdziwego data scientista są uznawane osoby mające „dobre podejście do rozwiązywania problemów", a samo DS za dziedzinę, która w odróżnieniu od matematyki i informatyki (*computer science*) „naprawdę" zajmuje się rozwiązywaniem problemów (Lowrie 2018).

Właściwie tak samo opisywano działalność naukową jako taką. Już we wczesnej pracy posługującej się ramą teoretyczną społecznych światów A. L. Straussa określano badania naukowe jako prace nad rozwiązaniem problemu (*problem-solving work*) (Gerson 1983:358–59). Sama A. E. Clarke uznała, że „praca naukowa koncentruje się na rozwiązywaniu problemów (*scientific work, then, centers on problem solving*)" (Clarke 1997:71), a data scientiści mogą postrzegać swoją pracę jako co do zasady pracę naukową, choć i tak niejasne rozróżnienie na naukę i inżynierię jest w przypadku DS mało istotne (Lowrie 2017, 2018).

Bez wątpienia badany przez nas świat poświęca wiele uwagi kategorii „rozwiązywania problemów", cokolwiek (i dla kogokolwiek) by to nie znaczyło, ale naszym zdaniem nie można w ten sposób ująć jego działania podstawowego. Proponujemy spojrzeć na omawianą kategorię jako z jednej strony praktykę legitymizowania świata DS w skali makro (wtedy, gdy komunikat o rozwiązywaniu problemów płynie na zewnątrz świata), z innej strony jako kryterium oceny działania podstawowego i osoby je podejmującej (czyli „tak powinno" się wykonywać działanie podstawowe, by rozwiązywać realne problemy, a osoba, która tak działa, może być uznana za prawdziwego data scientista). Skoro jest to kryterium oceny działania uczestników świata społecznego, to przede wszystkim traktujemy tę koncentrację na rozwiązywaniu problemów jako wyraz wartości badanego świata (por. część 5.4). Zauważmy także w tym kontekście arenę dotyczącą wykonywania działania podstawowego – czy uczestnicy realizują działanie podstawowe żeby rozwiązać problem, czy zrobić model? (por. powyżej fragment {wywiad 18}). Niemniej ta arena jest zdominowana przez pozycję „rozwiązać problem", a przynajmniej ta pozycja jest zdecydowanie silniej widoczna dla nas, jako kogoś bardziej z zewnątrz niż z wewnątrz badanego świata DS.

Podczas konsultacji Rozmówcy wskazywali nam także, że pisanie kodu nie jest w DS przez cały czas niezbędne, że kod jest tylko narzędziem, że powstają rozwiązania z interfejsem graficznym, które pozwalają wyklikać cały proces, jak np. oprogramowanie Orange[154]. Twierdzimy jednak, że **pisanie kodu postrzegane jest jako jedyny właściwy sposób wykonywania działania podstawowego** w badanym świecie (por. rys. 23. oraz Mason 2010; Wickham 2018d). Postulat używania kodu m.in. dla powtarzalności wykonanych operacji wskazuje na wartości badanego świata, jak efektywność (por. 5.4.). Nie znaczy to, że osoba, która użyje oprogramowania analitycznego z interfejsem graficznym zostanie napiętnowana czy poddana ostracyzmowi ze strony innych uczestników świata DS w Polsce. Uczestnicy używają takich narzędzi jak arkusz kalkulacyjny czy oprogramowanie do wizualizacji danych doraźnie i w sposób zgodny z wartościami badanego świata, np. kiedy jest to postrzegane jako bardziej efektywne niż pisanie kodu:

> R: Yyy, w pewnych sprawach Python jest lepszy, w pewnych sprawach R jest lepszy. (...) Poza tym, to też nie jest tak, że obowiązkowo trzeba korzystać tylko z tych dwóch narzędzi. Do takiej wstępnej analizy danych jak chce się popatrzeć na zmienne, to **żadne** z tych narzędzi nie jest dobre. To trzeba by użyć na przykład Tableau albo na przykład jakiegoś klikadełka typu QlickView, ewentualnie napisać sobie własny algorytm, który jest uniwersalny i pasuje do wielu zbiorów danych. {wywiad 5}
>
> R: ja korzystam z Excela, ale do takich rzeczy nie wiem jak gdzieś mam, muszę wyjechać, policzyć koszty eee [pauza] no bo tak naprawdę jeśli, można by popatrzeć do czego mógłbym używać Excela. I ja tak naprawę teraz w tym momencie nie jestem w stanie żadnego sensownego wykorzystania Excela znaleźć, to jest [pauza] typowo taki menadżerski, narzędzie, bo jeżeli chcę coś sobie zwizualizować, to nie używam, na szybko to użyję Power BI, Tableau, eee albo jakiegoś każdego innego narzędzia, no Power BI jest chyba najbardziej sensowny, więc yyy [krótka pauza] albo nawet jak mam po prostu jakiś problem to wizualizuję po prostu w Pythonie tak, typowo tak, ale jeżeli chcę coś szybko utworzyć to nie ma wstydu żadnego w używaniu Excela (…) {wywiad 12}

Kwestia arkuszy kalkulacyjnych jest bardziej złożona, bo choć „nie ma wstydu żadnego" to podczas obserwacji zauważyliśmy, że słowa Excel czy pivot[155] powodują uśmiechy zgromadzonych (co rozwiniemy w części 6.2.3). Chcemy w tym miejscu powiedzieć jedynie, że uczestnicy badanego świata nie posługują się wyłącznie kodem, ale doświadczenie w pisaniu kodu do opisanych już operacji na danych jest warunkiem

154 https://orange.biolab.si/, dostęp 19.04.2019 r.; zbliżone rozwiązania to np. Caffe, WEKA, Rattle, KNIME, RapidMiner. Pewne rozwiązania mają także duże firmy tworzące oprogramowanie, jak obecny dostawca całej gamy produktów SPSS firma IBM, firmy SAS Institute, Microsoft czy SAP, oraz inne, np. Dataiku, Alteryx, DataRobot (Gualtieri i in. 2018; Muenchen 2019). W polskim świecie data science część z nich jest postrzegana jako narzędzia innych światów (IBM, SAS, SAP, Microsoft), inne są używane marginalnie i nie traktuje się ich stosowania jako właściwego sposobu wykonania działania podstawowego.

155 Czyli tabela przestawna w arkuszu kalkulacyjnym, typowy sposób podsumowywania danych w takich narzędziach

niezbędnym bycia uczestnikiem. Mówimy o doświadczeniu, bo zauważyliśmy, że **uczestnicy nie kwestionują przynależności do badanego świata osób doświadczonych, które w danym momencie nie zajmują się własnoręcznym pisaniem kodu lub robią to bardzo rzadko**. Chodzi np. o osoby na stanowiskach kierowniczych, właścicieli firm lub utytułowanych naukowców, o ile oczywiście te osoby chcą być częścią świata DS i uczestniczą w jego dyskursie. Działania takich osób w sensie liczby godzin składać się mogą w dużej części ze np. spotkań, wystąpień publicznych, wykładów, pisania rozmaitych tekstów. Niemniej inni uczestnicy wiedzą, że te osoby są doświadczonymi koderami, mogą np. korzystać z ich pakietów, wykorzystywać fragmenty kodu z ich publicznie dostępnych repozytoriów (popularnym jest GitHub, por. część 5.3.4.), mogą czytać ich książki/artykuły z przykładami kodu, mogą oglądać ich wystąpienia publiczne także prezentujące fragmenty kodu, mogą przekazywać im swój kod do recenzji. Z drugiej strony jesteśmy pewni, że osoba, która wykonywała i wykonuje operacje na danych wyłącznie za pomocą interfejsów graficznych, (np. w zestawie Microsoft Office: Access, Excel i PowerPoint) nie będzie uznana za uczestnika świata DS.

5.2.4. Przykłady zastosowania data science

Teoretyczne ujęcie działania podstawowego ilustrujemy przykładami zastosowania DS w różnych sferach życia. Sądzimy, że pomoże to **zrozumieć, co robią (a czego nie) osoby zajmujące się DS**. Wiele przykładów zawarto już w części 2. tej pracy (szczególnie 2.2.3.), jednak tam miały one inny cel. Tutaj prezentujemy przykłady ikoniczne, wyraziste, najbardziej znane nie tylko w badanym świecie, ale także poza nim. W naszej opinii prześledzenie ich pokazuje nie tylko co robią (a czego nie) uczestnicy badanego świata, ale i **jak robią** – jakie są kryteria wartościowania określonych działań i powiązanych z tymi działaniami tożsamości.

Takie paralegendarne historie mają znaczenie dla:

> konstruowania wzorcowych tożsamości, bohaterów społecznego świata lub areny. Same wartości są wplecione w narracje i przekazywane w licznych opowieściach, których tematem jest kształtowanie określonego rodzaju tożsamości. Wartość pojawia się w kontekście opowieści i zawsze wiąże ze sobą jakiś wzór zachowania. W opowiadaniach o działaniach innych zawierają się konkretne oceny – osądy samych działań oraz powiązanych z nimi autoidentyfikacji. Funkcjonujące narracje wypełniają świadomość uczestników, stanowią tło moralne dla toczących się debat oraz opis porządku moralnego (Kacperczyk 2016:46).

Za każdym razem **mamy do czynienia z rozwiązaniem budowanym na bazie danych cyfrowych i operacji ich przetwarzania, analizy i modelowania**. Jak pokażemy, **ta część bazowa jest obudowywana zestawem innych technologii** (np. wtedy, gdy gotowy model uczenia maszynowego służy do poprawy obrazu z aparatu fotograficznego zainstalowanego w smartfonie).

> Ds team [tzn. zespół data science] robi tylko małą część systemu, który dostaje klient {notatka terenowa, obserwacja 36}

Tą **obudową badany świat nie zajmuje się w ramach działania podstawowego lub nie zajmuje się nią wcale**. Przykłady dzielimy na trzy kategorie: wgląd (*insight*), model, produkt.

W kategorii wglądu przykłady sięgają czasów popularności określenia data mining i to tutaj w kontekście wglądu mówiono o odkrywaniu wiedzy z danych (czy z baz danych). To „odkrywanie" mamy na myśli mówiąc o wglądzie (*insight*) – odnalezieniu w danych informacji, która nie była wcześniej znana, a może być użyteczna. Stosowane mogą być różnego rodzaju metody analizy i modelowania, także te najprostsze, jak podsumowanie danych za pomocą częstości i statystyk opisowych w grupach. Szeroko powtarzany był przykład, w którym amerykańska sieć sklepów wielobranżowych Target identyfikowała ciężarne klientki, w celu skierowania do nich oferty promocyjnej. Identyfikacji służyły takie informacje, jak kupowanie dużych ilości bezzapachowych produktów do nawilżania skóry, bezzapachowego mydła, dużych opakowań kulek bawełnianych czy suplementów cynku, magnezu. Andrew Pole, nazywany wówczas statystykiem firmy Target wskazywał, że można oszacować datę rozwiązania ciąży dla każdej ciężarnej klientki, i kierować do niej ofertę dostosowaną do danego stadium zaawansowania. Media powtarzały historię, w której ojciec nastolatki interweniował w jednym ze sklepów, pokazując kierownikowi placówki papierową gazetkę z promocjami produktów dla ciężarnych, zaadresowaną do jego córki. Mężczyzna, wściekły miał wykrzyczeć, że Target zachęca nastolatkę do zajścia w ciążę. Po kliku dniach ten mężczyzna telefonicznie przepraszał kierownika sklepu za zajście, mówiąc że w jego domu działy się rzeczy, o których nie wiedział. Ukuto w mediach hasło „Target wiedział o ciąży nastolatki wcześniej niż jej ojciec" (Hill 2012). Odniesienia do tej historii znajdują się np. w materiałach do nauki analizy i modelowania danych (Foreman 2017:229-312), zaś w amerykańskim środowisku DS bywała ona przykładem działalności wątpliwej etycznie (Brooks 2014:139; Gutierrez 2014:176, 321). Zauważmy, że w tym przypadku wgląd jako rezultat pracy data scientistów czy statystyków to informacja, z której skorzystanie zależy od

innych ról w organizacji – np. działu marketing przygotowującego ofertę, działu sprzedaży decydującego o promocyjnych cenach, działu obsługi klienta rozsyłającego listy itp. Sam wgląd bez dalszych działań nie ma wpływu na wyniki finansowe firmy.

Podobny jest przykład, w którym odkryto, że osoby kupujące wieczorem w supermarketach pieluszki dziecięce zazwyczaj kupują też piwo. Pierwsze wzmianki o tym, że jest to powszechnie powtarzana historia pojawiły się w 1996 r., w 1999 pisano już, że jest to historia najprawdopodobniej zmyślona, w 2006 r. nazwano ją mitem. Wyjaśniano odkrytą zależność tym, że przemęczeni, młodzi ojcowie zmuszeni do zakupu pieluch wieczorem wynagradzają sobie ten trud alkoholem do wypicia w domu. W różnych wersjach historii młody wiek i płeć męską miały wskazywać dane, a także zależność miała być widoczna w czwartki i piątki (Power 2002; Whitehorn 2006). Przykład ten występuje np. w materiałach do nauki modelowania reguł asocjacyjnych (Larose 2005:17, 180–81).

Na odkrywaniu połączeń pomiędzy produktami bazuje opatentowany system rekomendacji firmy Amazon (Linden, Jacobi, i Benson 2001), początkowo księgarni internetowej. Historia tego systemu posłużyła jak przykład zastosowania big data – i wsparcia tezy, że dane są lepsze od ekspertów – w wielokrotnie wcześniej wymienianej, entuzjastycznej pracy (Cukier i Mayer-Schönberger 2014). We wczesnych latach 90. polecenia książek w Amazon były recenzjami, pisanymi przez zespół krytyków literackich. Zwiększały one sprzedaż konkretnych książek, dążono jednak do spersonalizowania rekomendacji i sięgnięto do danych o transakcjach online. Początkowo badano wylosowaną, reprezentatywną próbę danych w celu wyłonienia kategorii klientów i dostosowania do nich rekomendacji. Okazało się, że rekomendacje te były bardzo powierzchowne i w efekcie wizyty w sklepie internetowym przypominały klientowi „zakupy w towarzystwie wiejskiego głupka”. Dokonano więc zmian. Zaczęto badać wszystkie zgromadzone dane, co zwane jest przez autorów podejściem N = all, szukano zaś powiązań nie pomiędzy klientami, a produktami. Badano relacje ilościowe pomiędzy wszystkimi książkami, bez względu na autora, czy gatunek literacki. Uzyskany model powiązań działał w czasie rzeczywistym, wyświetlając klientowi rekomendacje na stronie internetowej sklepu. Okazało się to metodą generującą sprzedaż większą niż recenzje krytyków. Kalkulacja kosztów doprowadziła do decyzji o zwolnieniu recenzentów i przejściu na rekomendacje na podstawie danych. Podano, że system ten generuje zakupy stanowiące ponad 30% zysków sprzedażowych Amazona, choć informacji firma nie potwierdziła oficjalnie (ibidem). Jak może ten system wyglądać z perspektywy użytkownika, mówił sam Zygmunt Bauman:

> Ilekroć wchodzę na stronę księgarni Amazon, wita mnie zestaw tytułów (…). Na podstawie historii moich wcześniejszych zakupów można założyć z dużym prawdopodobieństwem, że dam się skusić… I z reguły tak właśnie się dzieje! **Dzięki mojej sumiennej, choć mimowolnej współpracy serwery Amazona znają dziś moje preferencje i zainteresowania lepiej niż ja** [podkr. RŻ]. Nie uznaję już ich sugestii za reklamę; postrzegam je raczej jako przyjacielską pomoc, ułatwiającą wędrówkę przez dżunglę, jaką stał się rynek książki. I jestem im za to wdzięczny. A z każdym nowym zakupem uaktualniam swoje preferencje w bazach danych i niechybnie tworzę listy swoich przyszłych zakupów (Bauman i Lyon 2013:177–78).

Podkreślmy, że powyższy przykład to model działający „na produkcji" – klient wchodzi w interakcje z systemem informatycznym sklepu (czyli tzw. systemem produkcyjnym), rekomendacja za pomocą modelu wykonywana jest automatycznie, bez udziału pracowników. Surma zaznacza, że jest to zasadnicza, jakościowa różnica w porównaniu do znanych od dziesięcioleci systemów tzw. Business Intelligence (BI). BI daje wgląd w dane historyczne, jak np. zestawienia sprzedaży w podziale na kategorie produktów i regiony, co ma wspomagać proces decyzyjny ludzi. W omawianym przykładzie systemu rekomendacyjnego, a także dalszych przykładach, widoczne jest

> wyjście poza paradygmat klasycznych systemów BI w kierunku użycia tzw. złotej pętli decyzyjnej, gdzie wygenerowana propozycja nie trafia do menedżera, ale jest wykonywana automatycznie (Surma 2017:43).

Zauważmy także retorykę „algorytm jest lepszy niż człowiek": algorytm wie o ciąży córki wcześniej niż jej ojciec, algorytm rekomenduje książki lepiej niż recenzenci, algorytm zna moje preferencje lepiej niż ja sam. Obecnie Amazon sprzedaje mnóstwo innych niż książki produktów[156], do czego stosowany jest także system rekomendacji (Żulicki 2017:185).

Systemy rekomendacji znalazły zastosowanie w wielu sklepach internetowych czy platformach handlu internetowego (jak Allegro lub Ebay), a także w serwisach oferujących treści. Na modelach uczenia maszynowego bazują rekomendacje np. należącej do Google platformy wideo YouTube (Davidson i in. 2010), serwisów muzycznych iTunes i Spotify (van den Oord, Dieleman, i Schrauwen 2013), czy wypożyczalni seriali i filmów Netflix (Amatriain i Basilico 2012; Netflix 2019).

Przypadek Netflixa jest znany w świecie społecznym DS także dlatego, że algorytmy będące bazą rekomendacji powstały w rezultacie otwartego konkursu, ogłoszonego przez tę firmę w 2006. Netflix zaproponował milion dolarów dla autorów rozwiązania, które przewyższy o 10% dopasowanie prognozy oceny filmów wówczas działającego w firmie modelu Cinematch. Zwycięski zespół uzyskał rezultat lepszy o niecałe 8,5% dzięki

156 Także produktów i usług dla data science w postaci Amazon Web Services, czyli ogólnie rzecz biorąc platformy do obliczeń i przechowywania danych w chmurze. Więcej o tym w części 5.3.

kombinacji 107 algorytmów. Dział DS w Netflix wybrał z nich dwa o najwyższym dopasowaniu – SVD (*Singular Value Decomposition*; rozkład według wartości osobliwych, jest metodą służącą redukcji wymiaru danych, zatem rodzajem transformacji) i RBM (*Restricted Boltzmann Machines*; rodzaj modelu sieci neuronowej m.in. do zadań regresji). Kombinacja tych algorytmów została (jak mawiają uczestnicy badanego świata, po polsku i po angielsku) wdrożona na produkcję (*put into production* ewentualnie *deployed* lub *shipped*), czyli działała w czasie rzeczywistym w serwisie, zatem klienci Netflixa widzieli rekomendacje bazujące na działaniu tej kombinacji (Amatriain i Basilico 2012). Klienci mają możliwość samodzielnego wyszukiwania filmów, jednak około 75% nowych zamówień generowanych jest dzięki rekomendacjom (Cukier i Mayer-Schönberger 2014).

Omawiane systemy rekomendacji traktujemy jak przykład modeli (a językiem badanego świata – modeli wdrożonych, modeli na produkcji). Wytrenowana na danych, uzyskana w kolejnych iteracjach ostateczna wersja modelu uczenia maszynowego jest zapisywana do pliku np. w formacie .h5, który przechowuje architekturę modelu i wytrenowane wagi. Mówiąc potocznie, taki plik sam z siebie niczego nie zrobi. Żeby za pomocą tego modelu, automatycznie np. zarekomendować film klientowi logującemu się na swoje konto do serwisu internetowego, taki plik trzeba wmontować w infrastrukturę informatyczną serwisu. Wmontować tak, by działał ciągle, dla każdego klienta, dla wielu klientów równocześnie, odpowiednio szybko, w akceptowalnym przedziale kosztów utrzymania go, itp. Na przykładzie Netflixa zauważmy, że model zbudowany w środowisku eksperymentalnym to niekoniecznie model w środowisku produkcyjnym. Do zastosowania produkcyjnego wybrano dwa ze 107 działających razem modeli (*model ensemble*) zwycięskiego zespołu konkursowego (Amatriain i Basilico 2012). Konkurs kontynuowano, ale kolejnego zwycięskiego rozwiązania nie wdrożono na produkcję, ponieważ poprawa dopasowania modelu nie była warta wysiłku inżynieryjnego, koniecznego do wdrożenia. Dodatkowo zmienił się Netflix jako biznes – z działającej w USA internetowej wypożyczalni DVD (płyty przysyłano i zwracano pocztą!) w 2006, przez streaming wideo w 2008 w USA, Kanadzie i na wielu urządzeniach (strona www, Roku, Xbox), po 23 miliony klientów w 47 krajach i inne urządzenia (telewizory, smartfony iPhone i z systemem Android) w roku 2012. Tym trudniejsza była praca inżynieryjna, a także zmieniały się priorytety firmy[157] (ibidem).

157 Netflix, nie widząc perspektyw dalszej poprawy dopasowania modeli uczenia maszynowego zwrócił się po konsultacje do etnografa, Granta McCrackena. Etnograf zauważył (także pojawia się określenie *insight*), że klienci lubią długie sesje oglądania tego samego serialu (*binge-watch*), że sprawia im to przyjemność i nie czują się winni z powodu takich praktyk. Netflix przebudował rekomendacje i mechanizm autoodtwarzania tak, by uruchomienie kolejnego odcinka było dla klientów jak najwygodniejsze. Ten wgląd etnograficzny => c.d. s. 258

Podkreślmy, że w przypadku firm takich jak Amazon, Netflix, YouTube czy Spotify systemy rekomendacji są złożone, a modele uczenia maszynowego i inne algorytmy wykonujące operacje na danych są ich niewielką częścią. Poza tym wiele elementów serwisu jest personalizowanych w różny sposób, a te personalizacje powinny działać wspólnie – jeżeli klient Spotify kliknie, że nie podoba mu się piosenka w części „Radio", to ta piosenka nie powinna być mu proponowana w części „Odkryj w tym tygodniu" (Esh 2016). Przykładami innych, złożonych i częściowo bazujących na modelach uczenia maszynowego systemów rekomendowania czy tworzenia rankingów są, rzecz jasna, wyszukiwarka Google (Schwartz 2016) i newsfeed Facebooka (Jeffrey Dunn 2016). W przypadku Google przykłady wdrożenia modeli można by mnożyć: od słynnego i krytykowanego przez naukowców Google Flu Trends[158] przez popularnego Tłumacza Google[159], Gmail[160], do Asystenta Google, Google Home czy Google Duplex. O ostatnich powiemy później, ponieważ są to dla nas produkty z zastosowaniem DS, zatem to kolejna kategoria. Dodatkowo zarówno Google jak i Facebook, podobnie jak Amazon, oferują produkty dla DS – tzw. chmury obliczeniowe i do przechowywania danych oraz biblioteki TensorFlow i PyTorch do modelowania, o czym powiemy w części 5.3.

Jako jedno z największych osiągnięć w dziedzinie sztucznej inteligencji przedstawiane są sytuacje, w których model wytrenowany do grania w grę wygrywa z mistrzem tej gry, człowiekiem (por. Elish i Boyd 2018:59–63). Już w 1997 r. komputer IBM Deep Blue wygrał pojedynek szachowy z Garrim Kasparowem, podkreślmy jednak, że nie stosowano algorytmu uczenia maszynowego, a obliczanie i przeszukiwanie wszystkich możliwych kombinacji ruchów (*brute force*), co nie jest korzystaniem z danych i niekoniecznie jest nazywane AI (Press 2018). W roku 2010 inne dzieło IBM, Watson, zwyciężyło w teleturnieju „Jeopardy!"

=> kont. s. 257 przełożył się na poprawę wyników finansowych Netfixa i był przedstawiany jako sukces w integrowaniu wielkich danych z danymi gęstymi (*big data with thick data*, por. część 2.2.2.) przez twórczynię tego drugiego terminu (Wang 2016).

158 Serwis internetowy, który miał w czasie rzeczywistym przedstawiać rozprzestrzenianie się epidemii grypy na podstawie haseł wpisywanych w wyszukiwarkę Google. Od 2009 do 2010 r. system prognozował epidemię z dużą dokładnością, potem błędy zaczęły sięgać 140% rzeczywistego zasięgu. System nazywano *the poster child of big data* – ikoną big data, naukowcy wskazali jednak że model liniowy daje lepiej dopasowaną prognozę (Ginsberg i in. 2009; Lazer i in. 2014; Żulicki 2017:186–87).

159 Usługa początkowo działała na podstawie zaprogramowanych reguł językowych, później zastosowano uczenie maszynowe na podstawie dużej ilości nieuporządkowanych danych, co znacząco poprawiło wyniki tłumaczenia (Lewis-Kraus 2017). Zastosowano odmianę rekurencyjnych sieci neuronowych LSTM RNN, a także wyspecjalizowany hardware – procesor TPU z obniżoną precyzją obliczeń (Abadi i in. 2016; Wu i in. 2016).

160 M.in. oznaczanie wiadomości jako „ważna" czy, niedostępne dla języka polskiego, proponowanie kilku opcji odpowiedzi na otrzymaną wiadomość (*SmartReply*) (Kurzweil i Strope 2017). Poza tym filtry antyspamowe do poczty elektronicznej są jednym z najstarszych i najpowszechniejszych wdrożeń modeli uczenia maszynowego (Guzella i Caminhas 2009).

(Ferrucci i in. 2010). Serią osiągnięć szczyci się zespół DeepMind, należący do tej samej co Google spółki Alphabet: w 2015 r. ich model AlphaGo wygrał w chińską grę Go z mistrzem Fanem Hui. Podkreślano, że jest to rozwiązanie nie bazujące na zaprogramowanych regułach, i w małym stopniu na przeszukiwaniu możliwości ruchów, a składa się z sieci neuronowych częściowo trenowanych na danych historycznych, częściowo za pomocą uczenia ze wzmocnieniem (*reinforcement learning*), czyli w uproszczeniu model gra w grę przeciwko sobie. Później zespół na bazie tego rozwiązania przygotował bardziej ogólny model AlphaZero, który wykazywał bardzo dobre wyniki w grze w szachy, Go i chińskie shogi (Silver i in. 2016, 2017, 2018). Nowszym osiągnięciem w tej dziedzinie jest, przygotowany przez naukowców z Kanady i Czech przy dofinansowaniu z firmy IBM, model DeepStack, który wygrał w rodzaj pokera (*Heads-up no-limit Texas hold'em*, HUNL). Model w rozgrywkach po 3000 partii z jedenastoma pokerzystami Międzynarodowej Federacji Pokera uzyskał dla każdego przeciwnika wyniki lepszy, przy czym w 10 na 11 przypadków wynik był istotny statystycznie. Podkreślano, że HUNL ma zbliżoną do Go przestrzeń decyzji – na poziomie 10^{160} – ale poker jest grą o niedoskonałej informacji, zaś Go czy szachy to gry o informacji doskonałej. W grze o informacji niedoskonałej gracze nie mają pełnej informacji o historii ruchów swoich i przeciwnika. Tym samym DeepStack ma jakoby wykazywać się intuicją (Moravčík i in. 2017).

Innym zagadnieniem są próby generowania treści przez modele uczenia maszynowego, ale podobnym do powyższych przykładem było uczestnictwo w konkursie literackim powieści wygenerowanej przez model uczenia maszynowego, autorstwa naukowców japońskich z Uniwersytetu Hakodate. Powieść zgłoszona pod nazwiskiem prowadzącego projekt profesora Hitoshi Matsubara została zakwalifikowana przez sędziów do drugiego etapu konkursu, zatem określano ją jak nieodróżnialną od napisanej przez człowieka (Jozuka 2016).

Kategorię produktów z zastosowaniem DS zacznijmy od przywołanego wyżej poprawiania zdjęć z aparatu w smartfonie. Typowym zastosowaniem jest rozpoznawanie scenerii w kadrze zanim człowiek wciśnie spust migawki. Jest to rozszerzenie znanej dawno przed smartfonami funkcji auto w aparatach fotograficznych – automatycznego ustawiania np. ISO, balansu bieli, przesłony w zależności od otoczenia. Model machine learningowy klasyfikuje obraz z aparatu jako jedną ze zdefiniowanych scenerii (portret, krajobraz, zbliżenie itd.), i w połączeniu z informacją z innych czujników automatycznie dostosowywane są wszystkie ustawienia aparatu. Upraszczając nieco, wytrenowany model

zapisany do pliku jest w tym przypadku raczej instalowany na smartfonach, choć stosowane są też rozwiązania w których model wykonuje klasyfikację na serwerach dostawcy, a komunikacja ze smartfonem odbywa się przez internet. To pierwsze rozwiązanie jest jednak szybsze, co uznano za ważne dla użytkownika w przypadku robienia zdjęć. Producenci smartfonów, jak np. Huawei, podkreślają, że ich urządzenia są wyposażone w przeznaczone do „sztucznej inteligencji" procesory (*AI chip* lub *neural engine*), które mają przyspieszać proces działania modelu (Torres 2018). Zdjęcia już zrobione także mogą być poprawiane z użyciem modeli. Może być to np. wygładzanie skóry, zmiana kolorów. Część z tych elementów działa także dla nagrywania wideo. Tzw. aparaty ze sztuczną inteligencją (*AI camera*) są przeważnie we flagowych, czyli najdroższych i najlepiej wyposażonych smartfonach firm Samsung, Google, Apple, czy wspomnianej Huawei. Istnieją także aplikacje mobilne, które umożliwiają korzystanie z omawianych funkcji na urządzeniach nie posiadających ich fabrycznie (Gogoi 2019).

Wspomniany Google Duplex to także rozwiązanie na smartfony. Zaprezentowany w 2018 na corocznej konferencji Google I/O asystent osobisty Duplex wykonywał zlecenie głosowo – umawiał wizytę u fryzjera za pomocą inicjowanego przez Duplex połączenia telefonicznego, przetwarzając w czasie rzeczywistym mowę pracownika zakładu fryzjerskiego i generując komunikaty w języku naturalnym. Głos generowany przez Duplex określano jako nie do odróżnienia od głosu człowieka. Duplex w tej formie zautomatyzowanej rozmowy telefonicznej wzbudził szereg kontrowersji (Goode 2018; Leviathan 2018; Mandziejewski 2018). Na kolejnej konferencji I/O 2019 zaprezentowano Duplex dla internetu (*Duplex for the web*). Człowiek podaje zlecenie, jak np. rezerwacja wynajęcia samochodu, asystent wykonuje je przechodząc przez odpowiednie strony internetowe, wypełniając formularze, korzystając z innych usług Google (np. Gmail, Kalendarz Google), na koniec człowiek zatwierdza rezerwację. Usługa dostępna jest w USA od maja 2019 (Phandroid 2019).

Z Duplex i wielu innych funkcji na smartfonach z systemem Android można korzystać głosowo, tzn. nie wpisując poleceń na ekranie urządzenia, a mówiąc do mikrofonu w języku naturalnym, za pomocą Asystenta Google. Tego rodzaju asystentów oferuje także Samsung – Bixby, Apple – Siri, a dla urządzeń innych niż smartfony także Microsoft – Cortana, i Amazon – Alexa. Ogólnie rzecz ujmując, usługi te pozwalają na wykonywanie szeregu czynności w środowisku cyfrowym (np. obsługa skrzynki email, mediów społecznościowych, uzyskiwanie informacji o sporcie/polityce/pogodzie/itd., planowanie trasy dojazdu,

odtwarzanie muzyki). Mogą działać na laptopach, jak Cortana i Siri, ale także na tzw. inteligentnych głośnikach (np. Sonos One, Google Home, Amazon Echo) czy inteligentnych wyświetlaczach (np. Amazon Echo Show, Google Nest Hub, Lenovo Smart Display), stając się częścią tzw. inteligentnego domu. Możliwe jest zatem głosowe sterowanie np. ogrzewaniem, oświetleniem, alarmem itd. (Van Camp 2019; Jeff Dunn 2016; López, Quesada, i Guerrero 2018). Podkreślmy, że operacje na danych i rezultaty działania modeli uczenia maszynowego – czyli to, za co odpowiada DS – są niewielką częścią tych jakoby inteligentnych rozwiązań. Składa się na nie cały ekosystem, od infrastruktury informatycznej, przez elektronikę i obudowę samego urządzenia, przez zasoby naturalne zużywane przez centra danych „w chmurze" oraz na jego wykonanie, transport, przechowywanie, sprzedaż, po pracę ludzką na każdym poziomie wykwalifikowania i każdym elemencie łańcucha dostaw (Joler i Crawford 2018). Rozwiązania te wzbudzały zainteresowanie także m.in. z uwagi na doniesienia o przypadkach inwigilacji za ich pomocą (Sauer 2017; Socha 2018; Wiggers 2019) czy wpływu korzystania z nich na dzieci (Druga i in. 2018; Williams i in. 2018).

Zauważmy na tych przykładach, że mówi się o inteligencji niekoniecznie w sensie zadania wykonywanego przez maszynę, a wtedy, kiedy interakcja człowieka z urządzeniem odbywa się za pomocą języka naturalnego. Obecnie raczej języka w formie głosowej, tzn. mówienia do urządzenia i otrzymywania głosowych odpowiedzi, ale dawniej także w formie tekstowej. Za inteligentne uznawano pierwsze chatboty, jak np. opublikowany w 1966 r. na MIT o nazwie Eliza (Przegalińska 2016c; Sack 2018), choć bazują one na zaprogramowanych regułach, nie uczeniu maszynowym, co niekoniecznie dziś jest uznawane za sztuczną inteligencję. Przypomnijmy także, że słynny test Turinga polega na decyzji człowieka, czy partnerem interakcji, w którą wszedł on z użyciem komunikatów tekstowych, jest inny człowiek, czy komputer (Turing 1950).

> W dorocznym konkursie o Nagrodę Loebnera brązowy medal przyznaje się botowi czatowemu (konwersacyjnemu), który najlepiej przekona sędziów o swoim człowieczeństwie. Srebrny medal - kryteria jego przyznawania oparte są na oryginalnym teście Turinga - z całą pewnością nie został dotychczas przyznany. Podstawą do wręczenia złotego medalu jest komunikacja wizualna i słuchowa. Innymi słowy, SI [sztuczna inteligencja] musi posiadać przekonującą twarz i głos przekazywany za pośrednictwem terminalu i ludzki sędzia musi mieć wrażenie, że rozmawia z prawdziwą osoba przez wideotelefon (Kurzweil 2016:282).

Dalej Kurzweil twierdzi, że złoty medal może być o tyle łatwiej zdobyć, że sędziowie mogą mniejszą uwagę zwrócić na sam generowany tekst rozmowy, o ile symulacja będzie przekonująca wizualnie (ibidem). Komercyjnym symulatorem seksu czy towarzystwa, który poza głosem i twarzą mają sztuczne ciało, zatem wymiar interakcji dotykowej są tzw.

seksroboty (King i in. 2019:17–18; Realbotix 2018). Są to lalki o wyglądzie i wymiarach dorosłego człowieka. Jedyne ruchome elementy to szyja i twarz, co ma symulować mimikę, zatem interakcja dotykowa nie różni się zasadniczo od tej z innymi gadżetami erotycznymi, ruch wykonuje tylko człowiek. Użytkownicy mogą wchodzić w interakcję głosową z lalką za pomocą aplikacji na urządzenia mobilne – należy mówić do smartfona, lalka odpowiada głosowo, twarz ma mimikę, głowa obraca się. Modele uczenia maszynowego są odpowiedzialne przede wszystkim za rozpoznawanie i przetwarzanie głosu człowieka i częściowo za treść odpowiedzi, częściowo treść bazuje na ręcznie zaprogramowanych regułach. Użytkownicy mają możliwość ustawiania „osobowości" lalki za pomocą aplikacji, a także lalka ma uczyć się, „poznawać" użytkownika w toku kolejnych interakcji. Pomimo sensacji medialnych o końcu ery intymnych związków międzyludzkich badaczka interakcji człowiek-komputer wskazuje, że takie lalki pozostaną niszowym gadżetem (K. Devlin 2018).

Firma Amazon stworzyła bezobsługowe sklepy spożywcze Amazon Go:

> wplatając najbardziej zaawansowane uczenie maszynowe, wizję komputerową i AI w samą materię sklepu, żebyście nie musieli już nigdy czekać w kolejce. (…) Jak to działa? Wykorzystaliśmy wizję komputerową, algorytmy głębokiego uczenia maszynowego, czujniki, podobnie jak w samochodach samobieżnych (*self-driving cars*) (Amazon 2019a).

W grudniu 2016 uruchomiono testowy sklep wyłącznie dla pracowników Amazona, zaś w styczniu 2018 powstał pierwszy sklep ogólnodostępny. Klienci przy wejściu skanują swoje urządzenie mobilne z zainstalowaną aplikacją Amazon Go, w sklepie pakują wybrane produkty do torby i wychodzą. Płatność odbywa się automatycznie. Pracownicy sklepu przygotowują świeże kanapki lub sałatki, rozkładają towar na półkach i ewentualnie sprawdzają dowody osobiste osób sięgających po alkohol (Day 2018). Obecnie działa 13 sklepów – w Seattle, Chicago, Nowym Jorku i San Francisco (Amazon 2019).

Na koniec chcemy pokazać przykład traktowany jako historia sukcesu firmy Google. Jest on jednak inny niż powyższe, bo koncentruje się na osobie laika, który dzięki technologiom Google'a zbudował własny model, pomagający mu w pracy. Mowa o Japończyku Makoto Koike, byłym inżynierze z branży samochodowej, który zaczął pracować w gospodarstwie rolnym swoich rodziców. Sortowanie ogórków okazało się trudnym i czasochłonnym zadaniem, zatem Koike, który słyszał o możliwościach uczenia maszynowego na przykładzie Alpha Go, chciał zatrudnić do jego realizacji komputer. Dzięki temu, że Google właśnie udostępniło w wolnym dostępie (*open source*) bibliotekę Tensor Flow, mógł wykonać pierwszy prototyp. Wysoki poziom dopasowania modelu (żadnej liczby nie podano)

już na początkowym etapie prototypu miał dać Koike pewność siebie i poczucie, że jest on w stanie rozwiązać problem. Inżynier zbudował maszynę do sortowania ogórków: człowiek kładzie warzywo na półkę, miniaturowy komputer Raspbery Pi 3 wykonuje trzy cyfrowe zdjęcia i pierwszy, mniejszy model zainstalowany na komputerze klasyfikuje warzywo jako ogórek lub coś innego. Jeżeli to ogórek, zdjęcia są przesyłane przez sieć na serwer Google Cloud, gdzie na wirtualnej maszynie z systemem Linux działa drugi, większy model, który klasyfikuje ogórka do jednej z dziewięciu grup. Wynik klasyfikacji jest przesłany na Raspbery Pi 3, ten komputer steruje taśmą, po której jedzie ogórek i ramieniem, które wrzuca go do jednego z dziewięciu pudełek. Trenowanie modelu na domowym komputerze PC z systemem Windows zajęło Koike trzy dni, a dysponował on zaledwie siedmioma tysiącami własnoręcznie zrobionych i zaetykietowanych zdjęć ogórków, które były celowo zmniejszone. Użycie większych, dokładniejszych zdjęć poprawiłoby najpewniej działanie modelu – obecnie dokładność klasyfikacji na produkcji to około 70% – ale wymaga to znacznie większej mocy obliczeniowej niż domowy PC. Tu z pomocą znów spieszy Google. Serwer Google Cloud oferuje nową usługę Cloud Machine Learning, dzięki której można skorzystać z mocy wirtualnych maszyn w celu trenowania swoich modeli z użyciem TensorFlow. Usługa jest rzecz jasna niedroga, a użytkownik płaci tylko za realnie wykorzystany czas obliczeń. Koike nie mógł się już doczekać, aż ją wypróbuje (Sato 2016).

5.3. Opis technologii społecznego świata

> Bez względu na to, czy technologia zostaje 'odziedziczona', czy stanowi innowacyjny sposób prowadzenia podstawowego działania, zawsze jest uwikłana w budowanie i podtrzymywanie społecznego świata, wiedzę o stosownych miejscach i okolicznościach, w których można podejmować dane działanie, oraz o specjalistycznym sprzęcie, jakim 'należy' dysponować, aby wykonywać je prawidłowo. (Kacperczyk 2016:36-37).

Clarke i Star mówią zaś o infrastrukturze technologicznej jako o „zamrożonych dyskursach” (*frozen discourses*) (2008:115).

Pokażemy, że technologia umożliwiająca specyficzny sposób wykonania działania podstawowego jest dla uczestników bardzo ważnym czynnikiem decyzji dotyczących granic społecznego świata i subświatów. Związane jest to z procesami profesjonalizacji, tzn. specjalizowanie się w używaniu określonych technologii prowadzi do tworzenia kolejnych światów lub subświatów, czego jasnym przykładem są stanowiska pracy wyspecjalizowanej w wybranych częściach procesu DS. Technologie są nieodłączną składową procesów stawania się uczestnikiem świata społecznego DS, nieodłączną składową tożsamości

uczestnika (a zatem także praktyk zaświadczania o autentyczności uczestnika), nieodłączną składową procesów legitymizacji działań badanego świata. Technologie jako zamrożone dyskursy odzwierciedlają także wartości społecznego świata, co przedstawimy w kolejnej części (5.4.). Technologie są nieodłączną częścią procesów wyznaczania granic społecznego świata, jego legitymizacji i zaświadczania o autentyczności oraz segmentacji (6.1). W innej części (6.2) pokażemy także, że technologie są areną sporów o wartości, co wyraża się w dyskutowaniu sposobów wykonywania działania podstawowego.

Staramy się nie popełniać błędów merytorycznych, pomijając i upraszczając zagadnienia techniczne, do naszym zdaniem niezbędnego minimum – niezbędnego dla zrozumienia wywodu socjologicznego w tej rozprawie.

5.3.1. Języki programowania

Skoro działanie podstawowe społecznego świata DS w Polsce określamy jako pisanie kodu (do przetwarzania, analizy i modelowania danych), to **najważniejszymi technologiami do realizacji tego działania są języki programowania skryptowego**. Zdecydowanie najbardziej popularne obecnie języki to wielokrotnie przywoływane Python i R. Przez lata trwały dyskusje o zaletach i wadach tych języków w porównaniu do siebie (Dar 2018; DeZyre 2015; Junco 2017; Kromme 2017; Matloff 2017; Muenchen 2019; Olszewski 2017; Piatetsky 2018c; Sekercioglu 2018; Sharp Sight 2018; Silaparasetty 2018; Watson 2018), co później opisujemy jako arenę (w części 6.2.3).

W sensie ilościowym obecnie **najczęściej używany jest Python** (Nunns 2017; Piatetsky 2017b, 2018c; Strong 2018), tak na świecie, jak i w Polsce. Około 87% badanych w Polsce firm AI deklarowało, w odpowiedzi na pytanie wielokrotnego wyboru, używanie Pythona, zaś 38% używało GNU R (Borowiecki i Mieczkowski 2019:36–37). Występują zatem osoby używające obu języków programowania – około 12% badanych w międzynarodowej ankiecie KDnugetss (Piatetsky 2017b), zaś zgodnie z danymi Fundacji w Polsce odsetek ten wynosi co najmniej 25%[161].

Oba języki to otwarte oprogramowanie (*open source*). Każdy program open source jest (Kotuła 2014:13–86):

- swobodnie redystrybuowany (może być sprzedawany lub rozdawany nieodpłatnie);

161 Skoro w pytaniu wielokrotnego wyboru 87% firm podało używanie Pythona, a 38% R (Borowiecki i Mieczkowski 2019:36-37), to zakładając, że 100% używa Pythona lub R, odsetek używających obu języków wynosi 25% = 87% + 38% - 100%. Mowa o firmach, a nie osobach – z badań własnych wiemy, że w zespole jedna osoba może używać Pythona, inna R, a kolejna obu i jeszcze kilku języków programowania.

- swobodnie wykorzystywany przez kogokolwiek i w jakimkolwiek celu (komercyjnym, naukowym, rozrywkowym i każdym innym);
- dostarczany z kodem źródłowym, który można utrwalać na różnych nośnikach i który może być modyfikowany, a w konsekwencji na jego podstawie mogą być tworzone programy pochodne;
- wszystkie jego komponenty są również open source;
- może stanowić integralną i niezależną autorsko całość, wyraźnie odróżnioną od kolejnych wersji, modyfikacji i wprowadzonych poprawek;
- nie może być relicencjonowany;
- można rozpowszechniać go z programami komercyjnymi;

Języki Python i R są dystrybuowane na różnych licencjach[162], zatem w przypadku Pythona termin open source jest jedynym poprawnym, zaś dla R prawidłowa będzie też nazwa wolne oprogramowanie (*free software* lub *libre sofware*). Dla obu typów używa się szerszego określenia FLOOS (*Free/Libre and Open Source Software*). Ruch otwartego oprogramowania (*open source*) wydzielił się z ruchu wolnego oprogramowania (*free software*). Wskazuje się na różnice dotyczące filozofii działania i etyki tych podobnych ruchów (Szpunar 2008). Twórca wolnego oprogramowania uważa, że:

> Ruch wolnego oprogramowania broni wolności użytkowników komputerów. Uważamy, że niewolny program jest krzywdzący dla jego użytkowników. Obóz otwartego kodu źródłowego nie postrzega sprawy w kategoriach sprawiedliwości wobec użytkowników i opiera swą argumentację na korzyściach wyłącznie praktycznych (Stallman 2016).

Nie zauważamy jednak, by rozróżnienie na oprogramowanie otwarte (Python) i wolne (GNU R) było elementem dyskutowanym w polskim świecie społecznym DS. **Widoczne wyraźnie jest za to rozróżnienie na oprogramowanie otwarte** (*open source* czy ogólnie FLOOS)**, a oprogramowanie zamknięte** (*proprietary software*), nazywane też oprogramowaniem komercyjnym. Oprogramowanie zamknięte jest używane zdecydowanie rzadziej niż otwarte. Jest ono używane marginalnie, a jeżeli już, to obok rozwiązań open source. W badaniu Fundacji digitalpoland wskazywano wyłącznie na oprogramowanie MATLAB (12% wskazań w pytaniu wielokrotnego wyboru) oraz SAS (6%) dodając, że ten pierwszy jest stosowany przez osoby pracujące na uczelniach, zaś drugi jest znanym od dawna narzędziem analizy danych w korporacjach (Borowiecki i Mieczkowski 2019:36-37).

162 Dla Pythona od 2001 do dziś to licencja kompatybilna z GPL (*General Public License*), zaś dla R to przez cały czas licencje GNU GPL (kilka wersji dla różnych komponentów).

Podkreślmy, że oba te narzędzia umożliwiają pisanie kodu we własnych, zamkniętych językach oprogramowania (*proprietary programming language*), czyli językach MATLAB i SAS. Podobnie jest w przypadku IBM SPSS, i w toku badań spotkaliśmy się z marginalnym używaniem także tego oprogramowania w świecie DS. Piotr Migdał wskazał, że narzędzia open source są tak popularne z uwagi na ich elastyczność czy „zwinność", pożądaną w DS ze względu na szybki rozwój dziedziny:

> Obecnie najpopularniejsze [w data science] technologie są open-source, zarówno jeśli chodzi o języki, jak i poszczególne biblioteki. Można to łatwo wytłumaczyć szybkim tempem ewolucji w świecie uczenia maszynowego i głębokich sieci neuronowych. Dominują rozwiązania najbardziej zwinne (*agile*) i oferujące większość opcji. Niemniej najbardziej popularne rozwiązania (*frameworks*) łączą otwarty dostęp z bezpośrednim wsparciem giganta technologicznego. TensorFlow, Keras i TensorFlow.js są wspierane przez Google; PyTorch - przez Facebook (Borowiecki i Mieczkowski 2019:41).

Inne wymienione w raporcie języki programowania do operacji na danych to Scala (14% wskazań), Julia (6%), Octave (5%), Lua (3%) (Borowiecki i Mieczkowski 2019:36-37). W ogromnym skrócie Scala jest językiem open source, uznawanym za odpowiedni do pracy z bardzo dużą ilością danych, w dużych zespołach i generalnie we wdrażaniu operacji na danych na produkcję (Bugnion 2016). Scala częściej niż inne używane w DS języki jest wykorzystywany do pracy z technologiami tzw. big data (np. Spark/Hadoop, por. 5.3.4.) oraz z uczeniem głębokim (*deep learning*) (Piatetsky 2018d). Julia to także język open source, ma on łączyć m.in. zalety łatwości korzystania z Pythona w zakresie języka ogólnego zastosowania (*general programming*), R w zakresie statystyki, MATLABa w zakresie algebry liniowej, języka powłoki (*shell*) w zakresie sklejania różnych innych programów w całość (*glue together*), przy szybkości działania znanej z języka C. Julia jest szybsza od R kilkadziesiąt do kilkuset razy, ma być „banalnie łatwa do nauki, a jednocześnie zadowalać najpoważniejszych hakerów (*dirt simple to learn, yet keeps the most serious hackers happy*)" (J. White 2012). Octave jest wolnym (*free software*) odpowiednikiem MATLABa, znanym w badanym świecie z użycia w słynnym kursie uczenia maszynowego Andrew Ng (por. część 2.1. oraz Eaton 2019; Ng 2012). Lua to także język open source, uznawany za prosty i popularny wśród twórców gier komputerowych. W DS znany jest pakiet Torch w języku Lua do m.in. uczenia głębokiego (*deep learning*) na procesorach CPU i graficznych GPU (Collobert i in. 2019; Gutierrez 2014:56–57; Ierusalimschy, Celes, i de Figueiredo 2019).

W raporcie digitalpoland wskazano także na szereg języków programowania, które mają zastosowanie przede wszystkim do działań wspomagających lub działań nie związanych ze światem DS, w tym wdrożeń produkcyjnych modeli uczenia maszynowego. Połowa

badanych firm wskazała na używanie języków C/C++ lub C#, 38% używa JavaScript, 30% języka Java, 6% korzysta z Ruby, wymieniono także Prolog (2%) i Lisp (2%) (Borowiecki i Mieczkowski 2019:36-37). Nie wdając się w poważne różnice techniczne, C/C++ lub C# oraz Java to nie języki skryptowe (do których należą Python i R), a języki statyczne, kompilowane, do ogólnego zastosowania, używane do tworzenia oprogramowania (*software development*) (Chen 2010:7, 23–24). JavaScript i Ruby to języki skryptowe, interpretowane, dynamiczne, także ogólnego zastosowania, używane do tworzenia stron www czy aplikacji mobilnych (Paulson 2007). W data science C/C++ lub C# oraz Java stosowane mogą być do budowania oprogramowania[163], którego częścią są modele lub szerzej operacje na danych, zaś JavaScript i Ruby do budowania interfejsów dla użytkownika końcowego, np. w formie serwisu internetowego (por. Borowiecki i Mieczkowski 2019:36); ogólnie wszystkie te języki mają zastosowanie we wdrożeniach rezultatów projektów DS. Prolog jest językiem programowania logicznego powstałym w latach 70., stosowanym głównie do tzw. sztucznej inteligencji bazującej na zaprogramowanych ręcznie regułach (Colmerauer i Roussel 1993). Lisp to język zbliżony do Prologu, stworzony pod koniec lat 50. przez twórcę terminu sztuczna inteligencja, Johna McCarthy. W tym języku do celów nauczania zaimplementowano klasyczne programy tzw. sztucznej inteligencji, jak chatbot Eliza i inne, np. systemy ekspertowe czy programy grające w gry (Mccarthy 1960; Norvig 1992). Vladimir Alekseichenko, znany w badanym świecie data scientist, konsultant, przedsiębiorca i popularyzator DS tak podsumował wyżej prezentowane wyniki:

> **Python i R to języki używane zarówno do tworzenia prototypów, jak i do wdrażania rozwiązań AI. Python jest prosty**, pozwala użytkownikom zbudować funkcjonalność za pomocą kilku linii kodu i w efekcie **umożliwia skupienie się na rzeczywistym przypadku biznesowym**. Dlatego biznes lubi ten właśnie język. **Wielcy gracze, jak Google, Facebook czy Uber**, stworzyli wokół siebie społeczność użytkowników i z tego powodu **łatwo jest rozpocząć swoją podróż w dziedzinie uczenia maszynowego w języku Python. Język R** ma swoje zalety i wielu fanów, ale także pewne **wyzwania, które obejmują standaryzację kodu**. Ponadto w R może być **trudno wdrożyć uczenie głębokie** i jest do tych zastosowań **niewiele dobrych bibliotek** (przy czym nikt naprawdę nie pracuje nad rozwiązaniem tego problemu). Z perspektywy data scientista pracuje się z Pythonem lub R, a jeśli istnieje potrzeba większej wydajności, to **tylko niektóre części kodu są przepisywane w C / C ++ lub Javie** [podkreślenia org.]. Inne języki są właściwie nieważne (Borowiecki i Mieczkowski 2019:37-38).

Dlaczego akurat Python i R są powszechnie wykorzystywanymi językami programowania w DS, skoro istnieje tak wiele różnych języków? W badanym świecie społecznym zwraca się uwagę przede wszystkim na **prostotę czy łatwość ich używania** oraz

163 Sam język R jest w około 45% napisany w języku C (Wickham 2014a)

społeczność użytkowników, co wyraźnie ilustruje powyższa wypowiedź. Później pokażemy, że znaczenie ma także historia rozwoju tych języków programowania.

Co do **prostoty** w sensie technicznym, to mowa tutaj m.in. o tym, że są to języki interpretowane i dynamiczne, co można sprowadzić do tego, że osoba pisząca kod ma możliwość natychmiastowego uruchamiania każdej linii/fragmentu kodu po jej napisaniu i natychmiast widzi wynik tego działania. Nazywa się to interaktywną pracą z kodem i pozwala poprawiać na bieżąco błędy czy działanie niezgodne z zamierzonym. Podkreśla się także, że w językach takich jak Python czy R ilość pisanego przez człowieka kodu do wykonania danego zadania jest znacznie mniejsza niż w językach takich jak np. Java lub C++. Mówiąc ogólnie, w językach Python i R więcej czynności jest obsługiwanych domyślnie, i człowiek nie potrzebuje już myśleć np. o zaplanowaniu i zwolnieniu pamięci komputera na wykonywane operacje czy podawanie typu danych dla każdej przypisywanej zmiennej. A tego typu czynności muszą być zapisane przez człowieka np. w Javie. Wypisanie przez komputer słynnej frazy początkujących koderów „Witaj, świecie! (*Hello, world!*)" w Pythonie i R (a także w JavaScript, Lisp, Lua, Ruby) to jedna linia kodu. W Javie i C++ to sześć lub siedem linii, właśnie z uwagi na zapisanie w kodzie takich czynności jak podanie typu danych (Scriptol.com 2006). **Dla tego samego zadania kod w Pythonie zajmuje około 10% - 20% linii, potrzebnych na kod w Javie** (Paulson 2007:12). To jedna z różnic pomiędzy językami kompilowanymi a interpretowanymi, a jak zaraz wskażemy, także statycznymi a dynamicznymi. Nie wdając się w szczegóły techniczne powiedzmy, że kiedy człowiek pracuje z językiem kompilowanym, jak C#, to najpierw pisze większy fragment kodu bez możliwości wykonania go, zobaczenia wyniku i sprawdzenia błędów. Po zapisaniu wybranej całości kod należy skompilować i wtedy dopiero widać ewentualne błędy, np. składniowe, ale jeszcze nie wynik działania kodu. Wynik widać dopiero po wykonaniu kodu skompilowanego. Generalnie kompilacja to czynność, w której kod zapisany w języku zrozumiałym dla wykwalifikowanego człowieka, np. C#, tłumaczony jest na kod maszynowy. Komputer faktycznie wykonuje operacje zadane w kodzie C# na podstawie kodu maszynowego. Wykonywana jest całość operacji zadanych komputerowi przez ten kod maszynowy. Interpretacja czy parsowanie (termin pojawiał się już w części 5.2.), to wykonywanie przez komputer kodu np. Pythona kolejno linia po linii. Komputer sprawdza, co oznacza dana linia, wykonuje ją, a następnie przechodzi do następnej linii. Dzięki temu człowiek może uruchamiać kod linia po linii i widzieć jego działanie, może też uruchomić

cały skrypt (przy czym komputer faktycznie i tak wykonuje skrypt interpretując/parsując linia po linii).

Kompilacja i interpretacja są jak dwa sposoby czytania tekstu w nieznanym języku. Kompilacja jest jak całkowite przetłumaczenie artykułu z chińskiego na polski u tłumacza, a następnie przeczytanie całego artykułu. Interpretacja jest jak czytanie z tłumaczeniem sobie każdego słowa jeden po drugim, używając chińsko-polskiego słownika (Chen 2010:23).

> W jaki sposób wolałbyś się uczyć grać na pianinie: normalny, interaktywny sposób, w którym słyszysz każdą nutę, gdy tylko uderzysz w klawisz lub tryb „wsadowy" (*batch*), w którym słyszysz zagrane dźwięki dopiero po zakończeniu całej piosenki? Oczywiście tryb interaktywny ułatwia naukę gry na fortepianie, a także naukę programowania (Norvig 2001).

Zatem R i Python to języki interpretowane. Są to także języki dynamiczne, w przeciwieństwie do statycznych, np. Java czy C#. Oznacza to, że różnią się m.in. sposobem zapisywania typu danych. W R czy Pythonie człowiek nie musi zapisywać ręcznie typu danych, a także sam kod może być napisany w ten sposób, że typ danych może się zmieniać. Dodajmy, że językiem dynamicznym jest także np. wspomniany wcześniej Lisp (Paulson 2007:12–13). Przykładowo w R funkcja wykonująca obliczenie może być napisana w ten sposób, że przyjmuje za argumenty wartości o różnych typach danych – liczby całkowite (*integer*), jak np. 2, i zmiennoprzecinkowe (*floating-point*), jak 2,0001. To właśnie przykład tego, że typ danych może się zmieniać już po zapisaniu kodu. Wykonanie obliczenia zarówno dla 2, jak i 2,0001 zapewnia nie człowiek, a komputer. A dokładniej interpreter kodu, który sprawdza jaki jest typ danych dla konkretnej wartości w momencie wykonania tego kawałka kodu (Wickham 2014a).

Wymienianych jest szereg zalet języków takich, jak Python i R, czyli języków skryptowych, dynamicznych, interpretowanych. Wypowiadał się w tej sprawie m.in. twórca Pythona, Guido van Rossum, mówiąc o (Paulson 2007):

- elastyczności – kod może być napisany w ten sposób, że obsłuży różne typy danych (co omawialiśmy wyżej), poza tym Python i R działają na różnych systemach operacyjnych (Windows, macOS, dystrybucje Linuxa);
- efektywności – osiąganej dzięki mniejszej niż dla języków typu Java ilości kodu do wykonania zadania, oraz interaktywnej pracy metodą iteracji, a co za tym idzie szybkość dostarczenia działającego rozwiązania;

- kreatywności rozwiązań – dzięki odciążeniu człowieka z najbardziej szczegółowych i powtarzalnych zadań.

To, że w językach takich jak Python i R wiele czynności obsługiwanych jest domyślnie, i że możliwa jest interaktywna praca odbywa się kosztem szybkości działania kodu. Języki dynamiczne ogółem działają wolniej niż statyczne. Kod wykonujący to samo zadanie będzie zużywał więcej zasobów mocy komputera – pamięci i procesora (Paulson 2007). Jest to celowe:

> R nie jest szybkim językiem. To nie przypadek. R został specjalnie zaprojektowany tak, aby ułatwić analizę danych i statystyki Tobie, a nie ułatwić życie Twojemu komputerowi. Choć R jest powolny w porównaniu do innych języków programowania, to w większości przypadków jest wystarczająco szybki (Wickham 2014a).

Jak wskazaliśmy wyżej, w przypadku Pythona i R w DS postrzega się jako zaletę nie to, że kod jest szybko wykonywany przez komputer, lecz to, że jest szybko (bo łatwo) pisany przez człowieka – w porównaniu do kompilowanych języków programowania.

Zauważmy, że **zarówno Python jak i R są językami uznawanymi za proste lub łatwe, i odpowiednimi do pisania kodu przez osoby nie mające wykształcenia informatycznego**. Porównajmy je, po raz ostatni, z językami dla „pełnoetatowych programistów":

> Języki programowania, takie jak C ++ i Java, są zaprojektowane do profesjonalnego rozwoju (*development*) [oprogramowania] przez duże zespoły doświadczonych programistów, dla których ważna jest wydajność działania kodu w czasie. W rezultacie języki te mają skomplikowane części zaprojektowane dla tych okoliczności [tzn. dla wydajności]. Jesteś zainteresowany nauką programowania. Nie potrzebujesz tej komplikacji. Chcesz języka, który byłby łatwy do nauczenia się i zapamiętania przez pojedynczego, nowego programistę (Norvig 2001).

Co do różnic historycznych, R został stworzony przez statystyków dla statystyków. Jego nazwa pochodzi od imion twórców Rossa Ihaka i Roberta Gentlemana oraz jest nawiązaniem do nazwy języka S, na którym był wzorowany (Fox 2009; Ihaka i Gentleman 1996; Thieme 2018).

> Od początku język R był tworzony i rozwijany z myślą o osobach zajmujących się analizą danych. Z tego też powodu powodu przez informatyków często nie był traktowany jak pełnowartościowy język, ale raczej jak język domenowy (DSL, ang. *domain-specific language*). Z czasem jednak możliwości R rosły, pojawiło się coraz więcej rozwiązań wykraczających poza analizę danych i dziś R jest jednym z popularniejszych języków programowania (Biecek 2017a:3).

Python jest językiem ogólnego przeznaczenia (nie tylko do statystyki czy w ogóle pracy z danymi), a jego nazwa pochodzi od brytyjskiej grupy komików Monthy Python. Python był wzorowany na języku ABC przeznaczonym do nauki programowania dla początkujących, i jest uznawany za jeden z najlepszych języków do rozpoczęcia nauki programowania, oraz jeden z najlepszych dla ludzi używających kodu od czasu do czasu (*casual coders*) (ActiveState 2018; Hughes 1999; Nunns 2017; van Rossum 1997). Jednocześnie mówi się o „stromej krzywej uczenia (*steep learning curve*)" programowania w tych językach, ale odnosi się to do porównania wykonywania operacji na danych za pomocą programowania skryptowego z wykonywaniem ich za pomocą klikania w interfejsie graficznym (por. rys. 23., Biecek 2015a; Fox 2009:10; Muenchen 2014), a nie porównania do wykonania takich operacji w innych językach programowania. Przy tym ta względna trudność w korzystaniu z R w porównaniu z klikanymi pakietami analitycznymi (jak SPSS) bywa postrzegana jako to, co wzmacnia społeczność użytkowników (Muenchen 2014).

Wracając do naszego rozróżnienia z części 5.2.1., w DS raczej pisze się skrypty niż tworzy oprogramowanie, więc korzysta się z głównie języków skryptowych, dostępnych i odpowiednich dla nie-programistów. Osoby zajmujące się DS to zatem ludzie „potrafiący programować":

> W poprzednim akapicie celowo użyłem stwierdzenia "osoba, która potrafi programować" zamiast "programista". Jest to kolejny powszechny mit, że każda osoba która potrafi stworzyć program musi od razu zostać pełnoetatowym programistą. Pisanie programów jest narzędziem, które ma wspomóc twórcę w pewnym celu. Jednym z celów może być zostanie profesjonalnym deweloperem stron internetowych, aplikacji mobilnych, gier komputerowych, itd. Nie jest to jednak jedyny cel - programowanie może być, na przykład przydatnym narzędziem w analizie danych (Wiąże się to z popularnym na Zachodzie terminem data science, który łączy programowanie, analizę danych i wiedzę dziedzinową). Umiejętności programistyczne są wykorzystywane przez ekonomistów, biologów, geografów i osób z wielu innych dziedzin. Dodatkowo, podstawowe aspekty programowania są bardzo cenne w zawodach, w których ważna jest częsta współpraca z programistami (Nowosad 2019).

Prostota czy łatwość korzystania jako cechy języka programowania są związane na wiele sposobów ze **społecznością** ludzi korzystających z języka programowania. Społeczność (*community*, w Polsce oba słowa są używane zamiennie) bez wątpienia jest uznawana w badanym świecie za ważną cechę czyniącą korzystanie z narzędzi prostszym czy łatwiejszym (choć nie tylko). Przede wszystkim chodzi o **możliwość łatwego dostępu do rozwiązań wypracowanych przez innych członków społeczności dla własnych celów**. Zarówno Python jak i R są uznawane za języki, wokół których zbudowała się silna społeczność i łatwo jest uzyskać wsparcie w korzystaniu z nich. Dostępne są dla nich liczne rozszerzenia –

pakiety zawierające funkcje, dostępne są rozwiązania technologiczne ułatwiające korzystanie z tych języków, dostępna jest duża baza napisanego kodu, liczne materiały edukacyjne, wiele odpowiedzi na pytania, dostępne są spotkania twarzą w twarz z osobami korzystającymi z tych języków.

Zwrócenie uwagi na społeczność skupioną wokół danej technologii przy wyborze własnych narzędzi pracy jest typową poradą dla początkujących koderów, nie tylko w DS:

> Wielu ludzi pytało, od jakiego języka programowania zacząć naukę. (...) Użyj swoich znajomych. Na pytanie „jakiego systemu operacyjnego powinienem używać, Windows, Unix czy Mac?" zazwyczaj odpowiadam: „**używaj tego, czego używają Twoi znajomi**". Korzyść płynąca z uczenia się od znajomych zrównoważy wszelkie wewnętrzne różnice między systemem operacyjnym lub językami programowania. **Weź również pod uwagę swoich przyszłych znajomych: społeczność** (*community*) **programistów, których będziesz częścią**, jeśli będziesz kontynuować. **Czy wybrany język ma dużą, rosnącą społeczność, czy społeczność niewielką, umierającą?** [podkr. RŻ] Czy istnieją książki, strony internetowe i fora internetowe, na których można uzyskać odpowiedzi? Czy lubisz ludzi na tych forach (Norvig 2001)?

O pakietach[164] mówiliśmy już kilkukrotnie (por. część 2.1.). Przypomnijmy, że pakiet jest zbiorem funkcji, wykonujących określone zadania: „funkcjami możemy analizować dane, rysować dane, odsłuchiwać dane, wykonywać obliczenia, wysyłać maile, robić najróżniejsze rzeczy" (Biecek 2015a). Pakiet rozszerza wachlarz dostępnych w języku funkcji, co ułatwia wykonywanie zadań. Człowiek potrzebuje wykonać pewną operację na danych, i zamiast pracochłonnego pisania kodu na niższym poziomie abstrakcji szuka w pakietach gotowej funkcji, która zrealizuje tę operację. Przypomnijmy, że znani w DS twórcy narzędzi opowiadają, że podejmowali wysiłek tworzenia takich funkcji z powodu braku dostępnych rozwiązań, początkowo sami dla siebie (por. 2.1.).

Dla języka R największymi repozytoriami pakietów są CRAN (*The Comprehensive R Archive Network*), Bioconductor i GitHub (Fox i Leanage 2016:10). Do innych należy np. MRAN firmy Microsoft. CRAN to repozytorium ogólne, najstarsze, pod opieką R Foundation. Tam pobiera się także podstawową dystrybucję[165] GNU R. Bioconductor powstał w 2001 r. i jest repozytorium pakietów dla bioinformatyki. GitHub to publiczne repozytorium kodu dla dowolnych języków programowania, praktyka publikowania pakietów w tym miejscu jest najnowszą, pakiety nie są poddawane kontroli jakości jak w pozostałych

164 Używane są określenia: pakiet (*package*), biblioteka (*library*) i moduł (*module*), a w Polsce zdarza się określenie „paczka". Zasadniczo oznaczają one to samo (McClure 2019). Używamy w tej pracy słowa pakiet.

165 Sam język programowania jest abstrakcją, zaś dystrybucja to coś, co pozwala używać języka na komputerze (Wickham 2014a). Język R ma także dystrybucje oferowane przez Microsoft (Microsoft 2019b) i Anacondę (Anaconda 2019b)

repozytoriach (ibidem). Z naszych badań wynika, że pakiety z GitHuba mogą być traktowane jako nieoficjalne. Istnieje też praktyka publikowania tam wczesnych lub innych wersji pakietów, które są dostępne „oficjalnie" na CRAN. Na przykładzie CRAN zaobserwujmy rozwój liczby pakietów R: repozytorium powstało w 1997 r., pierwsze pakiety pojawiły się tam w 2001 r., zaś w 2016 było ich prawie 8 tys. Liczba pakietów rosła do 2011 r. nieco szybciej niż eksponencjalnie, zaś od 2011 do 2016 nieco wolniej (Fox i Leanage 2016:3–5). Na obecną chwilę – 16.07.2019 – CRAN zawiera ponad 14,5 tys. pakietów (R Foundation 2019). Przypomnijmy, że każdy z nich to, tak jak R, open source. Dla porównania SAS jako najpopularniejsze na świecie analityczne oprogramowanie zamknięte (*proprietary*) miał około 1,2 tys funkcji w 2015 r. Na CRAN w samym 2015 r. zamieszczono ponad 1,3 tys. pakietów, co oznacza około 27,6 tys. funkcji w jednym roku. Zatem społeczność R stworzyła więcej funkcji przez jeden rok niż firma SAS Institute przez cały okres istnienia, tzn. od połowy lat 70. (Muenchen 2019).

Dla Pythona koncentrujemy się wyłącznie na pakietach związanych z operacjami na danych, pamiętając, że jest to język ogólnego przeznaczenia i posiada pakiety do zupełnie innych zastosowań. Największe i najpopularniejsze jest repozytorium PyPI (*The Python Package Index*), z którego instaluje się pakiety komendą „pip", co pokazujemy poniżej. Dla dziedziny nauka/badania[166] obecnie w tym repozytorium jest ponad 10 tys. pakietów; dokładnej liczby nie podano (Python Software Foundation 2019). Później pokażemy, że przede wszystkim dla Pythona rozwijane są pakiety do modnego obecnie uczenia głębokiego (*deep learning*).

Skorzystanie[167] z pakietu może sprowadzać się w najłatwiejszym scenariuszu do wykonania dwóch linii kodu, dla każdego z omawianych języków. Dla podstawowych w DS pakietów „dplyr" i „NumPy" (por. część 2.1.) kod wygląda następująco:

```
# R
install.packages("dplyr")
library(dplyr)

# Python
pip install numpy
import numpy as np
```

166 Wybrano filtr „Intended Audience :: Science/Research".

167 Upraszczając dla przypadków, kiedy pakiet jest dostępny z CRAN dla R i PyPI dla Pythona. Dla obu języków pierwsza linia kodu instalująca pakiet potrzeba jest tylko wtedy, kiedy pakiet nie jest ściągnięty na komputer użytkownika, zatem przed pierwszym użyciem. Przy kolejnych użyciach należy jedynie załadować pakiet drugą linią. Część „as np" dla Pythona nie jest niezbędna, ale zwyczajowa, podyktowana m.in. wygodą. Dzięki temu elementowi korzysta się później z pakietu „numpy" pisząc krótsze „np".

Linie rozpoczynające się znakiem „#” to tzw. komentarze i nie są one wykonywane przez komputer. Pierwsze linie dla każdego języka to pobranie pakietu z repozytorium przez internet na komputer kodera i zainstalowanie pakietu w odpowiednim miejscu. Drugie linie to uruchomienie pakietu, tzw. załadowanie go do przestrzeni roboczej. Pakiety ładuje się przy każdej nowej sesji, tzn. kolejnym uruchomieniu środowiska, w którym piszemy kod, ponowna instalacja standardowo nie będzie potrzebna.

Jak zaraz pokażemy, to, że pakiety są publikowane w otwartym dostępie (*open source*) nie oznacza, że są one budowane tylko oddolnie, przez osoby fizyczne czy nieformalne grupy. Zasadniczo wszystkie używane w DS technologie open source mogą być budowane oddolnie, jak również w firmach i na uczelniach, we współpracy firm i uczelni, a także przy wsparciu finansowym z różnego rodzaju grantów (por. 2.1. i 5.1.3.). Niemałą rolę odgrywają wspierający rozwój open source giganci technologiczni (Google, Facebook, Microsoft, Amazon, por. część 6.2.1. rozprawy), nie tylko jeżeli chodzi o pakiety, ale o wiele innych narzędzi świata DS.

Istnieją także tzw. ekosystemy pakietów: dla Pythona np. SciPy, dla R np. tidyverse. W skład każdego z nich wchodzą pakiety zawierające funkcje do wykonania każdego z etapów procesu DS (przetwarzania danych, analizy, modelowania, komunikacji, por. rys. 24.), choć raczej nie wystarczają one do wykonania każdego projektu w całości (Wickham i Grolemund 2017). Mówiliśmy o tych ekosystemach więcej w części 2.1. Taki ekosystem nie jest tylko kolekcją pakietów – pakiety te mają podobną konwencję pisania kodu, podobne sposoby zapisywania obiektów (Bryan i Wickham 2017:785), krótko mówiąc są zintegrowane i dobrze ze sobą współgrają:

> (...) Tidyverse to zbiór pakietów R, które mają wspólną filozofię projektowania na wysokim poziomie (*high-level design*) oraz gramatykę i struktury danych na niskim poziomie (*low-level grammar and data structures*), tak aby nauka jednego pakietu ułatwiała naukę drugiego (Wickham i in. 2019:1).

Ma to pomagać w rozwiązywaniu złożonych problemów za pomocą prostych elementów, małych kroków, umożliwiających łatwe rekonfiguracje, a zatem iteracyjną pracę z danymi (Wickham 2018f). Podkreślmy, że korzystanie z pakietów w tych dwóch ekosystemach jest w badanym świecie uznawane za elementarz umiejętność koderskich.

Inne powszechnie znane pakiety to używane głównie do modnego obecnie uczenia głębokiego (*deep learning*): TensorFlow (TF; por. 5.2.4.) i pakiet PyTorch oraz tzw. nakładka Keras. W Polsce na używanie TF wskazało 69% badanych firm, Keras 49%, zaś

PyTorch/Torch – 38% (Borowiecki i Mieczkowski 2019:39–40). Korzysta się z nich zdecydowanie najczęściej przy pomocy Pythona, bardzo rzadko z R, lub innych języków programowania. Keras i TF wraz z językiem Python to trzy składniki najpopularniejszego, według ankiet portalu Kdnuggets, sześcioelementowego zestawu technologii (też nazwanego ekosystemem) open source dla DS / uczenia maszynowego (Piatetsky 2018d). TF jest to

> **system uczenia maszynowego działający na dużą skalę i w środowiskach heterogenicznych**. TensorFlow wykorzystuje grafy przepływu danych (*dataflow graphs*) do reprezentowania obliczeń, stanu współdzielonego i operacji, które zmieniają ten stan. Mapuje on węzły grafu przepływu danych **na wielu maszynach w klastrze oraz w maszynie na wielu urządzeniach obliczeniowych**, w tym wielordzeniowych procesorach, procesorach GPU ogólnego przeznaczenia i niestandardowych układach ASIC znanych jako jednostki przetwarzania tensorów (*Tensor Processing Units,* TPU). Ta **architektura daje elastyczność twórcy aplikacji** (...) TensorFlow **umożliwia twórcom eksperymentowanie** z nowymi optymalizacjami i algorytmami trenującymi. (…) **Kilka usług Google używa TensorFlow na produkcji, wydaliśmy go jako projekt open source i stał się on szeroko stosowany w badaniach nad uczeniem maszynowym**. W tym artykule opisujemy model przepływu danych TensorFlow i **przedstawiamy atrakcyjną wydajność, jaką TensorFlow osiąga w kilku rzeczywistych zastosowaniach** [podkr. RŻ] (Abadi i in. 2016).

Chodzi o to, że TF pozwala na wygodne i szybkie eksperymentowanie z różnymi parametrami modeli uczenia maszynowego/głębokiego, a także szybkie działanie modeli, w tym modeli na produkcji. Obliczenia mogą być wykonywane „na wielu maszynach w klastrze oraz w maszynie na wielu urządzeniach obliczeniowych" (ibidem). Upraszczając nieco maszyna to komputer. Klaster to zestaw komputerów działających razem. Urządzenia obliczeniowe w maszynie to procesor ogólnego zastosowania CPU, procesor graficzny GPU, ewentualnie wspomniane TPU – to wyspecjalizowany (*custom*) hardware o zmniejszonej precyzji obliczeń, dostępny wyłącznie na maszynach Google, nie jest w sprzedaży. Zatem obliczenia mogą być wykonywane z użyciem TF szybko i na dużych danych, bo można[168] użyć procesorów GPU lub TPU (urządzeń szybszych niż CPU do obliczeń w uczeniu maszynowych), a także można użyć wielu procesorów w wielu komputerach jednocześnie. PyTorch działa dość podobnie, także pozwala na korzystanie z szybkości GPU (ale już nie TPU, bo należy do Google, a PyTorch to pakiet wspierany przez firmę Facebook) (Ketkar 2017; Yegulalp 2017). PyTorch jest pythonowym portem do wspomnianej biblioteki Torch, napisanej w języku C z dostępem za pomocą języka Lua (Collobert i in. 2019; Ierusalimschy i in. 2019; Yegulalp 2017). Podkreślaną zaletą PyTorch jest dobra integracja z popularnymi pakietami[169] Pythona i językiem Python w ogóle (Paszke i in. 2017; Sopyła 2019). Spory o wady i zalety TF i PyTorcha (por. Sopyła 2019) lub Keras + TF a PyTorch (Agarwal 2019;

168 Za pomocą innych rozwiązań też jest to możliwe. Jednak jako zalety omawianych pakietów uznawany jest fakt, że można omawiane czynności wykonać relatywnie do innych rozwiązań łatwo, szybko, wygodnie.

Misal 2018) opisujemy dalej jako element areny w badanym świecie. Zgadzamy się z poglądem, że publikowanie w otwartym dostępie omawianych pakietów przez Google i Facebooka to element gry tych firm o pozycję w globalnej branży związanej z DS i uczeniem maszynowym (Arakelyan 2017), co staramy się ująć w pełnej sytuacji badania (por. Clarke 2005:73 i część 6.2. rozprawy).

Keras nazywany jest nakładką ponieważ

> [Keras] ma być szkieletem na poziomie modelu (*model-level framework*), dostarczającym zestaw „klocków Lego" do budowania modeli uczenia głębokiego w szybki i prosty sposób. Wśród platform uczenia głębokiego Keras jest zdecydowanie wysoko w hierarchii abstrakcji (Chollet 2015).

Zatem Keras jest narzędziem wyższego poziomu (*high-level API*) niż TF, nie można wytrenować modelu korzystając wyłącznie z Kerasa, potrzebne jest zaplecze (*backend*) w postaci np. TF. Można skorzystać z innego backendu, z których każdy to open source: Theano (pakiet autorstwa LISA Lab Université de Montréal), CNTK (autorstwa Microsoft), TF (Google) (Keras 2019). Na PyTorch nie ma takiej nakładki. Zarówno PyTorch jak i Keras są stosowane w szkoleniu początkujących (Migdał 2018; Sopyła 2019). Piszący te słowa (RŻ) uczestniczył w warsztatach dla początkujących prowadzonych właśnie z użyciem Kerasa i TF z językami R i Python {obserwacja 26, 35, 39, 43}.

Rozwiązania technologiczne ułatwiające korzystanie z omawianych języków to m.in. dla R środowisko programistyczne, tzw. IDE (*Integrated development environment*) o nazwie R Studio (RStudio Team 2016), zaś dla Pythona tzw. dystrybucja języka o nazwie Anaconda (Anaconda 2018) oraz notatnik programistyczny Jupyter Notebook (Perez i Granger 2015). Oba te rozwiązania umożliwiają także pracę z R. Dla Pythona ogółem istnieje także szereg IDE i innych edytorów kodu (Reitz 2018:26–31), ale nie ma w tym zakresie, tak jednoznacznie jak w przypadku RStudio, dominującego[170] dla DS wyboru (Piatetsky 2018b). Dla języka Python używanie notatnika Jupyter (w pytaniu wielokrotnego wyboru) wskazało

169 przede wszystkim NumPy z uwagi na korzystanie z podobnych obiektów zwanych tensorami, czyli rodzajem wielowymiarowej tabeli (*multidimensional array*)

170 Dla języka R istnieją także inne IDE, ale twierdzimy, że przynajmniej w Polsce są one bardzo rzadko lub w ogóle nie używane. Z pewnością nie można także w badanym świecie wskazać na podziały wśród użytkowników R z uwagi na używane IDE. Z języka R można korzystać także za pomocą notatnika Jupyter (nazwa pochodzi od języków programowania Julia, Python, R), jego szerszego odpowiednika Jupyter Lab, edytora RIDE (także dla Pythona i innych), platformy Radiant, rozszerzenia R Tools dla Visual Studio firmy Microsoft, IDE Architect, IDE Atom (dla wielu języków) za pomocą rozszerzenia rbox, również za pomocą rozszerzenia z edytora Vim (dla wielu języków). Dostępne są także graficzne interfejsy (GUI), które pozwalają wyklikać pewne operacje korzystając z R jako backendu: Rattle, RKWard, Tinn-R, BlueSky, Exploratory, Displayr, Stagraph, Deducer, The R Commander, R MLstudio, esquisse, ggraptR, R AnalyticFlow (dimlakgorkehgz 2019; karupakalas 2018).

57% badanych (ponad 1 900 respondentów z różnych kontynentów), z IDE PyCharm korzystało 35%, IDE Spyder 27%[171] (ibidem). Nie ma badań dotyczących wyboru IDE dla języka R, co wskazuje na jednoznaczną dominację RStudio. Mówiliśmy o RStudio i Anacondzie w części 2.1., koncentrując się na osobach ich twórców, przeważnie naukowcach. Tutaj chcemy pokazać podstawowe aspekty techniczne, oraz to, że są to rozwiązania open source – ale także z dodatkowymi, płatnymi opcjami. Niemniej „otwartość" oprogramowania uznawana jest w badanym świecie za przejaw siły społeczności użytkowników tych języków.

Jak wymienione technologie ułatwiają korzystanie z języków programowania? Za przykład IDE posłuży nam RStudio. Co do zasady to każde IDE ma zapewniać narzędzia do automatyzacji typowych, powtarzalnych czynności. Ma to na celu zwiększenie szybkości pracy osoby piszącej kod i zmniejszenie ilości popełnianych przez nią błędów (Muûlu i in. 2012:669). IDE jest oprogramowaniem, przy pomocy którego pisze się kod w danym języku, czy kilku językach. IDE ma funkcje pomocne w samym pisaniu kodu – jak autouzupełnianie, o czym dalej – a także funkcje pomocne w dodatkowych czynnościach, jak korzystanie z kontroli wersji, instalacja pakietów itp.

Na rysunku 25. widoczne są trzy okna: 25A to terminal języka R w systemie Windows – Rterm, 25B to interfejs domyślnego edytora kodu podstawowej dystrybucji języka R – RGui, 25C to IDE – RStudio. Posługujemy się tym samym przykładem kodu generującego wykres, co w opisie działania podstawowego (por. 5.2.1., rys. 23.). W oknie 25A oprócz kodu, rozpoczynającego się tzw. znakiem zachęty „>" widać komunikat powitalny używanej wersji języka R. Widoczny jest on zawsze po uruchomieniu konsoli, także w inny sposób (por. 25Bi, 25Cii), dla zaoszczędzenia miejsca prezentujemy go tylko raz.

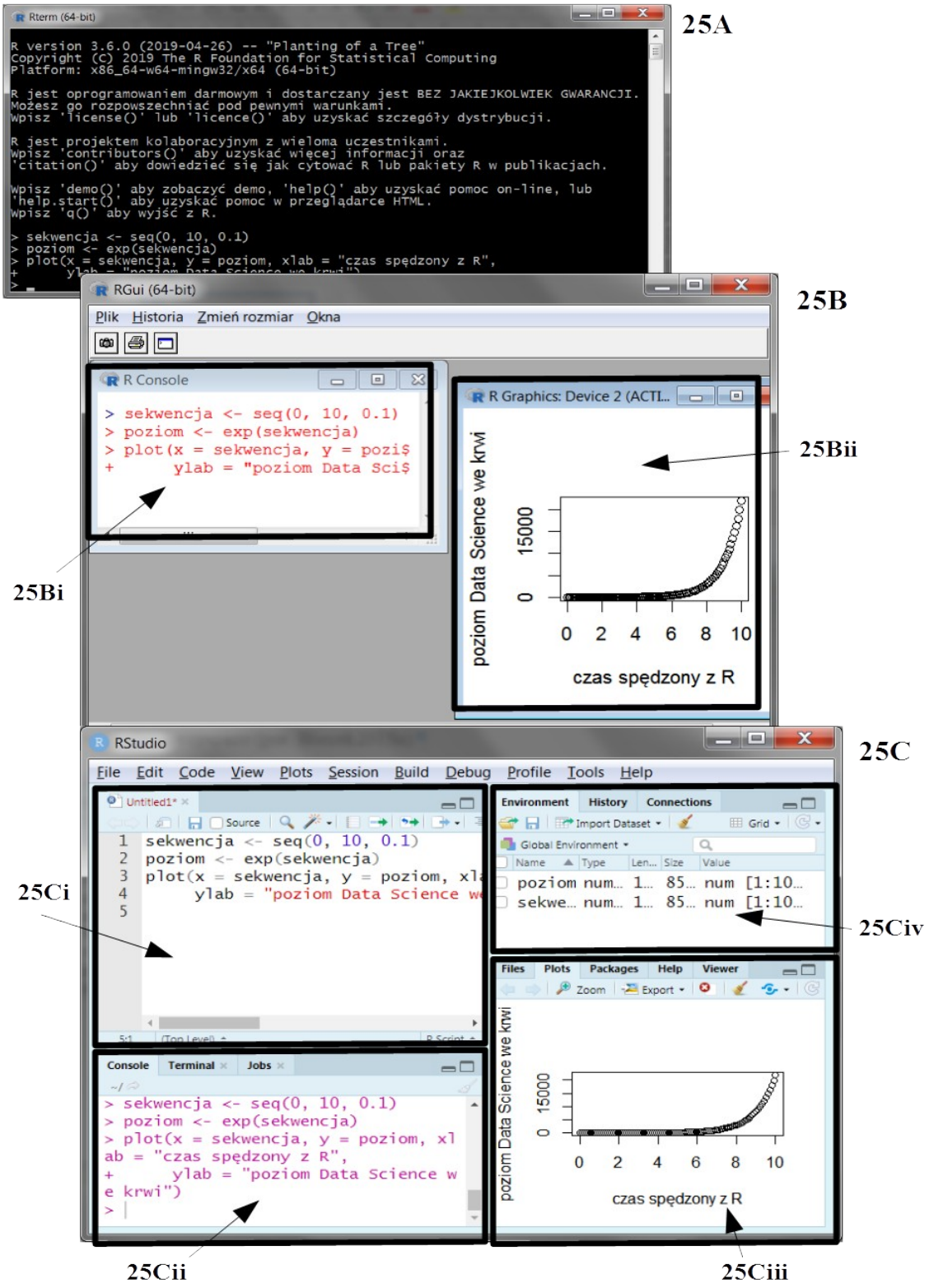
Konsola jest podstawowym sposobem pisania i uruchamiania kodu – konsolę widzimy korzystając z języka programowania na każdy z trzech sposobów (por. 25A, 25Bi, 25Cii). Dla Rterm jest to tylko konsola (por. 25A), dla RGui konsola (25Bi) jest już wewnątrz okna, które ma swoje menu (25B), dla RStudio konsola (25Cii) to jeden z wielu elementów aplikacji (25C). Właściwie wszystko, co nie jest konsolą pełni funkcje pomocnicze. W konsoli mamy możliwość pisania i natychmiastowego uruchamiania linii kodu. Zauważmy także, jak wyświetlany jest wynik działania kodu. Przy pracy w Rterm wykres rysowany jest

171 Wymieniono także: Visual Studio (21% wskazań), Sublime Text (12%), Atom (10%), Vim (8,5%), inne IDE (3,1%), Eclipse (3%), inny edytor (2,5%), Emacs (2,2%), gedit (1,3%), Rodeo (1,2%). W podziale na kategorię miejsca pracy ani w podziale na lokalizację geograficzną nie widać wyraźnych różnic w rozkładzie odpowiedzi – wszędzie dominuje notatnik Jupyter, zaś na drugim i raz na trzecim miejscu PyCharm (Piatetsky 2018b).

w osobnym oknie urządzenia graficznego, dlatego nie ujęliśmy go na rysunku. W RGui takie samo okno (25Bii) otwiera się wewnątrz okna (25B), w RStudio wykres zobaczymy w jednej z kliku kart (otwarta „Plots") panelu w prawym, dolnym rogu (25Ciii). W konsoli praca z kodem jest interaktywna – piszemy linię, i po wciśnięciu klawisza Enter jest ona wykonywana. W praktyce jest to sposób używany raczej do szybkiego sprawdzania działania pojedynczych linii czy nawet fragmentów linii kodu, nie do realizacji całego ciągu operacji na danych – właśnie dlatego, że konsola jest interaktywna. Typową strategią jest pisanie kodu w edytorze skryptów (25Ci). Taki edytor można otworzyć także w RGui, czego nie ujęto na rysunku. Podczas pisania skryptu możemy pisać wiele linii jedna pod drugą bez natychmiastowego uruchamiania, możemy swobodnie nawigować po liniach, kopiować i wklejać elementy. W konsoli nie jest to możliwe. Możemy jedynie wyświetlić historię uruchomionych linii klawiszami strzałek „góra" i „dół". W edytorze można także wygodnie zapisać tworzony skrypt – w przykładzie widać, że skrypt nie jest zapisany, co sygnalizuje czerwony kolor czcionki wyświetlającej domyślną nazwę „Untitled1". Linie kodu uruchamiane ze skryptu (25Ci) są przesyłane do konsoli (25Cii) i tam wykonywane. Zauważmy także dostępną wyłącznie w RStudio część interfejsu z kartą „Środowisko (*Environment*)" (25Civ). Wyświetlają się m.in. przypisane wcześniej zmienne – w naszym przykładzie „sekwencja" i „poziom" – oraz informacje o tych zmiennych (por. 25Civ). W innych sposobach pracy (25A i 25B) możemy zobaczyć listę przypisanych zmiennych w konsoli, wywołując ją w kodzie odpowiednią funkcją.

Zauważmy także, że w Rterm kod nie odróżnia się kolorem od innego tekstu (25A), w RGui kod w konsoli wyróżnia z okna jednolity kolor czerwony (25Bi); kod w skrypcie miałby jednolity kolor czarny, czego nie prezentujemy. Niemniej jest tutaj zaprojektowana różnica w wyglądzie kodu skryptu względem kodu w konsoli (co nie ma sensu w Rterm, gdyż nie ma narzędzia do pisania skryptów). W RStudio ta różnica także występuje (por. 25Ci versus 25Cii), a dodatkowo w skrypcie automatycznie kolorowana jest składnia, np. kolorem niebieskim wyróżniono liczby, kolorem czerwonym ciąg znaków (*string*). Silnie odczuwaną we własnoręcznym pisaniu kodu funkcją jest autouzupełnianie (*auto-completion*). W Rterm i RGui autouzupełnianie jest ograniczone. Można zobaczyć, jakie funkcje zaczynają się od liter np. „se", wpisując te litery i wciskając klawisz Tab. Lista funkcji zostanie wypisana w konsoli. W przypadku kiedy możliwa jest tylko jedna sytuacja np. wpiszemy litery „len", to nazwa funkcji zostanie automatycznie uzupełniona do „length" (ale przy wpisaniu „le" otrzymamy listę nazw w konsoli). W RStudio przy wpisaniu nawet jednej litery po wciśnięciu

Tab otrzymujemy rozwijaną listę podpowiedzi, wyświetlaną w oknie pomocniczym przy wpisanej literze, wraz z informacją, z jakiego pakietu pochodzi funkcja. Po tej rozwijanej liście można nawigować klawiszami strzałek, wyświetlają się automatycznie krótkie opisy funkcji i ich argumentów, zaś wciskając klawisz F1 wyświetlimy pomoc w panelu 25Ciii w karcie „Help". Jeśli po wciśnięciu Tab widzimy w podpowiedziach szukaną funkcję, nie trzeba niczego przepisywać, a jedynie podświetlić nazwę funkcji za pomocą strzałek i wcisnąć Tab ponownie. Ludzie pisząc kod używają autouzupełniania nie tylko, by przypomnieć coś sobie, ale także by zmniejszyć ilość uderzeń placami w klawisze. Przykładowo funkcja „install.packages()" dzięki autouzupełnianiu jest wpisywana za pomocą sześciu uderzeń: „inst" + dwukrotnie Tab. Ręczne jej wpisanie to 18 uderzeń w klawiaturę. W RStudio autouzupełnianie polega także na domykaniu nawiasów i cudzysłowów. Przykładowo po wpisaniu „inst" + dwa razy Tab widzimy wpisane „install.packages()", a kursor tekstu znajduje się pomiędzy nawiasami, zatem człowiek przechodzi od razu do wpisania argumentów funkcji, nie myśląc już o konieczności wpisania znaku „)" na koniec, ani nie musząc cofać kursora do środka nawiasu. Przy bardziej złożonym i kilkukrotnie poprawianym kodzie pomocna jest funkcja domyślnego podświetlania początku/końca nawiasu, co służy ewentualnemu ręcznemu domknięciu – bez domkniętych nawiasów kod nie zadziała. RStudio ma także szereg wbudowanych skrótów klawiaturowych, np. niewygodny do wprowadzania (będący pozostałością po niewystępującym obecnie na klawiaturach symbolu) symbol przypisania „<-" w RStudio można wprowadzić skrótem klawiaturowym „Alt + -". Symbol wprowadzany jest wraz z poprzedzającą go i następującą po nim spacją. Przykłady tego rodzaju udogodnień można by mnożyć: RStudio oferuje m.in. szereg narzędzi do wyłapywania i poprawiania błędów w kodzie (*debugging*), połączenia z systemem publikacji wyników do plików tekstowych lub interaktywnych , połączenie z systemami kontroli wersji Git i SVN (o czym więcej w części 5.3.4.), menadżer pobierania i aktualizowania pakietów (por. karta „Packages w panelu 25Ciii"). Wiele operacji pomagających pracować z kodem można dzięki RStudio znaleźć w jednym miejscu, a nawet wyklikać w interfejsie graficznym. Wiele funkcji tego IDE opisanych jest w oficjalnej ściągawce (*cheat sheet*) (RStudio Team 2017). Tę ściągawkę także wyklikamy z poziomu RStudio pod przyciskiem „Help" (por. 25C).



Rysunek 25: IDE jako narzędzie ułatwiające pracę z kodem na przykładzie RStudio w porównaniu do RGui i Rterm

Źródło: opracowanie własne na systemie Windows 7

Jeśli pisanie kodu w Rterm (rys. 25A) porównać do jazdy drezyną kolejową, to praca w RGui (25B) jest kierowaniem lokomotywą spalinową, zaś w RStudio (25C) nowoczesnym

składem międzynarodowego ekspresu. Zawsze jednak jedziemy po szynach – języku programowania, a dokładniej zainstalowanej na komputerze dystrybucji języka. Niemniej „drezyna", czyli praca w konsoli/terminalu/wierszu poleceń ma swoje szerokie zastosowania (por. część 5.3.5.), nie tylko do korzystania z języków R lub Python, ale w ogóle korzystania z komputera (szczególnie na Unixowych systemach operacyjnych – macOS i dystrybucjach Linuxa).

Anaconda to nie IDE, a tzw. dystrybucja, która zdaniem twórców jest

> najłatwiejszym sposobem na realizację data science i uczenia maszynowego w Pythonie/R na systemach Linux, Windows i Mac OS X. Z ponad 15 milionami użytkowników na całym świecie, jest to przemysłowy standard dla programowania, testowania i trenowania na pojedynczej maszynie, umożliwiający *indywidualnym data scientistom* [podkr. org.]: [1] szybkie pobranie ponad 1 500 pakietów dla data science w Pythonie/R; [2] zarządzanie bibliotekami (*libraries*), zależnościami (*dependencies*) i środowiskami (*environments*) za pomocą Conda; [3] programowania i trenowania modeli uczenia maszynowego i uczenia głębokiego za pomocą scikit-learn, TensorFlow i Theano; [4] skalowalne i wydajne analizowanie danych za pomocą Dask, NumPy, pandas i Numba; [5] wizualizację wyników z użyciem Matplotlib, Bokeh, Datashader, Holoviews (Anaconda 2018).

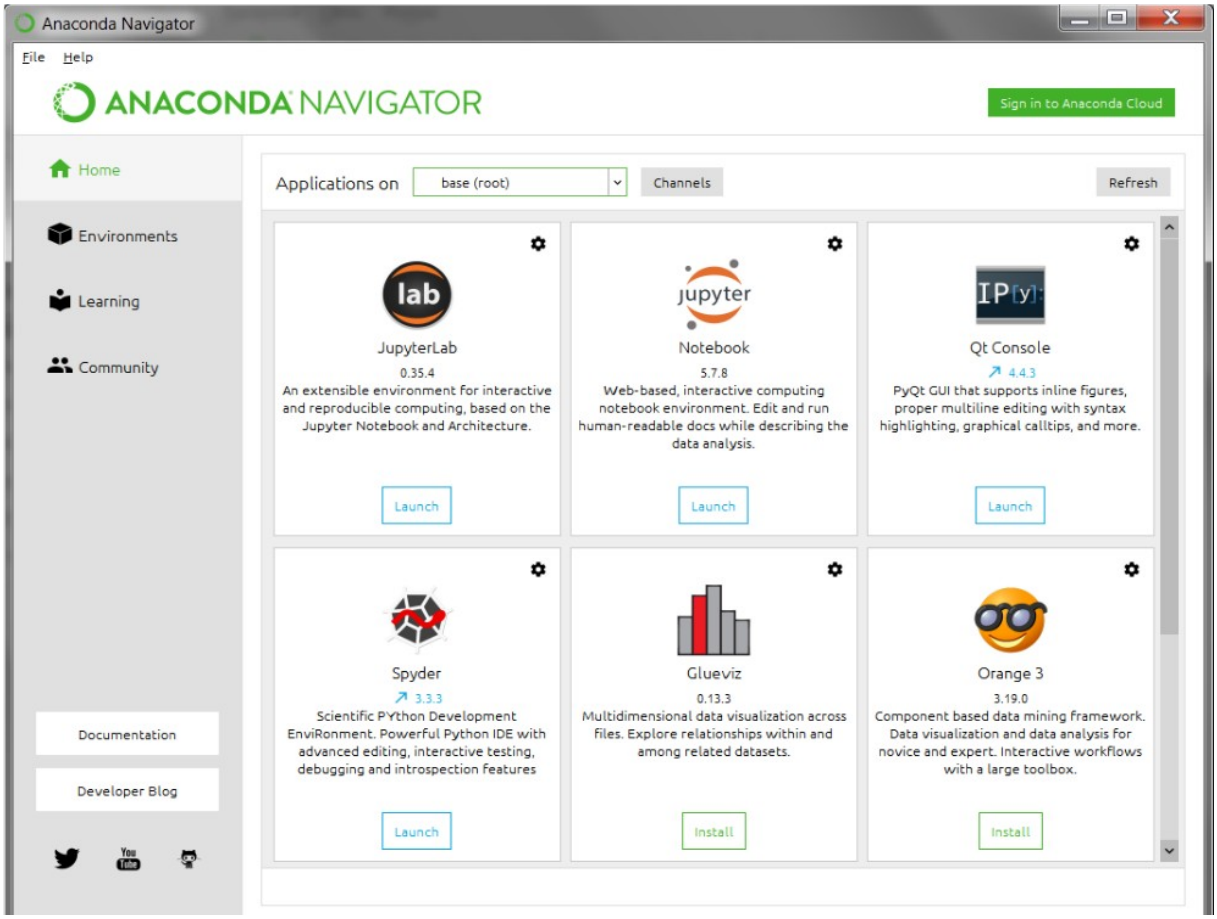
Słowo „dystrybucja" używane jest w odniesieniu do Anacondy jako całości, ale właściwie chodzi o dystrybucje języków Python i R. Dodatkowo na Anacondę składa się menadżer pakietów i środowisk Conda, oraz kolekcja ponad 1 500 pakietów dla DS. Około 200 z tych pakietów jest już domyślnie zainstalowanych, zatem nie trzeba robić tego samodzielnie. Inne pakiety można doinstalować z kolekcji pakietów, a jeszcze inne ze wspomnianego wcześniej ogólnego repozytorium PyPI. Podczas instalacji z kolekcji pakietów używane jest polecenie „conda install", a nie „pip install", jak dla repozytorium PyPI. To polecenie menadżera środowisk i pakietów Conda.

Menadżer Conda jest uznawany za ułatwienie pracy dlatego, że przy korzystaniu z narzędzi open source rozwijanych przez międzynarodową społeczność pojawiają się różne wersje samych języków programowania, pakietów i tzw. zależności (*dependencies*). Różne podmioty rozwijają różne części tej rozdrobnionej galaktyki technologii, działając trochę chaotycznie na zasadzie tzw. bazaru (Raymond 2001). Powstają nowe lub inne wersje rozmaitych narzędzi, a w konsekwencji różne elementy stają się niekompatybilne lub działają inaczej niż działały wcześniej. Jest to uznawane za wysoce problematyczne, ponieważ jedną z podstawowych zalet zapisywania operacji na danych za pomocą kodu ma być powtarzalność (*reproducibility*) czynności (Perez i Granger 2015; Wickham 2018f). Szczególnie dla języka Python widać tę różnorodność – używana jest zarówno wersja 2., jak i 3. samego języka[172].

172 Choć wersja 2. będzie utrzymywana tylko do końca 2020 r. (Reitz 2018:3–6)

Sama podstawa kodu może nie być dla tych dwóch wersji niekompatybilna, zatem tym bardziej dotyczy to pakietów. Menadżer Conda ułatwia tworzenie środowiska z konkretnymi wersjami technologii do realizacji konkretnych projektów. Jeden data scientist może za jego pomocą utworzyć wiele takich środowisk na jednym komputerze, tzw. maszynie, do różnych projektów. Conda ułatwia zarządzanie nimi, kontrolowanie wersji narzędzi składających się na środowiska, brak konfliktów pomiędzy tymi wersjami i ewentualne aktualizowanie narzędzi. Innymi narzędziami umożliwiającymi zamrożenie całego środowiska w sensie wszystkich wersji narzędzi w danej chwili są tzw. kontenery, jak np. Docker (Avram 2013) i technologia tzw. orkiestracji kontenerów Kubernetes (Lardinois 2015). Nie są to części Anacondy, są one technologiami używanymi głównie poza DS, m.in. w inżynierii danych (*data engineering*).

Jak wskazaliśmy Conda ułatwia także instalowanie pakietów za pomocą polecenia „conda install", i instalowanie pakietów w odpowiednich wersjach dla wybranych środowisk (Anaconda 2019a). Podkreślaną zaletą Anacondy i będącego jej częścią menadżera Conda jest to, że mogą one działać dla różnych języków programowania. Są one jednak kojarzone głównie z językiem Python poprzez związek z PyData i organizacją NumFOCUS, ale jak wspomniano umożliwiają pracę z R, a także zarządzanie innymi zależnościami (*dependencies*) np. narzędzia szyfrującego OpenSSL (Vanderplas 2016). Z naszych badań wynika, że Anacondą posługują się wyłącznie osoby używające Pythona. Mogą one używać także R i wielu innych narzędzi, ale nie ma praktyki korzystania z Anacondy przez ludzi w ogóle nie stosujących języka Python dla DS. Anaconda upraszcza także instalowanie innych narzędzi ułatwiających pracę z Pythonem, jak IDE Spyder i notatnik Jupyter. Omawiana dystrybucja pozwala użytkownikowi na obsługę komendami z poziomu terminala/wiersza poleceń, ale i na wyklikanie wielu operacji (jak właśnie instalacja IDE) w interfejsie graficznym Anaconda Navigator (rys. 26.). Z poziomu zakładki „Home" Anaconda Navigator użytkownik może za pomocą jednego kliknięcia zainstalować (zielony przycisk „Install") lub uruchomić (niebieski przycisk „Launch") wybrane aplikacje za pomocą skrótów. Zakładka „Environments" pozwala na dostęp do zbudowanych środowisk i ich komponentów. Zakładka „Learning" zawiera hiperłącza do dokumentacji, szkoleń, wideo i webinarów. Zakładka „Community" to odnośniki do konferencji, blogów, forów internetowych (w tym Stack Overflow). Poniżej znajduje się link do dokumentacji Anacondy, link do bloga DS firmy Anaconda Inc., jej konta Twitter, kanału YouTube i profilu GitHub (por. rys. 26.).



Rysunek 26: Interfejs Anaconda Navigator – zakładka "Home"

Źródło: opracowanie własne na systemie Windows 7

Zarówno RStudio jak i Anaconda są dostępne od 2012 r (Allaire 2012; Vanderplas 2016). Oba rozwiązania są rozwijane przez komercyjne podmioty, które oferują także ich płatne, rozszerzone wersje. Zasadniczo te wersje mają szereg udogodnień do pracy zespołowej i pracy na dużą skalę (Anaconda 2019b; RStudio 2018b).

Jak widać na powyższym rysunku (26., pierwszy wiersz druga kolumna skrótów do aplikacji), notatnik Jupyter to interaktywne środowisko obliczeniowe w formie notatnika. Działa ono w przeglądarce internetowej. Pozwala na edytowanie i uruchamianie dokumentów analitycznych, które są przystosowane do czytania przez ludzi (*human readable*). Twórcy Jupytera kierowali się chęcią stworzenia narzędzia do zespołowego tworzenia obliczeniowych narracji (*computational narratives*), które składają się z danych, kodu i narracji właśnie – równań, tekstu i wizualizacji (Perez i Granger 2015). Określili oni rozwiązany za pomocą tego notatnika problem jako „wspólne tworzenie powtarzalnych narracji obliczeniowych, które można wykorzystać w zakresie różnych odbiorców i różnych kontekstów" (ibidem). Notatniki można także w łatwy sposób publikować w formie interaktywnej i statycznej w

różnych formatach. Autorzy zaznaczyli wyraźnie, że choć zamknięte aplikacje komercyjne (*proprietary software*) – wymieniono MATLAB, Mathematica, SAS, SPSS i MS Excel – oferują zbliżone funkcjonalności, to fakt, że są zamknięte i kosztowne czyni je nieatrakcyjnymi dla otwartej i powtarzalnej nauki (*open and reproducible scientific research*) oraz dla DS (ibidem).

W prasie notatnik Jupyter przedstawiano jako nową jakość w komunikowaniu wyników badań naukowych. Publikowane obecnie artykuły w formie plików .pdf są zdaniem autora tylko przestarzałą symulacją kartki papieru, nie ma zasadniczej różnicy w porównaniu z XVII wiekiem. Obliczeniowy notatnik ma tworzyć nową jakość (Somers 2018). W artykule pojawia się nawiązanie do zwrotu obliczeniowego (*computational turn*) (por. część 2.2.2. oraz Boyd i Crawford 2011:3, 2012:665) w nauce – powołano się na słowa Stephena Wolframa (genialnego fizyka i twórcę oprogramowania Mathematica) z 2016 r., który napisał

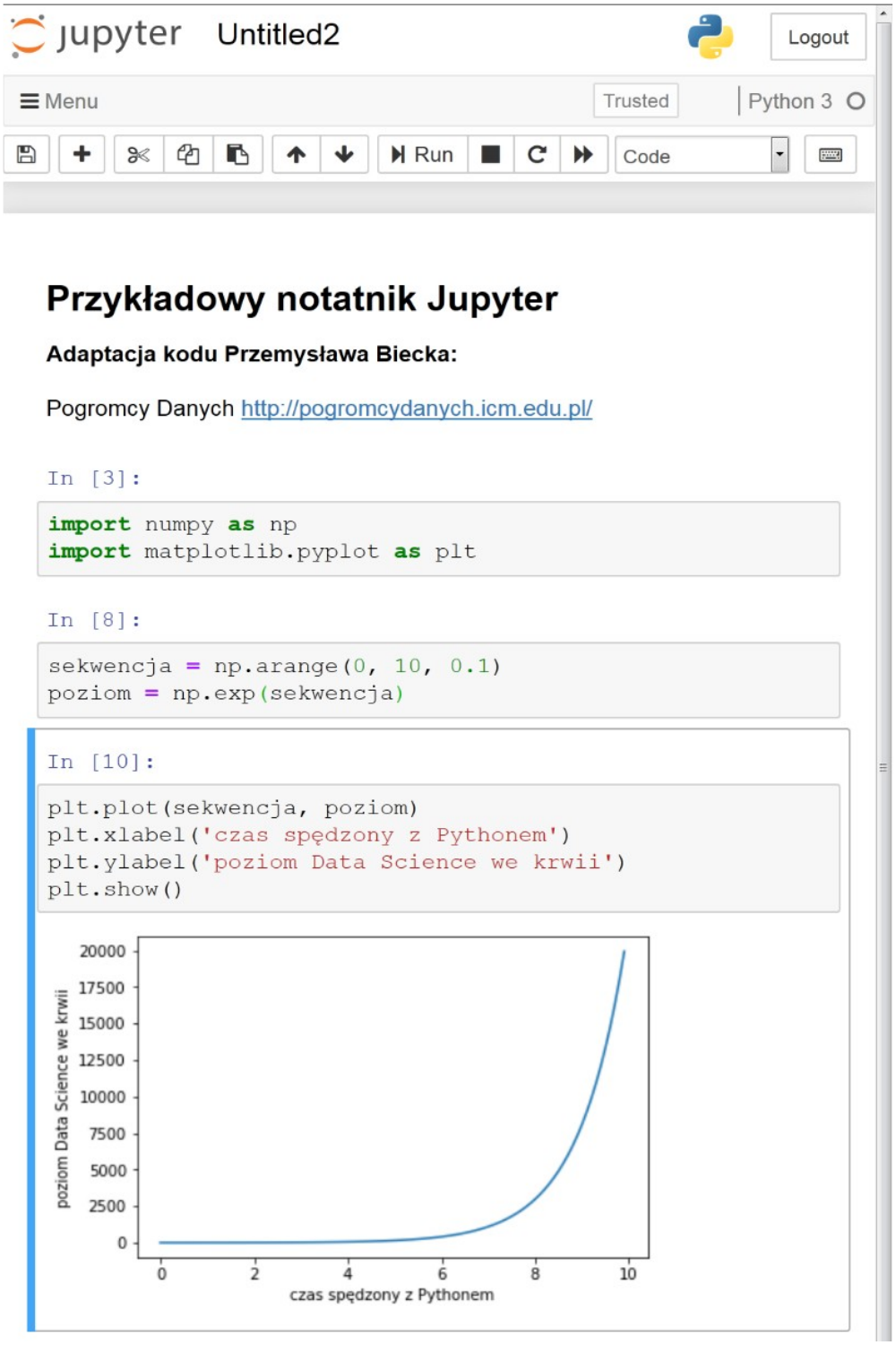
> Zauważyłem ciekawy trend. Wybierz dowolną dziedzinę nauki X, od archeologii po zoologię. Albo jest to teraz „obliczeniowa X" (*computational X*), albo wkrótce będzie. I jest to powszechnie postrzegane jako przyszłość tej dziedziny (Somers 2018).

Podkreślmy, że o tym samym znacznie wcześniej pisał Krzysztofek (2011:125): „wraz z coraz szerszym zakresem *computingu* w badaniach wszystkie nauki staną się poniekąd informatycznymi". Pérez z zespołem – twórcy Jupytera – wskazali prasie, że logo Jupyter (por. rys. 26.) jest wzorowane na rysunku księżyców Jowisza autorstwa Galileusza. Porównali oni swoją pracę nad notatnikiem Jupyter właśnie do Galileusza. On także zbudował sobie swój własny teleskop (Somers 2018). W tekście pojawia się także wątek odpływu naukowców do pracy w DS (ibidem), o czym powiemy w rozdziale 6. naszej rozprawy.

Przykładowy notatnik przedstawiamy na rysunku 27. Zauważmy, że notatnik składa się z komórek (*cells*), zarówno tekstowych jak i tych z kodem. Pierwsza komórka zawiera tekst „Przykładowy notatnik..." sformatowany w języku znaczników Markdown. Widoczny adres strony www jest działającym hiperłączem. Kolejne komórki zawierają kod w języku Python wersji 3. (numer wersji widoczny jest w prawej, górnej części poniżej przycisku „Logout"). Rezultatu działania kodu są także widoczne, w naszym przykładzie jest to wykres rysowany za pomocą kodu w komórce czwartej, ostatniej. W komórce drugiej kod ładuje odpowiednie pakiety, a w trzeciej wykonuje obliczenia i przypisuje je do zmiennych, więc nie ma żadnego wypisywanego do notatnika wyniku. Liczby widoczne przy komórkach, np. „In [3]:" wskazują kolejność uruchamiania komórek. Jeżeli każdą komórkę uruchomiono by tylko raz, w odpowiedniej kolejności, to widoczne byłyby cyfry 1, 2, 3. Byłoby tak dla w całości

bezbłędnie napisanego kodu. Jak widać (por. rys. 27.) ostatnie uruchomienie komórki ma numer 10. Podczas tworzenia tego przykładu elementy kodu nie działały zgodnie z naszym planem, wprowadzaliśmy więc poprawki i modyfikacje, po czym interaktywnie sprawdzaliśmy rezultat działania. Jest to typową strategią osób zajmujących się DS (choć tak prosty przykład bez trudu napisze w całości osoba posługująca się Pythonem na co dzień[173]).

173 Jak pokażemy w części 5.4. dotyczącej wartości społecznego świata data science, uczestnicy cenią efektywność, a nie perfekcję. Zatem to, czy kod działał za pierwszym czy trzecim—czwartym razem po błyskawicznym sprawdzeniu dokumentacji lub Stack Overflow i dokonaniu poprawki nie ma większego znaczenia. Co innego, gdyby poprawki trwały bardzo długo (cokolwiek to znaczy), a w najgorszym wypadku, gdyby kod nie działał.



Rysunek 27: Notatnik Jupyter – interfejs użytkownika z przykładowym tekstem, hiperlinkiem, kodem w języku Python i wizualizacją danych

Źródło: opracowanie własne na systemie Windows 7

Notatnik Jupyter ma wiele rozmaitych funkcji, wspiera łącznie około 40 języków programowania. Wiele poleceń można wyklikać, można także korzystać ze skrótów klawiaturowych. Istnieje szereg tzw. magicznych komend, poprzedzonych znakiem „%” dla

linii i dla komórek (*line magics*, *cell magics*). Pozwalają one np. na pisanie kodu w HTMLu lub JavaScriptcie, na korzystanie z wiersza poleceń, przekazywanie zmiennych pomiędzy notatnikami, uruchamianie gotowych skryptów Pythona czy notatników Jupyter z poziomu aktywnego notatnika (por. Rogozhnikov 2016).

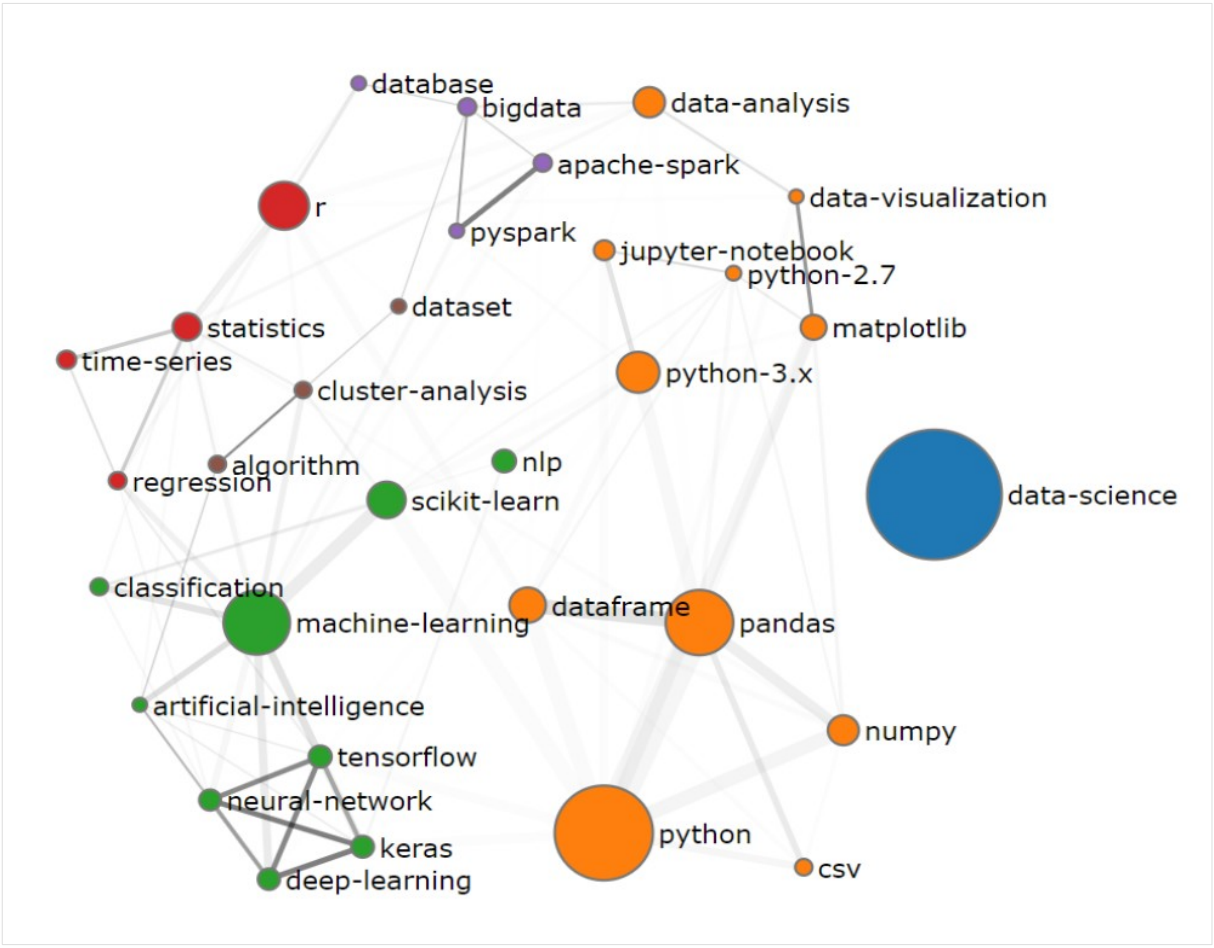
Notatnik Jupyter może być **uruchomiony lokalnie**, tzn. na własnym komputerze lub na serwerze wewnętrznym (maszyną w sieci lokalnej, bez połączenia przez internet), a także **uruchomiony w chmurze**, czyli na serwerze zewnętrznym (maszyną położoną gdziekolwiek na świecie, połączenie przez internet). Na bazie Jupytera zbudowany jest notatnik chmurowy Google Colaboratory, zwany Colab. Według oficjalnego stanowiska notatnik ten ma pomagać rozpowszechniać studia i badania w dziedzinie uczenia maszynowego (Google 2019a). Jest to narzędzie zbliżone do notatnika Jupyter i posiadające wiele z jego funkcji, a także dodatkowe ułatwienia – przykładowo jeżeli uruchomiona komórka z kodem zwraca błąd, uaktywnia się przycisk do wyszukania odpowiedzi na ten błąd na Stack Overflow. W odróżnieniu od Jupytera, Colab nie ma instalacji lokalnej. Z Colab można korzystać od razu po zalogowaniu się przez przeglądarkę internetową na swoje konto Google, także z urządzenia mobilnego (np. smartfona). Colab może działać z użyciem języka Python 2. i 3., domyślnie zainstalowanych jest szereg pakietów (np. TF, Keras). Dane są przechowywane i obliczenia wykonywane w chmurze Google Cloud, na tzw. maszynie wirtualnej (VM). Można korzystać z procesorów GPU i przeznaczonych specjalnie do uczenia głębokiego TPU. Całość rozwiązania jest bezpłatna (Carneiro i in. 2018). W trakcie naszych badań spotkaliśmy się z używaniem Colab podczas warsztatów {obserwacja 30, 39, 43}, raczej nie jest to narzędzie używane do celów innych niż szkoleniowe/dydaktyczne.

Duża baza napisanego kodu, liczne materiały edukacyjne, wiele odpowiedzi na pytania, łatwo dostępne spotkania twarzą w twarz z osobami korzystającymi z języków Python i R są także, jak wskazaliśmy wcześniej, uznawane za wskaźniki siły społeczności tych języków. O spotkaniach twarzą w twarz – konferencjach, meetupach, hackatonach – mówiliśmy już w części 5.1.1. Podkreślmy, że część z tych wydarzeń zorganizowana jest wokół tematu języków programowania, zaś na wszystkich zbierają się użytkownicy tych języków. Baza napisanego kodu, materiały edukacyjne i odpowiedzi na pytania to elementy społeczności języków programowania czy narzędzi w ogóle, które funkcjonują w internecie. Więcej o nich powiemy w części 5.3.5., omawiając technologie służące do komunikacji w badanym świecie. Wspomnijmy tylko, że najważniejszym ogólnodostępnym repozytorium kodu jest portal GitHub, powiązany z systemem kontroli wersjii Git, należący do firmy Microsoft.

Przestrzenią zadawania pytań, odpowiadania na nie, a przede wszystkim wyszukiwania odpowiedzi jest wielokrotnie wspominane przez nas forum Stack Overflow (SO). Jest to forum przeznaczone raczej do pomocy technicznej, dotyczącej języków programowania do wszelkich zastosowań. SO jest częścią portalu Stack Exchange, w skład którego wchodzą różne fora, także nie związane z technologiami. Istnieje tam także forum datascience.stackexchange.com, przeznaczone raczej do pomocy merytorycznej. Na forach Stack Exchange każde pytanie ma od jednego do pięciu tagów. Jest to folksonomiczne (Peters 2009), ale moderowane klasyfikowanie pytań do kategorii tematów, zbliżone do znanego z mediów społecznościowych. Na SO istnieje także tag „data-science".

Aktem samowiedzy nazywamy narzędzie TagOverflow, do analizy i wizualizacji tagów Stack Exchange, stworzone przez wielokrotnie wymienianego Piotra Migdała i Martę Czarnocką-Cieciurę. Na poniższym rysunku 28. przedstawiamy mapę sieci tagów współwystępujących z tagiem „data-science" na forum Stack Overflow, utworzoną właśnie przy pomocy tego narzędzia. Wierzchołki sieci (*nodes*) to 32 najczęściej występujące tagi. Powierzchnia wierzchołków – okręgów – reprezentuje ilość tagów na SO. Łącznie jest ich około 2,9 tys. (dokładniej ich liczby nie da się sprawdzić w TagOverflow). Liczbę około 2,9 tys. reprezentuje powierzchnia niebieskiego okręgu dla tagu „data-science". Wierzchołki (*nodes*) mogą być połączone krawędziami (*edges*). Szerokość krawędzi reprezentuje ilość pytań oznaczonych oboma tagami. Cień krawędzi reprezentuje siłę związku pomiędzy tagami. Im ciemniejsza krawędź, tym większa siła związku. Obliczono ją jako stosunek obserwowanych do oczekiwanych (*observed to expected ratio*, O/E)[174] (Czarnocka-Cieciura i Migdał 2015). Cała wizualizacja jest interaktywna, rysunek (rys. 28.) nie oddaje wielu jej funkcji – wyświetlania dodatkowych informacji po najechaniu kursorem myszy na element. Kolor wierzchołków zależy od siły związku między tagami, i pokazuje gęściej połączone podsieci (*densely connected subgraphs*), co autorzy nazywają wykrywaniem społeczności (*community detection*). Bazuje to na wynikach badań zrealizowanych w ramach pracy doktorskiej Migdała (Migdał 2014). Tag „data-science" nie jest połączony z innymi tagami, ponieważ współwystępuje z każdym z nich, i właściwie posłużył nam jako filtr tagów z całego SO (por. rys. 28.).

174 TagOverflow rysuje krawędź łączącą dwa tagi tylko wtedy, kiedy O/E jest większe od 1., czyli tagi współwystępują razem częściej niż przypadkowo (*random chance*). O/E obliczono jako: (ilość wszystkich pytań * ilość pytań z tagami A i B łącznie) / (ilość pytań z tagiem A * ilość pytań z tagiem B). O/E może być mniejsze od 1, co oznaczałoby, że tagi współwystępują rzadziej niż przypadkowo.



Rysunek 28: Mapa sieci tagów współwystępujących z tagiem "data-science" na Stack Overflow

Źródło: TagOverflow (Czarnocka-Cieciura i Migdał 2015), stan na 23.07.2019 r., https://p.migdal.pl/tagoverflow/?site=stackoverflow&size=32&tag=data-science, dostęp 23.07.2019 r.

Widoczne jest pięć grup tagów (rys. 28.). Kolorem pomarańczowym zaznaczono te dotyczące języka Python i podstawowych pythonowych narzędzi do analizy i wizualizacji danych. Są to tagi:

- „python" – jako język programowania ogółem;
- „python-3.x" i „python-2.7" – pytania dotyczące różnych wersji języka;
- „jupyter-notebook" – omawiany wyżej notatnik do obliczeniowych narracji;
- „pandas" – pakiet Pythona dostarczający struktur danych i narzędzi do ich analizy (The pandas project 2018);
- „numpy" – pakiet Pythona do pracy z danymi w formie tabelek oraz obliczeń naukowych i inżynieryjnych (Oliphant 2006);
- „csv" – popularny format pliku tekstowego z danymi tabelarycznymi, dosłownie oznacza wartości rozdzielane przecinkami (*Comma Separated Values*);

- „dataframe” – dosłownie ramka danych, jest to struktura danych w formie tabelki, podstawowa dla pakietu pandas. Ta sama nazwa oznacza podobną strukturę w języku R (widoczna krawędź łącząca tagi „dataframe” i „r” na rys. 28.);
- „data-analysis” – ogólnie analiza danych jako czynność;
- „data-visualization” – jw. wizualizacja danych;
- „matplotlib” – pakiet do wizualizacji danych w języku Python (Hunter 2007; Matplotlib 2018);

Kolorem czerwonym zaznaczono tagi związane z GNU R i jego typowymi zastosowaniami:

- „r” – język R ogółem;
- „statistics” – statystyka ogółem;
- „time-series” – dosłownie szeregi czasowe, oznacza dane, dla których kolejne obserwacje reprezentują kolejne momenty w czasie, przeważnie w stałym interwale;
- „regression” – regresja jako rodzaj tzw. nadzorowanego uczenia maszynowego (czyli takiego, gdzie trenujemy model na znanych wartościach zmiennej zależnej), polega na estymacji numerycznej wartości zmiennej zmiennej zależnej za pomocą zmiennych niezależnych. Mogą być użyte różne metody, np. od regresji liniowej przez regresyjne drzewa decyzyjne po sieci neuronowe (Larose 2005:12–13). Zauważmy krawędź łączącą ten tag z tagiem „machine learning”;

Kolor zielony ma grupa dotycząca uczenia maszynowego:

- „machine-learning” – uczenie maszynowe jako temat;
- „deep-learning” – jw. uczenie głębokie;
- „neural-network” – sieć neuronowa jako metoda;
- „artificial-intelligence” – sztuczna inteligencja jako temat;
- „tensorflow” – omawiany wyżej pakiet do uczenia maszynowego firmy Google;
- „keras” – omawiana nakładka ułatwiająca korzystanie z TF i Theano;
- „scikit-learn” – pakiet języka Python dla szerokiej gamy metod uczenia maszynowego (Pedregosa i in. 2011; scikit-learn 2018);

- „nlp” – skrót oznaczający dziedzinę przetwarzania języka naturalnego (*natural language processing*);
- „classification” – podobnie jak w przypadku regresji, podstawowa różnica to kategorialna zmienna zależna (Larose 2005:14–15);

Brązowy kolor oznacza najmniej spójną grupę, do której należą tagi:

- „dataset” – oznacza dosłownie zbiór danych;
- „algorithm” – dosłownie algorytm;
- „cluster-analysis” – w odróżnieniu od regresji i klasyfikacji jest to rodzaj nienadzorowanego uczenia maszynowego (model jest trenowany bez znanych wartości zmiennej zależnej). Chodzi o znalezienie struktury w danych, podział na klasy wyestymowane z danych na zasadzie podobieństwa. Używane są różne metody, np. k-średnich (*k-means*), sieci Kohonena (Larose 2005:16-17);

Fioletowy kolor to grupa tagów dotyczących pracy z dużą ilością danych, czyli big data w technicznym sensie:

- „bigdata” – jako temat;
- „database” – dosłownie baza danych;
- „apache-spark” – to powstała w 2010 r. technologia open-source rozwijana przez Apache Software Foundation. Jest to zunifikowany silnik do rozproszonego przetwarzania danych (*unified engine for distributed data processing*). Więcej powiemy o tym w części 5.3.4., na razie wskażmy jedynie, że jest to rozwiązanie do pracy z dużą ilością danych na wielu komputerach, szczególnie, jeśli dla takich danych ma być wykonane uczenie maszynowe. Zunifikowanie oznacza, że Spark łączy w jedną platformę funkcje znane z innych narzędzi do big data – tak jak smartfon łączy w jedno funkcje telefonu komórkowego, aparatu fotograficznego i GPSu (Zaharia i in. 2016:56–57);
- „pyspark” – to rozwiązanie pozwalające korzystać ze Sparka przy pomocy języka Python (Drabas, Lee, i Karau 2017).

Powyższa analiza (rys. 28.) jest niewątpliwie ogromnym uproszczeniem, i na pewno nie nagłaśnia cichych głosów (por. Clarke 2003:561; 569, 2005:85, 2015:108; Clarke i Star 2008:119, 124, 129). Jednak stosujemy ją , by **w ogólnym obrazie podobieństw pomiędzy**

językami Python i R (oba open source, skryptowe, interpretowane, relatywnie łatwe/proste, z silną społecznością) **zarysować różnice**. Zauważmy, że powoływanie się na społeczność użytkowników przy wyborze języka programowania jest przykładem tzw. bezpośredniego efektu sieciowego, wykorzystującego dodatnie sprzężenie zwrotne. Wartość przyłączenia do sieci (użytkowników danego języka/narzędzia) zależy jakoby od liczby już obecnych w sieci użytkowników (tego języka/narzędzia) (por. Surma 2017:62). Jak mówiliśmy, obecnie baza użytkowników Pythona w DS jest zdecydowanie większa niż użytkowników R (Borowiecki i Mieczkowski 2019:36–37; Nunns 2017; Piatetsky 2017, 2018c; Strong 2018). Nie sądzimy jednak, że język R obumiera. Raczej staje się niszą, związaną z określonymi zastosowaniami. Norvig (2001) w cytowanym wyżej fragmencie mówił, by używać technologii, których używają Twoi znajomi. Mówiąc tą retoryką, nieco inne grupy znajomych używają różnych języków i nie oznacza to, że jedna grupa z jakichkolwiek powodów całkowicie porzuci swoje narzędzie na rzecz narzędzia innej grupy, choć pewne zastosowania mogą być przedmiotem sporów.

Przy tym, **choć języki Python i R są w części zadań konkurencyjne, to w innej są komplementarne, zaś w jeszcze innej – rozłączne**. Inaczej mówiąc część zadań można wykonać w obu językach w dość zbliżony (w sensie wartościującym z punktu widzenia badanego świata) sposób, i tutaj trwają spory – które można opisać jako arenę – dotyczące tego, który język będzie „lepszy". Takie spory są wspierane tak argumentami ilościowymi, czyli głównie pomiarami szybkości działania kodu, jak i np. wygodą pisania, elegancją kodu, osobistą preferencją. To nazywamy konkurencją. Inne zadania są uznawane za zdecydowanie „lepsze" do wykonania w jednym języku niż w drugim. Jeżeli są to zadania związane ze sobą, tworzą proces w którym „dobrze" będzie pewne elementy wykonać w Pythonie, inne w R, to komplementarność. Rozłączność to specjalizacja, gdzie wykorzystywany jest przede wszystkim jeden lub tylko jeden z języków. To, jakie zadania są uznawane za konkurencyjne, komplementarne i rozłączne także jest przedmiotem sporów na arenie (por. część 6.2.3).

Podsumowując, **języki programowania są uznawane w badanym świecie za narzędzia niezbędne do prawidłowego wykonywania działania podstawowego**. Nie mogą być to jednak dowolnie dobrane języki programowania. Dominuje wykorzystanie języka Python. O ile do lat 2016-2017 użycie Pythona i R było zbliżone, to później jego popularność wzrosła wraz z rozwojem mody na używanie metod uczenia głębokiego. Uczenie głębokie w małym stopniu jest rozwijane w języku R. Oba języki mają szereg cech technicznych, traktowanych jako zalety w badanym świecie, i generalnie służą tutaj do pisania skryptów, nie

do tworzenia oprogramowania. Oba języki są uznawane za takie, wokół których wytworzyła się silna społeczność. Te społeczności są czymś innym niż świat społeczny DS, zatem osoby korzystające z tych języków mogą nie uważać się za część badanego świata. Język R uznajemy za technologię odziedziczoną ze statystyki akademickiej, zaś Pythona za język zaadaptowany z szeroko pojętego IT. Te historyczne różnice znajdują wyraz w różnicach dotyczących cech technicznych obu języków, jak i w społeczności osób korzystających z nich w DS.

5.3.2. Bazy danych

Działanie podstawowe świata społecznego DS jest zawsze nastawione na dane w formie cyfrowej. Co badany świat uznaje za dane? Zgadzamy się z Lowriem, że **osoby zajmujące się DS mogą traktować jako dane wszystko to, co zapisane jest cyfrowo** (Lowrie 2017:9). W całej naszej pracy pojawiają się pojęcia danych ustrukturyzowanych i nieustrukturyzowanych. Warto doprecyzować to zagadnienie. Typologię źródeł danych przedstawiamy w Tabeli 1., za zespołem zadaniowym HLG UNECE[175] (UNECE 2014), pracującym nad możliwościami wykorzystania tzw. big data w statystyce publicznej. Nie jest to zatem typologia wypracowana przez badany świat społeczny, ale naszym zdaniem pomaga w zrozumieniu kwestii struktury danych.

Tabela 1: Typologia źródeł danych w podziale na strukturę, pochodzenie i tzw. 4Vs of big data

Cecha	Dane generowane przez użytkowników	Dane z systemów biznesowych	Dane generowane przez maszyny
Struktura	Nieustrukturyzowane	Ustrukturyzowane	Ustrukturyzowane
Pochodzenie	Ludzie	Systemy informacyjne	Urządzenia/sensory
Szybkość pojawiania się (*Velocity*)	Szybko (różnice w zależności od zjawiska)	Jak dane z tradycyjnych źródeł	Bardzo szybko
Rozmiar (*Volume*)	Duży	Mniejszy	Bardzo duży
Różnorodność (*Variety*)	Wysoka	Niska	Bardzo wysoka
Wiarygodność (*Veracity*) zapewnia	Dostawcy/firmy i instytucje	Firmy	Technologia/dostawcy

Źródło: opracowanie własne na podstawie (UNECE 2014)

Dane nieustrukturyzowane scharakteryzowano w kolumnie „Dane generowane przez użytkowników" w tabeli 1. Są to, według autorów dokumentu, zapisy ludzkich doświadczeń, dawniej utrwalane w książkach bądź innych dziełach sztuki. Obecnie to różnego rodzaju

175 High-Level Group for the Modernisation of Statistical Production and Services (HLG), United Nations Economic Commission (UNCE)

teksty, fotografie, dźwięki i materiały wideo, wszystko w formie zdigitalizowanej. Dane, których źródłem są ludzie (*human-sourced*) przechowywane są w wielu miejscach: od dysków komputerów osobistych czy smartfonów po serwery różnych usług internetowych, np. mediów społecznościowych. Ich struktura jest „luźna", dane raczej nie są zapisywane w uporządkowany sposób (tabelarycznie). Do tego typu informacji przyporządkowano takie źródła, jak: portale społecznościowe (np. Facebook, Twitter, Tumblr), blogi i komentarze, prywatne dokumenty elektroniczne, portale z fotografiami (np. Instagram, Flickr, Picasa), portale z wideo (np. Youtube, Netflix), wyszukiwarki internetowe, urządzenia mobilne, wiadomości tekstowe, portale umożliwiające generowanie map przez użytkowników oraz poczta elektroniczna. Inaczej jest z danymi, których charakterystykę pokazano w kolumnie drugiej i trzeciej powyższej tabeli - tam dane są silniej ustrukturyzowane. Podstawową różnicą jest tutaj to, że dane o charakterze nieustrukturyzowanym są treściami generowanymi przez ludzi, a pozostałe czyli ustrukturyzowane - nie, choć mogą być rejestracją bądź pomiarem działań ludzkich. Wpisy i zdjęcia utworzone przez użytkowników portalu społecznościowego nie są przechowywane w formie relacyjnej bazy danych z określonymi atrybutami (zmiennymi). Dane pozyskane z systemów śledzących użytkowników internetu np. z Google Analytics (śledzenie zachowania na stronach internetowych bazujące na pliku w przeglądarce internetowej tzw. ciasteczku, *cookie-based tracking*) czy Facebook Pixel (śledzenie bazujące na koncie konkretnego użytkownika Facebooka, *people-based tracking*), dane o transakcjach, ruchu kursora na stronie czy kliknięciach taką ustrukturyzowaną formę przybierają.

Dane w formie ustrukturyzowanej przechowywane są od kilkudziesięciu lat w formie relacyjnych baz danych z dostępem przez SQL, zatem są uznawane za źródło tradycyjne (por. tab. 1.). Zanim jednak więcej o nich, dodajmy, że w tzw. big data (w sensie technicznym) również przy danych ustrukturyzowanych stosowane są inne, nierelacyjne architektury baz danych, m.in. z uwagi na problem tzw. skalowalności[176].

Wyjaśnienia wymagają jeszcze dwie sprawy: skąd się biorą dane, zanim zostaną gdzieś przechowane, oraz czy dane zawsze są przechowywane w bazach danych. Skąd się biorą pokazuje częściowo tabela 1., ale to tylko jedno z ujęć, poza tym nie jest wypracowane przez badany przez nas świat ani nie wskazuje na sposób pozyskiwania danych czy uzyskania dostępu do danych, ani w sensie technicznym, ani organizacyjnym.

176 Generalnie rozumianej jako możliwość rozbudowy wielkości i jednoczesnego zachowania odpowiedniej wydajności systemu, co może odnosić się tak do technologii, jak i np. całej organizacji

Z perspektywy firm, 78% badanych podmiotów w pytaniu wielokrotnego wybory wskazało, że **dane na potrzeby projektu zapewnia im klient**, 77% że **dane zapewniają oni sami – firma AI**, 73% wskazuje na **korzystanie z danych darmowych w wolnym dostępie** (*open access*), zaś najmniej (31%) deklarowało **zakup danych** (Borowiecki i Mieczkowski 2019:42–43). Autorzy podsumowali te wyniki stwierdzeniem, że polskie firmy AI korzystają z każdego dostępnego im źródła danych.

Wtedy, kiedy **dane zapewnia klient**, raczej będą to dane z bazy danych lub dane w plikach. W tym przypadku inaczej będzie, kiedy chodzi o zespół DS pracujący w ramach większej firmy, która jest tzw. klientem wewnętrznym tego zespołu, a inaczej przy pracy zespołu data scientistów dla tzw. klienta zewnętrznego. W pierwszym przypadku zespół DS raczej samodzielnie korzysta z wewnętrznych, firmowych baz danych, na potrzeby realizowanych projektów i eksperymentów. Członkowie zespołu DS mają nadane przez tzw. administratorów uprawnienia dostępu do danych. Uprawnienia mogą być różne w zależności od ról w zespole DS. Niemniej data scientiści raczej samodzielnie pobierają i łączą dane z różnych źródeł w firmie – samodzielnie, czyli pojedyncza osoba na swoim komputerze może napisać kod (np. w języku SQL), który wykonuje takie operacje pobrania czy łączenia danych z rozmaitych baz firmy, jakie ta osoba uzna za stosowne w danym momencie (w ramach swoich uprawnień). Udział zespołów IT, administratorów baz danych czy osób zajmujących się bezpieczeństwem danych, prawników, osób decyzyjnych jest w tym przypadku sporadyczny, znikomy w codziennej pracy pojedynczego data scientista. Przy pracy dla klienta zewnętrznego zespół DS może uzyskać zbliżony, samodzielny dostęp do baz danych na potrzeby realizacji konkretnego projektu, jednak wymaga to szeregu działań technicznych, prawnych i organizacyjnych. Pojawiają się tam kategorie **projektu jako tajemnicy firmy-realizatora**, **danych jako tajemnicy firmy-klienta**, oraz **strachu klienta przed udostępnieniem jego danych**:

> *B: OK, a dobra. Czy mógłbyś opowiedzieć o jakimś ostatnim projekcie czy projektach które robicie w zespole?*
> R: [cmoknięcie] Mogę w dość ogólny sposób, ponieważ to wiadomo obowiązuje mnie prywatność yyyy,
> *B: Wiem że są tajemnice wszelakie w tej branży*
> R: Tak tak tak tak. Znaczy powiem szczerze że to w ogóle, to co dotknąłeś jest **bardzo** ciekawym problemem który uważam że może też jest wart eksploracji, eee [pauza] my mówimy dobra, my jesteśmy od big daty, podchodzimy holistycznie, podchodzimy uczeniem maszynowym, tak? I firmy mówią **świetnie**! Bo nam jest bardzo trudno [cmoknięcie] nie wiem czy kiedyś pracowałeś albo miałeś kontakt jak działają duże przedsiębiorstwa, tak?
> *B: Mhm?*
> R: Wyobraź sobie że OK tutaj siedzimy w firmie w której pracuje mniej więceeeejjjj 200 000

> osób. Tak? Na całym świecie. Które, i około 30% z nich co roku odchodzi i przychodzi. I w firmie tego rozmiaru **bardzo** trudno jest [cmoknięcie] **wiedzieć na pewno** co się dzieje na niższych szczeblach. To jest ogromny problem, nie tylko tej firmy. Wszystkich firm globalnych. (...) No i teraz mówimy słuchajcie, my mamy technikę że my możemy to zrobić dla was, tak? Wziąć te dane, połączyć, stworzyć coś mądrego. No i oni mówią na początku no faktycznie, wow, no boże zróbmy to, nie? No a potem po roku pracy okazuje się że **ja**, po przeprowadzeniu dwóch lub trzech dużych projektów, zaczynam **baaardzo** dużo wiedzieć o tej firmie.
> *B: Mhm, o tym kliencie tak?*
> R: Tak. No i teraz na przykład spotykam się z kimś kto jest członkiem zarządu. I zajmuje się finansami. Globalnej firmy wartej miliardy dolarów. I po godzinnej rozmowie i prezentowaniu moich wyników on mówi [imię rozmówcy] słuchaj, tak mnie naszło [pauza] że ty wiesz więcej o tej firmie niż ja.
> *B: No tak.*
> R: Kurcze nie? I on mówi wiesz, jakbyś ty z tym **poszedł**, do gazety, to możesz mieć realny wpływ na wartość akcji, naszej firmy. To są to są **miliardy** dolarów. Iiii w wielu firmach yy ten koncept właśnie big data i data science uderzył w **ścianę,** [pauza] wiesz takie na początku wszyscy byli mega entuzjastyczni, potem się okazało kurcze, nie? Oni mają te dane, oni to wszystko wiedzą, modelują, no fajnie ale mogą są z tego potencjalne zyski ale też jest potencjalne duże ryzyko.
> *B: No tak.*
> R: I narzucono na nas baaardzo dużo regulacji. Dla mnie dzisiaj, w poważnym projekcie, który ja sprzedaję za milion dolarów powiedzmy, [pauza] zanim ja zobaczę dane, lub mój zespół, mija minimalnie osiem miesięcy. Od podpisania umowy. 8 do 12 miesięcy trwa [pauza] przeprowadzenie tego wewnątrz firmy klienta i wewnątrz firmy powiedzmy mojej która to oferuje takie usługi. Tak? No 6 miesięcy to jest **dobry** wynik tak? Oczywiście na początku wszyscy mówią że to będzie trwało 6 dni, nigdy nie trwa 6 dni. (...) potrzebujemy teraz we własnych zespołach ludzi, którzy nie znają się na analizie danych, ale znają się na tym jak działa korporacja i są w stanie przeprowadzić ten proces. Od stron prawno legislacyjno, przekonywania ludzii, stworzenia odpowiednich omów dla, klauzul ufności, czego tam jeszcze, żeby to się w ogóle mogło wydarzyć, tak? Dodatkowy zawód się wytworzył przy przy analizie danych [z uśmiechem]. {wywiad 6}

W książce, którą traktujemy jako akt samowiedzy świata DS w USA zwrócono uwagę na **dychotomię – dane są raczej zamknięte** (*proprietary*)**, traktowane jako tajemnica nawet wewnątrz jednej firmy. Technologie, kod i stosowane w DS metody przeciwnie, są raczej otwarte** (*open*) (Gutierrez 2014:217). Technologie, kod i metody są nierzadko szczegółowo prezentowane na konferencjach / meetupach / warsztatach, udostępniane w formie pakietów, publikowane na blogach / stronach www czy w czasopismach w wolnym dostępie. Natomiast nie spotkaliśmy się z prezentowaniem danych z projektu komercyjnego, często nie podaje się nawet nazwy firmy, dla której był realizowany projekt, ograniczając się do określenia w rodzaju „dla jednego z największych europejskich banków", „dla naszego klienta z branży FMCG". Nawet w wywiadach Rozmówczynie / Rozmówcy wielokrotnie mówili, że obowiązują ich formalne zobowiązania tajemnicy (np. w formie *non-disclosure agreement*, NDA) i nie mogą o czymś powiedzieć, badacz (RŻ) traktował podanie nazw klientów i podobne szczegóły jako wyraz zaufania wobec niego. Umowy typu NDA to ukonstytuowane zobowiązanie (*commitment*) uczestników świata DS wobec organizacji i

osób, które płacą za usługę. Niemniej sprawa otwartości technologii czy metod nie jest taka jednoznaczna. Z pozycji nauk społecznych czy prawników krytykuje się przecież same metody uczenia maszynowego jako tzw. czarne skrzynki. Z pozycji DS lub biznesowych rzeczników świata DS szczegóły techniczne i metodologiczne przedstawia się klientom – a właściwie nie przedstawia – jako magię. Powiemy o tym więcej w części 6.2.1. traktując modelowanie jako arenę.

Inne rozwiązanie, kiedy dane zapewnia klient, to przekazanie zespołowi DS plików z danymi. Ten scenariusz może zachodzić także przy pracy dla klienta wewnętrznego, w sytuacji, gdy jakaś część firmy z różnych powodów nie chce lub nie może dać zespołowi DS dostępu do baz. Wtedy zespół DS może być „odciążony" z samodzielnego pozyskiwania danych z baz, i otrzymać wstępnie przygotowany, pojedynczy plik z danymi do dalszej pracy. Może być to postrzegane przez data scientisów jako zaleta, bo jak wskazywaliśmy – przygotowanie danych jest raczej uznawane za nudą część projektów. Z drugiej strony, jeśli jakość danych w pliku nie pozwala na analizę lub dla tych danych modele nie działają w zadowalający sposób, brak możliwości korzystania z bazy danych ad hoc może być postrzegany jako i spowolnienie utrudnienie pracy. Brak dostępu do baz, niezależnie czy to klient wewnętrzny czy zewnętrzny, może być spowodowany także tym, że klient nie ma danych w bazie, a właśnie w plikach. Może tak być w przypadku danych w formie tekstów, obrazu lub dźwięku, co raczej nie budzi zastrzeżeń uczestników badanego świata. Inaczej jest w przypadku danych w formacie tabel, zapisanych w arkuszach kalkulacyjnych. Taki format danych może być z powodzeniem przechowywany w tzw. tradycyjnych, relacyjnych bazach danycvh, i takie przechowywanie jest w badanym świecie i jak sądzimy w szeroko pojętej informatyce także, uznawane za właściwe. Dane tabelaryczne w plikach mają potencjał bycia o wiele bardziej niespójnymi, nieuporządkowanymi, niezdefiniowanymi i zasadniczo trudniejszymi w pracy niż dane w relacyjnej bazie danych. Brak korzystania z relacyjnych baz danych – podkreślmy, technologii z kilkudziesięcioletnią historią – dla danych o charakterze tabelarycznym uznawany jest za wyraz tzw. **niskiej kultury danych**. Określenie to odnosi się także do innych aspektów korzystania z danych w organizacji: m.in. tego, czy ogólnie są one używane do wsparcia decyzji (tzw. bycie data-driven), właśnie sposobu przechowywania danych (także w bazach, bo i bazy relacyjne jak i nierelacyjne mogą być niespójne, chaotyczne, źle zaprojektowane, niekompletne), odpowiedzialności za dane, znajomości posiadanych danych, dostępu do danych, stosowanych narzędzi do pracy z danymi. Dlatego projekty DS czasami mają bardzo długą fazę przygotowania danych, ponieważ osiągnięcie

produkcyjnego wdrożenia modeli wymaga zmiany w organizacji – poprawy kultury danych w organizacji:

> R: (...) żeby dojść do tego wyniku my przeprowadziliśmy cały proces. Identyfikacji wiedzy, ludzi odpowiedzialnych za konkretne elementy tego procesu, **dane** które są w **opłakanym** stanie, to jest ogromna też jest ogromny temat. Dane najczęściej są w opłakanym stanie, nie ma do nich właścicieli, od lat są nie czyszczone, nie, standardy różne obowiązują, więc to wszystko się wyjaśnia, w jakiś sposób wiesz, na co dzień się o tym nie mówi. Ale my przychodzimy mówimy ej, czyje są te dane? No i nagle się robi ruch w firmie, ktoś mówi kurcze nie wiadomo, nie? Niczyje. Nikt nie jest za nie odpowiedzialny, ich jakość jest zerowa, więc wiesz my powodujemy jesteśmy powiem takim [cmoknięcie] stymulatorem że tą sprawę trzeba po prostu załatwić. Dlatego to trwa na przykład 8 miesięcy. {wywiad 6}

Z tego, że wykonanie zmiany w polityce organizacji wobec danych jest niezbędne dla realizacji projektu DS, firma-klient może nie zdawać sobie sprawy. Wiąże się to z tym, że uczestnicy badanego świata nierzadko postrzegają oczekiwania biznesu jako nierealne. Postaramy się pokazać ten złożony problem jako szeroką arenę dotyczącą modelowania i modeli jako obiektów granicznych (o czym w części 6.2.1.), tutaj pokażmy jedynie kwestię rozdźwięku między oczekiwaniem tzw. biznesu a kulturą danych w organizacji:

> R: (...) Się naczytają nasłuchają się tam na konferencjach że tam ten data mining i data science w ogóle jak Midas przyjdzie dotknie rękę i to wszystko się rozwiąże, natomiast bardzo rzadko na tych konferencjach mówi się o tym że **owszem,** tak może być, ale żeby to się wydarzyło, to i to trzeba wykonać. I to jest proces, i to są lata, i to są duże sumy, wchodzą, i tej świadomości często klienci nie mają. Mówią słuchajcie to jest problem, macie tutaj dane, w Excelu coś tam, i w ogóle za trzy miesiące niech to będzie wszystko automatycznie, i niech jeszcze odbiera telefony tak? Więc to jest coś czym trzeba umieć zarządzić i co też jest męczące. Tak? Wysokie oczekiwania, bardzo, tak jak mówię samo w byłoby super, natomiast nie ma tej drugiej, bardzo rzadko jest świadomość i chęć wtedy mówię super świetnie ale to będzie kosztować 10 milionów. [parsknięcie śmiechem] Nie ma sprawy. Ja to mogę zrobić. 10 milionów, tak? Yyyy no tak [ze śmiechem] {wywiad 6}

Dane przechowywane w wielu różnych, niespójnych plikach – „macie tutaj dane, w Excelu coś tam" {wywiad 6} – nie mają struktury, tzw. schematu, nie można ich jako całości agregować, przeszukiwać, generalnie wykonywać operacji analitycznych, które pozwalają zająć się semantyką danych. Najpierw, by było to możliwe, trzeba się zająć właśnie nadaniem struktury (Ceri 2018). To samo tyczy się źle utrzymanych czy zaprojektowanych baz danych – struktura przeszkadza data scientistom w dotarciu do semantyki.

Co oznacza, że **dane potrzebne do projektu zapewnia sama firma AI**? Przede wszystkim w zastosowaniach komercyjnych DS nie ma praktyk przeprowadzania badań w celu pozyskania danych. Nie spotkaliśmy się z praktyką inną niż korzystanie z szeroko pojętych danych zastanych. Także usługi takie, jak zainstalowanie śledzenia w środowisku

cyfrowym (np. Google Tag Manager, Google Analytics, Facebook Pixel) czy zainstalowanie kamer/czujników w świecie materialnym nie jest postrzegane jako to, czym zajmuje się świat DS, nawet w zakresie działań wspomagających czy pobocznych. To **zapewnienie dostępu do danych zastanych** jest tym, co robią tzw. firmy AI. Może chodzić np. o dostęp do danych z mediów społecznościowych, stron www, sklepów internetowych. Stosowane są m.in. metody tzw. scrapingu / webscrapingu, co dosłownie oznacza zdrapywanie lub zeskrobywanie danych z internetu. Nieco innymi, ale zbliżonymi metodami jest tzw. web harvesting (dosł. żniwa w sieci lub zbieranie plonów), mówi się także o skryptach realizujących scraping / harvesting jako tzw. pajączkach czy crawlerach (dosł. pełzaczach), pomijamy tutaj różnice między nimi. W formie porady dla naukowców społecznych zainteresowanych danymi z internetu przystępnie opisał scraping Jemielniak (2019:80-87). Zasadniczo są to metody pobierania danych ze środowiska cyfrowego, w których nie korzysta się z tzw. wtyczki czyli API zapewnionego po stronie źródła danych, a „zdrapuje" się dane bezpośrednio ze środowiska cyfrowego. Chodzi o to, by za pomocą komputera pobrać np. ceny książek ze strony Amazon, a nie przepisywać te ceny ręcznie jedna po drugiej (ibidem). Jemielniak pokazuje szereg narzędzi nie wymagających umiejętności pisania kodu, zaznacza tylko większą elastyczność scrapingu z użyciem Pythona lub R (ibidem). W badanym świecie korzysta się z narzędzi obsługiwanych kodem, czyli pakietów dla wybranych języków programowania. Przykładowo dla R elementem ekosystemu tidyverse jest to pakiet „rvest" (Wickham 2019b), zaś innym podejściem (symulacją działania użytkownika internetu) cechuje się zestaw narzędzi w pakiecie „Rselenium". Jest to rodzaj nakładki (*binding*) dla R na usługę Selenium WebDriver, która polega na sterowaniu przeglądarką internetową za pomocą kodu i może być obsługiwana przez wiele języków programowaniach w różnych celach, innych niż scraping (Harrison 2019).

Ponieważ scraping jest pobieraniem danych bezpośrednio ze środowiska cyfrowego, zatem odbywa się bez zgody i wiedzy podmiotów, które administrują tym środowiskiem, są jego właścicielami czy twórcami treści, zatem może to budzić wątpliwości etyczne lub prawne. Trudno czasem określić, czy konkretny element w sieci jest publiczny, i można traktować to jako zawartość mediów, czy prywatny (Shiab 2015). Spotkaliśmy się z opinią, że będąc data scientistem nie należy uczyć innych ludzi pisania tzw. pajączków, ponieważ naraża to nauczyciela na konsekwencje prawne {obserwacja 35}.

Niemniej osoby zajmujące się DS (i nie tylko one, bo np. naukowcy czy dziennikarze także) piszą pajączki. Na potrzeby analiz ilościowych prezentowanych w części 5.1. także

planowaliśmy napisać pajączka czyli skrypt z użyciem pakietu Rselenium, wzorując się na analizie meetupów przedstawionej na konferencji WhyR 2017 (Słomczyński 2017), w której badacz brał udział {obserwacja 9}. Nie udało nam się zaadoptować udostępnionego publicznie kodu pajączka (ibidem) do pobrania interesujących nas danych. Zauważyliśmy jednak, że witryna meetup.com wystawiła API, co uznawane jest za wygodniejszy i jednocześnie nie budzący wątpliwości etycznych ani prawnych sposób pozyskania danych ze środowiska cyfrowego. Dodatkowo znaleźliśmy pakiet umożliwiający dostęp do tego API z użyciem Pythona (Ferate 2016), a także tzw. tutorial – instrukcję z przykładem wykorzystania tegoż pakietu (Singleton 2018). W połączeniu z oficjalną instrukcją korzystania z API meetup.com (Meetup 2019) oraz wieloma nieudanymi próbami udało się badaczowi (RŻ) pobrać dane o 86 grupach (por. część 4.2 z opisem metodologicznym i 5.1. z wynikami). To doświadczenie nie różni się zasadniczo od doświadczeń osób zajmujących się DS, co prezentowaliśmy już pod koniec części 5.3.1. we fragmencie „w eRze nie było nic gotowego do tego, do obsługi tego, więc coś tam w Pythonie po prostu napisałem wtedy" {wywiad 9}. Przypomnijmy także, że wokół danych pozyskanych z API Facebooka obracała się tzw. afera Cambridge Analytica, co pokazaliśmy w części 2.2.3.4. Zapewnianie danych przez firmę ankietowane przez Fundację digitalpoland podmioty mogły zrozumieć także jako znalezienie przez firmę danych w wolnym dostępie lub zakup przez nią danych, a może dokładniej – skorzystanie na potrzeby projektu z danych w wolnym dostępie lub zakupionych, które to dane były już w zasobach firmy.

Dane uzyskane dzięki scrapingowi lub przez API mogą mieć postać plików różnego rodzaju. Mogą być to obrazy zapisane w rozmaitych formatach (jak png, jpg), a w przypadku danych tekstowych np. pliki html (czyli podstawowy format tekstowy zapisu stron www), pliki JSON lub XML (ustrukturyzowane formaty wymiany danych, mogą być odczytywane np. przez przeglądarki internetowe czy inne aplikacje). Parsowanie plików JSON na tabele do pracy w R wykonywaliśmy na potrzeby analizy danych z meetup.com (część 5.1.), do omawianych zadań istnieje szereg pakietów (Couture-Beil 2018; Ooms 2014).

Firmy AI ankietowane przez Fundację digitalpoland wskazywały także na korzystanie z danych w wolnym dostępie (*open access*). Wracając do przykładu naszych analiz, mogą to być np. dane z sondaży (u nas Kaggle i Stack Overflow oraz konferencji PyData i WhyR, por. część 5.1.), i właściwie z dowolnych badań, naukowych czy innych, dostępnych w publicznych repozytoriach. Podkreślmy, że jest szereg powszechnie znanych w badanym świecie zbiorów danych w wolnym dostępie, np. stronie Wydziału Informatyki (*computer*

science) Uniwersytetu w Toronto. Zbiory te[177] wykorzystywane są obecnie głównie w nauczaniu metod analizy danych czy machine learningu, a także w badaniach nad metodami/algorytmami – służą jako punkt odniesienia do tzw. benchmarku. Źródłem zbiorów danych może być sam portal Kaggle (publikuje dane do konkursów), rozmaite strony rządowe czy statystyki publiczne, znana w badanym świecie jest także wyszukiwarka do zbiorów danych[178] firmy Google.

Jednocześnie podkreślmy, że **w DS nie chodzi o samo zdobycie informacji**, np. o porównywanie danych z publicznie dostępnych raportów o kondycji finansowej spółek. Niemniej np. publicznie dostępne dane pogodowe w połączeniu z danymi „wyskrobanymi" z mediów społecznościowych i danymi o sprzedaży z firmowej bazy danych mogą posłużyć do modelowania interesującego firmę-klienta zjawiska. W przypadku uczenia głębokiego (*deep learning*) można także wyuczyć model na publicznym zbiorze danych i zastosować do danych klienta, ewentualnie douczając model na jego danych. Nazywa się to transfer learning (dosł. uczenie przeniesione), i oznacza skorzystanie z właściwości modelu wyuczonego przeważnie na łatwo dostępnym, dużym zbiorze danych (np. CIFAR-100) do zbliżonego zadania, dla którego klient ma mało danych (np. klasyfikacji zdjęć medycznych) (Ferrucci i in. 2010:76; Manyika i in. 2017:140; Ng 2018b:71; Nicolaus i in. 2016:11). Czymś nieco innym, lecz też traktowanym jako transfer learning, jest skorzystanie z gotowego, wyuczonego już modelu w otwartym dostępnie, np. ResNet50 (He i in. 2015), i douczenie z ewentualnymi zmianami w architekturze na własnych danych docelowych {obserwacja 39}. To robią raczej osoby zajmujące się DS, nie tylko do celów komercyjnych. Na marginesie dodajmy, że można na podobnej zasadzie próbować korzystać z już wyuczonych modeli do własnych zadań, w ogóle nie budując ani nie trenując na żadnych danych własnego modelu. Istnieją platformy zwane marketplace (dosł. plac targowy), gdzie można kupić gotowy model i jego implementację np. w chmurze giganta technologicznego (Amazon 2019b; Microsoft 2019a). Z takich modeli korzystają raczej firmy-klienci, nie spotkaliśmy się z praktyką kupowania modeli na markeplace'ach przez uczestników badanego świata.

177 Sądzimy, że do najbardziej znanych należą m.in.: MINST (obrazy – odręcznie pisane cyfry, stosowane do zadań klasyfikacji), CIFAR-10 (obrazy – obiekty w dziesięciu kategoriach, także do zadań klasyfikacji; jest także np. CIFAR-100), Titanic Survivor (tabela – dane o każdym pasażerze Titanica wraz z informacją przeżył/nie przeżył, zadania klasyfikacji binarnej), Boston Housing (tabela – ceny mieszkań w Bostonie, zadania regresji), a przede wszystkim Iris (tabela – pomiary kwiatów trzech gatunków irysa, zadania klasyfikacji) Ronalda Fishera z 1936 r. Dane te łatwo pobrać z internetu, powołuje się na nie wiele kursów/tutoriali a także artykułów amatorskich i naukowych, dostępne są także jako komponent popularnych pakietów Pythona czy R.

178 https://toolbox.google.com/datasetsearch, dostęp 7.04.2020 r.

Najrzadziej wskazano na **zakup danych** – podczas, gdy pozostałe scenariusze to dla każdego ponad 70% wskazań, to zakup danych deklarowało 31% ankietowanych firm (Borowiecki i Mieczkowski 2019:43). Zakup danych oferują różne podmioty, zarówno komercyjne (np. Acxiom, Experian, ChoisePoint; por. Crain 2018:88) jak i publiczne (Główny Urząd Statystyczny 2019). Widzimy trzy powiązane wzajemnie, potencjalne wyjaśnienia mniejszego zainteresowania kupowaniem danych niż danymi z innych źródeł. Po pierwsze, zakup danych to dodatkowy koszt w projekcie, a to przeważnie będzie postrzegane jako wada. Po drugie, wśród osób zajmujących się DS może panować przekonanie, że są w stanie samodzielnie poradzić sobie z pozyskaniem potrzebnych danych, dzięki własnej dociekliwości i umiejętnościom koderskim. Zatem zakup danych może być postrzegany jako pójście na łatwiznę lub zbędny wydatek. Po trzecie, działalność tzw. brokerów danych – czyli firm zbierających dane i sprzedających je rozmaitym odbiorcom – była przedmiotem krytyki etycznej i prawnej (por. Crain 2018), więc korzystanie z ich usług może być postrzegane jako nieetyczne, ryzykowne w sensie potencjalnych konsekwencji prawnych lub niekorzystne dla wizerunku firmy / pojedynczych osób zajmujących się DS. W części 2.2.3.3. mówiąc o krytyce wobec nieprzejrzystego działania systemów bazujących na modelach uczenia maszynowego przywołaliśmy historię Helen Stokes. Odmawiano tej osobie wynajęcia mieszkania z powodu informacji o jej dawnych aresztowaniach, znajdujących się w zbiorach danych zakupionych od brokerów. Informacje te zostały dawno wykreślone z oficjalnych rejestrów, ponieważ kobieta nie została skazana. O'Neil stwierdziła jednak, że dla brokerów nie miało to znaczenia, ponieważ „osoby takie jak Helen Stokes nie są ich klientami. Są produktem, a reagowanie na ich skargi kosztuje" (O'Neil 2017a:208).

Podkreślmy, że **w jednym projekcie**, szczególnie jeżeli mówimy o działalności komercyjnej DS, korzysta się z wielu źródeł danych różnego pochodzenia – **łączy się wiele zbiorów danych** pochodzących od klienta, zapewnionych przez firmę realizującą projekt, pozyskanych z wolnego dostępu i zakupionych.

Przechodząc do kwestii technologii **relacyjnych baz danych i języka SQL** zaznaczmy, że są to narzędzia powszechnie używane w informatyce, zdecydowanie nie tylko w związku z analizą czy modelowaniem danych. Mają one zastosowanie w wielu dziedzinach działalności człowieka, np. w systemach bankowych, sklepach internetowych, systemach bibliotecznych, kadrowych, czy magazynowych. „Teoria relacyjnych baz danych była jak dotąd najbardziej udana próbą opanowania danych, przechowywanych w formie elektronicznej" (Kriegel i Trukhnov 2011:xxv). Relacyjne bazy danych i język SQL są abstrakcjami, które mają liczne

dystrybucje, tzw. systemy baz danych (w Polsce używany jest skrót SZRBD – Systemy Zarządzania Relacyjnymi Bazami Danych). Do popularnych należą m.in. MySQL, PostgreSQL (oba open source) i Oracle, Microsoft SQL Server (oba zamknięte, komercyjne). Istnieją także aplikacje z interfejsem graficznym do zastosowań biurowych, np. Microsoft Access (por. Kriegel i Trukhnov 2011:xxvi).

Znajomość relacyjnych baz danych i SQL może być uznawana za elementarz umiejętności osoby zajmującej się DS (Amie Ismail i in. 2016; Chen i Jiang 2018; J. Devlin 2018; Muenchen 2019; Sharp Sight 2016; Więcko i in. 2016), choć w DS są to technologie **niemodne**, a także **pomijane, bo ich znajomość bywa uznawana za oczywistą**. Nie zdecydowaliśmy się opisać ich w kategorii wiedzy milczącej (*tacit knowledge*) (Polanyi 2005:96), której poświęcamy część 5.3.5., ale podkreślamy, że są to technologie dla przeważającej liczby osób zajmujących się DS, szczególnie komercyjnie, niezbędne do pracy. Są one jednak niemodne w badanym świecie. Za modne, „seksowne" technologie bazodanowe uznaje się ewentualnie[179] te kojarzone z tzw. big data (np. Hadoop, Spark, technologie NoSQL, o czym dalej), zaś w przypadku modelowania to, jak mówiliśmy wielokrotnie, technologie uczenia głębokiego. Ilustruje to poniższy cytat:

> Dlaczego ktoś, chcący pracować z danymi ma poświęcać czas na naukę „antycznego" języka [SQL-a]? Dlaczego nie przeznaczyć całego czasu na na opanowanie Pythona/R, lub nie skoncentrować się na „seksownych" umiejętnościach – uczeniu głębokim (*deep learning*), języku Scala lub technologii Spark? (…) ignorowanie SQL-a sprawi, że będzie znacznie trudniej znaleźć pracę związaną z danymi (J. Devlin 2018).

W tym samym tekście (ibidem) wskazano, że:

- SQL jest wszędzie – do tzw. kwerend czy „odpytywania" (*query*) baz danych używają go najbardziej znane firmy stosujące DS w swoich produktach, jak Uber, Netflix, Facebook, Google, Amazon;
- istnieje popyt na pracowników znających SQL – powołując się na wyniki analizy danych z portalu Indeed, zawierającego oferty pracy, wskazano, że znajomość SQL-a była najczęściej wymienianą umiejętnością dla stanowisk pracy z danymi (dla różnych przekrojów wymieniano ją w około 36% do 56% ofert pracy);
- SQL jest językiem o ustabilizowanej pozycji – wskazując na wyniki ankiet Stack Overflow Developer Survey 2017 i 2018 stwierdzono, że popularność SQL-a jest

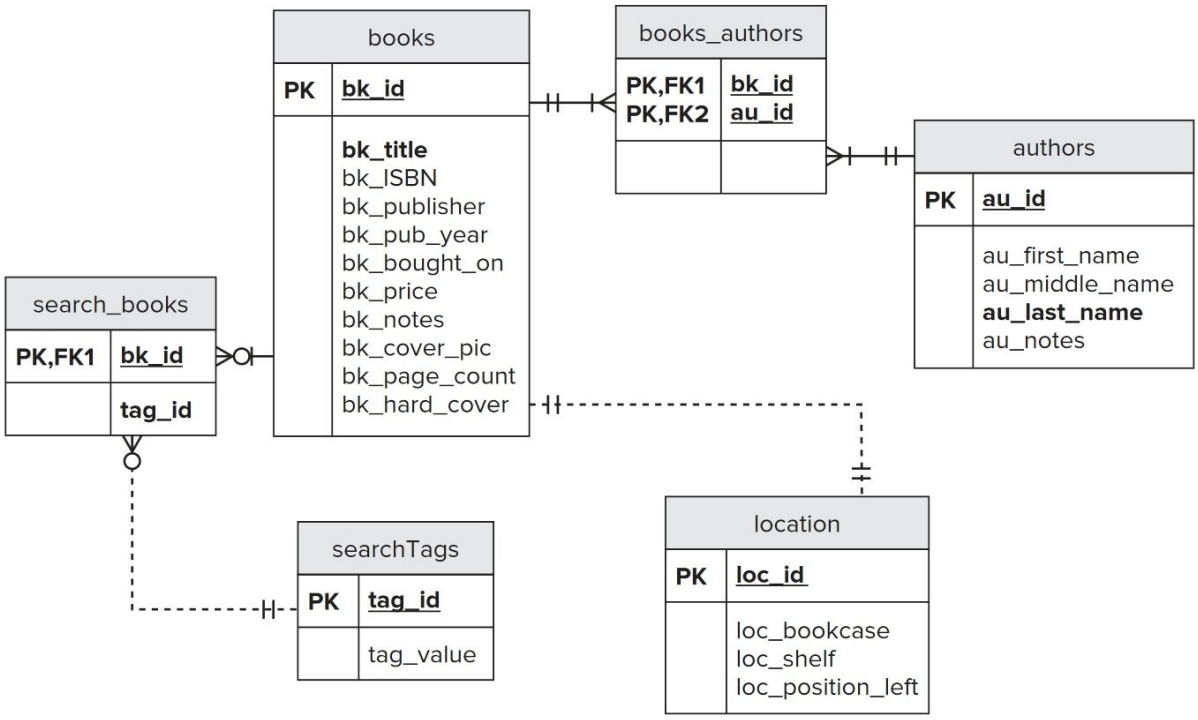
179 Ewentualnie, bowiem świat data science raczej nie zajmuje się, a na pewno nie w ramach działania podstawowego, problemami baz danych jako takich. Istnieje inny świat inżynierii danych (*data engineering*), który współpracuje z data science w zakresie udostępniania zbiorów danych, o czym więcej w rozdziale 6.

stabilna, a ponadto wyższa niż języków R i Python; ta sytuacja występuje zarówno dla ogółu ankietowanych, jak i wyłącznie dla osób zajmujących się DS.

Podkreślmy, że osoby zajmujące się DS używają omawianych technologii w początkowym etapie projektów – służą im one przede wszystkim do importu danych, a jak już mówiliśmy cały początkowy etap przetwarzania danych (którego pierwszym krokiem jest import, por. rys. 24) jest w badanym świecie uznawany raczej za nudny lub żmudny, choć niezbędny.

W rozmowie nieformalnej tłumaczono nam kwestię mody na technologie właśnie na przykładzie SQL. Osoba osadzona w badanym świecie, której badacz (RŻ) powiedział, że prowadzi badania socjologiczne „środowiska data science w Polsce" stwierdziła, że języki R/Python i data science są teraz bardzo modne, co widać po dużej ilości kobiet na konferencji (co oceniła na ponad 40% zgromadzonych), w której uczestniczyliśmy. Osoba mówiła to w porównaniu do jednej z największych konferencji poświęconych SQL, w której wcześniej brała udział – tam jej zdaniem była 1 kobieta na 100 mężczyzn.

Koncepcja relacyjnych baz danych powstała na początku lat 70. (Ceri 2018:68–69), a jej pomysłodawcą był Edgar F. Cood z firmy IBM (Codd 1970). Jej podstawę matematyczną stanowi teoria zbiorów. Ważnym elementem tej koncepcji jest unikanie tzw. redundancji (czyli, upraszczając, powtarzania wielokrotnie tej samej informacji w wielu miejscach) (Codd 1970:385-386) za pomocą przechowywaniu danych nie w jednej tabeli, a wielu tabelach połączonych ze sobą za pomocą tzw. kluczy (*keys*). Stosowanym w literaturze przykładem dydaktycznym jest biblioteka: potrzebuje ona informacji o książkach, ale jeżeli przechowywać by całość danych w jednej tzw. płaskiej (*flat*) tabeli (dwuwymiarowej, czyli mającej wiersze i kolumny), to prowadzi to do wielokrotnego powtarzania informacji, np. o autorach, dla różnych egzemplarzy książki. Przykład schematu relacyjnej bazy danych dla biblioteki (Kriegel i Trukhnov 2011) prezentujemy na rysunku 29.



Rysunek 29: Schemat relacyjnej bazy danych dla biblioteki do celów dydaktycznych

Źródło: Figure 3.1 (Library data model) (Kriegel i Trukhnov 2011:85)

Różne tabele dwuwymiarowe (zwane za Coddem relacjami – w powyższym przykładzie to: books, books_authors, authors, search_books, searchTags, location) połączone są ze sobą nawzajem za pomocą kluczy czyli jednoznacznych identyfikatorów unikatowych wpisów (bk_id, au_id, itd.). Na rys. 29. pod nazwami tabel jako pierwsze pole tabeli zapisano skróty: PK – *primary key*, klucz główny; FK – *foreign key*, klucz obcy). W bazach relacyjnych możliwe są trzy typy połączeń: jeden-do-jednego, jeden-do-wielu i wiele-do-wielu. Omawiamy tabele dotyczące książek (books) i autorów (authors). Złączonoe są one ze sobą na połączeniem wiele-do-wielu. Oznacza to, że jedna książka może mieć jednego lub więcej autorów, a autor może napisać jedną lub więcej książek. Takie połączenie wymaga zastosowania przynajmniej jednej tabeli pośredniej, łącznikowej, tutaj books_authors, w której każdy wiersz jest unikatowym wpisem, czyli połączeniem jednego autora z jedną książką (Kriegel i Trukhnov 2011:85–86). Można zatem przechowywać informację np. o imieniu autora tylko raz, w odpowiednim polu – czyli kolumnie au_first_name – tabeli authors, a nie przy każdym egzemplarzu książki tegoż autora. Ułatwia to także usuwanie, uaktualnianie, poprawianie czy dodawanie nowych wpisów, bo wykonuje się to tylko w jednym wierszu w kolumnie odpowiedniej tabeli, a nie wiele razy w wielu wierszach.

DDL (*Data Definition Language*, język definiowania struktury danych) to komponent języka SQL (*Structure Query Language*, strukturalny język zapytań), służący do definiowania

baz danych i zmian w tych definicjach. Osoby zajmujące się DS w ramach działania podstawowego korzystają z drugiego komponentu SQL-a, jakim jest DML (*Data Manipulation Language*, język operacji na danych), ponieważ pozwala im on na import danych z bazy do środowiska pracy w wybranym języku programowania (por. rys. 24.), przy czym za pomocą DML możliwe jest także wprowadzanie, aktualizowanie i usuwanie danych (Powell i McCullough-Dieter 2005:14–15). Dane importowane są z bazy za pomocą zapytania (*query*) zapisanego w języku SQL. Język SQL jest przeznaczony wyłącznie do pracy z bazami danych i na zbiorach danych, nie ma żadnego innego zastosowania. Podobnie jak R czy Python, SQL jest językiem tzw. wysokopoziomowym, deklaratywnym. Oznacza to w uproszczeniu, że człowiek piszący kod wydaje komputerowi instrukcje, co ma być zrobione, ale nie jak ma to być zrobione – ten niższy poziom wykonania zadania jest obsługiwany domyślnie (Kriegel i Trukhnov 2011:9-10).

Dane w postaci tabeli o dwóch kolumnach – nazwisku autora i tytule napisanej przez niego książki – można by „wyciągnąć" z bazy o przykładowym schemacie (rys. 29.) za pomocą następującego zapytania w języku SQL (Kriegel i Trukhnov 2011:95):

```
SELECT authors.au_last_name, books.bk_title
FROM books
JOIN books_authors ON (books.bk_id = books_authors.bk_id)
JOIN authors ON (books_authors.au_id = authors.au_id);
```

Wszystkie operacje na relacjach (czyli tabelach)w języku SQL dają w wyniku zawsze relacje. Polecenie „SELECT" dotyczy tego, jakie pola (czyli kolumny) tabel mają znajdować się w tabeli wynikowej. Jest to zapisane w składni „nazwa_tabeli_źródłowej.nazwa_pola". Klauzula „FROM" dotyczy tabeli źródłowej. „JOIN" to polecenie złączenia tabel, zaś „ON" dotyczy tego, za pomocą jakich kluczy ma być wykonane złączenie (ibidem).

Tego rodzaju zapytania – w praktyce mogą być znacznie bardziej skomplikowane[180] – są przez osoby zajmujące się DS realizowane raczej w środowisku Pythona / R, czyli tam, gdzie wykonywane są dalsze kroki projektu. Istnieją do tego celu odpowiednie pakiety. Dla języka Python jedynym z bardziej popularnych jest pakiet „SQLAlchemy" (Bayer 2012), zaś dla R to np. „odbc" i „sqldf" (Grothendieck 2017; Hester i Wickham 2018). Drugi z wymienionych pakietów pozwala stosować zapytania w SQL dla tabel z danymi już zaimportowanymi do środowiska R. Można używać tzw. kwerend / zapytań (*query*) w języku SQL za pomocą innych środowisk, ale w badanym świecie istnieje tendencja do realizowania

180 Pokazany przez nas przykład to cztery linie kodu, zaś „prawdziwe" (*real-world*) zapytania mogą zajmować np. 45 linii (J. Devlin 2018).

możliwie całości zadań wewnątrz środowiska, w którym używa się Pythona / R, gdyż to uznawane jest za najwygodniejsze i co za tym idzie najbardziej efektywne.

Relacyjne bazy danych i język SQL są technologiami, które zostały wprowadzone do użytku pod koniec lat 70. XX w., i w podobnym kształcie przetrwały do końca drugiej dekady wieku XXI. Pierwsze standardy ISO dla języka SQL określono już w roku 1986 (Melton 1996:143). Zauważmy, że relacyjne bazy danych były źródłem danych charakterystycznym dla tzw. data mining z lat 90. Mówiono przecież o KDD (*Knowledge Discovery in Databases*) – odkrywaniu wiedzy nie z danych, ale z baz danych (Azevedo i Santos 2008; Vedder 1999). W podręczniku do data mining mowa o przygotowaniu danych do analizy – czyszczeniu i transformacjach – wyłącznie w kontekście danych z baz danych (*database*) (Larose 2005:27–28).

Zatem mówimy tutaj o danych ustrukturyzowanych, przechowywanych w z góry zaprojektowanych, połączonych ze sobą tabelach, o zdefiniowanych polach. Na etapie projektowania bazy danych jest definiowane to, jakie wartości przechowywane są przez konkretne pola (kolumny), zarówno w sensie semantycznym (np. kwota miesięcznego wynagrodzenia w USD) jak i technicznymi (np. nie może być pusta, musi być większa od 0, jest liczbą w zakresie od 1 do 100 000, jest liczbą z maksymalnie dwoma miejscami po przecinku). Przy tym nie oznacza to, że dane te są zawsze gotowe do dalszych etapów projektów DS, jak analiza czy modelowanie. Mogą zawierać braki, elementy przestarzałe, nadmiarowe, wartości odstające, wartości niezgodne z definicją zmiennej bądź ze zdrowym rozsądkiem (np. wartość „Janek” dla kodu pocztowego lub „-5” dla wieku klienta), czy wartości przechowywane w formacie nieodpowiednim do dalszych czynności (np. kwota transakcji jako tekst „1 253,46 zł”) (ibidem). Niemniej twierdzimy, że obecnie w DS taki scenariusz, tzn. pozyskanie danych z relacyjnej bazy (używane są też określenia jak hurtownia danych, *data warehouse*, *data mart*, istnieją między nimi różnice znaczeniowe w sensie technicznym, ale pomijamy je tutaj) i dalej praca z danymi w formie tabelki jest uznawana za przypadek typowy, relatywnie prosty, bo znany od dawna i wielokrotnie opisany. Jak wspominaliśmy już w części 1.1.4., „data minerzy i tak wyszli już z mody” (Taylor 2016). Zauważmy, że metody uczenia maszynowego znane z podręczników do data mining – np. drzewa decyzyjne, lasy losowe, maszyny wektorów nośnych (*support vector machines*, SVM), tzw. najbliżsi sąsiedzi (*k-nearest neighbors*), metoda k-średnich (Anand i Jeffrey 2011; Biecek i Trajkowski 2011; Larose 2005), które dostępne np. za pomocą pakietu „scikit-learn” (Pedregosa i in. 2011) – są obecnie w świecie DS nazywane „tradycyjnymi”

metodami (por. Borowiecki i Mieczkowski 2019:28, 39), podczas, gdy metody uczenia głębokiego bazujące na sieciach neuronowych nie są łączone z takim przymiotnikiem.

Osoby, które pracują głównie z użyciem języka SQL, bądź z użyciem SQL i arkuszy kalkulacyjnych nie są uznawane za uczestników badanego świata DS. Głównie SQL-em, oraz np. szerszym językiem proceduralnym PL SQL (Feuerstein 1996) posługują się administratorzy lub architekci/projektanci baz danych. Są to dość ustabilizowane zajęcia kojarzone jednoznacznie z informatyką, nie z DS. Takie osoby poza zapytaniami do bazy danych w większym stopniu zajmują się administracją czyli np. nadawaniem uprawnień do dostępu do baz, lub architekturą czyli projektowaniem baz, zarządzaniem migracją, obsługą serwerów i generalnie infrastruktury technicznej umożliwiającej pracę baz itp. W odróżnieniu od nich, osoby pracujące z SQL i arkuszami kalkulacyjnymi, ewentualnie także innym, zamkniętym oprogramowaniem analitycznym, nazywane są np. analitykami danych, osobami zajmującymi się raportowaniem, ewentualnie BI (*business intelligence*), choć w tym ostatnim przypadku raczej mówi się dodatkowo o narzędziach automatyzujących raporty w formie wizualizacji (np. Tableau, Power BI, Qlick View). Takie osoby używają zapytań w SQL, by wyciągnąć z bazy danych zasoby, które następnie przetwarzają i analizują za pomocą innych narzędzi. Mogą również za pomocą samego języka SQL dokonywać podsumowania danych typu np. obliczenie średnich w podziale na grupy.

Jak pokażemy w rozdziale 6., niektórzy analitycy danych korzystają z języków Python lub R, zatem mogliby zostać uznani za uczestników. Jednak raczej nie uznaje się za uczestników świata DS osoby, które wykonują „tylko" analizę danych (np. średnia w grupach, tabele, wykresy), a w ogóle nie zajmują się modelowaniem, a szczególnie modelowaniem za pomocą metod uczenia maszynowego – nawet, gdy takie osoby wykonują analizy za pomocą kodu w Pythonie czy R. Niemniej powtórzmy, że język SQL bywa uznawany za niezbędny w pracy data scientista, za elementarz jego umiejętności, a jego dobra znajomość może być traktowana jako zaleta:

> R: No i zaczęłam w firmie informatycznej, która powiedzmy z analizą na początku nie miała za wiele wspólnego. Ale uważam że to było super fajne doświadczenie na początek, pracowałam z bazami danych. I takim moim pierwszym narzędziem to był SQL [pauza] no ale to jest o tyle fajnie, że mimo wszystko fajnie jak analityk umie się dosyć sprawnie w takich bazach danych poruszać. Może nie, nie ma potrzeby budowania tej, tej bazy, ale musi to rozumieć tak te struktury danych. To też, to się łapie na tym że jak mówię że na przykład tu są jakieś miary tu są wymiary, tu mamy taką strukturę to ludzie nie rozumieją o co mi chodzi. A ta wiedza oczywiście się przydaje. No i tak zaczęłam właśnie pracowałam przy hurtowni danych właśnie operatora telekomunikacyjnego w Polsce, ten projekt trwał 2 lata i potem z polecenia koleżanki trafiłam na zupełnie inny projekt (…) {wywiad 18}

Data science jako takie bywa łączone z pracą na danych nieustrukturyzowanych, jako danymi bardziej współczesnymi (w sensie – nieco później powstały metody składowania, przetwarzania i modelowania danych danych tego rodzaju), relatywnie trudniejszymi w pracy, i co sygnalizowaliśmy we wcześniejszym akapicie, raczej kojarzonymi z tzw. big data i bardziej modnymi metodami uczenia głębokiego. Niemniej **nie ma w badanym świecie zgody co do tego, że za osobę zajmującą się DS można uznawać wyłącznie taką, która pracuje z danymi nieustrukturyzowanymi** (jak obraz, tekst czy dźwięk):

> R: (…) często się mówi tak, że data scientist to jest ktoś kto potrafi pracować na danych nieustrukturyzowanych. Czyli potrafi pracować na tekście, potrafi na dźwięku na obrazach i tak dalej, z drugiej strony jest bardzo wielu data scientistów, którzy w takich klasycznych biznesach typu banki, ubezpieczenia, departamenty CRM w tych biznesach, robią super rzeczy nie dotykając danych nieustrukturyzowanych. No i w sumie **nie wiem** jaka definicja jest właściwa i tak trochę tym się nie przejmuję (…). {wywiad 5}

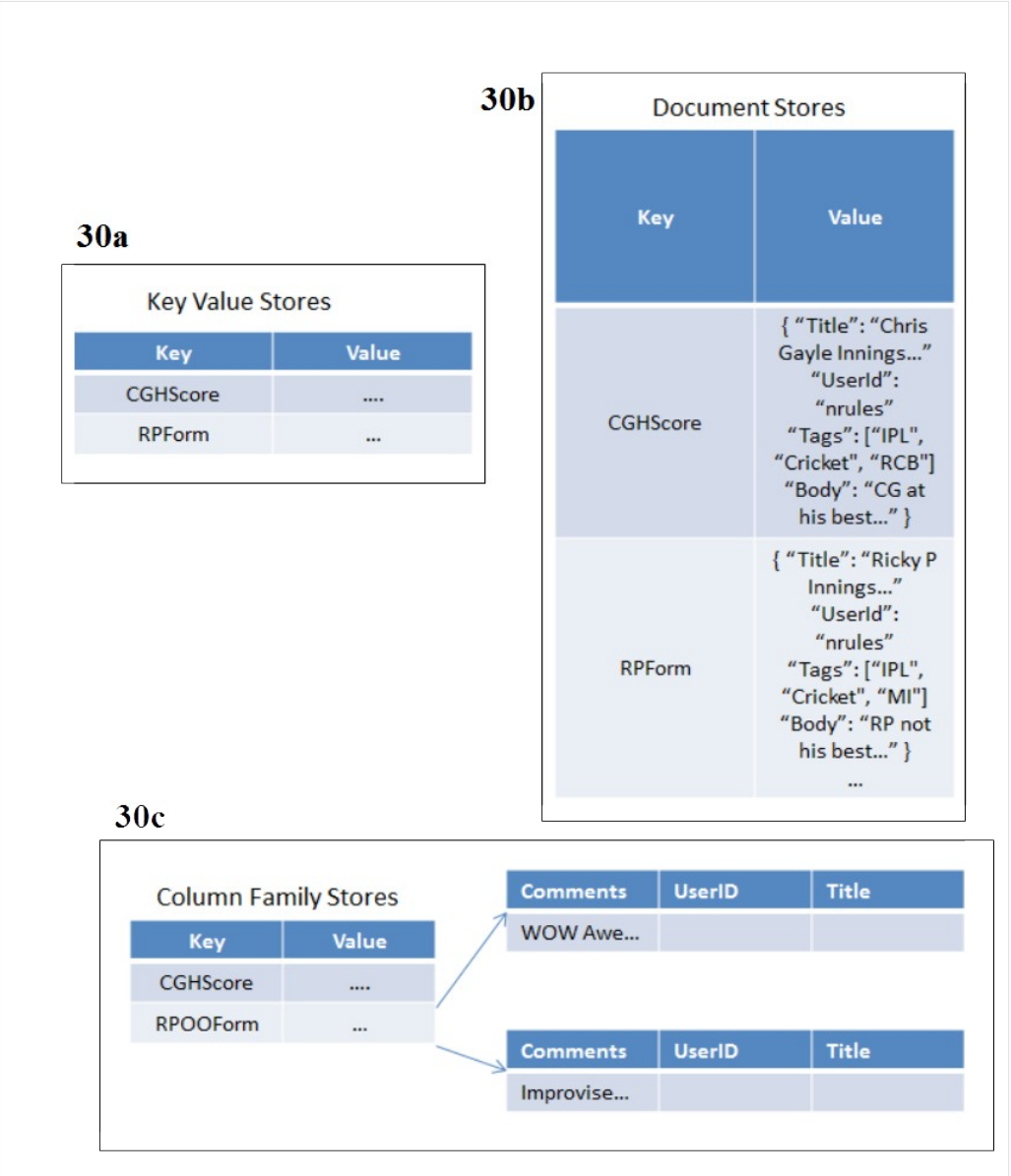
Niemniej doświadczenie w pracy z różnymi typami danych jest traktowane jako legitymizacja autentyczności uczestników (por. 6.1.2.2.).

Dane nieustrukturyzowane mogą być przechowywane w formie plików na komputerach, bez wykorzystania żadnej bazy danych, ale jak mówiliśmy jest to rozwiązanie uznawane za wyraz tzw. niskiej kultury danych, i rozwiązanie niewygodne oraz nieefektywne. Dane nieustrukturyzowane, szczególnie jeżeli mają one duży rozmiar, mogą być przechowywane w bazach o innej architekturze niż wyżej omawiane relacyjne bazy danych. Ogólnie mówi się tutaj o nierelacyjnych bazach danych (Feinberg 2017). Używane jest określenie NoSQL, co ma oznaczać „nie SQL" (*no SQL*), lub „nie tylko SQL" (*not only SQL*), jednak bywa ono uznawane za niewłaściwe, bowiem nierelacyjne bazy danych mogą umożliwiać dostęp za pomocą języka SQL (ibidem).

Bazy nierelacyjne czy NoSQL to ogólne określenie na różne, alternatywne do baz relacyjnych architektury przechowywania i dostępu do danych cyfrowych – kluczy-wartości (*key-value*), kolumnowe (*wide-column*, *column family* lub *column-oriented*), dokumentów (*document*), grafowe (*graph*) (Chen, Mao, i Liu 2014:186–89; Feinberg 2017). Umożliwiają one przechowywanie i dostęp do danych innych niż te pasujące do schematu relacyjnego, czyli przede wszystkim do danych nieustrukturyzowanych. Niemniej stosuje się je także dla danych ustrukturyzowanych, ale takich, dla których podejście relacyjne uznawane jest za nieodpowiednie (ibidem). Przykładem mogą być dane generowane przez maszyny (por. tab. 1.) czy tzw. Internet Rzeczy, gdzie dane napływają w dużej ilości w bardzo małych odstępach czasu, niemniej mają formę dwuwymiarowej tabelki. Zatem relacyjne bazy danych to tylko

jedno z możliwych rozwiązań do przechowywania danych ustrukturyzowanych, a właściwie konkretnej ich odmiany, która była stosowanym przez dziesięciolecia standardem.

Architektury kolumnowe i dokumentów są odmianami architektury kluczy-wartości (*key-value*). Idea podstawowa jest taka sama – dane przechowywane są w formie wierszy, a każdy wiersz posiada unikalny klucz (*key*) oraz dowolną wartość (*value*), co tworzy unikalną parę klucz-wartość (*key-value pair*). Popularną dystrybucją bazy klucz-wartość jest Redis. Na przykładzie mediów społecznościowych, wartością wiersza może być profil użytkownika, pojedynczy komentarz lub cała dyskusja. Nie ma zdefiniowanego schematu danych, jak w bazach relacyjnych – pole wartości może zawierać dane różne w sensie semantycznymi jak i technicznym. Dla baz dokumentów, np. popularnej MongoDB, wartości są dokumentami, które mogą mieć swoją wewnętrzną strukturę, zbliżoną do plików JSON. Dla baz kolumnowych, jak popularna Cassandra, wartości są przechowywanie w postaci tzw. rodziny kolumn (*column family*), w której każda kolumna stanowi drugi poziom pary klucz-wartość (Bhatt 2013). Ilustruje to rysunek 30. Część 30a prezentuje schematycznie architekturę bazy klucz-wartość, 30b bazy dokumentów, zaś 30c bazy kolumnowej.



Rysunek 30: Schematy architektur baz danych nierelacyjnych: klucz-wartość, dokumentów, kolumnowa

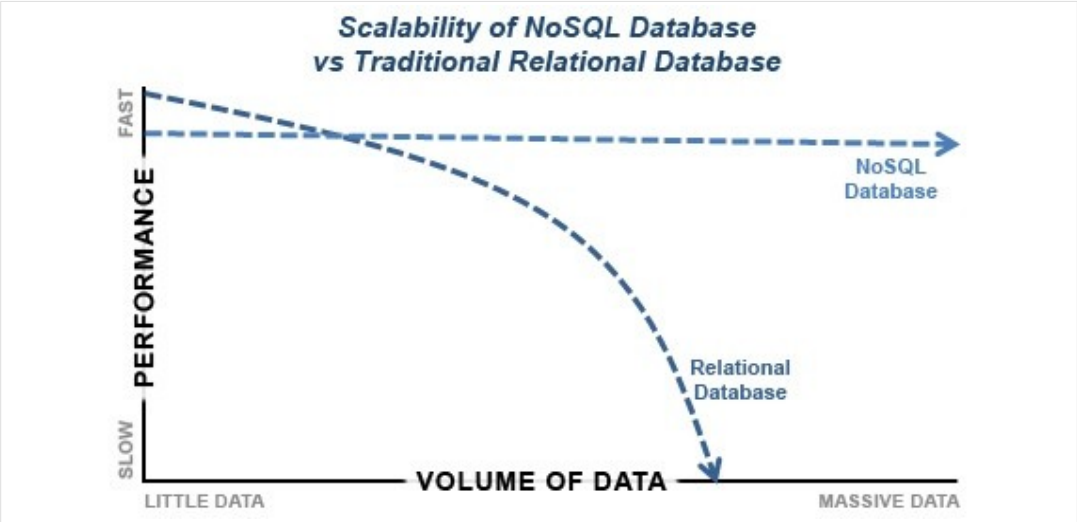
Źródło: Big Data, NoSQL and MapReduce (Bhatt 2013)

Architektury grafowe oparte są na teorii grafów, m.in. pojęciach krawędzi (*edges*) i wierzchołków (*nodes*) oraz ich własnościach. Stosuje się je w przypadkach, kiedy połączenia między danymi są uznawane są za tak samo ważne, jak same dane. Mogą być do dane dotyczące np. sieci społecznych, informacji w internecie czy połączeń między genami, białkami i enzymami (Miller 2013:142–43; Vicknair i in. 2010). Są to architektury o zastosowaniach węższych niż inne podejścia nierelacyjne. Do znanych dystrybucji bazy grafowej należy np. Neo4j. Architektur grafowych używają m.in. Google (Knowledge Graph), Facebook (Open Graph), Twitter (FlockDB). Największe firmy technologiczne

wykorzystują i udostępniają także omawiane wcześniej typy baz nierelacyjnych – BigTable to system nierelacyjnych baz danych stworzony i wykorzystywany przez Google, Cassandra to rozwiązanie open source Facebooka, Dynamo jest rozwiązaniem używanym przez Amazon, zaś Linkedin używa systemu o nazwie Project Voldemort (Vicknair i in. 2010). Jednym z najwcześniejszych systemów nierelacyjnych baz danych jest właśnie BigTable firmy Google (Chang i in. 2006).

Według rankingu portalu DB-Engines najpopularniejsze na sierpień 2019 r. systemy zarządzania bazami danych to: na miejscu pierwszym Oracle (bazy relacyjne, ale także dokumentów, grafowe, RDF czyli odmianie bazy grafowej), na drugim MySQL (relacyjne, także dokumentów), na trzecim Microsoft SQL Server (relacyjne, także dokumentów i grafowe), czwartym PostgreSQL (relacyjne, także dokumentów), kolejno MongoDB (dokumentów), IBM Db2 (relacyjne, dokumentów, RDF), Elasticsearch (wyszukiwarka dla danych nierelacyjnych, baza dokumentów), Redis (klucz-wartość, także dokumentów, grafowe, wyszukiwarka, baza szeregów czasowych), Microsoft Access (relacyjne), i na miejscu dziesiątym Cassandra (kolumnowa) (DB-Engines 2019).

Nierelacyjne bazy danych mają zastosowanie w sytuacjach, kiedy można spodziewać się potrzeby przechowywania danych w tabelach o wielu kolumnach, z których znaczna część będzie pusta, kiedy występuje dużo relacji wiele-do-wielu, kiedy dane mają charakter hierarchicznej struktury drzewa (*tree-like*), kiedy można spodziewać się potrzeby częstych zmian schematu danych (Vicknair i in. 2010). Zatem upraszczając, ideą baz nierelacyjnych jest po pierwsze, brak jednoznacznego, uprzednio zaprojektowanego schematu czy definicji nałożonej na dane, zaś po drugie, tzw. skalowalność (nierzadko określana jako horyzontalna skalowalność), czyli możliwość niemal nieograniczonej rozbudowy wielkości przechowywanych danych czy ilości użytkowników bazy i jednoczesnego zachowania odpowiedniej wydajności (*performance*) systemu (Bugnion 2016; Sumbul 2014). Ta skalowalność jest osiągana zarówno dzięki architekturze baz, jak i rozpraszaniu przechowywania i przetwarzania danych na dużą ilość komputerów (ibidem), o czym więcej powiemy w kolejnej części, na przykładach technologii Hadoop i Spark. Pojęcie skalowalności pojawia się w badanym świecie, a także w szeroko pojętej informatyce, w różnych kontekstach, zatem na przykładzie porównania relacyjnych i nierelacyjnych baz danych prezentujemy ten koncept na rysunku 31.



Rysunek 31: Porównanie skalowalności relacyjnych i nierelacyjnych baz danych – wydajność w stosunku do rozmiaru danych

Źródło: Figure 2: Scalability (Sumbul 2014)

Na rysunku 31. oś x reprezentuje rozmiar danych (*volume of data*), zaś y to wydajność bazy danych (*performance*). Relacyjne – zauważmy określenie „tradycyjne" w oryginalnym tytule rysunku – bazy danych (*relational*, linia ciemna) są przedstawione jako nieco bardziej wydajne niż bazy nierelacyjne (*NoSQL*, linia jasna) tylko dla małego rozmiaru danych. Wraz ze wzrostem rozmiaru danych spada wydajność baz relacyjnych. Dla baz nierelacyjnych wydajność ma pozostawać taka sama, bez względu na rozmiar danych. To właśnie oznacza, że bazy nierelacyjne są skalowalne, używa się także określenia „horyzontalnie skalowalne", lub bardziej potocznie „dobrze się skalują".

W odniesieniu do systemów, które mają zapewniać dostęp do wszystkich dostępnych w organizacji danych – ustrukturyzowanych i nieustrukturyzowanych – **używane jest określenie data lake** (dosł. jezioro danych, nie spotkaliśmy się w toku badań z tłumaczeniem na j. polski). Za jego twórcę uznaje się Jamesa Dixona z firmy Pentaho, który napisał o koncepcji stojącej za, wówczas nową, architekturą przeznaczoną dla big data w następujący sposób:

> Jeśli pomyślisz o tematycznej hurtowni danych (*data mart*), że jest ona magazynem wody butelkowanej – oczyszczonej, zapakowanej i uporządkowanej w celu łatwego jej spożycia – to jezioro danych (*data lake*) jest dużym zbiornikiem wody w bardziej naturalnym stanie. Dane płyną ze źródeł strumieniami (*stream*), aby wypełnić jezioro, a różni użytkownicy jeziora mogą przyjść, aby zbadać wodę, zanurkować w niej lub pobrać jej próbki. (Dixon 2010)

Dixon przedstawia data lake w odróżnieniu od data mart, zatem najpierw przybliżmy czym jest to drugie. Tematyczna hurtownia danych (*data mart*) to składowa hurtowni danych (*data*

warehouse). Są to rozwiązania zorientowane na odczyt danych historycznych w celach analitycznych, stosowane dla relacyjnych baz danych. Określane są one skrótem OLAP (*online analytical processing*), zaś relacyjne bazy danych zorientowane na zapis i przetwarzanie w celach wykonywania bieżących operacji nazywane są OLTP (*online transaction processing*). Zatem hurtownie danych (*data warehouse*) to bazy typu OLAP, a tematyczne hurtownie danych (*data marts*) to składowe hurtowni, zawierające dane o określonym temacie czy przeznaczone dla pewnych użytkowników (Kriegel i Trukhnov 2011:93–94).

Krytyka hurtowni i tematycznych hurtowni danych, która stała za koncepcją data lake, dotyczyła po pierwsze tego, że w hurtowniach tworzą się oddzielone od siebie silosy z danymi, np. na poziomie odrębnych działów w firmie lub odrębnych lokalizacji firmy. W wyniku tego nie możliwe jest przeprowadzanie analizy czy modelowania danych na poziomie całej organizacji. Po drugie, krytyka dotyczyła tego, że dane w hurtowniach / hurtowniach tematycznych to nie są dane w formie surowej (*raw*), a dane przynajmniej częściowo zagregowane i wybrane do celów analitycznych. Ten wybór i agregacja była podyktowana względami technicznymi i ekonomicznymi, czyli z uwagi na charakter techniczny relacyjnych baz danych uznawano za opłacalne – ze względu kosztów i wydajności – by zapewniać dostęp dla celów analitycznych do danych w tylko formie pogrupowanej tematycznie, tylko wybranych części i częściowo zagregowanych. To miał na myśli Dixon, mówiąc o tematycznych hurtowniach danych jako magazynach wody butelkowanej. Takie podejście krytykowano, jako mające pozwalać na udzielenie odpowiedzi wyłącznie na już znane z góry pytania (Dixon 2010; Feinberg 2017; Miloslavskaya i Tolstoy 2016; Nicolaus i in. 2016).

Data lake ma być zatem rozwiązaniem, które jakoby pozwoli na dostęp do surowej postaci aktualnej całości danych w organizacji. Nie chodzi o żadną konkretną technologię, choć właściwie koncepcja data lake zakłada, że składowanie i przetwarzanie danych będzie rozproszone na wiele komputerów (ibidem), np. za pomocą technologii Hadoop, o czym później. To, co uznawano za zalety data lake – łatwość implementacji, elastyczność względem rodzajów danych, przechowywanie danych w nieprzetworzonej postaci – nierzadko w praktyce prowadziło do powstawania chaotycznych repozytoriów, z których trudno było uzyskać dane do zastosowań analitycznych. Pojawił się termin data swamp (dosł. bagno danych), który oznacza właśnie takie zaniedbane repozytorium (Brackenbury i in. 2018:1–2).

> Pojawił się Hadoop, tanio można wrzucić dużo danych, jest MapReduce, więc można robić wielowątkowo. Miało być data lake, a jest data swamp i to ma teraz wiele firm. {notatka terenowa, obserwacja 13}

Podkreślmy, że koncepcja data lake, oraz bezpośrednio powiązane z nią technologie nierelacyjnych baz danych i rozpraszania ich składowania oraz przetwarzania na wiele komputerów są tym, co w sensie technicznym określa się jako big data. Zauważmy, że słynna definicja „3Vs of big data" (por. 1.1.1. tej pracy) – wielkość danych (*Volume*)[181], ich różnorodność (*Variety*)[182], i szybkość pojawiania się nowych danych (*Velocity*)[183] – odnoszą się właśnie do takiej charakterystyki danych, dla której koncepcja data lake i powiązane z nią technologie zostały stworzone.

Ponadto podkreślmy, że w badanym świecie **osoba zajmująca się DS nie jest osobą zajmującą się big data**. Uczestnik badanego świata może pracować z danymi o charakterze big data, choć jesteśmy przekonani, że uczestnicy ewentualnie byliby skłonni do określenia „dane nieustrukturyzowane", czy „dane o dużym rozmiarze". Niemniej, jak wskazywaliśmy wcześniej, praca z tzw. **big data nie jest w badanym świecie uznawana za element konieczny, by być uznanym za uczestnika świata społecznego DS**.

> *B: Dobra, a big data? Też jest jakby wewnątrz czy obok [data science]?*
> R: Big data jest pewnym konceptem **ściśle** związanym, wręcz myślę nierozerwalnie [pauza] natomiast big data jest bardziej konceptem [cmoknięcie] samego problemu ilości i analizy danych niż tego w jaki sposób będzie się z nich osiągało wyniki. Tak? Czyli koncept big data bardziej polega na tym, że [pauza] głównym paradygmatem jest to że danych jest baaardzo dużo, są bardzo szerokie i są w bardzo różnym stanie. Tam te cztery fał, volume,
> *B: Tak,*
> R: I tak dalej. Więc jest to koncept pokrewny, bo my jako data scientiści zajmujemy się też, pracujemy na **dużych** zbiorach, czyli te koncepty big datowe używamy, natomiast to nie jest definicja per se uważam że można robić data science, nie używając narzędzi big datowych i trzymając się **średnich** danych, tak, tu nawet niewielkich. Więc, natomiast na przykład z kolei jeśli mówimy już oooo typowym uczeniu maszynowym, to faktyczniee koncept uczenia maszynowego zakłada że jest wystarczająco **duża pula** danych na których będziemy pracować, tak? Więc wiesz, to się trochę tam krzyżuje. Natomiast na pewno to nie jest jeden do jednego, i są ludzie którzy zajmują się big data i nie mają nic wspólnego z data science, tak? Na przykład nie wiem, bank potrzebuje przechowywać sprawnie i mieć szybko dostępne, petabajty danych, i niekoniecznie chce robić na nich data science i coś zmieniać. Potrzebuje po prostu szybkiego taniego rozwiązania no i wtedy big data świetnie się sprawdza koncept Hadoopa, rozproszenia obliczeń, rozproszenia przechowywania plików, świetnie się sprawdza, tak? Są ludzie którzy są programistami, projektują bazy danych, zajmują się informatyką bardziej może, z data science nie mają nic wspólnego i też używają, też pracują na big data. Więc to są dwa niezależne od siebie byty, natomiast faktycznie często się, często się przenikają. {wywiad 6}

181 Rozmiar danych; bywa rozumiany jako zbiory większe niż 1 terabajt, a także takie, których nie da się przetwarzać za pomocą tradycyjnych narzędzi (relacyjnych baz danych i arkuszy kalkulacyjnych).

182 Zróżnicowanie; różne struktury, formaty i charakter danych.

183 Szybkość pojawiania się nowych danych; ciągły ich napływ, co powoduje, że potrzebne są metody umożliwiające wydobywanie informacji z danych w czasie rzeczywistym.

Rozmówca jednoznacznie wskazuje, że można zajmować się big data, czyli być, ogólnie rzecz biorąc informatykiem zajmującym się składowaniem i przetwarzaniem danych uznawanych za big data, i jednocześnie nie mieć nic wspólnego z DS, tzn. nie analizować ani nie modelować danych.

Praca z użyciem tzw. big data może być jednak przez część uczestników świata DS uznawana za ciekawszą lub trudniejszą niż praca wyłącznie z danymi z relacyjnych baz danych, szczególnie dlatego, że modne w badanym świecie metody modelowania za pomocą uczenia głębokiego (*deep learning*) są stosowane do pracy z dużą ilością danych nieustrukturyzowanych. Może być to zatem powiązane z większym prestiżem osób pracujących z danymi nieustrukturyzowanymi o dużym rozmiarze. Niemniej w badanym świecie pojawia się przekonanie, że „dobry" czy „prawdziwy" data scientist potrafi poradzić sobie zarówno z danymi ustrukturyzowanymi jak i nieustrukturyzowanymi, zarówno z dużym jak i małym wolumenem, na wszystkich etapach projektu: przetwarzania, analizy, modelowania danych. Wszytko to wiąże się z obecnym w badanym świecie postulatem wykonywania projektu w sposób właściwie dopasowany do celu czy rozwiązywanego problemu, a w pracy komercyjnej także do potrzeb klienta.

> *B: A co z takimi rzeczami, z tego co akurat ja słyszałem, które są od big daty tak zwanej, czyli jakieś tam,*
> R: Sparki
> *B: Sparki, Hadoopy, czy to też jest coś czego pan używa?*
> R: Tak
> *B: Czy to już jakiś inny zawód, jakaś inna działka?*
> R: Trochę inny zawód, to jest zazwyczaj zajmuje się tym inżynier danych, eee, natomiast, eee [pauza] po pierwsze Sparki i inne takie narzędzia mają jakieś szczątkowe modele zaimplementowane już w sobie czyli jak się chce robić jakieś proste rzeczy działające na ogromnej ilości danych tooo czasami to nie jest taki głupi pomysł, bo to jakby że model będzie prosty, eee, OK ale za to rekompensujemy to tym że ilość danych jest ogromna więc w sumie wyniki nie musi być jakiś totalnie zły, mmm, natomiast też czasami robimy tak że bierzemy sobie takie duże dane, dane które przekraczają już zdolności rozumienia i je na przykład filtrujemy, jakoś bierzemy subset[184] tych danych i wtedy modelujemy już naszymi narzędziami, wtedy to też dobrze wychodzi. Ale jasne że big dejtę robimy, no bo **klienci** mają, tak? Nie możemy sobie pozwolić na to żeby przyszedł klient i powiedział chciałbym coś zrobić ale mam dużo danych a my wtedy powiemy o nie! Masz za dużo danych, my tak nie umiemy. {wywiad 2}

Zauważmy, że Rozmówca sam podpowiada nazwę narzędzia Spark, zatem jednoznacznie kojarzy je z tzw. big data. Nieco dalej mówi też, że Spark ma zaimplementowane „szczątkowe" modele, zatem lepszy rezultat może dać pobranie próbki danych i praca „swoimi" narzędziami – chodzi o język Python bez użycia Spark. Zatem swoje

184 Dosłownie podzbiór danych, próbka.

narzędzie, narzędzie data scientista to dla Rozmówcy Python i odpowiednie pakiety zawierające już nie-szczątkowe modele, bez użycia Sparka.

Podkreślmy przy tym, że osoby zajmujące się DS nie pracują nad problemami infrastruktury do przechowywania danych – z ich punktu widzenia chodzi przede wszystkim o początkowy import danych do dalszych elementów projektu. Tutaj może się pojawiać współpraca z rolą inżyniera danych (*data engineer*), nazwa ta pada w powyższym fragmencie {wywiad 2}. Styczność z infrastrukturą do przechowywania danych może pojawiać się także w końcowym etapie pracy, czyli wdrożeniu, gdzie pojawia się współpraca z rolą inżyniera uczenia maszynowego (*machine learning engineer*) lub ogólnie rzecz ujmując działem IT. Więcej o tych rolach stykających się z data scientist powiemy w części 6.

Sądzimy, że w badanym świecie posługiwanie określeniem „big data" jest traktowane jako wskaźnik, że dana osoba nie jest uczestnikiem świata DS, a nawet więcej – że jest tzw. osobą nie-techniczną. Określenie „big data" jest uznawane za niejasne, przereklamowane, a obecnie nieco przestarzałe, i charakterystyczne dla dyskursu nietechnicznego: marketingowego, prawniczego, dziennikarskiego, ewentualne nauk społecznych[185]. Z punktu widzenia osób zajmujących się DS mówi się w kontekście big data głównie o technologii Spark, bo to jest narzędzie charakterystyczne dla badanego świata, o czym więcej w kolejnej części naszej pracy.

5.3.3. Obliczenia lokalne, w chmurze, rozproszone

Wszelkie operacje na danych cyfrowych przeprowadzane są na komputerach. Słowa „komputer" używamy tutaj w ogólnym sensie, ponieważ ma ono zastosowanie do każdego ze scenariuszy – operacji wykonywanych lokalnie, wykonywanych w chmurze, rozproszonych i nierozproszonych.

Różnicę między komputerem lokalnym, a tym w chmurze wskazaliśmy już omawiając notatnik Jupyter. Kiedy praca wykonywana jest **lokalnie**, oznacza to, że człowiek korzysta z hardware'u, który fizycznie znajduje się przy nim lub względnie blisko niego. Zatem pracuje na własnym komputerze lub na serwerze wewnętrznym (maszynie w sieci lokalnej, bez połączenia przez internet). W przypadku pracy **w chmurze** praca wykonywana jest na serwerze zewnętrznym (maszynie położonej gdziekolwiek na świecie, przy połączeniu przez internet). Przy tym praca z danymi na serwerze, bez względu czy jest to serwer lokalny czy w

185 Wczesne prace autora niniejszej rozprawy zawierały, rzecz jasna, określenie „big data" już w tytule tekstu (Żulicki 2016, 2017).

chmurze, wykonywana jest za pomocą komputera osobistego, tzn. laptopa lub stacjonarnego. Różnica względem pracy na własnym komputerze polega na tym, że do wykonywania operacji używane są podzespoły i moc obliczeniowa serwera. Komputer stojący na biurku służy wtedy jedynie jako urządzenie pozwalające na dostęp do innej maszyny, a jego podzespoły i moc obliczeniowa służą np. do uruchomienia przeglądarki internetowej czy terminala, a nie do operacji na danych.

Podkreślmy, że w przypadku pracy lokalnej, człowiek realizuje operacje na danych w całości na sprzęcie należącym do niego samego lub do organizacji, której jest on częścią. Zatem człowiek / organizacja kupiła ów sprzęt, skonfigurowała go i utrzymuje jego działanie na co dzień. W przypadku tzw. chmury operacje realizowane są na żądanie – sprzęt należy do innej organizacji, a podmiot korzystający z niego płaci za wykorzystane zasoby sprzętowe i czas, zatem jest to wypożyczanie, a nie posiadanie. Porównajmy to do sytuacji, kiedy firma posiada ciężarówkę (to praca lokalna) – konkretny samochód z danego rocznika, o pewnym przebiegu, z takim a nie innym silnikiem, z określoną przestrzenią bagażową i ogólnie specyfikacją techniczną. Firma dba o naprawy, przeglądy techniczne, płaci ewentualne mandaty itp. Kiedy firma wypożycza ciężarówkę na godziny od innej firmy, która ma ich wiele (to praca w chmurze), może dobrać samochód do jednorazowego zadania, posiadanego w tej chwili budżetu itd.

Komputer, na którym napisana jest niniejsza praca ma jeden czterordzeniowy procesor CPU (procesor ogólnego zastosowania), procesor GPU (procesor graficzny), 8 gigabajtów (GB) pamięci RAM i 256 GB przestrzeni dyskowej. Naszym zdaniem w badanym świecie DS taki sprzęt byłby uznany za przeciętny zestaw do pracy biurowej. Na takim komputerze, tj. używając jego podzespołów i mocy obliczeniowej, można pracować jedynie ze zbiorami danych rzędu maksymalnie kilku GB. Dlaczego nie 100 GB? Dzieje się tak[186], ponieważ w języku R generalnie (zaś w języku Python np. dla popularnego w zastosowaniach DS pakietu „pandas") całość zbioru danych jest wczytywana do pamięci RAM i tam wykonywane są wszystkie operacje. Efektywna praca możliwa jest dla zbiorów danych rzędu 10% – 20% pamięci RAM konkretnej maszyny, zaś dla zbiorów wielkości powyżej 50% RAM praca jest niewykonalna (Kane, Emerson, i Weston 2013:2–3). Jednym z rozwiązań jest pobranie próbki danych (*subset*) na własny komputer, innym zastosowanie pakietów, które umożliwiają bardziej efektywne wykorzystanie zasobów własnego komputera (ibidem). Wickham i

186 Pomijamy dla uproszczenia kwestie tego, ile przestrzeni dyskowej zajmuje system operacyjny i zainstalowane aplikacje, oraz ile wykorzystują one pamięci RAM.

Grolemund mówią o tych dwóch rozwiązaniach, dodając możliwość użycia przetwarzania danych rozproszonego czy równoległego (*parallel*), o czym więcej powiemy później.

> Ta książka z dumą koncentruje się na małych zbiorach danych mieszczących się w pamięci [RAM] (*in-memory datasets*). Od tego trzeba zacząć, ponieważ nie możesz poradzić sobie z dużymi zbiorami danych, jeśli nie masz doświadczenia z małymi danymi. Narzędzia, których nauczysz się w tej książce [czyli ekosystemu „tidyverse"], z łatwością poradzą sobie z setkami megabajtów danych, a przy odrobinie staranności możesz zwykle użyć ich do pracy z 1-2 GB. Jeśli pracujesz zwykle z większymi danymi (powiedzmy 10-100 GB), powinieneś dowiedzieć się więcej o [pakiecie] „data.table". Ta książka nie uczy „data.table", ponieważ ma on bardzo zwięzły interfejs, co utrudnia naukę, ponieważ oferuje mniej wskazówek językowych [niż „tidyverse"]. Ale jeśli pracujesz z dużymi danymi, osiągana wydajności jest warta dodatkowego wysiłku wymaganego do nauczenia się tego pakietu. Jeśli Twoje dane są jeszcze większe, dokładnie zastanów się, czy problem z dużymi danymi (*big data problem*) może być ukrytym problemem z małymi danymi (*small data problem*). Chociaż pełne dane mogą być duże, często dane potrzebne do odpowiedzi na określone pytanie są niewielkie. Być może uda Ci się znaleźć podzbiór lub agregat, które mieszczą się w pamięci i nadal pozwalają odpowiedzieć na pytanie, które Cię interesuje. (...). Inną możliwością jest to, że problem z dużymi danymi jest w rzeczywistości dużą liczbą problemów z małymi danymi. Każdy problem może mieścić się w pamięci, ale masz ich miliony. (...). Na szczęście każdy problem jest niezależny od innych (konfiguracja, która jest czasami nazywana niezręcznie „równoległą" (*parallel*)), więc potrzebujesz tylko systemu (takiego jak Hadoop lub Spark), który pozwala wysyłać różne zestawy danych do różnych komputerów w celu przetworzenia. (Wickham i Grolemund 2017)

Można także zadbać o mocniejszą konfigurację komputera personalnego. O ile nie stosuje się więcej niż jednego procesora CPU, a rzadko więcej niż jeden GPU, to jakość tych procesorów może być różna. Niektórzy uczestnicy badanego świata inwestują w personalne komputery, np. instalują 64 GB RAM w laptopie[187], inni korzystają ze sprzętu „biurowego". Na praktyki dbania o swoje komputery zwrócił uwagę Lowrie, do czego jeszcze wrócimy, mówiąc o umieszczaniu naklejek na pokrywach laptopów (Lowrie 2016). Istnieje także praktyka inwestowania, indywidualnie lub na poziomie organizacji, w wysokiej jakości karty graficzne (GPU)[188]. Podkreślmy, że GPU nie ma nic wspólnego z wielkością zbioru danych, na którym są wykonywane operacje, a wyłącznie z szybkością trenowania modeli, szczególnie tych z zastosowaniem uczenia głębokiego (*deep learning*), które może być realizowane np. w Tensor Flow czy PyTorch. Zauważmy, że w GPU mogą inwestować także tzw. gracze w Kaggle, którzy uczestniczą w konkursach modelarskich i – o ile granie w Kaggle to ich jedyna aktywność związana z uczeniem maszynowym – raczej nie są uznawani

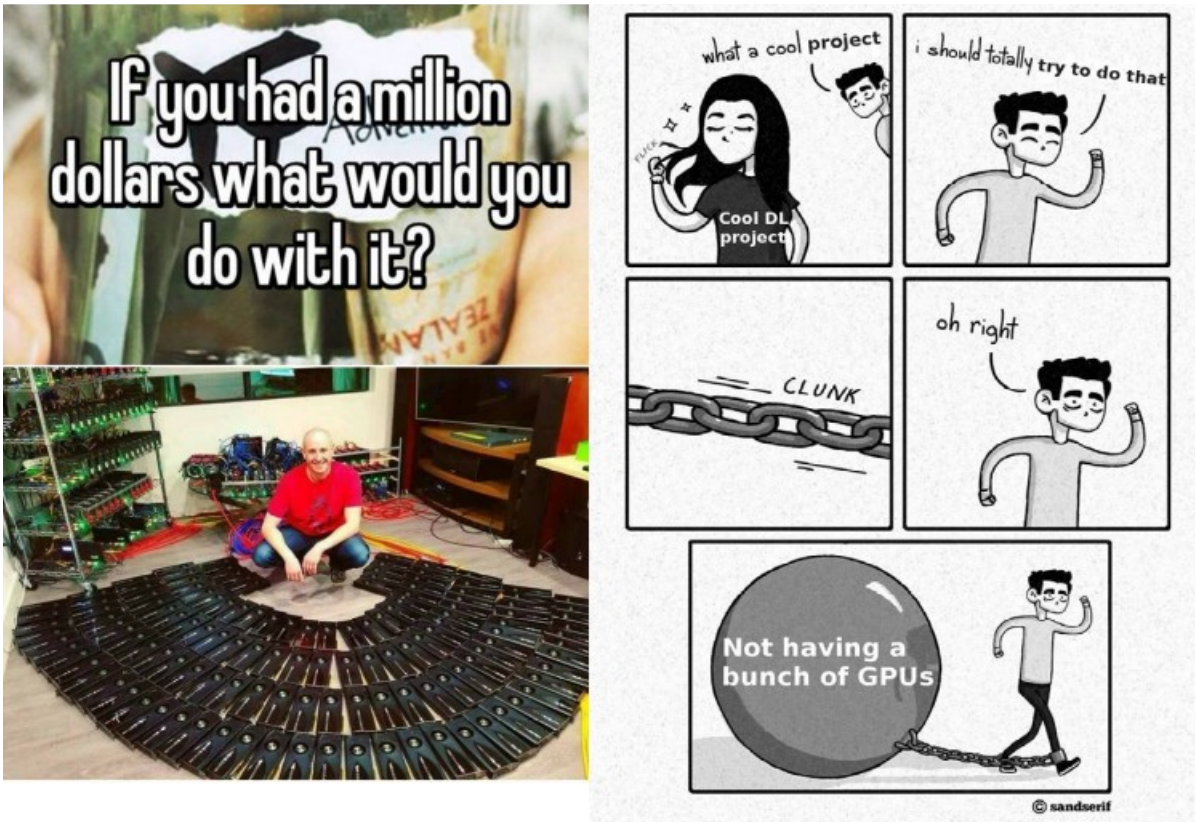
187 W takim przypadku są to przeważnie tzw. gamingowe laptopy, czyli maszyny przeznaczone dla osób grających w gry komputerowe, które do płynnego działania rozgrywek wymagają wysokiej mocy obliczeniowej (w tym wydajnej karty graficznej – GPU).

188 Przykładowa oferta komputera stacjonarnego, skonfigurowanego do data science to cztery GPU firmy NVIDIA, 128 GB RAM, jeden CPU Intel Xeon W-2135, 1 TB dysk SSD, zainstalowana dystrybucja Linuxa (Ubuntu 18.04 albo CentOS 7.x) oraz narzędzia m.in. Docker, TensorFlow, PyTorch, Anaconda (Exxact 2019).

za uczestników badanego świata, bowiem nie podejmują działania podstawowego w sensie realizowania całego projektu DS.

Co do GPU, powstają na ten temat memy internetowe[189], publikowane w kanałach mediów społecznościowych, które śledzi część uczestników badanego świata. Przykładowe memy ze strony facebookowej „Artificial Intelligence Memes for Artificially Intelligent Teens" (dosł. memy o sztucznej inteligencji dla sztucznie inteligentnych nastolatków) prezentujemy na rysunku 32.

189 Zwracamy uwagę na memy, ponieważ pokazywano nam je w trakcie wywiadów, używa się ich podczas prezentacji np. na meetupach czy konferencjach. RŻ, po tym jak wykazał się zrozumieniem fragmentów świata data science prezentując wybrane wynik swoich badań na jednym z meetupów {obserwacja 33}, był testowany przez znajomego uczestnika naszych badań w rozmowie prywatnej pod kątem tego, czy „też to czuje" tzn. czy śmieszy go żart ukuty w jednym z zespołów data science (nie śmieszył, co rozmówca skwitował porównaniem RŻ do sieci neuronowej, czyli „jesteś użyteczny bez generalizacji", oznacza to mniej więcej „potrafisz powtórzyć rzeczy, których się nauczyłeś, ale nie rozumiesz naszego świata tak jak my"). Zgadzamy się zatem z Jemielniakiem (2019:132) – "osoby uprawiające etnografię mawiają, że prawdziwe zrozumienie danej kultury zostaje ostatecznie potwierdzone, jeśli badacz lub badaczka zaczynają rozumieć żarty swoich rozmówców, a więc posiadają zbliżony do nich kapitał kulturowy".



Rysunek 32: GPU w modelowaniu metodami uczenia głębokiego – memy internetowe

Źródło: https://www.facebook.com/artificialintelligencememes/, dostęp 17.05.2019 r.

Lewa strona: [mem kolorowy] Co być zrobił/zrobiła z milionem dolarów? [zdjęcie uśmiechniętego mężczyzny kucającego w pokoju wypełnionym sprzętem komputerowym, przed nim na pierwszym planie rozłożone na podłodze kilkaset kart graficznych]

Prawa strona: [mem czarno-biały] Ale fajny projekt uczenia głębokiego (DL, *deep learning*)! Zdecydowanie powinienem spróbować go zrobić. [szczęk łańcucha] A, tak. [napis na kuli u nogi: brak kliku GPU]

Podkreślmy, że mem ma charakter humorystyczny i satyryczny (por. Jemielniak 2019:139–41). Jest przerysowaną, ironiczną formą wypowiedzi, dostosowaną do zasad ekonomii uwagi (por. Harris 2012) w mediach społecznościowych.

Wracając do zagadnień technologii, przede wszystkim jednak, kiedy brakuje zasobów sprzętowych czy mocy obliczeniowej, zarówno w odniesieniu do wielkości zbioru danych jak i trenowania modeli, można jednak skorzystać z innej, mocniejszej maszyny – serwera, lub wielu połączonych ze sobą maszyn, tzw. klastra (*cluster*).

Serwer to generalnie komputer lub oprogramowanie, przeznaczone do wykonywania usług lub udostępniania zasobów dla innych komputerów w sieci, tzw. klientów. Słowo serwer pochodzi od angielskiego *serve* (służyć). Tutaj skupiamy się na serwerze jako komputerze – komputerze raczej o większej niż komputer personalny mocy obliczeniowej i zasobach sprzętowych, przystosowanym do nieprzerwanej pracy (bez wyłączenia / restartu

przez wiele miesięcy), mogącym posiadać wiele procesorów CPU, zazwyczaj pracującym pod kontrolą systemów operacyjnych z rodziny Linux/Unix. W przypadku pracy lokalnej osoby zajmujące się DS komercyjnie korzystają z serwera przynajmniej w zakresie importu danych z baz danych, tzn. bazy danych są przechowywane na serwerach, a data scientist na personalnym komputerze pisze zapytanie do bazy danych i importuje dane do swojego środowiska pracy (jak Python / R). Nawet w przypadku, gdy zbiór danych mieści się w pamięci RAM komputera personalnego, stosowana jest czasem praktyka wykonywania operacji na serwerze, gdyż znacznie może to skrócić czas obliczeń (a jest to w badanym świecie traktowane jako zaleta), szczególnie na etapie trenowania modelu uczenia maszynowego, czy porównywania wielu modeli ze sobą. W poniższym fragmencie Rozmówca opowiada o wykorzystywanej infrastrukturze sprzętowej, i częściowo o oprogramowaniu, co naszym zdaniem podsumowuje części 5.3.1. i 5.3.2., oraz stanowi ilustrację tego, jak omawiane tam technologie łączą się w pracy data scientista z technologiami omawianymi tutaj:

> R: (...) Czyli pracujemy na laptopie, na tym laptopie mamy SQL Microsoft coś tam Studio Server, czy jakkolwiek tam to się nazywa, eee w każdym razie korzystamy z tego SQLa Microsoftowego eeem [pauza] no i z eRa[190], a jeżeli ktoś ma ochotę to każdy ma lokalnie admina, akurat ja mam zainstalowaną też Anacondę, czy jakieś tam inne frameworki analityczne w Pythonie, czy w czymkolwiek się na prawdę chce. (...) I mamy też dostęp serwera do takiego analitycznego, potężnego [pauza] gdzie mamy 80 rdzeni iii [krótka pauza] jedną kartę graficzną którą tak jakby na próbę sobie na razie kupiliśmy, yyy jakąś tam Nividię (...). Więc też mamy możliwość jakby pracy z takim sprzętem większego kalibru (...) po prostu jesteś w stanie zapodać jakiś algorytm optymalizacyjny na noc i on nam sprawdzi milion wersji różnych modeli i to jest bardzo fajne akurat (…) na tym serwerze uyy jest, postawiony **Debian**, ale wszystkie inne serwery, w firmie w zasadzie chodzą na Windowsie, tak samo systemy operacyjne **nasze**, są windowsowe, no więc na tym serwerze obliczeniowym to się logujemy przez tam PuTYY, yyy m chyba że tam zainstaluję jakiegoś basha na tym, na kompie, ale no raczej jest to po prostu takie logowanie się przez jakieś tam SSH, yyy no, a reszta to mamy z poziomu Windowsa (…).
> *B: Ok. A czy taki serwer, czy w ogóle jakieś takie rzeczy mmm sprzętowe też robią ludzie z zespołu data science, czy macie jakiś oddzielny zespół IT, czy może nie wiem jeszcze jakoś inaczej?*
> R: Mamy oddzielny zespół IT, ale też dużo rzeczy robimy sami, no bo, IT to nasze głównie się zajmuje utrzymywaniem infrastruktury takiej [krótka pauza] **operacyjnej** nazwijmy to, czyli ta część firmy która posługuje się własnym oprogramowaniem, no to oni utrzymują to oprogramowanie głównie, nie, i całą infrastrukturę taką twardą, czyli właśnie te wszystkie serwery na których trzymamy dane, my mamy swój serwer taki, raportowy gdzie [krótka pauza] powstaje kopia danych produkcyjnych codziennie i na podstawie tego robimy swoje obliczenia, także jest to oddzielone od systemów produkcyjnych, żeby im tam nie muliło. Kiedyś tego nie

190 Zapisujemy nazwę języka R w ten sposób, by oddać stosowaną polską wymowę, jak „w eRrze", „do eRa", „z eRem". Bardzo rzadko zdarza się wymowa anglojęzyczna „aR", „w aRze" (chyba, że osoba wypowiada się w całości po angielsku, co nie jest rzadkie). Nazwę Python zapisujemy za to anglojęzycznie, ponieważ stosowana jest wymowa anglojęzyczna z polską odmianą, zatem „Pajton", „w Pajtonie", „do Pajtona", co zapisujemy jako „w Pythonie" itd. Niezwykle rzadko zdarza się wymowa „Pyton" bez anglosaskiego akcentu, naszym zdaniem ma to charakter żartobliwy lub ironiczny.

> wiedzieliśmy i w jednym kraju, eem to był ten sam serwer, tylko tam wirtualnie został podzielony, no i się okazało że jak my tam zapodaliśmy jakąś analizę to [krótka pauza] to pracownicy tego kraju nie mogli w ogóle nic zrobić, bo system im nie reagował na klikanie. No ale to się tylko raz zdążyło nam takie coś. Generalnie po tym przypadku, już wszystkie kraje przeszły na **podwójne** serwery. Eee no, i tak sobie właśnie liczymy, no i część z tych, część administracji tymi serwerami analitycznymi jest po naszej stronie, yy też nasz departament analityczny zajmuje się utrzymaniem hurtowni danych, gdzie w ogóle łączymy te źródła [krótka pauza] yyy jakby w jedno wielkie źródło, yym iii i obliczamy jakby i w ogóle już z pominięciem tych serwerów raportowych, obliczamy to po naszej stronie, głównie te właśnie yyy sprawy raportowe, emm no i my mamy ten, my jako nasz zespół ten analityczny to administrujemy tym, tym serwerem, właśnie tym kolosem takim co ma te GPU i te 80 rdzeni {wywiad 15}

Zauważmy, że w omawianym przykładzie „każdy [w zespole data science] ma lokalnie admina" {wywiad 15}, co znaczy, że na personalnym komputerze może instalować to, co uważa za stosowne, i korzystać np. częściowo z R, częściowo z Pythona. Każdy ma także oprogramowanie firmy Microsoft, umożliwiające import danych z bazy za pomocą zapytania w SQL-u. Dane podzielone są na serwery operacyjne i analityczne (czyli OLTP i OLAP, por. 5.3.2.), zbudowane tak, by zespół DS pracował na kopiach danych, nie zaburzając pracy operacyjnej firmy – zwracamy uwagę na historię, gdzie „my tam zapodaliśmy jakąś analizę to pracownicy tego kraju nie mogli w ogóle nic zrobić" {wywiad 15}. W tej firmie osobnym bytem jest hurtownia danych, czyli „jedno wielkie źródło [danych]" a jeszcze innym serwer analityczny, „kolos" o 80 rdzeniach procesora CPU i z GPU kupionym „na próbę" {wywiad 15}. Serwer analityczny pracuje pod kontrolą wolnego (*free software*) systemu operacyjnego Debian, będącego dystrybucją Linuxa. Człowiek z laptopem pracującym pod kontrolą systemu Windows, czyli tzw. klient, łączy się z serwerem przy pomocy otwartego (*open source*) oprogramowania PuTTY, które emuluje terminal i zapewnia połączenie z użyciem protokołu SSH (*secure shell*, upraszczając jest to szyfrowany standard komunikacji klient-serwer). Rzecz wygląda następująco: człowiek na laptopie uruchamia program PuTTY, widzi interfejs graficzny, wprowadza numer IP serwera, z którym chce się połączyć. Następnie widzi okno terminala (podobne np. do Rterm, por. 25A na rys. 25.), wprowadza login i hasło do serwera. Jeżeli są poprawne, połączenie zostaje ustanowione i można poprzez terminal PuTTY korzystać z serwera, używając komend systemu operacyjnego serwera. Rozmówca mówi także o „instalowaniu basha na kompie" jako alternatywie. Bash to także rodzaj terminala[191], typowego dla systemów Linux / Unix (w tym macOS, który także jest systemem unixowym). Terminal systemu Windows nie jest kompatybilny z bashem, stąd instalowanie

191 Bash (*The Bourne-Again Shell*), to domyślna tzw. powłoka (*shell*) systemów Linux/Unix, a także język programowania skryptowego, którego się w niej używa. Określenia: terminal / powłoka / bash / wiersz poleceń są używane nierzadko jako synonimy, zasadniczo chodzi o program do kontroli innych programów i systemu operacyjnego za pomocą interfejsu tekstowego, a nie graficznego (GUI), o czym więcej w części 5.3.5.

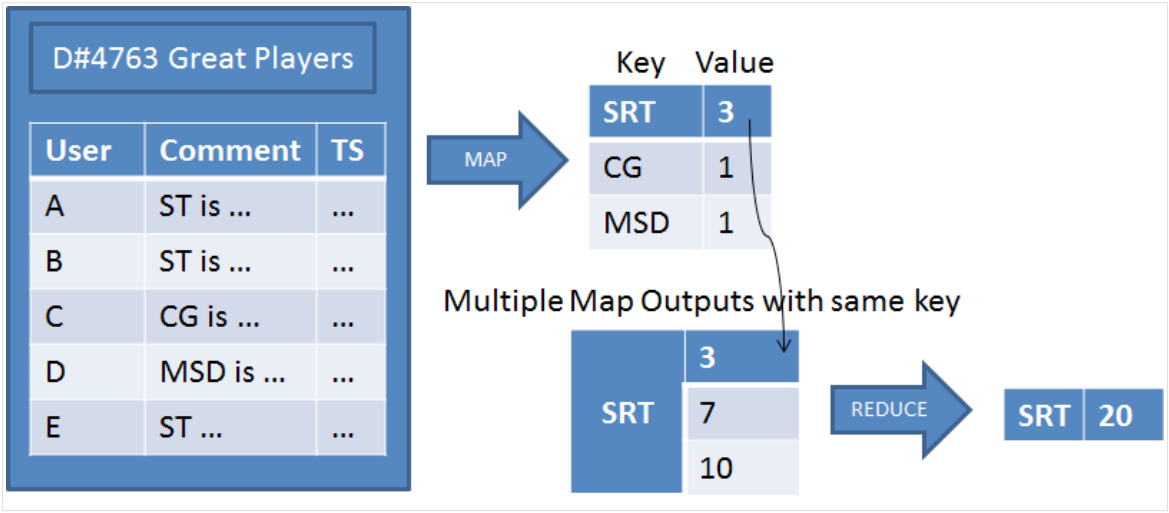
na Windowsie takich emulatorów, jak właśnie PuTTY czy inne, do różnych zadań (np. korzystania z kontroli wersji Git, o czym w kolejnej części rozdziału).

W powyższym fragmencie widzimy, że to zespół DS zajmuje się utrzymaniem infrastruktury do celów analitycznych, zaś zespół IT infrastruktury operacyjnej. **Osoby odpowiedzialne za infrastrukturę w zespołach DS raczej nie są w badanym świecie uznawane za osoby zajmujące się DS**, ta rola określana jest przeważnie jako inżynier danych (*data engineer*), czemu przyjrzymy się bliżej w części 6.

Rozmówca cytowany wyżej {wywiad 15} mówi o 80 rdzeniach serwera analitycznego. Niestety nie zapytaliśmy w rozmowie, czy jest to jeden czy więcej procesorów, gdyż ma to znaczenie. Podkreślmy, że **rozpraszanie czy tzw. równoległe operacje na danych wykonywane są na wielu maszynach –** a właściwie procesorach lub dyskach – **równocześnie**. Już dekadę temu znacznie taniej można było kupić więcej słabszych komputerów niż jeden o równoważnej mocy: osiem serwerów, każdy z ośmioma CPU i 128 GB RAM było bardziej opłacalne niż jeden serwer z równoważnymi zasobami czyli 64 procesorami CPU i 1 TB RAM (Jacobs 2009:43). Tutaj ponownie pojawiają się pojęcia wierzchołków (*nodes*) i krawędzi (*edges*) z teorii grafów – jako wierzchołki są traktowane pojedyncze maszyny (tzn. serwery, ewentualnie procesory), zaś krawędzie to połączenia między nimi. Taka sieć współpracujących maszyn nazywana jest klastrem (*cluster*). Rozproszone może być zarówno składowanie danych, jak i operacje na nich, o ile te operacje nie wymagają komunikacji pomiędzy wierzchołkami (maszynami), czyli mogą być wykonywane niezależnie, równolegle na różnych częściach zbioru danych (ibidem). O tym właśnie mówili Wickham i Grolemund (2017) we wcześniej przytaczanym fragmencie – mając bardzo wiele niezależnych problemów w dużym zbiorze danych, do wykonania obliczeń „potrzebujesz tylko systemu (takiego jak Hadoop lub Spark), który pozwala wysyłać różne zestawy danych do różnych komputerów".

Hadoop ma swoje początki w firmie Nutch, gdzie rozpoczęto prace nad systemem umożliwiającym rozproszone obliczenia (*distributed computing*) na potrzeby otwartej (*open source*) wyszukiwarki internetowej. Projektem zainteresowała się firma Yahoo!, co doprowadziło do oddzielenia się zespołu pracującego nad rozproszonymi obliczeniami od Nutch, i nazwania nowego projektu Hadoop (T. White 2012:xv). Autorzy wskazują (ibidem), że pomocna dla nich była publikacja firmy Google, prezentująca wypracowany na wewnętrzne potrzeby Google'a paradygmat rozproszonych obliczeń MapReduce (Dean i Ghemawat 2008). Bez zagłębiania się w szczegóły techniczne idea jest następująca: pierwszy

krok to mapowanie (*map*), polegające na przetworzeniu par klucz-wartość w celu wygenerowania pośrednich kluczy-wartości, zaś drugi to redukcja (*reduce*) czyli złączenie wielu wartości z kluczy pośrednich do jednej dla wszystkich wystąpień klucza pośredniego (Dean i Ghemawat 2008:107). Jest to wyrażenie innymi słowami tego, co wcześniej za Wickhamem i Grolemundem (2017) nazwaliśmy równoległym przetworzeniem wielu niezależnych problemów na różnych częściach zbioru danych i różnych komputerach. Obrazuje to rysunek 33.



Rysunek 33: Schemat działania paradygmatu MapReduce na przykładzie bazy danych o architekturze klucz-wartość

Źródło: Big Data, NoSQL and MapReduce (Bhatt 2013)

Rozproszone na wiele maszyn jest także samo przechowywanie danych. Stosowana jest technologia HDFS – powiedzmy jedynie, że ma ona umożliwiać to, co nazywane jest horyzontalną skalowalnością baz danych (por. rys. 31.), przy czym chodzi o wolumen danych na poziomie całego internetu (*storage and analysis at internet scale*) (White 2015). Nie zagłębiamy się w szczegóły HDFS, ponieważ problemy infrastruktury do przechowywania, a nawet przetwarzania danych (bez wątpienia w przypadku celów innych niż analityczne) nie leżą w obszarze zainteresowania świata społecznego DS. Przynależą one do szeroko pojętych ról informatyków zajmujących się big data, w sensie technicznym, ewentualnie do inżynierów danych (*data engineer*), których świat może przenikać się ze światem DS (a na pewno występują tutaj przecięcia, bo data engineerowie nierzadko współpracują z data scientistami).

Spark jest technologią, która realizuje odmienne niż Hadoop MapReduce podejście do paradygmatu MapReduce, i obecnie wypiera rozwiązanie Hadoopowe (tylko w zakresie operacji na danych, bowiem w zakresie przechowywania danych Spark nie jest rozwiązaniem

w żaden sposób przydatnym, ani tym bardziej konkurencyjnym wobec Hadoopowego systemy rozpraszania przechowywania plików HDFS). W sensie technicznym Spark rozszerza MapReduce o koncept *Resilient Distributed Datasets*, co prowadzi do tego, że Spark może wykonywać wiele rodzajów zadań przetwarzania rozproszonych danych (tworzenia zapytań zbliżonych do SQL-owych, uczenia maszynowego i przetwarzania danych w reprezentacji grafowej), do czego wcześniej potrzebne było kilka innych narzędzi (Zaharia i in. 2016:56). Zatem Spark to wywodzący się z pracy trwających od 2009 na Uniwersytecie w Kalifornii „zunifikowany silnik do przetwarzania big data" (ibidem).

Zasadniczą różnicą – w sensie możliwości wykonywania operacji na danych – między Hadoopem a Sparkiem, jest to, że Hadoop pracuje wyłącznie w trybie wsadowym (*batch*), zaś Spark zarówno w trybie wsadowym, jak i strumieniowym (*stream*). W trybie wsadowym przetwarzany jest zbiór danych historycznych, np. wszystkie transakcje z dużej platformy aukcji internetowych z jednego dnia, a następnie zwracany jest wynik, np. suma kwot tych transakcji w podziale na kategorie produktów. W trybie strumieniowym dane przetwarzane są w czasie rzeczywistym (*real-time*), np. każdy przelew wykonywany za pomocą bankowości elektronicznej danego banku jest oceniany jako poprawny albo oszustwo (*fraud*). W drugim z omawiany przykładów przetwarzanie całości danych z wybranego okresu byłoby bezsensowne, z punktu widzenia bezpieczeństwa w banku. Spark realizuje przetwarzanie strumieniowe w sposób inny niż tradycyjne systemy do tego rodzaju zadań – za pomocą dzielenia danych wejściowych na bardzo małe wsady co 200 milisekund (Zaharia i in. 2016:60). Z uwagi właśnie na możliwość przetwarzania strumieniowego, Spark jest stosowany także do produkcyjnego wdrażania modeli uczenia maszynowego.

Spark jest obecnie technologią szeroko wykorzystywaną, szczególnie w dużych firmach. Autorzy podkreślają szybkość jego działania, w porównaniu do Hadoopa. Chiński gigant internetowy Tencent używał Sparka do przetwarzania około 1 PB[192] danych dziennie, z użyciem klastra złożonego z 8 000 wierzchołków, co według autorów stanowiło największy opublikowany przypadek zastosowania tej technologii (Zaharia i in. 2016:61). Ze Sparka można korzystać za pomocą języków Scala, Python, R i Java, dla Pythona i R istnieją odpowiednie do tego celu pakiety. Rozproszone uczenie maszynowe z użyciem Sparka jest możliwe, ale dostępnych jest 50 typowych algorytmów (ibidem), co w badanym świecie może być uznawane za ograniczenie:

192 PB, czyli petabajt to biliard bajtów, czyli 10^{15} B. Dla porównania 1 TB to 10^{12} B, 1 GB to 10^{9} B, 1 MB to 10^{6} B, 1 KB to 10^{3} B. Kilkanaście stron tekstu i kolorowej grafiki (np. artykuł naukowy) w pliku .pdf zajmuje około 1 MB. Zatem 1 PB to około 10^{9} = 1 000 000 000 takich zapisanych w .pdf artykułów.

> *B: Mhm, OK. A coś takiego jak rozpraszanie obliczeń, to właśnie do tego jest Hadoop chyba czy czy jak to?*
> R: Tak, ale to **niezbyt** dobrze działa. Znaczy zależy co się chce policzyć. Jeżeli chce się policzyć proste rzeczy w stylu daj mi średnią z czegoś albo pogrupuj mi po czymś tam i policz minimum maksimum czy tego typu rzeczy to Hadoop czy tam Spark, eee, w szczególności Spark to nie jest złe rozwiązanie, tak? One sobie z tym całkiem dobrze radzą **chociaż** trzeba się przyzwyczaić do innej specyfiki pracy czyli zadajemy pytanie i już za 3 godziny dostajemy odpowiedź, także trzeba **dobrze** myśleć co się zadaje jakie się zadaje pytanie. Natomiast rozpraszanie obliczeń takich gdzie modelujemy rzeczy zazwyczaj wychodzi przynajmniej na razie dosyć średnio czyli na przykład modelujemy sieciami neuronowymi [pauza]
> *B: Mhm*
> R: No to **w zasadzie** nie znam jeszcze architektury która działa **dobrze** do takiego rozpraszania. Niestety zazwyczaj wygląda to w ten sposób że kupuje się **dobry** komputer, wkłada się do niego cztery najdroższe karty graficzne na jakie cię stać i to jest twoja twoja przestrzeń do modelowania więcej mieć nie będziesz. Bo inaczej nie działa. Znaczy działa bardzo słabo to rozpraszanie. Myśmy co prawda zrobili parę projektów gdzie rozpraszaliśmy obliczenia na na przykład na 2000 serwerów naraz, ale [pauza] ale musiały być proste rzeczy, tak? To wtedy działa. {wywiad 2}

Podkreślmy, że zdaniem Rozmówcy technologia Spark spełnia swoje zadania do analizy danych typu statystyki opisowe w grupach, zatem do podsumowań, a to uznawane w badanym świecie jest za proste. Nie sprawdza się ona jednak w modelowaniu, szczególnie przy użyciu uczenia głębokiego. Jak później pokażemy, używanie zaawansowanych – cokolwiek to nie oznacza – metod uczenia maszynowego, a szczególnie uczenia głębokiego, może być tym, co z perspektywy uczestników badanego świata odróżnia data scientistów od data analystów / analityków danych. Tym samym twierdzimy, że Spark nie jest w badanym świecie postrzegany jako narzędzie charakterystyczne dla DS, a raczej ma zastosowanie na początkowych etapach projektów (może dochodzić do współpracy data scientistów z data engineerami lub generalnie działami IT) oraz na etapach wdrożenia (gdzie może zachodzić współpraca data scientistów z machine learning engineerami lub działami IT). Niemniej **Spark został na podstawie ankiet portalu KDnuggets uznany za składową najpopularniejszego, sześcioelementowego ekosystemu technologii open source do zastosowań w DS i uczeniu maszynowym** (*6 tools form the modern open source Data Science / Machine Learning ecosystem*). Ten ekosystem to (Piatetsky 2018d):

- **Python** – język programowania;
- **Anaconda** – dystrybucja Pythona do zastosowań DS;
- **scikit-learn** – pakiet metod tzw. tradycyjnego uczenia maszynowego;
- **TensorFlow** (TF) – pakiet stosowany głównie do metod uczenia głębokiego;

- **Keras** – tzw. nakładka ułatwiająca korzystanie z TF (te pięć narzędzi omówiliśmy w części 5.3.1.);
- **Apache Spark** – omówiona wyżej zunifikowana technologia do rozproszonego przetwarzania danych.

Do zadań przetwarzania danych rozproszonych na klaster można wykorzystywać wiele innego rodzaju technologii, nieco bardziej wyspecjalizowanych niż Spark lub współpracujących ze Sparkiem, np. uznawane za część – to określenie pojawia się kolejny raz w odniesieniu do zestawu komplementarnych narzędzi – ekosystemu Apache Hadoop: Storm (przetwarzanie strumieniowe), Impala czy Hive (zapytania zbliżone do SQL), Pig (platforma analityczna, która służy do kodowania np. zadań w Hadoop MapReduce lub Sparku), Crunch (podobna do Pig platforma do kodowania zadań w MapReduce) (White 2015; Zaharia i in. 2016). W wielokrotnie przywoływanej ankiecie skierowanej do polskich firm AI autorstwa Fundacji digitalpoland (Borowiecki i Mieczkowski 2019) nie zadano żadnego pytania o tego rodzaju technologie big data.

Tak zwana „chmura obliczeniowa" to wiele klastrów, których zasoby i moc obliczeniową można wypożyczyć od właścicieli. Model biznesowy tego typu usług wydawał się nam początkowo tym, co określa się jako ekonomię współdzielenia (*sharing economy, peer-to-peer economy, mesh, collaborative economy, collaborative consumption*), czyli

> system społeczno-gospodarczy zbudowany wokół podziału zasobów ludzkich i materialnych. Obejmuje on wspólną kreację, produkcję, dystrybucję, handel oraz konsumpcję dóbr i usług przez różnych ludzi, a także organizacje (Zgiep 2014:193).

Niemniej nie zachodzi tutaj praktyka współdzielenia czy dzielenia się posiadanymi zasobami przez, jak sądzimy względnie równorzędnych w sensie gospodarczym partnerów – **w toku badań nie spotkaliśmy się w świecie DS z korzystaniem z chmur dostawców innych niż największych firm technologicznych: Microsoft Azure, Amazon Web Services (AWS) i Google Cloud Platform (GCP)**. Według raportu Fundacji digitalpoland z obliczeń w chmurze (wymieniono wyłącznie Azure, AWS, GCP) korzystało 76% badanych firm (pytanie umożliwiało wielokrotny wybór), 67% firm używało własnych komputerów personalnych, 36% własnych serwerów, 33% wykonywało obliczenia na hardware zapewnionym przez swoich klientów, 20% wskazało na wynajmowanie wyspecjalizowanego (*dedicated*) hardware'u, zaś 3% na jeszcze inny sprzęt komputerowy (Borowiecki i Mieczkowski 2019:43). Choć za ideą korzystania z chmury obliczeniowej stoi słynne zdanie

„nie potrzebuję mieć wiertarki, a tylko dziurę w ścianie", to **nie jest to współdzielenie**, jak wspólny przejazd prywatnym samochodem kierowcy umówiony przez BlaBlaCar (*hitchhiking/carpooling* – kierowca jedzie własnym samochodem we własnej sprawie, a jeżeli inne osoby nie dołączą, to przejazd odbędzie się z pustymi miejscami dla pasażerów), **a wypożyczenie czy wynajęcie na żądanie** (*on-demand* – kurs odbywa się tylko, jeżeli jest zamówiony, jak w przypadku taksówek czy usług firm Uber, Lyft) (Frenken i Schor 2017:5).

Zatem za pomocą chmury na żądanie dostępne są zasoby i moc obliczeniowa dużych klastrów komputerowych. By skorzystać z chmury po raz pierwszy, przeważnie konieczne jest – poza czynnościami w rodzaju utworzenia konta użytkownika, akceptacji regulaminów, dodania numeru karty kredytowej – skonfigurowanie tzw. **maszyny wirtualnej**, na której będzie wykonywana praca. Od mocy i konfiguracji tej maszyny zależeć będzie koszt korzystania z chmury. Maszyna wirtualna to wynajęta „część" chmury. Użytkownik konfiguruje taki zdalny, wirtualny komputer, wybierając jego procesory (CPU, GPU), rodzaj i wielkość dysku, rodzaj i wielkość pamięci RAM, system operacyjny. Maszyna wirtualna wykorzystywana w praktyce może mieć dużą moc. Na przykładzie GCP może być to np. 80 rdzeni CPU architektury Intel Broadwell i 1,88 TB RAM (konfiguracja o nazwie n1-ultramem-80), jednak do celów dydaktycznych konfiguruje się małe, tanie maszyny, np. 2 rdzenie CPU Intel Skylake i 7,5 GB RAM (konfiguracja n1-standard-2) oraz 10 GB przestrzeni dyskowej i praca pod kontrolą systemu operacyjnego Ubuntu 16.04 LTS {obserwacja 35}. Konfiguracja n1-ultramem-80 to koszt około 8,826 USD za godzinę pracy, zaś n1-standard-2 za ten sam czas kosztować będzie 0,067 USD[193]. Mówimy tutaj o jednej maszynie wirtualnej, zwanej pojedynczą instancją. Taka pojedyncza maszyna będzie służyć raczej jako platforma obliczeniowa do trenowania modeli uczenia maszynowego, ewentualnie do produkcyjnego wdrożenia modelu. Możliwe jest także skonfigurowanie wielu instancji, identycznych lub różnych, i połączenie ich w klaster, a następnie używanie klastra do obliczeń lub przechowywania danych w systemie rozproszonym / równoległym. Taki klaster raczej będzie służyć jako infrastruktura dla baz danych.

Podsumowując, przedstawiamy fragment własnej notatki z obserwacji uczestniczącej przeprowadzonej na warsztatach online. Prowadzący wyjaśnił krótko, czym jest chmura i dlaczego się z niej korzysta, oraz przeprowadził uczestników przez proces założenia konta i konfiguracji maszyny wirtualnej na GCP:

193 Przy tej samej przestrzeni dyskowej 10 GB i systemie Ubuntu 16.04 LTS. Stan na 28.08.2019 r., informacja pochodzi z prywatnego konta RŻ Google Cloud Platform, założonego podczas badań własnych {obserwacja 35}.

> o chmurze - pewne rzeczy nie zmieszczą się na naszym żelazie (tak na hardware mówią geekowie) więc trzeba wziąć większy komputer
> chmura to cudzy komputer i przemiana wykonała się na naszych oczach, stać teraz na niego nie tylko duże instytucje, ale praktycznie każdego i możesz mieć np. 1 - 2 TB RAMu tylko wtedy, kiedy tego potrzebujesz.
> to kosztuje tak mało, że nawet student, decydując czy zje dziś obiadek czy wynajmie komputerek na dwie godziny może sobie na to pozwolić – [prowadzący mówi] „myślę, że decyzja tu będzie zrozumiała"
> 3 największe [chmury to]: Google, Amazon, Microsof Azure
> [prowadzący] podkreśla, że nie pracuje w Google (ale byłoby mu miło), sam go używa, jest rok za darmo, można się pouczyć
> „polecam ci to co najlepsze wiem, jak wiesz coś lepszego to daj mi znać" (...)
> [by skorzystać z GCP] trzeba się zarejestrować, klikać OK i akceptować regulaminy
> konfigurujemy maszynę w chmurze, "zaczyna się robić ciekawie",
> wybieramy lokalizację, [prowadzący] mówi że jak masz biznes w UE to nie opłaca się i nie jest legalne wybieranie serwera w USA z uwagi na to że dane są wtedy poza UE, ale dla uczenia się samodzielnie tańszy będzie ten ze stanów i ten bierzemy (...)
> na chmurze działamy tak - jak potrzebujemy to włączamy, jak nie potrzebujemy to wyłączamy
> {notatka terenowa, obserwacja 35}

Jak widać dostęp do mocy obliczeniowej na żądanie, „tylko wtedy, kiedy tego potrzebujesz" jest uznawany za tak tani, że „nawet student (…) może sobie na to pozwolić" {obserwacja 35}. Ponadto GCP dla nowych użytkowników zapewnia bezpłatne korzystanie z usług do wartości 300 USD przez jeden rok. W formie memów i anegdot powtarzane są żarty typu „budzę się w nocy i boję się, że zapominałem wyłączyć maszynę na AWSie" {np. obserwacja 40}. Opłaty pobierane są bowiem przez dostawców za czas włączenia instancji, bez względu na to, czy maszyna wykonuje obliczenia czy nie. Włączenie i wyłączenie wykonywane jest ręcznie przez człowieka.

Używanie maszyny wirtualnej w chmurze jest określone w powyższym fragmencie jako ciekawe. Przypomnijmy (por. część 5.2.1.), że korzystanie ze sprzętu innego niż komputery personalne lub specyficzne wykorzystanie sprzętu komputerowego (chodzi o używanie GPU do obliczeń) może być postrzegane w badanym świecie, także przez doświadczonych użytkowników właśnie jako ciekawe, fajne, czy jako zabawa:

> R: (…) Jak projekt data science jest źle poprowadzony to wtedy te rzeczy **przygotowawcze** zżerają prawie cały czas projektu i to co jest najfajniejsze o czym się mówi [pauza] **modelowanie**, właśnie ta **zabawa,** z dużymi komputerami, zabawa z GPU, wiele procesorów, to na to zostaje 10% czasu {wywiad 5}

W część 5.4. pokażemy ciekawość – w różnych znaczeniach – jako jedną z kluczowych wartości świata społecznego DS. Prześledzimy także to, co i jak zostaje uznane w badanym świecie za ciekawe. W powyższych fragmentach widzimy, że np. ciekawe może być to, co – upraszczamy celowo – uznawane jest za trudne i wymagające wiedzy matematycznej

(modelowanie), jak i to, co jest wyjątkowe, specyficzne, a jednocześnie mające dużą moc (chmura, GPU, obliczenia równoległe).

„**Chmura to cudzy komputer**" {obserwacja 35}, to skrócone tłumaczenie popularnego – nie tylko w świecie DS ale w ogóle w IT – powiedzenia t*here is no cloud, it's just someone else's computer* (dosł. nie ma chmury, jest tylko cudzy komputer). Twierdzimy, że przez część badanego świata, a nawet przez część osób zajmujących się IT w sensie technicznym, posługiwanie się określeniem „chmura" może być postrzegane jako wskaźnik bycia tzw. osobą nie-techniczną. Sądzimy, że to określenie nie jest aż tak jednoznacznie kojarzone, przez środowiska techniczne, z nie-technicznym dyskursem jak termin „big data". Niemniej jesteśmy przekonani, że przynajmniej przez część uczestników świata DS termin „chmura" (*cloud*) jest kojarzony głównie z komunikatami marketingowymi, kierowanymi z badanego świata do klientów, kandydatów do pracy lub ewentualne wannabesów. Osoby zajmujące się DS w komunikacji między sobą raczej posługują się nazwami konkretnych usług w chmurze, np. „to było postawione na SageMakerze AWSowym". Słowo „postawione" oznacza zrobione, wdrożone, działające.

Reprezentujące nauki humanistyczne Lisa Portmess i Sara Tower (2015) zauważyły, że chmura jest jedną z retorycznych metafor związanych z tzw. big data (autorki używają terminu w sensie szerokim, nie-technicznym). Ich wnioski są zbliżone do tego, co omówiliśmy powyżej – chmura nie jest bezcielesna, jest po prostu cudzym i do tego dużym komputerem. Przy tym autorki słusznie wskazują, że metafora chmury kierowana jest do klientów (to mogą być zarówno osoby zajmujące się DS, jak i ich klienci) i ma wywołać wrażenie, że chodzi o coś niesamowitego, niematerialnego, nieuchwytnego i nadzwyczajnego, podczas gdy mowa o wielkich budynkach – halach z rzędami serwerów, gigantycznymi systemami chłodzącymi, licznymi zabezpieczeniami zapewniającymi energię elektryczną – strzeżonych przez pracowników ochrony[194]. Twierdzą one, że **wrażenie braku materialności tzw. chmury wywołuje** (chyba u osób korzystających z chmury) **poczucie wyzwolenia także z ciężaru moralnego**, jak rozumiemy związanego z tym, jakie dane i do jakich celów są w tej chmurze przetwarzane:

> Chociaż mamy system opisywania coraz większej ilości danych, tylko wyobraźnia może rozróżnić terabajty i petabajty, zettabajty i zebibajty. Big Data to ilość bez ustalonej materialności, chociaż nie oznacza to, że całkowicie nie ma materialności. Wręcz przeciwnie. Choć przetwarzanie w chmurze sugeruje inaczej, Big Data nie jest przechowywana w

194 Zachęcamy czytelniczki / czytelników do obejrzenia wideo, np. Google przygotowało film reklamowy w technologii 360º, prezentujący centrum danych GCP w miejscowości The Dalles w Oregonie, USA . Film dostępny jest na https://www.youtube.com/watch?v=zDAYZU4A3w0, dostęp 7.01.2020 r.

troposferze, ale opiera się na jawnej fizyczności, wypełniając ogromne stodoły danych (*data barns*) i farmy serwerów położone na odległych polach kukurydzy, w porzuconych młynach, zmrożonej tundrze i wszędzie tam, gdzie energia elektryczna jest tania i łatwo dostępna. Metafora chmury obliczeniowej przeciwstawia się wyobrażaniu sobie danych jako uziemionych (*grounded*), a zamiast tego przywołuje wyobrażenie nieprzejrzystej i wyzwolonej (*liberated*) masy, nieobciążonej twardą mechaniką, siłą grawitacji i ciężarem moralnym, które wiążą konwencjonalne [czyli nie-chmurowe] narzędzia i technologie. Nie zdajemy sobie sprawy z tej [materialnej] rzeczywistości za każdym razem, gdy odwołujemy się do transcendentnej amorficznej chmury, ostatecznego wyzwolenia (*the ultimate liberation*) z naszej okablowanej, mechanicznej przeszłości [tłumaczenie własne RŻ] (Portmess i Tower 2015:6).

Autorki zwracają także uwagę na **problem wykorzystywania zasobów naturalnych** – prądu, wody, przestrzeni fizycznej – przez to, co nazywane jest chmurą, a co one określają ogólnie centrami danych, materialnym wymiarem big data:

Zużywając wiele miliardów kilowatogodzin rocznie, centra danych stanowią znaczne obciążenie dla sieci elektroenergetycznych i możliwości generowania energii w społecznościach lokalnych, w których się osiedliły. [Centra danych] Zużywają setki tysięcy galonów wody rocznie, aby przepłukiwać swoje przemysłowe systemy chłodzenia, i wymagają użycia wielu zasilanych dieslem generatorów zapasowych lub banków akumulatorów ołowiowo-kwasowych, w celu ochrony przed awarią zasilania. Zaskakujące statystyki dotyczące zużycia energii przez centrum danych i zasobów zrzucają powszechne złudzenia co do eleganckiej wydajności i przyjazności przemysłu informatycznego dla środowiska. Większość centrów danych zużywa ułamek energii elektrycznej zasilającej ich serwery do wykonywania zadań w danym momencie; reszta służy do utrzymywania serwerów w stanie gotowości na wypadek nagłego wzrostu [przepływu danych]. Zapotrzebowanie na energię doprowadziło do zwiększonego zainteresowania wydajnością energetyczną, zielonymi technologiami i lokalizacjami z dużą ilością energii przy budowie nowych centrów danych [tłumaczenie własne RŻ] (Portmess i Tower 2015:6).

O wykorzystaniu zasobów naturalnych przez systemy bazujące na sztucznej inteligencji pisał także AI Now Institute (2018:6-7, 33-34). Wskazano na problem kosztów, jakie ponosi środowisko naturalne z tytułu istnienia centrów danych. Ma to stanowić, obok problemu pracy rzeszy niewykwalifikowanych osób (tzw. klikaczy, *clickworkers*) z krajów postkolonialnych i rozwijających się, wgląd w pełen łańcuch dostaw (*full stack supply chain*) systemów bazujących na AI (ibidem).

Nie spotkaliśmy się w badanym świecie z dyskutowaniem problematyki wykorzystywania zasobów naturalnych przez, ogólnie rzecz biorąc, infrastrukturę obliczeniową. Uczestnicy świata DS raczej nie zwracają uwagi na materialne aspekty infrastruktury obliczeniowej (poza osobami samodzielnie rozbudowującymi komputery, np. o wysokiej klasy GPU), bowiem nie należy to do ich obowiązków służbowych, raczej nie mają fizycznego kontaktu z materialnością tej infrastruktury (tzn. dotykają głównie klawiatury komputerowej, nie podłączają kabli w serwerowniach) i właściwie całość swoich zadań

wykonują w środowisku cyfrowym. Z zagadnieniami kosztów energii elektrycznej mogą stykać się data scientiści na stanowiskach kierowniczych bądź właściciele firm, choć twierdzimy, że przeważnie będą oni zwracać uwagę na koszt użycia usług chmurowych jako taki, w sensie ceny za godzinę przy uwzględnieniu specyfiki danego projektu. Niemniej z krytyki wobec zużycia zasobów naturalnych przez centra danych zdają sobie sprawę dostawcy usług chmurowych, jak Google, które na swoich stronach www publikuje np. raport Greenpeace'u z 2017 r. Greenpeace dokonał oceny globalnych firm internetowych (*global internet companies*) pod kątem wydajności energetycznej. Na całkowitą ocenę składa się przejrzystość (20%); polityka zaangażowania i lokalizacji w zakresie energii odnawialnej (20%); efektywność energetyczna i łagodzenie emisji gazów cieplarnianych (10%); zamówienia na energię odnawialną, w tym obecny miks energetyczny (30%); rzecznictwo (20%) (Cook i in. 2017:42). Google uzyskało najwyższą kategorię „A", podobnie jak Facebook i Apple. Inni najwięksi dostawcy usług chmurowych otrzymali kategorię „B" - Microsoft i „C" – Amazon (ibidem:8). Greenpeace stwierdził, że cały sektor IT konsumuje 7% energii elektrycznej świata, a przenoszenie obliczeń do tzw. chmury faktycznie zwiększyło światowe zużycie węgla (ibidem:7).

5.3.4. Komunikacja lub współpraca

Omawiamy narzędzia stosowane w badanym świecie do komunikacji lub współpracy pomiędzy uczestnikami badanego świata. Część ma zastosowanie także w komunikacji z osobami nie należącymi do świata. Są to:

- Systemy kontroli wersji, głównie Git;
- Repozytorium kodu bazujące na kontroli wersji – np. serwisy GitHub, BitBucket;
- Narzędzie do pracy grupowej Slack;
- Poczta elektroniczna oraz komunikatory tekstowe i głosowe, np. Skype, Google Hangouts / Meet, FaceTime, MS Teams;
- Media społecznościowe, każde ma odrębną specyfikę: Facebook, Linkedin, Twitter;
- Portal meetup.com;
- Strony internetowe, np. KDnuggets, Towards Data Science (na medium.com);
- Blogi, szczególnie na medium.com, mogą być czytane z użyciem aplikacji agregującej newsy, jak np. Feedly;

- Fora internetowe, ze szczególnym wskazaniem na StackExchange i jego część Stack Overflow;
- Repozytorium preprintów artykułów naukowych Arxiv.

Spośród wszystkich wymienionych wyżej narzędzi, **Rozmówczynie/Rozmówcy wskazywali ewentualne tylko na Git. Pozostałe nie są traktowane jako narzędzia osoby zajmującej się DS**. Niemniej są one powszechnie używane, a wskazanie sposoby ich użytkowania w badanym świecie uważamy za ważny element dla opisania zachodzących w tym świecie społecznym procesów, jak i wartości tego świata.

Zanim jednak opiszemy wymienione narzędzia, podkreślmy, że osoby zajmujące się DS raczej pracują bez ściśle określonych godzin, mogą pracować zdalnie, a współpraca z osobami z innych miast czy międzynarodowa nie należy do rzadkości. Stąd narzędzia do komunikacji i współpracy na odległość są przez wiele osób wykorzystywane codziennie. W artykule będącym aktem samowiedzy osób zajmujących się DS w firmie IBM podkreślano, że „data science to sport zespołowy", i praca w DS polega nie tylko na współpracy osób technicznych – za takie uważają się data scientiści – ale także technicznych z nie-technicznymi, np. managerami, ekspertami dziedzinowymi (Zhang i in. 2020:2–4). W naszych badaniach wyraźnie obecna była także kwestia komunikacji z klientami, zarówno nielubiana i unikana jak i uznawana za bardzo ważną.

Jak mówiliśmy (por. 5.2.2.) działanie może być realizowane przez tę samą osobę w ramach sfery komercyjnej, akademickiej, hobbystycznej równolegle lub jednocześnie, możliwe są także wszystkie inne połączenia tych sfer działalności w ramach działania jednej osoby. Za uczestnika świata DS (ale raczej nie za data scientista) uznana będzie zatem np. osoba, która jest zatrudniona na uczelni wyższej jako asystent, przygotowuje doktorat z biologii w obszarze genetyki roślin, korzystając z języków R, Python i zaawansowanych metod statystycznych, była uczestnikiem konferencji PyData w Warszawie, a w czasie wolnym współorganizuje meetup DS w miejscu zamieszkania oraz z poznanymi na konferencji ludźmi przygotowuje projekt z zastosowaniem uczenia głębokiego, dotyczący rozpoznawania kwiatów na fotografiach (który chce przedstawić na meetupie, na kolejnej konferencji PyData, ewentualnie w przyszłości zmonetyzować, czyli zamienić w produkt komercyjny, np. aplikację na urządzenia mobilne, gdzie użytkownik mógłby rozpoznawać żywe kwiaty za pomocą kamery w smartfonie). Ta hipotetyczna osoba będzie korzystać np. z systemu Git i portalu GitHub zarówno w swoim projekcie badawczym – będzie śledzić

zmiany, które sama wprowadziła, prezentować kod i wyniki promotorom – jak i w projekcie hobbystycznym – będzie śledzić zmiany swoje i zespołu, wskazywać na błędy czy proponować poprawki do czyjegoś kodu. Na GitHubie będzie także szukać inspirujących repozytoriów innych ludzi. Przeszukuje Stack Overflow, kiedy jej kod, pisany w dowolnym celu, nie działa zgodnie z oczekiwaniami; czasami zakłada nowy wątek z prośbą o poradę, rzadziej odpowiada na takie prośby innych użytkowników. Będzie korzystać ze Slacka, by rozmawiać ze swoim zespołem hobbystów, planować pracę, udostępniać sobie nawzajem różne treści. Inny *workplace* (dosł. miejsce pracy) na Slacku będzie służyć tym samym celom w działaniu zespołu organizującego meetup. Będzie korzystać z komunikatorów tekstowych i wideo do rozmów z pojedynczymi osobami. Na Facebooku będzie ona uczestniczyć np. w grupie Data Science PL i śledzić strony z memami poświęconymi DS, na LinkedIn np. zamieści certyfikat ukończenia kursu MOOC lub informację o kolejnym spotkaniu swojego meetupu (wraz z linkiem do odpowiedniej strony na meetup.com), na Twitterze będzie śledzić konta osób uznawanych za autorytety w dziedzinie uczenia głębokiego. W mediach społecznościowych będzie śledzić profile wybranych strony www, jak KDnuggets, zaś w aplikacji Feedly zobaczy codzienne najnowsze wypisy z wybranych stron i blogów. Arxiv i inne repozytoria artykułów naukowych będą dla niej pomocne zarówno w szukaniu materiałów do pracy doktorskiej, jak i w szukaniu najbardziej skutecznej metody optymalizacji hiperparametrów w sieci neuronowej, tworzonej w projekcie hobbystycznym.

Narzędzia do komunikacji i współpracy są niezbędne w pracy komercyjnej, zarówno do kontaktu z klientem, w ramach zespołu DS jak i działu/firmy DS. Osoby zajmujące się DS komercyjnie pracują w małych zespołach, tworzonych ad hoc na potrzeby projektów. W takim, np. trzyosobowym zespole, przełożony może wyznaczyć kogoś do roli lidera projektu. Bywa tak, że pojedyncze osoby zajmują się projektem dla konkretnego klienta, są odpowiedzialne za całość jego realizacji. **Data scientiści raczej najwięcej czasu spędzają na pracy indywidualnej przy komputerze**. Rzadziej komunikują się z członkami zespołu projektowego, dominuje komunikacja drogą elektroniczną. Jeszcze mniej czasu zajmuje komunikacja z całym działem/firmą (chodzi o małą, kliku-kilkunastoosobową firmę zajmującą się tylko DS, tzw. start-up), co raczej odbywa się w formie krótkich spotkań, np. codziennie o godzinie 10 dział/firma spotyka się i każdy pracownik ma dwie minuty na wypowiedź: co robi, co zostało wykonane a co nie, wyniki eksperymentów, ewentualne problemy techniczne/metodologiczne/organizacyjne, następne kroki. Data scientiści na takich spotkaniach konsultują się sobie wzajemnie, co nazywane bywa „dzieleniem się wiedzą” czy

„dawaniem inspiracji” (por. poniżej fragment {wywiad 10}). Konsultowanie się wzajemne odbywa się także w sytuacjach nieformalnych, jak wyjścia na kawę lub posiłek.

Część data scientistów komunikuje się ze „swoimi” klientami – dla których realizują właśnie projekt – część w ogóle nie prowadzi takiej komunikacji. Ten drugi scenariusz występuje raczej w większych organizacjach, gdzie istnieje dział DS, lub w bardziej dojrzałych start-upach. Generalnie w takich działach/fimach, gdzie istnieją względnie stabilne, zdefiniowane role odpowiedzialne za taki kontakt, a czasami za pozyskanie klienta – może być to kierownik działu/zespołu, menedżer projektów, osoba na stanowisku sprzedażowym. Innym scenariuszem jest to, że z klientami pracuje, a także pozyskuje ich właściciel małej firmy. Czasami kontaktu z klientami nie mają osoby z małym doświadczeniem zawodowym w DS, tzw. stanowiska juniorskie, a już stanowiska średniego doświadczenia (zwane np. regular lub specialist) i wysokiego doświadczenia (tzw. senior) pracują z klientami. Komunikacja z klientami ma różne formy: email, telefon (zaplanowane rozmowy to tzw. calle, od słowa *call*), komunikatory wideo, spotkania.

W rozmowie nieformalnej wskazywano nam, że **istnieją praktyki izolowania pewnych data scientistów od kontaktu z klientem**. Osoby uznawane za bardzo introwertyczne, mające trudności z komunikacją nie-techniczną, a nawet w ogóle z komunikacją werbalną, z kontaktem wzrokowym podczas rozmowy z drugą osobą, mogą być tymi izolowanymi od klientów przez swoich przełożonych lub kolegów. Może także występować praktyka celowego unikania kontaktu z klientem. Słyszeliśmy także żartobliwe wypowiedzi w rodzaju „takiego to lepiej klientowi nie pokazuj, może co najwyżej na callu posiedzieć”. **W opisie takich izolowanych od klientów osób – ale nie tylko takich – data scientiści mogą stosować kategorię nerda**. Nerd jest złożoną i absolutnie nie traktowanym w badanym świecie jako negatywna kategorią, z którą identyfikuje się część uczestników świata. Data scientist w ogóle bywa przez osoby nie-techniczne stereotypowo kojarzony z nerdem, jako zaniedbanym, mrukliwym mężczyzną, spędzającym samotnie długie godziny przy komputerze. Do nerdów powrócimy jeszcze wielokrotnie.

Nie spotkaliśmy się z praktyką dołączania klientów do komunikatorów do pracy grupowej typu Slack, ani udzielania im dostępu do repozytoriów kod. Czasami w pewnych przypadkach klient może otrzymywać fragmenty kodu, ale raczej nazywa się to raportami, ponieważ chodzi o komunikowanie obecnego stanu projektu – są to dokumenty zawierające prezentacje wybranych wyników analiz np. w Jupyter Notebook lub R Markdown. Zdecydowanie częściej jednak uznaje się, że klienci nie-techniczni nie są w stanie

skonsumować takiej formy wyników, zatem raporty przyjmują formę arkuszy Excela, dokumentów tekstowych w pdf-ach, w przypadku spotkań np. slajdów. Na późniejszych etapach projektów może dochodzić do tego prezentacja prototypu, np. budowanej aplikacji internetowej – dashboardu z automatycznie aktualizowanymi wizualizacjami danych, metrykami (tzw. KPI: *Key Performance Indicators* czyli kluczowe wskaźniki efektywności), tabelami.

Wracając do pracy zespołowej i indywidualnej data scientistów, organizowanie tej kwestii w start-upie ilustruje poniższa wypowiedź:

> *B: Albo właśnie, może, jak to jest zorganizowane? W jakieś zespoły czy*
> R: Tak, tylko że to są zespoły, uwarunkowane tym dla jakiego klienta akurat się pracuje. Są zespoły na przykład pięcioosobowe do tego klienta przyporządkowane i trzyosobowe do tego klienta i tak dalej, więc to nie jest na stałe.
> *B: Ok, zależnie od projektu po prostu, tak? Łączycie się w zespół zależnie od projektu?*
> R: Tak, znaczy do tej pory było tak [pauza] było trochę z przypadku, że akurat dwie osoby doszły akurat rozpoczął się projekt, no to te dwie nowe osoby szły do nowego, nie było takiego przydziału że aha wy jesteście dobrzy w tym to idziecie tam. Trochę za młody jest ten start-up chyba jeszcze na to żeby tak było.
> *B: Aha, ok. Czyli tak w, tak właściwie to zawsze pracujecie zespołowo? Y czy też indywidualnie?*
> R: Znaczy nie ma za bardzo tak żeby jedna osoba tylko pracowała nad projektem dla danego klienta. Natomiast już sama forma pracy **jest** bardzo indywidualna, no bo pogadamy sobie z pół godziny, ale później każdy siada przy komputerze i robi swoje, nie?
> *B: Aha, ok.*
> R: Więc nie ma ciągle pogaduch. Tak bynajmniej było parę miesięcy temu, więc mam nadzieję że się nie zmieniło [z uśmiechem]. No bo nie pracuję z nimi teraz tylko tylko siedzę w korporacji, od tych paru miesięcy, ale wydaje mi się że [krótka pauza] nie no trzeba jednak swoje odsiedzieć przed komputerem, swoje gdzieś tam, napisać kodu i tak dalej, później znowu się można zgadać, ale generalnie to jest raczej indywidualna praca. W sensie każdy ma podział obowiązków i robi swoje nie? Tylko bardziej się później chcemy dowiedzieć od kogoś, bo może ktoś mógł, zainspiruję go po prostu czymś. {wywiad 10}

Rozmówczyni zaznacza, iż zespoły do nowych projektów tworzyły się nie z uwagi na kompetencje członków zespołu, ale na to, że w ogóle przyjęto nowe osoby – zapewne przyjęto właśnie z uwagi na pozyskanie nowego klienta. W takim „młodym" start-upie każdy pracownik, piszący kod do operacji na danych (w sensie działania podstawowego) pełni szeroką rolę o nazwie data scientist. W bardziej „dojrzałych" organizacjach pewne obowiązki takich data scientistów wykonują osoby o rolach nie-technicznych (np. menedżer projektów), oraz osoby o węższych rolach technicznych (np. data engineer, machine learning engineer, data analyst) lub innych rolach technicznych (szeroko pojęci informatycy).

W części poświęconej wartościom społecznego świata (5.4.) powiemy więcej o wielokrotne wspominanej, a tak cenionej w DS efektowności, a także o samodzielności, obu

widocznych w powyższej wypowiedzi. Zauważmy, że Rozmówczyni „ma nadzieję, że nic się nie zmieniło" w kwestii braku „pogaduch" w pracy. To wskazuje na uznawanie kolejnego wymiaru efektywności za wartość – pogaduchy, czyli rozmowy towarzyskie nie są wartościowe, a konsultacje już tak. Samodzielność widzimy w tym, że „każdy ma podział obowiązków i robi swoje" {wywiad 10}, zatem ludzie sami pilnują wykonania swoich zadań. Ponadto **konsultacje z kolegami następują raczej po wykonaniu czegoś samodzielnie, nie w trakcie ani przed**. Związane jest to z iteracyjnym charakterem procesu wykonywania projektu DS – żeby coś poprawić w kolejnej iteracji, trzeba najpierw skończyć pierwszą iterację, mieć jakikolwiek wynik czy efekt zakończonego etapu.

Zauważmy, że tak rozumiana **samodzielność jest kluczowym elementem posługiwania się forum przeznaczonym do zadawania pytań i udzielania odpowiedzi związanych z kodowaniem – Stack Overflow**. Nie możemy zatem uznać jej za charakterystyczną tylko dla DS, a dla szerokiej społeczności osób piszących kod w ogóle. Najpierw należy poprawnie przygotować pytanie (m.in. podać stosowane wersje narzędzi, system operacyjny komputera), najlepiej dołączając działający fragment kodu lub kodu i danych wraz z wyjaśnieniem jaki jest rezultat osiągany, a jaki oczekiwany (Skeet 2010). Istnieją pakiety wspomagające przygotowanie takich działających, powtarzalnych przykładów (*reproducible example*), na potrzeby zamieszczania na Stack Overflow, zgłaszania tzw. issue na GitHubie i komunikacji na Slacku (Bryan i in. 2019). W pytaniach na Stack Overflow nie należy używać zwrotów grzecznościowych, gdyż może to być uznawane za rozpraszające i niepotrzebnie zajmujące miejsce, zatem stratę czasu – a to wbrew zasadzie efektywności. Zdecydowanie nieakceptowane są generalnie pytania na zasadzie „jak zrobić X od zera?" – akceptowana forma to „próbowałem A (działający przykład), które ma dać mi X (opis), niestety uzyskuję Y (działający przykład), jak osiągnąć X?". Zauważmy przy okazji to, co jest przedstawiane jako kolejna zaleta pisania kodu do operacji na danych w porównaniu do używania GUI (czyli klikania) – kod to po prostu tekst, więc można np. skopiować go i wkleić, wyszukiwać go w internecie, można go wysłać emailem, czyli generalnie łatwiej skomunikować się z innymi ludźmi w kwestii tego, co dokładnie chcemy zrobić i łatwiej uzyskać pomoc w rozwiązaniu swojego problemu (Wickham 2018f).

Systemy kontroli wersji, jak najpopularniejszy Git, zostały stworzone po to, by wspomóc pracę zespołów programistów (*software developers*) nad dużymi projektami dotyczącymi budowy oprogramowania. Git lub inny system służy do zarządzania i śledzenia ewolucji całego zbioru plików cyfrowych, które składają się na projekt programistyczny.

Można to porównać do funkcji śledzenia zmian, znanej z edytorów tekstu, jak Microsoft Word. Jednak kontrola wersji ma znacznie bardziej rozbudowane możliwości, śledzenie zmian jest bardziej rygorystyczne, i może odnosić się do więcej niż jednego pliku. Kontrola wersji, a szczególnie Git, została zaadaptowana przez osoby zajmujące się DS do ich specyfiki pracy. Jest zatem używana do zarządzania zbiorem plików, które zazwyczaj składają się na projekt DS: pliki z oczyszczonymi danymi, wykresy (pliki obrazów), raporty (dokumenty tekstowe) i kod (Jennifer Bryan 2018:2–3). Sam Git jest oprogramowaniem na wolnej licencji (*free/libre software*) GNU GPL, dostępnym od 2005. Inne, starsze, a obecnie mniej popularne rozwiązania to SVN i CVS, na tej samej licencji (Spinellis 2005:4).

Projekt DS zlokalizowany w konkretnym folderze na dysku komputera to, w terminologii kontroli wersji, repozytorium[195] (*repository*, w skrócie *repo*). Pod kontrolą jest właśnie takie repo i wszystkie zlokalizowane w nim pliki. Po wykonaniu zmiany, jak napisaniu kolejnych linii kodu, człowiek zapisuje plik z kodem, tzw. skrypt. Następnie wykonuje *commit* (dosł. zatwierdzenie, nie stosuje się polskiego tłumaczenia), tworząc odpowiednik „fotografii" repozytorium w konkretnym punkcie czasu. Zestaw różnic w zmienionym pliku nazywany jest *diff*. Ta funkcjonalność umożliwia dokładne prześledzenie zmian w każdym pliku, i to nie tylko pomiędzy następującymi po sobie commitami, a pomiędzy dwoma dowolnymi (Jennifer Bryan 2018:10). Tego rodzaju podejście do śledzenia zmian pozwala zachować kontrolę nad kolejnymi wersjami plików nawet pojedynczej osobie, której łatwo zagubić się w nieformalnej strategii bazującej na zmianach nazw plików, jak np. art1_final.doc, art1_final_poprawki.doc, art1_popr_2 itd. Sprawa staje się znacznie bardziej skomplikowana przy współpracy zespołu, który np. przesyła sobie kolejne wersje kolejnych części tekstu za pomocą emaili. W kontroli wersji Git każdy commit musi mieć wiadomość, opisującą krótko, co zostało wykonane, istnieje także możliwość dodania czytelnego tagu do wersji pliku, którą chce się wyróżnić (Jennifer Bryan 2018:3–4, 11–12). Git nie jest przeznaczony do pracy z dużymi plikami binarnymi, tzn. z plikami w formatach biurowych, jak .pdf, .xlsx, .docx i odradza się korzystania z nich, m.in. z powodu nieczytelnego śledzenia w nich zmian oraz niedziałającego mechanizmu automatycznego łączenia wersji (*merge*) (Jennifer Bryan 2018:15–16).

W przypadku pracy zespołowej, podkreślaną zaletą Gita jest to, że jest on zdecentralizowaną kontrolą wersji. Oznacza to, że każdy ze współpracowników ma pełną

195 Przypomnijmy, że także cały GitHub nazywa się repozytorium, podobnie mówi się o repozytorium pakietów, jak np. CRAN.

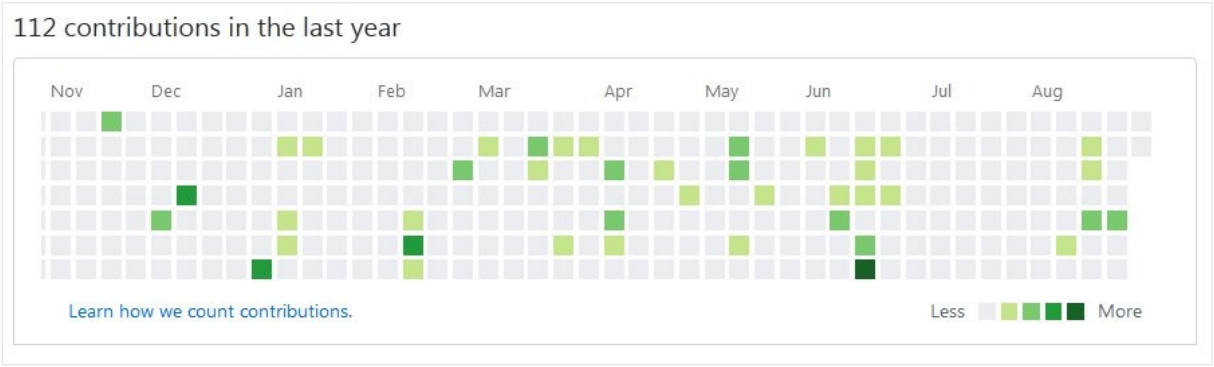
kopię repozytorium i historii jego zmian. Mogą oni pracować jednocześnie lub nie, a zmiany przez nich wprowadzone można prześledzić, a także automatycznie złączyć w całość (*merge*) i kontrolować kłopoty ze złączeniem (*merge conflict*). Istnieje także możliwość tworzenia *branch* (dosł. odgałęzienie, nie tłumaczone), czyli części projektu, które pozostają w repozytorium, ale czasowo są rozwijane oddzielnie, a później złączone z gałęzią główną (ibidem:16-18, 20).

GitHub jest narzędziem zarówno ułatwiającym pracę grupową, bowiem repozytorium na GitHubie jest centralnym w, z założenia zdecentralizowanym, systemie Git (Jennifer Bryan 2018:17) i np. umożliwia ściągnięcie najnowszej wersji repozytorium wraz z całą historią, w przypadku dołączenia do zespołu nowej osoby czy awarii czyimś na komputerze, oraz **narzędziem służącym upublicznianiu swojej pracy**. Serwis GitHub, obecnie należący do firmy Microsoft, umożliwia darmowe tworzenie publicznych repozytoriów, repozytoria prywatne są płatne. GitHub to zatem serwis internetowy, zapewniający hosting dla repozytoriów Git. Można porównać GitHub do bardziej zaawansowanej i sprzężonej z Gitem wersji tzw. dysków w chmurze, jak Dropbox czy Google Drive. Na GitHubie przechowywana jest zsynchronizowana kopia repozytorium lokalnego.

Na poziomie pracy lokalnej, na komputerze osobistym, człowiek, który wykonał jeden lub więcej commitów może poleceniem *push* (dosł. pchać, nie tłumaczone) dodać zmiany także do repozytorium w chmurze, czyli na GitHubie. W ten sposób synchronizuje repozytorium „centralne" z lokalnym. Polecenie *pull* (dosł. ciągnąć, także nie tłumaczone) służy do lokalnego pobrania zmian wprowadzonych z innej maszyny do repozytorium chmurowego, „centralnego". Repozytoria na GitHubie można wygodnie oglądać w przeglądarkach internetowych, bez umiejętności pisania kodu ani nawet posiadania konta w tym serwisie. Istnieje także wykorzystywana w pracy grupowej, wspomniana wyżej funkcja *issue* (dosł. problem lub sprawa, nie tłumaczone), która pozwala na zgłaszanie błędów, tworzenie list zadań, przy czym te wiadomości mogą być przydzielone do konkretnych osób oraz oznaczane tematycznymi tagami (Jennifer Bryan 2018:7–8).

Dla użytkowników GitHuba dostępnych jest szereg mechanizmów, zbliżonych do funkcjonowania mediów społecznościowych. Istnieje możliwość obserwowania konkretnych autorów, więc i bycia obserwowanym. Istnieje podobny do newsfeedów, np. na Facebooku, panel przedstawiający repozytoria obserwowanych przez nas użytkowników. Można wyróżniać repozytoria gwiazdką (*star*), a gwiazdki są traktowane jako wyraz uznania dla pracy autorów repozytorium (por. część 2.1. tej rozprawy). Można skopiować całość

cudzego repozytorium i pracować nad nim niezależnie od twórców, przy czym informacja o tym skopiowaniu (*fork*) jest zachowywana. Powiązane jest to z funkcją *pull request*, czyli mechanizmem proponowania zmian do repozytorium. Z pull requesta korzysta się za pomocą utworzenia brancha, albo za pomocą forka. Ten drugi mechanizm, tzn. człowiek znajduje publiczne repozytorium i jest nim zainteresowany, następnie kopiuje je w całości z uznaniem autorstwa (*fork*), potem wprowadza zmiany i proponuje ich dodanie do oryginalnego repozytorium (*pull request*), jest akceptowanym w środowisku open source sposobem włączania się nowych osób w rozwój otwartego oprogramowania, w tym np. wielu pakietów w R (Jennifer Bryan 2018:20). Ponadto GitHub jest traktowany jako część publicznego portfolio osób piszących kod, nie tylko zajmujących się DS. Publicznie dostępne są np. wykresy, obrazujące wkład (*contributions*) w czasie każdego użytkownika GitHuba (por. rys. 34).



Rysunek 34: Wkład użytkownika w czasie – GitHub Remigiusza Żulickiego
Źródło: https://github.com/zremek, dostęp 02.09.2019 r.

Tego rodzaju repozytoria mogą być przeglądane przez osoby zajmujące się rekrutacją do pracy bądź poszukujące osób do współpracy. Użytkownicy GitHuba mogą także na podstawie repozytoriów tworzyć profesjonalnie wyglądające strony www lub pokazy slajdów w przeglądarce internetowej za pomocą usługi GitHub Pages. Jennifer Bryan podkreśla wagę łatwego upubliczniania swojej pracy, szczególnie, że GitHub (bez użycia GitHub Pages) w przyjazny dla oka sposób wyświetla pliki sformatowane w języku znaczników Markdown. Przypomnijmy, że np. RStudio umożliwia wygodne przygotowywanie dokumentów, mogących zawierać kod jak i rezultaty jego działania dzięki pakietowi R Markdown. Taki dokument możemy upublicznić w repozytorium na GitHubie, będzie mógł być on odczytany przez każdego za pomocą przeglądarki internetowej na dowolnym urządzeniu, i będzie wyglądać jak element zwyczajnej strony www (Jennifer Bryan 2018:12–14).

Zauważmy, że zdaniem tej, pracującej w firmie R Studio, autorki opisywany sposób kontrolowana wersji projektu DS za pomocą Git i Github jest powszechnie uznawany za najlepszą obecnie praktykę. Przy tym, nie jest to sposób pracy promowany przez statystyków akademickich, co jej zdaniem jest jednym z przykładów na to, że DS to nie tylko statystyka (ibidem:21). Z GitHuba jako źródła danych etnograficznych próbowano korzystać w badaniach inżynierii oprogramowania (Turner 2019).

Byliśmy ciekawi, jak uczestnicy świata DS oceniają przejęcie GitHuba przez korporację, jaką jest Microsoft i czy nie kłóci się to jakkolwiek z ideami open source. Jeden z Rozmówców wskazywał, że przejęcie tego repozytoriom przez tak wysokobudżetową firmę przyczynia się do utrzymania jego najbardziej pożądanej funkcji, czyli możliwości dzielenia się otwartym kodem. Przy tym użytkownik GitHuba nie jest produktem sprzedawanym reklamodawcom, jego dane są raczej bezpieczne, jeśli chce może wykupić pakiet obejmujący m.in. prywatne repozytoria (w wersji podstawowej można tworzyć wyłącznie publiczne). Rozmówca wskazywał, że tak pojęta opieka finansowa Microsoftu może pomóc w zachowaniu niekomercyjnego, przyjaznego użytkownikom charakteru GitHuba. Jako przykład platformy, która zapewniała podobne funkcje, ale „zeszła na psy" Rozmówca podał repozytorium SourceForge:

> R: [rozmowa dotyczy przejęcia GitHuba przez Microsoft] swoją drogą to też nie jest dla mnie jasne, że GitHub sam też byłby, dobry. Byśmy na przykład stwierdzili że im się kończy kasa, jakieś reklamy w ten czy inny sposób dają, albo twoje dane, w ten czy inn sposób. Było takie, najczęściej kiedyś stosowana platforma open source'owa Source SourceForge,
> *B: OK,*
> R: Było takie pierwsze, wiesz, wielkie, wielka sieć open source w ogóle, to wiesz. Zeszło na psy w ogóle. Totalnie na psy w sensie że coraz więcej reklam, coraz więcej takich rzeczy, coraz więcej jakichś projektów z **wirusami**,
> *B: Na czymś próbują zarobić.*
> R: **Z wirusami!** Które po prostu instalują ci jakieś reklamy. Rozumiesz.
> *B: Masakra.*
> R: **Masakra**. Rozumiesz, już takie rzeczy totalnie shady[196]. (...) Więc po prostu takie **nie**. Takie totalnie totalnie nie, i to nie tylko to że zostało przejęte przez Google czy Microsofta, ale dlatego że tak to wyszło. Więc to też jest dla mnie takie pytanie co stanie się chodzi o inne rzeczy ze społeczności może przyniesie ale GitHub jest **pięknym** miejscem i w ogóle takie miejsca i w ogóle takie miejsca to wiesz, ostatnim takim miejscem które to były to biblioteki Aleksandrii jak GitHub.
> *B: To co było, jeszcze raz?*
> R: Ostatnim takim miejscem jak GitHub to była biblioteka w Aleksandrii. I jeszcze taka społeczność, no jedynym dla mnie porównywalnym miejscem jest Wikipedia. (...) {wywiad 26}

Z korzystaniem z publicznych repozytoriów, jak GitHub i ewentualne podobnym BitBucket spotykaliśmy się podczas badań wielokrotnie, na niemalże wszystkich meetupach,

196 Dosł. ciemne, szemranie interesy, działania nielegalne lub nieetyczne.

warsztatach, w bardzo wielu prezentacjach konferencynych. Konfiguracja Gita i GitHuba jest jedną z pierwszych lekcji na popularnym kursie MOOC, poświęconym narzędziom do DS {obserwacja 8}. Podkreślmy jednak, że są to narzędzia zaadaptowane do DS z zespołów programistycznych i choć są uznawane za przydatne, to raczej nie spełniają swojej roli, jeżeli chodzi o zadania charakterystyczne dla DS – kontrolowania wersji modeli uczenia maszynowego i wersji zbiorów danych. **Ponieważ projekty DS mają charakter iteracyjny, w wyniku eksperymentowania powstaje wiele wersji modeli i zbiorów danych.** Rozmówczyni mówi o testowaniu w swoim zespole DS narzędzia, które ma służyć właśnie do kontroli wersji modeli:

> *B: A czy, czy jakieś yy narzędzia takie, no nie wiem, kontrola wersji na przykład, czy to ci się też przydaje?*
> R: Yyy, jeżeli chodzi o kontrolę wersji [kaszel] przepraszam, mm
> *B: Czy może nie?*
> R: Wiesz co było, i bardzo mi się, bardzo przydawało. Zwłaszcza to się przydaje gdy rzeczywiście pracuje się w zespole, i kiedy parę osób pracuje nad tym samym ma to jakiś, jakiś system choćby nie wiem wspólne miejsce na folderze razem z mmm opisanym, opisaną metodą do tego kontrolowanie wersji jest niezbędne. [pauza] Mmm, no i rzeczywiście jak już się ma jakieś finalne produkcyjne rozwiązanie tworzy, a my **w tej chwili** yy w zespole mamy tak zbudowaną infrastrukturę że **do** samego jako repozytorium samego kodu używamy GitHuba, aa jako, jeśli chodzi o kontrolę wersji to jeszcze jesteśmy w trakcieeeee **wybierania** narzędzia. Bo było, tak naprawdę że tak powiem szef nam wskazał narzędzie, ale też open source'owe które testowaliśmy o nazwie Acumos. Yyy to jest takie narzędzie które ma służyć jako marketplace dla modeli ma też właśnie tam trzymać wersjonowanie ma tam trzymać dokumentację, problem w tym że to jest totalnie świeże narzędzie i tam połowa rzeczy nie działała. [śmiech] Eee, mieliśmy się z nim zapoznać, mieliśmy je potestować a efekt jest taki że no wygląda to na papierze pięknie, bardzo fajny pomysł bo nie ma w ogóle takiego rozwiązania innego na rynku, open source'owego ale po prostu nie działa, ale może za rok albo za dwa zacznie, albo jak wprowadzą jakąś dobrą wersję, to może ta wersja będzie działała [śmiech] {wywiad 18.}

Acumos to platforma open source, która ma nie tylko wspomagać kontrolowanie wersji modeli, ale także ich wdrażanie. Zauważmy, że **data scientiści projektują i testują modele w rodzaju laboratorium, odrębnym od środowiska operacyjnego/produkcyjnego firmy**. To laboratorium może być po prostu laptopem z zestawem narzędzi do DS. Acumos może zamrozić parametry danej wersji wytrenowanego modelu i sklonować całe środowisko pracy do „paczki", którą później można uruchomić. Taka paczka nazywana jest kontenerem (*container*). Data scientiści mają wówczas swobodę w budowaniu i trenowaniu modeli za pomocą ulubionych narzędzi. Kontener z gotowym modelem i środowiskiem, w którym model powinien działać tak, jak w laboratorium, jest zamieszczany na tzw. marketplace, gdzie mogą go użyć inne zespoły firmy, np. programiści zajmujący się wdrożeniem na produkcję. Taki kontener z modelem może być zatem używany przez osoby nie mające nic wspólnego z uczeniem maszynowym, ponieważ jest po prostu działającą, czarną skrzynką – dostaje dane

na wejściu i zwraca rezultat. Podkreślmy, że jest to rozwiązanie zapewniające bardziej powtarzalne działanie modelu niż tylko skorzystanie z jego zapisu w pliku .h5 (por. część 5.2.4.), ponieważ w omawianym rozwiązaniu kontener zawiera także odpowiednie środowisko do działania modelu. Acumos oferuje także izolację na poziomie modelu. W pewnych sytuacjach data scientiści chcą trenować wiele modeli do różnych zadań na jednym zestawie danych. Dzięki tej izolacji różne zespoły mogą pracować nad tym niezależnie (Zhao i in. 2018).

Innym rozwiązaniem, z którym spotkaliśmy się na jednej z konferencji {obserwacja 26} jest DVC (*data science version control*). To także narzędzie open source, bazujące na Git i chmurowych usługach przechowywania danych. Git służy w nim standardowo do kontroli kodu (a właściwie plików z kodem – skryptów), zaś połączenie z chmurą zapewnia kontrolę „laboratoryjnych" zbiorów danych i plików z zapisem wytrenowanego modelu (typu .h5 lub zbliżonych). Ponadto trzeci komponent DVC to kontrola *pipeline* (dosł. rurociąg, nie tłumaczone), czyli całego iteracyjnego procesu, na jaki składa się budowa kolejnych wersji modelu uczenia maszynowego. W historii kolejnych commitów zachowywane są ponadto metryki z ewaluacji modelu, określające np. jego dopasowanie do danych (Petrov 2017; Vazquez 2017). DVC to zatem narzędzie lepiej niż sam Git i GitHub dostosowane do specyfiki projektów DS (por. rys. 24). Twórca narzędzia mówił o tym następująco:

> Przede wszystkim musimy zrozumieć, jaka jest różnica między inżynierią oprogramowania a uczeniem maszynowym. **Uczenie maszynowe oparte jest na eksperymentach, problem polega na tym, że masz wiele eksperymentów, masz tysiące, czasem setki eksperymentów i musisz komunikować ich wyniki ze swoimi kolegami**. Zasadniczo także ze sobą, ponieważ jutro nie będziesz pamiętać, co się dzisiaj stało. Za miesiąc nie będziesz w stanie zapamiętać, co zrobiłeś dzisiaj i dlaczego ten pomysł przynosi tak kiepski wynik, musisz zatem śledzić wszystko. Możesz być w sytuacji, gdy jeden z twoich kolegów przyjdzie do twojego biura i powie - „hej, wiesz co? Spędziłem dwa dni próbując tego pomysłu. Wiesz co, to nie działa". A ty: „tak, próbowałem tego samego pomysłu dwa tygodnie temu. Wiem, że to nie działało". Trudno jest przekazać te pomysły, ponieważ masz ich tylko kilkadziesiąt, czasem setki. I to jest różnica. Potrzebujesz ram do komunikowania się za pomocą ogromnej liczby pomysłów, ogromnej liczby eksperymentów. (...) **Mogę powiedzieć jedno kontrowersyjne stwierdzenie - w inżynierii oprogramowania prawie nigdy nie odniesiesz porażki. Jeśli zdecydujesz się zaimplementować jakiś problem** (*issue*)**, zaimplementować jakąś funkcję lub naprawić błąd, utworzysz branch. W 9 przypadkach na 10 zostanie wykonany merge z odgałęzieniem głównym. Możesz ponieść porażkę w sensie jakości wyprodukowanego oprogramowania, możesz ponieść porażkę co do budżetu, ponieważ myślałeś, że projekt zajmie jeden dzień, a zajmuje dwa tygodnie. Ale w końcu merge zostanie zrobione. W data science wypróbowujesz 10 pomysłów, jeden działa bardzo dobrze, a reszta – nie działa wcale. Ale nadal musisz je wszystkie zakomunikować swojemu zespołowi** [podkr. RŻ, tłumaczenie własne RŻ] (Vazquez 2017).

To, że DS polega w dużej mierze na eksperymentach, najczęściej nieudanych, było podkreślane przez Rozmówczynie/Rozmówców wielokrotnie, nierzadko właśnie w porównaniu do pracy programistów czy inżynierów oprogramowania. Więcej o tych różnicach powiemy w części 6.

Kończąc wątek narzędzi dla DS do kontroli wersji zaznaczmy, że o ile **Git/Github i podobne rozwiązania są w badanym świecie powszechnie używane, a ich znajomość jest uznawana za elementarz umiejętności data scientistów, to narzędzia jak np. DVC używane są bardzo rzadko**. Zespoły DS stosują rozmaite chałupnicze metody kontrolowania projektów i wyników eksperymentów. Mogą być to np. notatki w arkuszach kalkulacyjnych, wymiana emaili czy wiadomości na Slacku[197] lub innych komunikatorach, a w przypadku osób pracujących w komercyjnych zespołach DS prawie zawsze codzienne, krótkie spotkania z wzajemną wymianą informacji. W zależności od firmy zespoły mogą pracować tzw. metodykami zwinnymi, zbliżonymi do Scurm czy Agile, mogą mówić o inspirowaniu się podobnymi metodykami lub odcinać się od nich, uznając je za schematy odpowiednie do pracy zespołów programistycznych, nie DS. Zdarzają się żarty z takich metodyk:

> Co jeszcze trzeba wiedzieć: Scrum, [osoba notuje w Google Docs, widać na projektorze] mówi „o napisałem scum[198]" - śmiech na sali {notatka terenowa, obserwacja 38}

W małych firmach, nazywanych start-upami, system organizacji pracy w ogóle – nie tylko kontrolowanie pojedynczego projektu i wyników eksperymentów w ramach niego – jest znacznie mniej sformalizowany i ustrukturyzowany niż w dużych firmach, tzw. korporacjach. Wyraźnie widzimy to w narracji nieco wyżej cytowanej Rozmówczyni, która pracuje w start-upie DS i tę firmę uważa za swoją, jednak w momencie wywiadu realizuje projekt dla korporacji w ten sposób, że fizycznie pracuje właśnie w korporacji, jest tymczasowym członkiem zespołu projektowego w firmie-kliencie swojego start-upu. Korporacji nie uważa ona za swoją, a także kwestionuje to, czy obecny projekt może uznać za DS:

> *B: Yhm, ok. A czy masz jakiś taki typowy dzień pracy, który byś mogła mi opisać?*
> R: [pauza] Właśnie nie. A to dlatego że [krótka przerwa] przez jeden tydzień mogę się jedną rzeczą zajmować, przez drugi inną, więc jedyne co pozostaje stałe to to że przychodzę loguję się do komputera, ewentualnie idę później na kawę, herbatę i zawsze jestem prawie że przy komputerze. Jedyne co, to pracuje w korporacji większej, po prostu mamy co jakiś czasy spotkania, takie bardziej oficjalne, w stylu jakieś groomingi, planingi i tak dalej, stand upy, a w

197 Częściej jest to praca ręczna, ale podejmowane są próby automatyzacji tych chałupniczych metod. Są osoby używające Slacka za pomocą języka Python, pakiet „Slacker", łącząc się z usługą przez API. W ten sposób mogą np. przesłać wyniki ewaluacji modelu uczenia maszynowego wykonanego w Kerasie automatycznie jako wiadomość na Slacku, a nie przeklejać je ręcznie ze środowiska programistycznego (Koehrsen 2018).

198 *Scum* dosł. szumowina, potocznie męt / ścierwo / syf (nie jest to słowo wulgarne, ale poniżające).

małym, w małym start-upie, nie ma tych powiedzmy wielkich planingów, groomingów, no ale nadal to jest praca przy komputerze i pójście na kawę. I czasem jakieś spotkanie. Czasem spotkanie z jakimś klientem, ale to bardzo rzadko.
B: Ok. A co to są groomingi? Bo bo bo nie wiem
R: Grooming to jest takie, przygotowanie do planingu. Czyli jak przychodzisz w poniedziałek rano i robicie sobie plany na dwa tygodnie w przód, co kto robi, w jakim czasie zrobi i tak dalej, to w piątek przed tym poniedziałkiem dobrze jest mieć grooming, czyli tak jakby wyczesanie wszystkich tematów i sprawdzeni co właściwie chcemy przez przyszłe dwa tygodnie robić.
B: Yhm, ok.
R: I opisanie dokładnie tych tych celów, jeszcze bez przyporządkowania ich do osób tak ogólnie. Czyli typowo korporacyjne właściwie rzeczy. Nie wiem czy w tym jest w ogóle data science, właściwie nie jest, po prostu typowo programistyczne rzeczy. No ale data science jest, tak jak w korporacji jakiejś większej, współpracą z programistami, więc musimy się do ich trybu dostosować, więc tutaj bardzo przejmujemy właściwie te same role, i styl życia chyba, tej pracy.
B: To jest też coś, o co myślałem żeby cię zapytać, o różnice pomiędzy data scientistami a programistami.
R: No właśnie to bardzo dobrze widać na przykładzie tego co było yyy, w start-upie małym, w którym są sami data scientiści, a co jest w korporacji gdzie pracujemy z programistami. I to idealnie widać, jakby, właśnie to wszystko co powiedziałam, że jak jest, yym praca programisty, to oni zazwyczaj robią to co wiedzą mniej więcej jak zrobić, w sensie takim że ok, może ktoś nie wie jak zrobić, ale wie że jest ktoś kto wie, bo wie że ma kogo zapytać, albo wie że to rozwiązanie jest w Internecie, bo ktoś kiedyś musiał to zrobić. A w data science? Właściwie nie wiesz do końca co robisz, nie wiesz co ci to da, to jest takie, R&D właściwie czystsze, że, no i to wpływa że nie masz tak ustrukturyzowanego trybu pracy, że nie możesz przewidzieć na dwa tygodnie wprzód co będziesz robił, bo możesz jutro zrobić eksperyment który sprawi że musisz porzucić cała ścieżkę. Dlatego trochę ciężko się pracuje w takim korpo gdzie to jest jednak, powinna być jakaś struktura, a my jednak robimy taki research, że to jednak ciężko jest ubrać w strukturę. {wywiad 10}

Przypomnijmy, że Lowrie (2017, 2018) zaproponował postrzeganie DS jako dziedziny nierozerwalnie łączącej inżynierię i naukę. Choć uznajemy to za dość trafne (por. 2.2.4.) to twierdzimy, że **uczestnicy świata DS postrzegają siebie zdecydowanie bardziej jako poszukujących badaczy, eksperymentatorów niż jako inżynierów**.

Niemniej twierdzimy, że nawet w korporacjach narzędzia typu DVC są używane bardzo rzadko, a jeśli już, to raczej używają ich osoby odpowiedzialne za wdrożenie lub tzw. utrzymanie modeli uczenia maszynowego na produkcji. Utrzymaniem określa się ogólnie serwis i naprawy wdrożonego modelu. Osoby zajmujące się takimi wdrożeniami i utrzymaniem nazywane są np. machine learning engineer, DataOps, DevOps dla DS. Opisujemy te i inne, specjalizujące się role w kontekście procesów profesjonalizacji w świecie społecznym DS w rozdziale 6.

Slack jest unikalnym i popularnym narzędziem do pracy grupowej – pojedynczą platformą dla wysyłania wiadomości, plików i połączenia z innymi narzędziami (Woodgate 2019). Jest to oprogramowanie zamknięte (*proprietary*), z którego w ograniczonej wersji

można korzystać bezpłatnie. Slack nie jest narzędziem charakterystycznym tylko dla DS, używają go bardzo różne środowiska i grupy, także nie związane z technologią ani pracą zawodową.

> Slack wypełnia niezręczną lukę w nowoczesnej komunikacji biznesowej i zespołowej, łącząc nieformalny czat, prywatne wiadomości, przesyłanie plików i zbiorową pamięć o tym, co się wydarzyło. Pozwala on na skupienie grupy wokół tematu - dedykowanej przestrzeni konwersacyjnej zwanej kanałem - a także, w razie potrzeby, na tworzeniu czysto prywatnych dyskusji. Możesz przesyłać obrazy, dodawać sformatowany tekst i łącza w zewnętrznych usługach chmurowych. Korzystanie ze Slacka może czasem przypominać grupowy czat tekstowy, ale dzięki zagnieżdżonym dyskusjom (wątkom), wyszukiwaniu w kanale i obszarze roboczym (*workplace*) oraz integracji aplikacji innych firm, **Slack jest znacznie bardziej wydajny i użyteczny [niż czat]. Wyróżnia się on jako pojedyncze repozytorium mądrości (i danych) dla grupy** [podkr. RŻ]. (...) Na tym polega kłopot dla wielu użytkowników Slacka: nie zdecydowaliśmy się na używanie Slacka, ale musimy to robić, aby uczestniczyć w interakcjach niezbędnych do pracy zawodowej, w działaniu zespołu sportowego, a także rodzicielskiej grupy „Przyjaciele Kapeli" w liceum naszych dzieci (Fleishman 2019:8).

Slack może zatem służyć osobom zajmującym się DS do współpracy i komunikacji w projekcie. Bywa także elementem chałupniczego systemu rejestrowania i kontrolowania eksperymentów z modelami uczenia maszynowego, o czym pisaliśmy wyżej. Slacka można połączyć z innymi narzędziami do komunikacji i współpracy, także tymi związanymi z pisaniem kodu, jak np. GitHub i Stack Overflow; w sumie z około 1,5 tys. różnych innych narzędzi (Woodgate 2019). Slack obsługuje także język znaczników zbliżony do Markdown, zatem można za jego pomocą programowalnie formatować wpisywany tekst, jak również zamieszczać w wiadomościach fragmenty kodu, zachowując ich charakterystyczny dla IDE czy terminala wygląd (czcionka, wcięcia itd.) i wyraźnie odróżniając od innego tekstu.

Twórcy Slacka sądzą, że do sukcesu narzędzia przyczyniła się jego wizualna odrębność od podobnych narzędzi do pracy grupowej, które wyglądają „jak tani kostium na bal z lat 70. – wszędzie przygaszony błękit i szarości" (ibidem). Slack zaprojektowano na wzór żywej kolorystyki gier wideo, użyto zaokrąglonej czcionki bezszeryfowej, przyjaznych ikon i dużej ilości emotikonów, by nie kojarzył się z aplikacją do pracy w przedsiębiorstwie (*enterpise*) (ibidem). **Slack wygląda[199] nieformalnie, swobodnie, a to jak sądzimy zdecydowanie zgadza się z dominującym i uznawanym za odpowiedni wizerunkiem w świecie społecznym DS**. W rozmowach na Slacku nie zaczyna się wiadomości od „Szanowny Panie" i generalnie rozmawia się w sposób zdecydowanie nieformalny, a samo „slack" oznacza dosłownie coś luźnego.

199 Wersja demonstracyjna narzędzia dostępna jest na https://slackdemo.com/, dostęp 7.04.2019 r.

Slack i inne komunikatory tekstowe, mogą być w badanym świecie uznawane za wygodne i co za tym idzie sprzyjające efektywności, ponieważ komunikując cokolwiek całemu zespołowi czy pojedynczej osobie nie oczekuje się natychmiastowej odpowiedzi i nie postrzega się takich wiadomości tekstowych jako odrywających od właśnie wykonywanego zadania – powtórzymy, że zazwyczaj zadania wykonywanego samodzielnie przy komputerze, nierzadko z słuchawkami na uszach. Komunikacja werbalna, gdy druga osoba pracuje wpatrzona w ekran z słuchawkami na uszach, jest raczej postrzegana w badanym świecie jako niemiła, rozpraszająca, a więc kontrproduktywna. Przeszkadza w byciu efektywnym, a to działanie niezgodne z wartościami (por. 5.4.) badanego świata.

Slack jest dostępny przez przeglądarki internetowe, jako aplikacja na komputery z systemami Windows i macOS oraz na urządzenia mobilne z Androidem i iOS, nie ma aplikacji na systemy Linux. Jako wada, szczególnie w komercyjnych zespołach DS z uwagi na formalne umowy dotyczące tajemnicy projektów, bywa postrzegane to, że całość danych ze Slacka znajduje się na serwerach Slacka, a nie firmy DS. Dodatkowo serwery Slacka to faktycznie maszyny firmy Amazon, gdyż usługa działa w chmurze AWS, zatem w niektórych projektach komercyjnych nie używa się tego narzędzia z przyczyn prawnych.

Obszary robocze (*workspace*) mogą być zamknięte lub otwarte. W tych pierwszych mogą dołączać jedynie osoby zaproszone, w drugich każdy posiadacz konta Slack. Istnieje szereg międzynarodowych workspace'ów, na których komunikują się osoby zainteresowane DS, uczeniem maszynowym, AI, big data itp. (Blumenkrantz 2018; Formulated.by 2018). Znamy tylko jeden polskojęzyczny workspace tego typu, ale jest on zamknięty. Tworzą go osoby, uczestniczące w bezpłatnym kursie, tzw. wyzwaniu Data Workshop autorstwa V. Alekseichenko {obserwacja 35, 39, 43}. W polskim świecie DS zdecydowanie bardziej popularne w ogólnodostępnej komunikacji cyfrowej niż otwarte workspace'y Slacka są platformy GitHub, forum Stack Overflow, a komunikację polskojęzyczną zagospodarowuje niemal w całości grupa facebookowa Data Science PL.

Nie opisujemy szerzej komunikatorów tekstowych i głosowych, np. Skype, Google Hangouts, FaceTime, bowiem sądzimy, że są one znane Czytelniczkom/Czytelnikom z doświadczenia. Na ich przykładzie **wracamy do kategorii nerda**. Na portalu LinkedIn pojawił się następujący post, autorstwa osoby silnie osadzonej w badanym świecie:

> Lista „wyzywam Cię (*I dare you*)" dla Data Scientists:
> - napisać wiadomość e-mail do klienta - 10 punktów.
> - napisać przyjazną wiadomość e-mail do klienta - 50 punktów.

- odbyć telekonferencję grupową z klientem, tylko audio - 100 punktów.
- przeprowadzić rozmowę na Skype z klientem, przy włączonych kamerach - 500 punktów.
- wykonać rozmowę telefoniczną jeden-na-jeden z klientem - 1.000 punktów.
- porozmawiać twarzą w twarz z klientem - 10.000 punktów.
- porozmawiać twarzą w twarz z klientem, z kontaktem wzrokowym - 20 000 punktów.
- zacząć pogawędkę (*small talk*) podczas sesji networkingowej na konferencji - 100 000 000 punktów (Kuncewicz 2019b).

W komentarzach do powyższego postu jedna z wypowiedzi wygląda zaś następująco:

> Robię te rzeczy cały czas. (…) Samo to, że ktoś jest ekspertem od ML lub wie, jak kodować, nie czyni go data scientistą. Zdaję sobie sprawę, że należę do mniejszości i że firmy musiały wymyślić tytuł „Analytic Translator" [dosł. tłumacz analityki], ponieważ nikt, kogo zatrudnili jako „data scientist", nie mógł / nie chciał się zaangażować w kontakt z klientami i (powiedzmy, że tak nie jest) uczyć się od klientów, ale data scientist powinien mieć znacznie szerszy obszar ekspertyzy (*expertise*), zainteresowania i doceniać sprawy inne niż algorytmy i kodowanie (i nie, nie mówię o architekturze komputera). Mówię o marketingu, budowaniu zaangażowania klientów, sprzedaży, wydajności sieci, zwrotach kosztów (*reimbursement*), cyklu przychodów i wszystkim innym, czego potrzeba, aby być konkurencyjnym w swoim segmencie rynku. Najlepszym sposobem na nauczenie się tych rzeczy są ludzie, którzy są ekspertami, a często to właśnie Twoi klienci! (ibidem: komentarz użytkownika Henry Zeringue, PhD)

Przygotowana przez Ł. Kunczewicza, żartobliwa lista „wyzwań" ułożona jest w kolejności rosnącej punktacji[200] za wykonanie konkretnych aktów komunikacji. Każdy kolejny punkt listy jest aktem w coraz mniejszym stopniu zapośredniczonym przez medium cyfrowe i uznawanym za trudniejszy – w sensie umiejętności komunikacji werbalnej i niewerbalnej, umiejętności interpersonalnych czy towarzyskich. Zauważmy, że lista kończy się „wyzwaniem" dotyczącym nie kontaktu z klientem, ale z innymi, nieznajomymi osobami zajmującymi się DS. Być może Kuncewicz ocenił to nawiązanie pogawędki na konferencji tak wysoko (100 mln – to 5 tys. razy więcej niż punktacja za rozmowę twarzą w twarz z klientem, z kontaktem wzrokowym), bo osoba próbująca to zrobić napotka nie tylko trudności wewnętrzne, dotyczące jej samej (a jak sądzimy tylko takie założył on w kontakcie z klientem), ale i zewnętrzne, dotyczące potencjalnych partnerów pogawędki. Ci partnerzy – inni data scientiści – będą bowiem w podobny sposób nietowarzyscy czy niekomunikatywni, co inicjator rozmowy. Pamiętajmy, że cała lista to żart i post w medium społecznościowym, zatem uważamy ją za wypowiedź jednowymiarową, przejaskrawioną, nastawioną na przyciągnięcie uwagi i wywołanie emocji, a dalej reakcji w postaci polubienia / polecenia / komentarza itp.

200 O'Neil (2017a:80) sądzi, że posługiwanie się ilościowymi wskaźnikami oceny jest charakterystyczne dla ludzi zajmujący się data science – jakoby prymusów, absolwentów najlepszych uczelni: „koncentrują się na zewnętrznych wskaźnikach takich jak punktacja SAT [odpowiednik matury w USA] czy testy wstępne na uczelnię. To stanowi całe ich życie".

Niemniej omawiany żart w oczywisty sposób nawiązuje do stereotypowego postrzegania nerdów jako osób introwertycznych, nietowarzyskich, racjonalnych i logicznych, nie zainteresowanych nie-technicznymi pogawędkami (czyli błahymi, rytualnymi rozmowami z nieznajomymi, w których treść konwersacji jest nieistotna). **Identyfikacja nerdów z introwertycznymi, racjonalnymi typami osobowości[201] może być nie tylko stereotypem** – wskazywano to za najważniejszą, obok wysokiego IQ, składową definicji prawdziwego nerda dla osób uważających siebie za nerdów. Te charakterystyki są uznawane za podstawę do umiejętności myślenia racjonalnego, logicznego, a przy tym abstrakcyjnego, typowego dla matematyki i informatyki. **Wśród osób uznających się za nerdów może zatem panować przekonanie o tym, że są one elitą intelektualną, wąskim gronem ludzi myślących racjonalnie i logicznie w ogromnej rzeszy nieracjonalnych i nielogicznych przeciętniaków, których interesują towarzyskie gierki** (Tocci 2009:225–28). Niektórzy wiążą nerdów, a także hakerów z osobami w spektrum autyzmu lub konkretnie z zespołem Aspergera, czyli osobami niekomunikatywnymi czy nietowarzyskim, ponadprzeciętnie inteligentnymi i wykazującymi zdolności matematyczne i techniczne z uwagi na jednostkę diagnostyczną w sensie medycznym[202] (Tocci 2009:228–32; Zaród 2018:229, 350). Przegląd literatury na temat opublikował jeden z najbardziej znanych w Polsce data scientistów, Piotr Migdał (Krawczyk i Migdał 2011). W polskim świecie społecznym DS tego rodzaju elitarność intelektualna, tzn. racjonalność / logika są uznawane za wartość (por. część 5.4.), przy czym nie można tego utożsamiać z kategorią nerdów.

Sądzimy, że w badanym świecie data scientist, który lubi kontaktować się z klientami, stara się robić to jak najlepiej, jest osobą pokazującą swą wysoką sprawność w komunikacji werbalnej / niewerbalnej, czy osobą towarzyską, może być przez część uczestników świata postrzegany jako ktoś, kto poświęca swój czas na nieistotne sprawy. Nie znaczy to, że takich osób nie ma w badanym świecie – wręcz przeciwnie. Niemniej naszym zdaniem jest to mniejszość, a przez to tego rodzaju umiejętności interpersonalne (wraz z chęcią ich użycia w kontaktach z klientami czy ogólnie osobami nie-technicznymi) mogą być postrzegane, szczególnie w komercyjnym DS, jako cenna zaleta. O data scientiście jako unikalnej hybrydzie, łączącej umiejętności techniczne z nie-technicznymi pisano bardzo wiele (Conway 2010). Taką pozycję akcentującą znaczenie umiejętności interpersonalnych reprezentuje wyżej pokazany komentarz H. Zeringue.

201 INTJ (Introversion, Intuition, Thinking, Judgment) lub INTP (Introversion, Intuition, Thinking, Perception)
202 Od 1992 r. zespół Aspergera jest w Międzynarodowej Klasyfikacji Chorób ICD-10

Co do mediów społecznościowych, to w polskim świecie DS na **Facebooku** zdecydowanie najważniejszą i właściwie jedyną platformą ogólnopolskiej komunikacji jest grupa Data Science PL. Omówiliśmy ją w części 5.1.4. Dodajmy, że w badanym świecie bardzo dobrze znany jest tzw. skandal Cambridge Analytica i inne kontrowersje związane z tym, co dzieje się z danymi przekazywanymi Facebookowi przez jego użytkowników. Świadomość uczestników uważamy w tych tematach za bardzo wysoką, rzecz jasna nie tylko w odniesieniu do działań firmy Facebook. Cześć uczestników świata zwraca szczególną uwagę na swoją prywatność w internecie, np. nie korzysta z mediów społecznościowych i stosuje alternatywy wobec usług Google'a (jak np. wyszukiwarka internetowa DuckDuckGo i szyfrowana poczta elektroniczna w domenie protonmail.com), niektórzy używają szyfrowania połączeń z użyciem VPN. Inna część uważa, że ich prywatność nie jest zagrożona i korzysta z różnego rodzaju usług. W rozmowach nieformalnych spotkaliśmy się z osobą, która pokazała autorowi (RŻ) jak wyłączyć profilowanie reklam Google na jego telefonie z systemem Android, dodając, że sama ma włączone profilowanie „ze względów zawodowych" – chce widzieć, jakie treści są do niej dopasowywane. Nawet u osób szczególnie chroniących swoją prywatność w internecie może przy tym pojawiać się przekonanie, że **muszą korzystać z mediów społecznościowych**. Także piszący te słowa był przekonywany, że na potrzeby realizacji badań do niniejszej pracy, powinien ponownie[203] założyć, przynajmniej fikcyjne, konto na Facebooku**:**

> R: (...) Natomiast ty podchodzisz do tematu globalnie eee czyli różni data scientiści w Polsce, to na pewno musiałbyś dotrzeć też do tych bardziej za, yyy nie wiem czy zamkniętych, no w każdym razie, no niepotrzebujących tych interakcji nie, chociażby w postaci meetupów iii, i tak dalej. Więc wydaje mi się że tutaj wymagałoby dużego doprecyzowania to co chcesz osiągnąć, żeby wiedzieć czy warto się tam udzielać czy nie. A jeżeli byś chciał dotrzeć, to nie obejdziesz się bez social mediów.
> *B: Mhm, masz rację, masz rację. To jeszcze pomyślę nad tą strategią, w takim razie.*
> R: Nooo. No bo powiem ci że ja na Facebooku no do teraz nie mam zdjęcia, a jak zakładałam grupę to nawet nie miałam imienia ani nazwiska, tylko jakieś tam wiesz yyy **literki** wskazujące moim znajomym że ja to ja, natomiast Facebook mnie bardzo szybko zablokował po prostu, bo [niezrozumiałe] dosłownie moja aktywność tak? Bo gdzieś tam też było dla mnie ważne, powiedzmy trzymanie się swojej prywatności, no ale potem się okazało że coś za coś. No nie mam wiarygodności zupełnie, jako taki anonim, więc już mam imię i nazwisko, no ale nadal nie mam zdjęcia, no bo co ich to obchodzi. [ze śmiechem] No ale w tym sensie, że rozumiem twoje obawy, natomiast ja już się przekonałam że bez tego yyy po prostu nie, nic się nie zdziała jeśli chcesz mieć kontakt z ludźmi pracującymi z nowymi technologiami, nie? Że [pauza] że może faktycznie warto się zastanowić pod tym względem {wywiad 21}

203 Autor (RŻ) przez wiele lat korzystał z Facebooka, jednak w toku badań i lektur, w maju-czerwcu 2017 skasował konto i stopniowo stał się osobą chroniącą swoją prywatność w sieci. Wciąż korzysta z LinkedIn, ResearchGate i wybranych usług Google (poza wyszukiwarką).

Zauważmy, że jest to kolejne odniesienie do kategorii nerda – osoby nietowarzyskiej, mało komunikatywnej, która nie będzie np. przychodziła na meetupy. Kilkoro Rozmówców wskazywało nam, że nie mamy szansy dotarcia z wywiadami do takich właśnie ludzi, a w powyższym wywiadzie Rozmówczyni jako remedium na ową lukę zaproponowała dotarcie do takiego grona za pośrednictwem Facebooka.

Osoby zajmujące się DS lub zainteresowane DS mogą na Facebook śledzić wiele stron i grup związanych z tą tematyką. Z uwagi na profilowanie treści przez Facebooka otrzymują one w newsfeedach różnego rodzaju reklamy i oferty związane z DS. Bez wątpienia możemy wskazać na dwie kategorie takich treści, czyli oferty edukacyjne (np. kursy, szkolenia, studia podyplomowe) i oferty pracy (mamy na myśli także programy stażowe, praktyki studenckie). Jedna z Rozmówczyń wskazała, że **obecnie pracuje w firmie, której reklama praktyk studenckich wyświetliła się jej na Facebooku**:

> R: (...) No i w trakcie studiów robiłam statystykę iii rachunek prawdopodobieństwa i to był przedmiot który, na dość wysokim poziomie był nas uczony i trzeba było się do niego bardzo przyłożyć i mi się to spodobało, zaczęłam więcej o tym czytać, zaczęłam w domu to zgłębiać, czytać podręczniki dodatkowe. Iii jakimś szczęściem yym Facebook nie wiem, wyczuł to co, wyczuł czym się zaczęłam interesować, bo dostałam reklamę że szukają osób na praktyki, które znają znają to mniej więcej co uczyłam się na studiach, więc eee stwierdziłam że, hej lubię lubię to robić, te rzeczy się wydają ciekawe czemu, czemu by nie spróbować robić tego zawodowo, zobaczyć jak to wygląda od, od tej strony przemysłowej, więc tak się zaczęła moja, moja przygoda, no i jak już zaczęłam pracować to wciągnęłam się bardzo mocno w data science. {wywiad 20}

Twitter jest platformą mało popularną w polskim świecie DS. Spotkaliśmy się z przypadkami obserwowania przez uczestników badanego świata kont twitterowych o międzynarodowej sławie, kont najbardziej znanych w badanym świecie (przykładowo na dzień 5.09.2019 z Twittera korzystają Piotr Migdał i Dominik Batorski, zaś nie korzysta Przemysław Biecek), także kont firmowych.

LinkedIn to medium społecznościowe przeznaczone do kontaktów zawodowych i kreowania profesjonalnego wizerunku.

> LinkedIn jest portalem umożliwiającym autoprezentację zawodową danej osoby i prezentację profilu zawodowego (...). W portalu można opisać swoje dotychczasowe doświadczenie zawodowe oraz przedstawić obecne stanowisko. Ważnym elementem jest krótkie podsumowanie na temat swoich zainteresowań lub dotychczasowej kariery. Portal umożliwia też dodawanie linków do prezentacji, elektronicznych książek, filmów wideo, zdjęć, blogów. Na portalu zamieszcza się też dane kontaktowe, takie jak: adres e-mailowy, telefon do pracy, linki do stron internetowych swoich i firmy, w której się pracuje. Dodatkowe informacje, o które prosi portal, to udział w: wolontariacie, projektach, kursach, oraz opis nagród i wyróżnień. LinkedIn może służyć do zdobywania referencji (...). Umożliwia również połączenie się z

> innymi osobami oraz dołączenie do różnych grup. **Im większa sieć kontaktów i liczba referencji, tym kandydat wydaje się cenniejszy dla obecnych i przyszłych pracodawców**. LinkedIn jest największym portalem biznesowym, wykorzystywanym w procesie rekrutacji (...). Korzystanie przez pracodawców z tego portalu jako narzędzia rekrutacyjnego jest bardzo popularne, ponieważ jest łatwe w użyciu i jest rozwiązaniem niedrogim (...). Uruchomiony 5 maja 2003 roku, obecnie liczy ponad 546 milionów użytkowników w ponad 200 krajach. Od 2006 roku należy do firmy Microsoft, a jego użytkowanie jest bezpłatne. Dochody osiąga z subskrypcji członków, sprzedaży reklam i rozwiązań rekrutacyjnych (...). **Portal szybko stał się popularnym narzędziem selekcji kandydatów do pracy wśród specjalistów z działów zarządzania zasobami ludzkimi (...) oraz osób poszukujących pracy. Umożliwia dotarcie do kandydatów do pracy z całego świata. Pozwala znaleźć ludzi na podstawie różnych kryteriów, na przykład definiując konkretną branżę, zawód, wielkość firmy lub grupy, w których dana osoba jest członkiem** [podkr. RŻ] (Paliszkiewicz 2018:84–85).

Jest to medium popularne w badanym świecie, a **wśród tej części jego uczestników, którzy biorą udział w konferencjach branżowych, posiadanie konta na LinkedIn jest przyjmowane za pewnik**. Stwierdzenie J. Paliszkiewicz (2018:84) „im większa sieć kontaktów i liczba referencji, tym kandydat wydaje się cenniejszy dla obecnych i przyszłych pracodawców" częściowo podzielane jest przez tę grupę uczestników:

> R: (...) Szkoda żeś wcześniej nie powiedział [że szukasz kolejnych osób do rozmowy – przyp. RŻ], bo ta konferencja u [imię prowadzącego] była dla ciebie idealna, a ta konferencja była ile, dwa tygodnie temu? [nazwa konferencji] była właśnie nastawiona na networking, przed wszystkim, nie? Że no pod kątem LinkedIn miałem plus 100 kontaktów, nawet jeśli mi się nie przydadzą, no to przynajmniej mam tych ludzi, może kiedyś akurat będzie pierwszy krok do tego, do jakichś rozmów. {wywiad 13}

Zbieranie nowych kontaktów na LinkedIn wśród osób zajmujących się DS jest raczej postrzegane jako cenne dla przyszłej kariery zawodowej, ponieważ może się wiązać z nowymi propozycjami zatrudnienia, udziału w projektach czy uzyskania zlecenia i tak samo pozyskania osób do współpracy. Na pewno jednak nie chodzi tylko o ilość kontaktów – bo popyt na pracowników z doświadczeniem w pisaniu kodu do operacji na danych jest wysoki i są oni oblegani przez rekruterów – ale o jakość kontaktów.

Jak wskazaliśmy w części 4.3., LinkedIn służył nam często do umawiania się na rozmowy z osobami poznanymi osobiście, a czasami do nawiązywania kontaktu po raz pierwszy. Kilkoro Rozmówców, w tym tacy poznani poprzez LinkedIn, wskazywało nam, że jest to prawidłowa strategia. Jak sądzimy proponowanie rozmowy do naszych celów profesjonalnych (badania do pracy doktorskiej), proponowanej z uwagi na profesjonalny obszar działania osoby (komercyjne i rzadziej akademickie DS), było postrzegane jako zgodne z charakterem LinkedIn. Pojawiły się przy tym ostrzeżenia: **to, że ktoś na LinkedIn**

tytułuje się „data scientist" nie oznacza, że przez uczestników badanego świata jest uznawany za data scientista.

> R: (...) czasami to jest tak że ktoś zawsze był statystykiem, pracował na próbach kilkuset elementowych a teraz sobie, eee, nadaje nazwę data scientista [pauza] tak się tytułuje na LinkedInie, no i uważa, że będzie robił nie wiadomo jakie projekty na złożonych danych. (...) W zasadzie można próbować patrzeć na ludzi na LinkedInie na ich doświadczenie, na ich opisy i po prostu ich łapać na LinkedInie, to jest dobra strategia, no bo teraz pytanie co by pan chciał uzyskać? Czy zobaczyć, co się dzieje na meetupach czy porozmawiać z ludźmi? Jak spotka pan człowieka na meetupie to nie wie pan kto to jest, i tak trzeba by go sprawdzić żeby on pasował do grupy, którą pan analizuje. Czy pan chce mieć doświadczonych czy mniej doświadczonych czy takich którzy się uważają za data scientistów, czy takich, aaa, którzy rzeczywiście coś takiego robią, więc łatwiej jest łapać na LinkedInie ewentualnie, aaa, łapać ich po łapać prowadzących na meetupach, no ale będzie dużo jeżdżenia. {wywiad 5}

Rozumiemy powyższą wypowiedź tak, że **zdaniem Rozmówcy czytając czyjś profil na LinkedIn można zorientować się, czy ktoś rzeczywiście zajmuje się DS, czy tylko tak się tytułuje, tylko uważa się za data scientista, ale naprawdę nim nie jest**. Rozmówca wskazuje, że osoba spotkana na meetupie dotyczącym DS może nie być kimś, kto zajmuje się DS, więc poleca sprawdzić profil takiej osoby na LinkedIn, by stwierdzić, czy pasuje ona do naszej grupy badanej. W części 6. opisujemy osoby tego rodzaju jako tzw. **wannabesów**, **czyli aspirujących do uczestnictwa w badanym świece**. Tacy ludzie mogą chętnie brać udział w meetupach, jako dość łatwo dostępnej[204] i bezkosztowej możliwości poznania osobiście osób osadzonych w badanym świecie. Inną kategorię stanowią ludzie, określani czasem jako **fake data science**, to jest tacy, **którzy sami siebie określają jako data scientist, a niekoniecznie są za takich uznawani przez uczestników badanego świata**. Kategoria fake data science jest nierzadko stosowana wobec ofert pracy, a nie wobec konkretnych ludzi. Więcej o tym powiemy w rozdziale 6., na razie zaznaczmy, że w badanym świecie **fake data science nie jest stroną w sporze o bycie prawdziwym/autentycznym uczestnikiem społecznego świata a praktyką celowego podszywania się pod uczestnika**.

Wracając do bycia obleganym przez rekruterów na LinkedIn, każda z Rozmówczyń/Rozmówców zapytanych o to odpowiadała twierdząco.

> R: (...) wydaje mi się że jest duża dynamika teraz jest **duże** zapotrzebowanie rynku na takie osoby, no wystarczy wejść tam w LinkedIna czy gdzieś albo chociaż jak ktoś ma profil na LinkedInie związany z danymi to na pewno ma mnóstwo zapytań od rekruterów, co wygląda na to że jest duże zapotrzebowanie, yyy, mam takie doświadczenie że za każdym razem jak pytam jakiegoś rekrutera jakie są widełki to każdy podaje. Od razu. Bez żadnej rozmowy wstępnej, co oznacza że ten rynek jest na tyle [ze śmiechem] na tyle po prostu **potrzebujący** żeeee [pauza] to jest rynek **pracownika**, to **naprawdę jest** rynek pracownika nie to co mówią nam w telewizji że w ogóle mamy rynek pracownika.

204 Ale tylko w dużych miastach, por. część 5.1.1.

> *B: W ogóle dla wszystkich w Polsce*
> R: Tak! Co w ogóle nie jest prawdą, bo tyle ludzi w [nazwa miasta] nie pracowałoby za 2000 na rękę, a taką mamy sytuację niestety wciąż, prawda? Gdzie koszty życia potrafią być wyższe niż niż te 2000. Także no to jedynie aspekt finansowy.
> *B: Czyli popyt jest wysoki.*
> R: Wydaje mi się, że jest bardzo wysoki.
> *B: Czy do ciebie też się odzywają rekruterzy?*
> R: Oczywiście że tak.
> *B: Próbują cię zwerbować?*
> R: Każdego tygodnia, każdego tygodnia. To jest, to też daje pewność siebie, uważam że, yy [śmiech] **łatwiej** się rozmawia w momencie kiedy ma się pracę i co tydzień ktoś dopytuje czy nie chciał byś zmienić pracy. Co prawda wydawałoby się [pauza] z punktu widzenia rekruterów bądź firmy która rekrutuje wydało by się że [pauza] może być trochę **ryzykowne** zatrudnianie ludzi bądź pytanie ludzi którzy przepracowali krócej niż **rok** w danej firmie i proszenie ich pytanie czy nie chcieli by przejść. Booo, nie wiem [pauza] wydaje mi się, że przechodzenie do firmy częściej niż raz w roku z firmy do firmy, yyy, może powodować no różne komplikacje no. {wywiad 3}

Wskazywano nam czasami, że bycie nagabywanym przez rekruterów – pojawiało się określenie „rekruterski spam" – na LinkedInie stało się dla pewnych osób z badanego świata męczące, zatem wolą się z nami kontaktować np. za pomocą emaila.

Na przykładzie cytowanej wyżej listy „wyzwań" dla data scientista widać, że na LinkedIn publikowane są różnego rodzaju posty, jak właśnie tego rodzaju żarty, anegdoty krótkie artykuły czy wypowiedzi. Stosowana jest także praktyka udostępniania informacji czy krótkiej relacji z konferencji branżowej, w której bierze się udział lub w której weźmie się udział, szczególnie w przypadku osób w roli prelegentów.

Inną praktyką jest potwierdzanie umiejętności innych osób, np. w zakresie posługiwania się językami programowania R/Python. Posiadanie wielu potwierdzeń, szczególnie jeśli chodzi o te języki programowania, jest uznawane za wskaźnik profesjonalizmu i osadzenia w badanym świecie. Dzieje się tak m.in. dlatego, że **w badanym świecie nie ma uznanych sposobów formalnego potwierdzenia umiejętności**. Więcej o tym powiemy w rozdziale 6., na razie podkreślmy, że (ograniczając się do informacji mogących być prezentowanymi na LinkedIn) potwierdzenia umiejętności od innych uczestników badanego świata – szczególnie tych doświadczonych, silnie osadzonych[205] w świecie DS – są znacznie bardziej cenione niż certyfikaty ukończenia kursów/szkoleń (w tym MOOC), a także raczej bardziej cenione niż ukończenie studiów podyplomowych i studiów wyższych.

205 Podkreślmy, że w LinkedIn taki mechanizm społeczny jest zapisany w funkcjonalności portalu, która pozwala zobaczyć, że umiejętność jednej osoby została potwierdzona przez inną osobę, która „posiada tę umiejętność na wysokim poziomie".

Co do strony **meetup.com**, to czym są meetupy przybliżyliśmy w części 5.1. Na pewno meetupów jako spotkań ani samej strony meetup.com nie należy uważać za narzędzia do współpracy, ale do komunikacji już tak. Powtórzmy, że meetupy to przede wszystkim spotkania. Komunikacja na meetup.com to w zdecydowanej większości strony prezentujące grupy meetupowe i informacje o kolejnych meetupach. Inne funkcje tego medium są mało używane. Choć istnieje możliwość komentowania konkretnych spotkań, to nie zauważyliśmy, by ten mechanizm był w badanym świecie popularny. Ewentualnie prelegenci mogą w komentarzu do swoich wystąpień przesyłać linki do używanych prezentacji, kodu (często będzie to link na GitHub), sporadycznie pojawią się lakoniczne komentarze uczestników spotkań. Istnieje także możliwość prowadzenia indywidualnych rozmów tekstowych na meetup.com, niemalże wcale nie używana. Organizatorzy meetupów przeważnie korzystają z wysyłania powiadomień i innych informacji do osób zapisanych do grup za pomocą emaili grupowych. Są wysyłane zarówno informacje o meetupach, jak i konferencjach czy hackatonach, nierzadko wraz z możliwością uzyskania zniżki na biletowane imprezy dla „członków społeczności".

Niektóre meetupy są okazją do spotykania się grup znajomych i konsultowania się, tzw. inspirowania czy dzielenia się wiedzą w sposób podobny, jak na spotkaniach w zawodowych zespołach DS. Różnicą jest to, że na meetupach spotykają się osoby z różnych firm, a także nie pracujące w komercyjnym DS, co może być w badanym świecie uznawane za jeszcze bardziej cenną okazję do dzielenia się wiedzą niż we własnym zespole, ponieważ grono jest zróżnicowane. W tym zakresie wymieniane są informacje i poglądy dotyczące rozmaitych rozwiązań technicznych i metodologicznych. Często będzie to rozmowa o zrealizowanych lub trwających projektach, a także o wykorzystywanych pakietach w Pythonie/R. Rozmówca, organizator meetupu, wskazywał, że dla niego i „stałych bywalców" **spotkania są przyjemnością i nie można ich traktować jako dokształcania się ani rozmów o pracy**. To, że rozmowy dotyczą DS, porównuje on do wspólnego zastanawiania się nad interesującą zagadką, w dziedzinie będącej hobby czy pasją dla spotykających się:

> *B: (...) jak to jest możliwe, że ludzie którzy zajmują się analizą danych, potem po pracy, w swoim wolnym czasie sami z siebie spotykają się prywatnie i dalej rozmawiają o analizie danych. Czy mógłbyś mi to wyjaśnić?*
> R: No w pewnym sensie, znaczy to już nawet analiza danych, właśnie moim zdaniem data science może być hobby, czy pasja tak? No, a jeśli mówimy o pasji no to, robisz to wolnym czasie, tylko ewentualnie coś bardziej odmóżdżającego, w sensie nie musisz rozmawiać o tym co robisz w pracy, tylko totalnie o innych działkach. Więc myślę że to tak właśnie działa, że część ludzi przychodzi bo to jest ich hobby, jakieś takie ich zainteresowanie, no bo z przyjemnością się, to trochę działa na zasadzie takich zagadek, problemów do rozwiązania,

> które często ciebie jakby meczą i chciałbyś je rozwiązać, tak? Więc ludzie przychodzą jakby, rozwijają swoje zainteresowania w tym kierunku, jest też część ludzi którzy, być może chcieliby kiedyś robić takie rzeczy, więc sobie słuchają, jakby budują sieć kontaktów [pauza] więc powiedzmy, to jest, jejku [pauza] Może tak, organizatorzy i właśnie taka 20% ludzi znajduje ten czas, ponieważ to jest jakby coś związanego z przyjemnością. To nie jest na zasadzie, że ja sobie chodzę i chcę się dokształcać, tylko to jest bardziej meetup, czyli spotkanie, pizza, jakieś tam kontakty, być może czasami inspiracje, tak, no ale to zawsze pod tym kątem. No część osób która tam przychodzi, no to są studenci, którzy po prostu jeszcze szukają odpowiedzi na pytania co mam robić po studiach, więc się jakby orientują, szukają, też jakby no, też chcą może zbudować sieć kontaktów, ale raczej tak przychodzą żeby tylko posłuchać. Więc tak na prawdę no nie za bardzo jakby rozumiem ten, w sensie że ten meetup nie należy traktować jak ekstra wykłady, tak? To jest totalnie co innego, to jest bardziej spotkanie towarzyskie, gdzie no, musi być jakaś prelekcja, żeby był powód do spotkania, może tak.
> B: *Mhm, mhm no OK.*
> R: Przynajmniej dla tych **stałych** bywalców, tak? Może na takiej zasadzie. {wywiad 13}

Meetupy mogą odbywać się w pubach, większość uczestników pije piwo podczas i po prezentacjach, nierzadko piwo i pizza po prelekcji są zakupine dla obecnych przez firmę-sponsora meetupu. Rozmowy o charakterze konsultacji, a także towarzyskie obywają się raczej po prelekcjach, czasami przed. Meetupy mogą także spotykać się w uczelnianych aulach wykładowych. Na takich spotkaniach nie spożywa się alkoholu, sponsorowana pizza i piwo może być poczęstunkiem w pobliskim pubie, po zakończeniu prelekcji. Przy takich meetupach konsultacje i rozmowy towarzyskie mają miejsce prawie wyłącznie po prelekcjach, czasem nawet bez wizyty w pubie, na korytarzu, schodach, w okolicy wejścia do auli.

Rozmówca mówił także, że część osób na meetupach:

> R: (…) jest też część ludzi którzy, być może chcieliby kiedyś robić takie rzeczy, więc sobie słuchają, jakby budują sieć kontaktów (...) to są studenci, którzy po prostu jeszcze szukają odpowiedzi na pytania co mam robić po studiach, więc się jakby orientują, szukają, też jakby no, też chcą może zbudować sieć kontaktów, ale raczej tak przychodzą żeby tylko posłuchać. {wywiad 13}

Ludzie, o których mówi Rozmówca to właśnie tzw. wannabesi. Sądzimy, że **być może już od 2018 meetupy stały się przede wszystkim miejscem spotkań wannabesów z osobami osadzonymi w badanym świecie**. Osoby osadzone w badanym świecie, zatrudnione w firmie sponsorującej meetup mogą zajmować się organizacją meetupów w ramach swoich obowiązków zawodowych, a celem głównym takiego meetupu jest budowanie wizerunku firmy-pracodawcy (*employer branding*) wśród osób zainteresowanych wejściem do świata DS i rozpoczęciem kariery zawodowej w DS (czyli przeważnie tzw. wannabesów). **Uczestnicy badanego świata raczej nie są zainteresowani prelekcjami na meetupach** (jedynie ewentualnie w roli prelegenta)**, a jeżeli już biorą udział w spotkaniach, to będą**

zainteresowani rozmowami z innymi uczestnikami świata (nie z wannabesami) **przed i po prelekcjach**. Osoby osadzone w badanym świecie może zatem przyciągnąć na meetup możliwość rozmowy z rzadko widywanymi lub nieznanymi osobiście osobami o podobnym lub głębszym stopniu osadzenia, czasami także względy towarzyskie.

Podkreślmy, że niemalże wszystkie wydarzenia – meetupy, konferencje, hackatony – w badanym świecie są sponsorowane przez firmy, które zatrudniają data scientistów. Jest to zdecydowanie dominująca w badanym świecie praktyka, choć spotkaliśmy się z opiniami, że **bycie sponsorowanym nawet przez uczelnię może być ograniczające** i dla meetupu będzie lepiej, jeżeli piwo postawi organizator z prywatnych pieniędzy (szczególnie, że posiada wysokie dochody). V. Alekseichenko promuje swoją konferencje m.in. jako nie posiadającą sponsorów podkreślając, że jego wydarzenie w odróżnieniu od większości obecnie organizowanych nie jest „festiwalem roll-upów" (Alekseichenko 2019c), czyli stojących wszędzie banerów z logami sponsorów, wszechobecnej rekrutacji i sprzedawania usług analitycznych sponsorów.

Ponadto **bycie organizatorem meetupu może być postrzegane jako cenny element budowania własnego wizerunku profesjonalnego i pozycji w świecie społecznym DS** – wcześniej pokazaliśmy, ze Rozmówca wskazuje na organizatorów meetupów jako osoby, z którymi powinniśmy porozmawiać w ramach badań etnograficznych, bowiem będą to prawdziwi data scientiści, a nie ci, którzy tylko się tak tytułują. W rozmowie nieformalnej wskazywano nam także, że podanie informacji o byciu organizatorem meetupu przez kilka lat w CV i podczas rozmowy rekrutacyjnej wywarło silne, pozytywne wrażenie na osobach rekrutujących. Świadczy to o wspomnianym wyżej postrzeganiu połączenia umiejętności technicznych z interpersonalnymi (a bycie organizatorem meetupu może być traktowane jako wskaźnik tych drugich) jako cennych i unikalnych w pracy data scientista.

Uczestnicy badanego świata śledzą strony internetowe, związane z tematyką DS, np. KDnuggets, Towards Data Science, oraz różnego rodzaju blogi, raczej anglojęzyczne, szczególnie na medium.com. Właśnie na platformie medium.com znajduje się strona Towards Data Science. W Polsce blogi związane ze światem społecznym DS to m.in. „SmarterPoland"[206] prowadzony przez Przemysława Biecka, blog Łukasza Prokulskiego[207], blog „Szychta w danych"[208] autorstwa Piotra Sobczyka, blog Mateusza Grzyba[209], blog Piotra

206 http://smarterpoland.pl/, dostęp 19.04.2019 r.

207 https://blog.prokulski.science/, dostęp 19.04.2019 r.

208 http://szychtawdanych.pl/, dostęp 19.04.2019 r.

209 http://mateuszgrzyb.pl/, dostęp 19.04.2019 r.

Migdała[210], „Big Data Passion”[211] Radosława Szmita i Marcina Wojtczaka, „R addict”[212] Marcina Kosińskiego, najnowszy z wymienionych „Uczymy maszyny”[213] Konrada Łydy, a także podkasty: „Data Science po polsku”[214] Szymona Drejewicza (publikacje w latach 2017-2018, później w większym zespole pod nazwą „Data Evangelist”, obecnie nieaktywny) oraz „Biznes Myśli”[215] Vladimira Alekseichenko. Istnieją także liczne blogi firmowe. Autor bloga „Deep Drive PL”, poświęconego częściowo pokrywającej się z data science tematyce autonomicznych samochodów, mobilnej robotyki i sieci neutonowych, sporządził spis polskich blogów o „sztucznej inteligencji i analizie danych” (Majek 2020). Karol Majek wymienił także blogi „nietechniczne, bez kodu czy przykładów”, których autorzy nie są uznawani za uczestników świata DS – jak blog sztucznainteligencja.org.pl. Niektórzy uczestnicy badanego świata mogą śledzić np. strony szerokiej branży IT jak bulldogjob.pl czy poświęcony bezpieczeństwu niebezpiecznik.pl. Trudno jest nam wskazać więcej szczegółów, co do konkretnych stron i blogów, z których korzystają uczestnicy, niemniej jesteśmy pewni, że **w badanym świecie istnieje silna presja bycia na bieżąco**. Śledzenie nowo pojawiających się pakietów i innych narzędzi, a nawet nowych metod (to w tym celu wykorzystywane jest czasem repozytorium preprintów artykułów naukowych Arxiv), generalnie śledzenie rozmaitych osób, firm, mediów i newsów związanych z danymi i nowoczesnymi technologiami w ogóle jest, naszym zdaniem, powszechną w badanym świecie strategią.

Chęć pozostawania na bieżąco wraz z efektywnością korzystania z treści to idee stojące za narzędziami takimi jak np. Feedly, których używa przynajmniej część badanego świata. Jest to agregator źródeł, bazujący na dość starej technologii czytnika RSS (*Rich Site Summary*). Użytkownik Feedly może w jednym interfejsie przeglądać najnowsze nagłówki i opisy z wielu źródeł informacji w internecie, w każdym momencie może przejść do materiału źródłowego. Użytkownik nie widzi żadnych elementów rozpraszających, tylko tekst. Można utworzyć różne kategorie treści, zachowywać treści do przeczytania później, udostępniać je przez inne media np. społecznościowe lub email. Dostęp do swojego Feedly na różnych urządzeniach, za pomocą przeglądarki internetowej lub aplikacji, ma zapewniać możliwość

210 https://p.migdal.pl/, dostęp 19.04.2019 r.
211 http://bigdatapassion.pl/, dostęp 19.04.2019 r.
212 http://r-addict.com/, dostęp 19.04.2019 r.
213 https://uczymymaszyny.pl/, dostęp 19.04.2019 r.
214 https://www.youtube.com/channel/UC7DGutbNdAjEFbOV-Al8aRA/featured, dostęp 19.04.2019 r.
215 https://biznesmysli.pl/, dostęp 19.04.2019 r.

produktywnego zapoznawania się z rozwijającymi treściami podczas przestojów, jak np. dojazd do pracy (Wray 2016:185).

Jak wspominaliśmy, **komunikacja między uczestnikami badanego odbywa się także poprzez sam kod**. Zarówno Python jak i R bywa określany, niekoniecznie przez uczestników badanego świata, jako *lingua franca* data science (ActiveState 2018; Barry 2016; Nunns 2017; Thieme 2018), zatem Python / R to także języki służące komunikacji pomiędzy ludźmi. Wyraźnym przykładem tego jest np. forum Stack Overflow, gdzie zamieszczane fragmenty kodu służą objaśnianiu zarówno pytań jak i odpowiedzi. W posługiwaniu się kodem komunikacja z komputerem jest najważniejsza – pozostając przy Stack Overflow najbardziej cenione są pytania i odpowiedzi zawierające kompletny, działający kod. Wyjątkowym narzędziem współpracy i komunikacji, zorientowanym głównie na użytkowników konkretnego języka programowania, są wielokrotnie wspominane pakiety (*packages*). Widzimy je jako wyjątkowy przypadek podkreślanego w badanym świecie dzielenia się wiedzą. Te zestawy funkcji są chętnie prezentowane w badanym świecie, szczególnie na wydarzeniach nastawionych na konkretny język programowania {obserwacja 9}. Podkreślmy, że pakiety składają się nie tylko z kodu – czytelny opis prozą, przez użytkowników R zwany winietką (*vignette*), jest uznawany za niezbędną część dobrego pakietu. Pakiety mogą być narzędziem ogólnodostępnym przez publiczne repozytoria (co omówiliśmy w części 5.3.1.), ale istnieje praktyka budowania pakietów na użytek wewnętrzny firmy, zespołu DS czy pojedynczej osoby zajmującej się DS. Takie prace mogą ewoluować w kierunku publicznych pakietów i innych narzędzi, co widać na przykładzie najbardziej znanych w badanym świecie twórców (por. część 2.1.), którzy wskazują, że początkowo budowali narzędzia dla samych siebie.

5.3.5. Przemilczane technologie

Zainspirowani słynnym pojęciem wiedzy milczącej (Polanyi [1958]2005:96) opisujemy dwie technologie społecznego świata DS: **język angielski i tzw. terminal** wraz z właściwym mu językiem programowania. Traktujemy pojęcie Polanyi'ego w sposób redukcyjny, nie pisząc tutaj o wiedzy w ogóle, a koncentrując się na tych przemilczanych technologiach badanego świata, których znajomość tak oczywista i powszechna, że przezroczysta, tzn. nie werbalizuje się jej (np. nie zapisuje w ofertach pracy, nie problematyzuje w materiałach edukacyjnych). W przypadku posługiwania się terminalem znajdujemy rzadkie przykłady problematyzowania tej technologii, zaś język angielski jest w badanym świecie technologią

krystalicznie przezroczystą. Powiemy także krótko o całym zestawie przemilczanej wiedzy i umiejętności, który składa się na coś, co nazywamy domyślną w badanym świecie biegłością technologiczną.

W dalszej części pracy (por. rozdz. 6.) powiemy więcej o przemilczanych aspektach wykonywania działania podstawowego, wiązanych przez uczestników z tym, że DS jest w ich ocenie bliskie sztuce lub eksperymentom naukowym. Nie ma zatem w badanym świecie jednoznacznych ani tym bardziej sformalizowanych procedur postępowania w projekcie, a rozmaite mikrodecyzje (np. dotyczące zmiany parametrów modelu uczenia maszynowego) są przez uczestników podejmowane na podstawie intuicji czy doświadczenia. Spotkaliśmy się w tym zakresie także z odwołaniami do kreatywności, twórczości, myślenia nieszablonowego – to uznawane jest przez uczestników badanego świata za kolejną cechę, odróżniająca ich od programistów lub nawet inżynierów w ogóle.

Język angielski jest bez wątpienia domyślnym językiem badanego świata. Przypomnijmy, że w działalności komercyjnej zdecydowanie więcej firm zajmujących się tzw. AI w Polsce pracuje dla klientów zagranicznych niż dla polskich (por. 5.1.3.), a w komunikacji z nimi prawie zawsze będzie wykorzystywany język angielski. Wydarzenia w polskim świecie DS nierzadko odbywają się głównie w języku angielskim, co jest określane jako wydarzenie otwarte na osoby nie znające polskiego (czasami z takimi osobami się spotykaliśmy). W zespołach DS w Polsce mogą pracować osoby nie znające języka polskiego, ponieważ zespół pracuje w całości po angielsku. Niemniej to tylko małe elementy wszechobecności języka angielskiego w badanym świecie.

Twierdzimy, że **nie znając angielskiego nie można wykonywać działania podstawowego badanego świata**. Na języku angielskim bazują wszystkie wykorzystywane w DS języki programowania. Bez znajomości języka angielskiego dostęp także do całej komunikacji w badanym świecie byłby niezwykle ograniczony; sądzimy, że poza granice umożliwiające bycie uczestnikiem. Dlatego traktujemy ten język jako technologię świata społecznego. Tak samo jak języki programowania jest on abstrakcją, z naszego punktu widzenia nie ma znaczenia, że jest to technologia nie-cyfrowa.

Angielski jest, nie tylko w badanym świecie, uznawany za międzynarodowy język technologii, nauki i biznesu, choć nie jest to rodzimy język dla około 95% światowej populacji (Guo 2018:1). Badania mają wskazywać, że uczenie się programowania jest trudniejsze dla osób, dla których angielski nie jest językiem rodzimym (*native*), bowiem

korzystają one głównie z anglojęzycznych materiałów i języki programowania bazują na języku angielskim (Guo 2018). Publicystyczne określenie „programowanie jest dla każdego, pod warunkiem że zna angielski" (McCulloch 2019) jest w naszej opinii trafne. Na przykładzie Pythona wskazywaliśmy opinie, że jest to język czytelny, jak *plain English* (dosł. prosty angielski, por. Dodge 2018; Garner 2018; Orsini 2014; Reitz 2018). „Czytelny" znaczy „czytelny dla osób znających angielski". Przypomnijmy przy tym, że dla twórcy Pythona, Guido van Rossum, Holendra, język angielski nie jest językiem rodzimym.

Zauważmy, że choć pewne elementy narzędzi programistycznych są tłumaczone na inne języki – tłumaczone czyli źródłowo były zapisane po angielsku – to istnieje praktyka powrotu do języka angielskiego. Przykładowo, choć RStudio jest instalowane domyślnie w języku lokalnym, to korzystanie z tego ustawienia może być uznawane za niewygodne i co za tym idzie nieefektywne, bowiem człowiek piszący kod uzyskuje komunikaty o błędach częściowo w języku lokalnym. To z kolei utrudnia szukanie pomocy względem nie zrozumiałego błędu np. na Stack Overflow. Sam Wickham doradził początkującym koderom korzystającym z RStudio:

> Jeśli Twoje komunikaty o błędach nie są po angielsku, odpal linijkę kodu 'Sys.setenv(LANGUAGE = "en")'; masz większe prawdopodobieństwo otrzymania pomocy dla komunikatów w języku angielskim (Wickham i Grolemund 2017).

W toku własnych doświadczeń z pisaniem kodu nauczyliśmy się szukać w internecie pomocy w tym zakresie tylko i wyłącznie w języku angielskim.

Mówiliśmy już o angloamerykańskim rodowodzie (por. 3.1.2.) data science (zauważmy – ten termin tytułowy i rzesza innych nie jest tłumaczona w badanym świecie), niemniej naszym zdaniem taki rodowód mają technologie komputerowe (*computing*) w ogóle. Fakt, że języki programowania bazują na języku angielskim, jak słowach „if", „then", „break", „list", „return", ma tylko kulturowe uzasadnienie. Równie dobrze, w sensie technicznym, języki programowania mogłyby działać bazując na dowolnym zestawie symboli, nawet nie będącym ludzkim językiem (McCulloch 2019). Istnieje np. język programowania Whitespace, (dosł., oczywiście po angielsku, biały znak), który składa się wyłącznie ze znaków niedrukowanych, jak znak spacji i tabulatora (ibidem). McCulloch słusznie zauważyła, że jedyne powszechnie wykorzystywane narzędzia z elementami pisania kodu, dla których istnieje akceptowana przez użytkowników praktyka tłumaczenia samego kodu na języki lokalne, są Excel i Wikipedia (ibidem). Formuły Excela są przecież tłumaczone w całości, np. funkcja „AVERAGE()" to w polskiej wersji językowej „ŚREDNIA()", choć dodajmy, że w badanym

świecie formuły arkusza kalkulacyjnego raczej nie byłby uznane za kod w ogóle. Excel jest w badanym świecie zdecydowanie uznawany za narzędzie niewłaściwe do wykonywania działania podstawowego. Podsumowując, **człowiek bez znajomości język angielskiego co najmniej na poziomie B2 nie może być uczestnikiem świata społecznego DS w Polsce**. Bez znajomości angielskiego nie będzie mógł wykonywać działania podstawowego, ani uczestniczyć w pełni w komunikacji badanego świata.

Terminal to nieco mniej niż język angielski przezroczysta dla badanego świata technologia, zatem choć nie jest całkowicie przemilczana, to **jej znajomość jest przyjmowana za pewnik i niemalże tożsama z umiejętnością posługiwania się komputerem w ogóle**. Wzmianki o niej pojawiają się sporadycznie np. w materiałach edukacyjnych związanych z DS dla osób nie mających wykształcenia informatycznego, np. studentów statystyki. W takiej właśnie książce wyjaśniono:

> Powłoka (*shell*) to program na twoim komputerze, którego zadaniem jest uruchamianie innych programów. Pseudo-synonimy to „terminal" (*terminal*), „linia poleceń" (*command line*) i „konsola" (*console*). Różnicom pomiędzy nimi poświęcony jest cały wątek na StackExchange[216], ale nie wydaje mi się, żeby był wybitnie pouczający. (...) Wielu programistów spędza dużo czasu w powłoce, w przeciwieństwie do GUI [czyli interfejsów graficznych], ponieważ powłoka jest bardzo szybka, zwięzła (*concise*) i wszechobecna w odpowiednich środowiskach komputerowych. Zanim otrzymaliśmy myszkę i GUI, w powłoce była wykonywana cała praca. Najczęściej używaną powłoką jest *bash*, którego nazwy czasami używa się jako proxy dla „powłoki", podobnie jak „Coca-cola" i „Kleenex" są proxy dla napojów typu cola i chusteczek[217] (Jenny Bryan 2018).

Określenia terminal i konsola pochodzą z czasów, gdy urządzenie wyposażone w monitor i klawiaturę do komunikacji człowieka z systemem operacyjnym – czyli konsola / terminal – było czymś odrębnym od komputera, czyli urządzenia wykonującego obliczenia, zajmującego oddzielne pomieszczenie (Bugalski 2019).

Terminal w pracy osób zajmujących się DS służy wielu celom. Przykładowo od uruchamiania skryptów wykonujących operacje na danych w językach Python / R, przez korzystanie z Git, korzystanie z technologii kontenerowych (jak Docker), do czynności, które w pracy biurowej wykonywane są za pomocą interfejsów graficznych: uruchamiania i zamykania programów, otwierania plików, tworzenia / przenoszenia / kopiowania / kasowania / nazywania / zmiany nazw plików i folderów, sprawdzania zawartości folderów / plików, instalowania / usuwania oprogramowania (w tym pakietów), aktualizowania

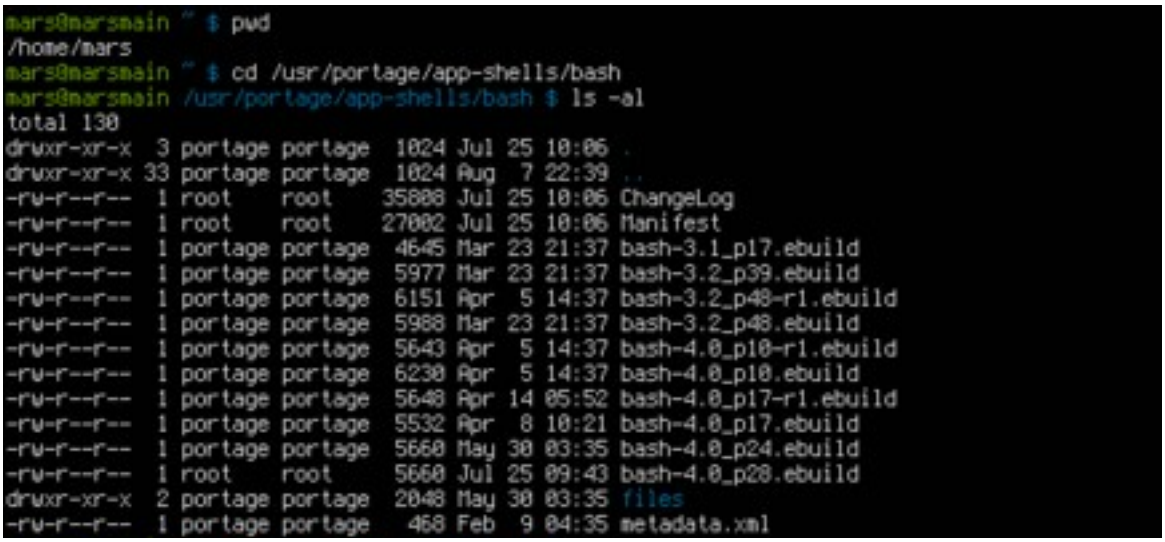
216 https://askubuntu.com/questions/506510/what-is-the-difference-between-terminal-console-shell-and-command-line, dostęp 13.10.2019 r.

217 Dla Polaków bardziej zrozumiały będzie przykład: bash to konsola, tak jak adidasy to buty sportowe.

oprogramowania i systemu operacyjnego. Terminal jest intensywnie używany niemal zawsze, kiedy praca (w sensie faktycznego użycia sprzętu i mocy obliczeniowej) wykonywana jest na maszynie innej niż komputer personalny, czyli przy pracy na lokalnym serwerze lub w tzw. chmurze. Praca na takiej maszynie wykonywana jest w ten sposób, że na komputerze personalnym uruchamia się okno z terminalem tej maszyny i całość czynności wykonuje się przez terminal.

W części 5.3.1. porównaliśmy pracę w języku programowania R z użyciem terminalu do jazdy drezyną, a korzystanie z RStudio do kierowania składem nowoczesnego ekspresu. Ogólne korzystanie z komputera za pomocą terminalu a za pomocą GUI oddaje także porównanie do gotowania z podstawowych produktów a podgrzewania gotowych potraw ze sklepu. Gotując samodzielnie (czyli korzystając z terminalu) człowiek decyduje o każdym kolejnym kroku, jest to scenariusz wymagający bardziej zaawansowanych umiejętności i narzędzi, dający relatywnie większe możliwości, swobodę, kontrolę; można przygotować większą ilość potrawy na kilka dni, przygotować półprodukty do wykorzystania na później. Odgrzewając gotową porcję (korzystając z GUI) człowiek potrzebuje mniej zaawansowanych umiejętności i narzędzi, ma relatywnie mniejsze możliwości, po prostu wygodnie korzysta z tego, co zostało przygotowane przez kogoś innego.

Interfejs terminalu to jednokolorowe okno z krótką informacją tekstową o zalogowanym do komputera użytkowniku i hoście, następnie widoczny jest znak zachęty (domyślnie w systemach Unix/Linux znak „$” dla użytkownika, „#” dla administratora) i migający kursor. Nie ma możliwości korzystania z GUI, zatem i myszki. Wygląd terminalu prezentujemy na rys. 35.



Rysunek 35: Terminal i przykładowe polecenia języka bash

Źródło: „Happy Git and GitHub for the useR” (Jenny Bryan 2018)

Widoczne na rys. 35. pierwsza komenda to „pwd”, wyświetlająca lokalizację roboczą (*print working directory*), czyli – mówiąc językiem użytkownika biurowego – folder, w którym użytkownik w tej chwili coś robi. Na konsolę wypisana jest lokalizacja „/home/mars”. Następnie poleceniem „cd” z podaną dalej ścieżką „/usr/...” wykonywana jest zmiana lokalizacji roboczej (*change directory*). Nie ma w tym przypadku wypisania na konsolę, efekt działania komendy widoczny jest w kolejnej linii jako lokalizacja („/usr/...”, kolor niebieski) przed znakiem zachęty („$”), a po informacji o użytkowniku i hoście („mars@marsmain”, kolor zielony). Ostanie polecenie „ls” to wyświetlenie listy plików w lokalizacji (*list files*) z dodanymi dwiema opcjami: „a” (wyświetlenie ukrytych plików i lokalizacji) oraz „l” (długi format opisu zawartości lokalizacji, każdy w osobnej linii), poprzedzonymi znakiem minusa (zatem całość to „ls -al”). W powyższym przykładzie zaprezentowana została interaktywna praca w języku bash za pomocą konsoli, tj. człowiek wpisuje komendę, wciska Enter i komenda jest wykonywana. Tak samo jak w Pythonie czy R, w języku bash można także zapisać skrypt. Skrypty służą głównie automatyzacji powtarzalnych czynności:

> Korzyści [pisania skryptów w bashu] są ogromne, mam nadzieje dostrzeżesz to bardzo szybko, bynajmniej postaram się Tobie to wkrótce pokazać. Jeśli **mamy pewne operacje powtarzające się i używamy konsoli do wykonywania tego zadania, to czemu nie ułatwić sobie życia i nie napisać skryptu, który zrobi to za nas**? Dajmy na to przykład następujący, w katalogu znajduje się 10000 plików (na przykład tekstowych) z rozszerzeniem. Chcemy aby każdy plik trafił do katalogu o takiej samej nazwie jak nazwa pliku, tylko bez rozszerzenia. Analizując po kolei musimy najpierw zobaczyć jak nazywa się plik, następnie utworzyć katalog o takiej nazwie bez rozszerzenia, a ostatnią czynnością jest przeniesienie danego pliku do tego katalogu. **Jak myślisz jak długo robiłbyś to sam? A jak myślisz jak szybko zrobi to komputer za nas** [podkr. RŻ]? (Pyszczuk 2011:6)

Terminal nie jest w żadnym razie narzędziem charakterystycznym wyłącznie dla DS. Niemniej twierdzimy, że jest to **narzędzie służące efektywności – kluczowej wartości społecznego data science**. Korzystają z niego różnego rodzaju ludzie związani z szeroko pojętą informatyką, a także osoby, które ogólnie i nieprecyzyjnie nazywamy zaawansowanymi użytkownikami komputera. **Używanie terminalu ma pomagać w uczynieniu pracy na komputerze szybszą, bardziej efektywną, produktywną, płynną** (Jenny Bryan 2018; Goldstein 2019; Pyszczuk 2011). Terminal, a jak sądzimy interfejsy tekstowe w ogóle, są uznawane w tej nieprecyzyjnie zdefiniowanej grupie za **narzędzia potężniejsze niż interfejsy graficzne** (GUI) (Groen 2019), dające człowiekowi możliwość bardziej bezpośredniej interakcji z komputerem i dzięki temu **większą niż przez GUI kontrolę nad komputerem** (Goldstein 2019:41). Zauważmy, że osoby już biegłe w korzystaniu z domyślnej na systemach Unix/Linux powłoki bash mogą problematyzować jej „potęgę" i efektywność, a zatem korzystać z innych powłok, np. zsh, fish lub rc. Groen pisze także o znudzeniu starym, dobrze znanym terminalem (ibidem).

Charakterystyczne dla DS narzędzia ułatwiające pisanie kodu – Jypyter Notebook, Jupyter Lab, RStudio – umożliwiają korzystanie z komend bashowych bez konieczności otwierania kolejnego okna, co jest uznawane za wygodne. W notatniku Jupyter komendy wpisuje się bezpośrednio do komórki (*cell*)[218], w pozostałych dwu wymienionych uruchamia się okno terminala otwartego w lokalizacji roboczej wewnątrz aplikacji. Te rozwiązania mają sprzyjać efektywności pracy.

Interfejs **terminalu jest odbierany przez osoby niedoświadczone w posługiwaniu się tym narzędziem jako nieprzyjazny i onieśmielający**, gdyż w drugiej dekadzie XXI w. użytkownik wykonujący na komputerze jedynie prace biurowe lub używający go do spraw prywatnych czy rozrywki ma styczność wyłącznie z GUI (Goldstein 2019). Niewątpliwie także z tego powodu **korzystanie z terminala może być uznawane za prawdziwe użytkowanie komputera**, jak napisaliśmy wyżej – wręcz tożsame z umiejętnością korzystania z komputera w ogóle – właśnie dlatego, że jest ono trudniejsze i daje więcej możliwości niż GUI. Zatem korzystanie z terminala odróżnia użytkowników biurowo-prywatno-rozrywkowych, czyli komputerowych laików od tych zaawansowanych. Uczestnicy świata społecznego data science to zdecydowanie ci drudzy. Niektórzy mogą z przyjemnością żartować na temat popkulturowych skojarzeń z siedzącym w piwnicy nerdem lub groźnym, zakapturzonym hakerem, jakie widok okna terminala budzi wśród laików:

218 Poprzedzając je znakiem „!" lub poleceniem zwanym *cell magic* „%%bash"

B: Jeszcze, jeszcze mam jedno pytanie, widzę tutaj w notesie. Yyy chyba się wcześniej wstydziłem je zadać, albo jakoś je przeoczyłem. Bo niektórzy mówią mi o tym że, znaczy no pada słowo nerd, czasami, w wywiadach
R: [śmiech]
B: Śmiejesz się, dobrze. [śmiech] I ludzie mówią, jedni mówią, ja to jestem totalnym nerdem, inni mówią, ja to nie mam nic wspólnego z nerdami, w ogóle, nie mówmy o tym, inni mówią jeszcze co innego. No to jak to jest z tobą? Jesteś [krótka pauza] nerdką? Nie wiem jak to powiedzieć.
R: [śmiech] Myślę że, że jestem, ja jeszcze zanim robiłam data science to zajmowałam się e-sportem i byłam częścią drużyny e-sportowej[219], więc poświęcałam dużo czasu na granie na komputerze, więc, tak myślę że ciągle jest coś we mnie, we mnie nerdowskiego. Staram się grać w wolnym czasie, nie mam go za dużo, więc [krótka pauza] yyy tak, gry komputerowe to trochę moje moje uzależnienie [śmiech] Oczywiście nie, nie jestem uzależniona, tylko poświęcałam dużo czasu. W moim teamie bym powiedziała że jest 50 na 50, nerdów. Część, część ludzi jest z daleka od tej kultury, a część ludzi, część ludzi jest jej bardzo blisko. I zależy też jak definiujemy nerda, czy to jest ktoś, kto jest bardziej introwertyczny, nie lubi z domu wychodzić, to tak to jestem ja [śmiech].
B: To znaczy ja nie mam jakiejś ścisłej, ścisłej definicji, bo też widzę że ludzie to różnie, różnie rozumieją. Ale czyli dla ciebie, dla ciebie to są gry tak, i raczej takie domatorstwo? Nie wiem, czy coś jeszcze?
R: Tak. Tak. Gry, domatorstwo, yyy ogólnie interesowanie się komputerami i programowaniem, znaczy no niektórzy by do tego dodali jakieś specyficzne dziedziny popkultury, ale [krótka pauza] dla mnie to to nie jest, nie jest na tyle ważne. Ja nie jestem wybredna jak chodzi o popkulturę [ze śmiechem]
B: Mhm, mhm, ok. Yyy, to jeszcze jak, jak przy tym jesteśmy to, czasami, to już powiem że bardzo rzadko, ale czasami też ktoś coś mówi o hakerach. Czy ty jesteś hakerką?
R: Yyy nie. Nie jestem hakerką.
B: To w ogóle ma coś wspólnego z data science?
R: Czasami, na przykład takie sprytne rozwiązania programistyczne, się mówi że coś jest hakiem, albo coś się rozwiązuje hakersko, jak się coś rozwiązuje w taki, eee nie wiem, sprytny sposób, jakieś łatwe obejście, to może w takim sensie tak. Ale jak chodzi, jak mówimy o hackerstwie w takim sensie, tradycyjnym czyli związanym z bezpieczeństwem, ze złamaniem zabezpieczeń, eee nie wiem, z konsultingiem dotyczącym bezpieczeństwa, to ja się, ja się tym, ja się tym nie zajmuję, nie mam nic wspólnego, eee bardziej hakowanie w sensie, yee jakieś sprytne rozwiązania, to wtedy tak.
B: Mhm, ale to rzeczywiście używasz takiego słowa, czy czy czy twoi koledzy używają takiego słowa, że to jest hak?
R: Eee, bardziej ludzie online mam wrażenie, nie słyszałam żeby ktoś w teamie mówił, ale ja tak. Ja mówię że, oo znalazłam tutaj taki **hak** że można coś zrobić szybciej, albo znalazłam, albo czy znasz jakiś hak żeby to lepiej poszło. To, ja ja tego używam, niektórzy też używaj, ale to właśnie w takim znaczeniu, eee w znaczeniu jakiegoś tricku.
B: Mhm, mhm, rozumiem. Nie bo też pytam pytam o to, bo, no yyy ludzie yyy którzy zupełnie nie mają najmniejszego pojęcia o technologiach, mówią że ktoś jest hakerem jak, no nie wiem włączył terminal tak, to już jest hakerem, yyy także
R: [śmiech] Ja często
B: Tak, tak?
R: Ja często właśnie sobie żartuje, bo mam czarne tło na terminalu i zielone litery i zawsze sobie żartuję że czuję się jak haker jak otwieram terminal.
B: Aha, taki styl styl wizualny Matrixa?
R: Tak, tak. {wywiad 20}

219 E-sport, także sport elektroniczny, oznacza rywalizacje w grach komputerowych.

Rozmówczyni identyfikuje się z nerdami, wskazując na swoje zamiłowanie do gier komputerowych i introwertyzm, natomiast z hakowaniem identyfikuje swoje działania tylko w sensie czegoś sprytnego[220], sztuczki zapisanej w kodzie, szybkiego rozwiązania. Zauważmy, że Rozmówczyni spersonalizowała wygląd swojego terminala, ustawiając czarne tło z zielonymi literami, dopasowując wygląd tego narzędzia do obrazów znanych z popkultury[221].

Nie jest jednak tak, że terminal jest w badanym świecie uznawany za jedyny właściwy sposób korzystania z komputera. Na przykładzie Gita:

> Nikt nie przyznaje odznak Git Nerd. **Korzystaj z Git w taki sposób, abyś był jak najbardziej efektywny** [podkr. RŻ]. Zachęcam do zmiany podejścia w miarę upływu czasu lub stosowania różnych podejść do różnych zadań. Nikt nie odróżni, czy korzystasz z wiersza polecenia, czy GUI, gdy przegląda historię Git lub repozytorium GitHub. Czasami spotykam ludzi, którzy uważają, że „lepiej" używać wiersza poleceń Git, ale z bardzo źle określonych powodów. Ci ludzie mogą czuć, że powinni pracować w powłoce, nawet jeśli prowadzi to do unikania Git, częstych błędów lub ograniczania się do niewielkiego zestawu ~ 3 poleceń Git. To jest przeciwproduktywne (Jenny Bryan 2018).

Unikać myszki i dążyć – nawet kosztem efektywności – do używania terminala dla wszelkich zadań mogą raczej początkujący uczestnicy badanego świata, chcący dowieść swojej biegłości technologicznej. Z drugiej strony spotykaliśmy podczas obserwacji prelegentów (niekoniecznie osoby osadzone w świecie DS), stylizujących tytuł slajdu o sobie na komendę bashową „$ whoami"[222]. Zatem na pewno manifestowanie korzystania z basha jest strategią autoprezentacji i legitymizowania siebie jako (co najmniej) zaawansowanego użytkownika komputera. Twierdzimy, że efektywność jest jednak w świecie DS wartością cenioną bardziej niż stosowanie konkretnego narzędzia. Jeżeli korzystanie z interfejsu tekstowego ogranicza efektywność a także wpływa negatywnie na jakość pracy, należałoby użyć GUI, a także zmieniać podejścia i iteracyjnie, podobnie jak same modele uczenia maszynowego, szukać najbardziej optymalnego sposobu pracy (tzw. *workflow*).

---

220 „(...) do dzisiaj obserwowaną cechą hakerów pozostaje zamiłowanie do żartów oraz samo słowo 'hack'. Pomimo dystansu czasowego i geograficznego, wciąż oznacza ono sprytne, niekonwencjonalne rozwiązanie techniczne, obejście istniejącego problemu" (Zaród 2018:20).

221 Czarne tło z zielonymi rzędami pionowych zer i jedynek, jako nawiązanie do utrwalonego w popkulturze przez filmy „Matrix" skojarzenia z hakerstwem, pojawiło się np. w grafice opublikowanej na okładce pisma oraz stronie www The Guardian, na której umieszczono zdjęcia Roberta Mercera, Nigela Farage'a i Donalda Trumpa. Był to jeden z pierwszych artykułów prasowych (Cadwalladr 2017) o tzw. aferze Cambridge Analytica, którą opisaliśmy w części 2.2.3.4.

222 Komenda pochodzi od „Who am I?" (dosł. „kim jestem?"), wypisuje ona w konsoli informację o aktualnie zalogowanym użytkowniku komputera.

W terminalu działa opisywana już funkcja autouzupełniania wpisywanego tekstu – po wpisaniu pierwszych liter i wciśnięciu Tab słowo jednoznaczne będzie uzupełnione przez komputer. Autouzupełnienie nie zadziała jednak dla nazw plików czy folderów, w których istnieją spacje. Sądzimy, że dla komputerowych laików ten potencjalny kłopot jest niedostrzegalny, ponieważ z poziomu GUI np. w Windowsie taki plik/katalog zadziała. Mogą pojawić się także inne problemy techniczne z taką nazwą pliku/katalogu, nie tylko przy korzystaniu z terminalu, zatem stosowana jest np. konwencja *snake case* „nazwa_pliku" lub *camel case* „nazwaPliku"[223]. Wśród użytkowników Linuxa pojawił się na Twitterze komentarz „przerwa (*space*) w nazwie pliku to przerwa w czyjejś duszy" (Jenny Bryan 2018). Zatem bez wątpienia zasady nazywania plików są elementem wiedzy milczącej badanego świata DS.

Podkreślmy, że z terminalu korzysta się inaczej na różnych systemach operacyjnych komputera. **Zasadnicza różnica występuje pomiędzy Windowsem a systemami Unix/Linux**. „Windows jest wyjątkowy… ale nie w dobrym tego słowa znaczeniu" (Jenny Bryan 2018), ponieważ domyślnie nie ma on powłoki bash. W domyślnej powłoce Windowsa nie ma możliwości używania np. Gita czy Anacondy, podstawowe komendy czasem różnią się[224] od bashowych, zatem czasami użytkownicy tego zdecydowanie najpopularniejszego na świecie[225] systemu operacyjnego instalują dodatkowe terminale, np. Git Bash czy będący częścią dystrybucji Anacondy dla Windowsa terminal Anaconda Prompt. Jeszcze inny terminal to opisywany już Rterm (por. 5.3.1.), kolejny może służyć korzystaniu z technologii kontenerowych, jak Docker. **Windows może być zatem uznawany za najmniej wygodny i najmniej efektywny system operacyjny do data science**, a nawet w ogóle do obliczeń naukowych i programowania:

> Windows nie jest idealną platformą do obliczeń naukowych (*scientific computing*) i tworzenia oprogramowania (*software development*). Wiele funkcjonalności będzie poszarpanych (*janky*) i uwiązanych/ograniczonych (*strapped on*). Po prostu tak jest. Istnieją nie mniej niż 4 możliwe powłoki, w których możesz skończyć. Dopóki sam nie wiesz lepiej, prawie na pewno chcesz być w powłoce Git Bash (Jenny Bryan 2018).

Nie oznacza to, że Windows uznawany jest za nie nadający się zupełnie do data science, raczej traktowany jest w badanym świecie jako nie najlepszy wybór. Część uczestników

223 Przy czym do nazw plików używane jest raczej snake_case, zaś przy nazwach zapisywanych w kodzie (np. zmienne, funkcje) trwają rozmaite dyskusje o wadach i zaletach różnych konwencji (Sharif i Maletic 2010; Wickham i Grolemund 2017).

224 Przykładowo pokazane wyżej bashowe „pwd" (rys. 35.) to w powłoce Windowsa „dir".

225 W sierpniu 2019 pod kontrolą Windowsa pracowało około 78% komputerów na całym świecie, drugi w kolejności macOS to już zaledwie 13%, zaś Linux niecałe 2% (StatCounter 2019).

świata DS korzysta z Windowsa, szczególnie na komputerach służbowych, a także po to, by uprościć sobie komunikację z klientami korzystającymi z tego systemu i aplikacji Microsoft Office. Wiele z tych osób instaluje obok Windowsa dodatkowo (tzw. *dual boot*) dystrybucję Linuxa. Sądzimy, że część uczestników badanego świata nie przykłada większej wagi do systemu operacyjnego na komputerze personalnym, część może preferować system macOC i komputery Apple, a jeszcze inni różne dystrybucje Linuxa.

Komputery Apple są to faktycznie maszyny z systemem Unixowym, zatem domyślnie można korzystać z funkcjonalności powłoki bash, a opinia o nich jako niezawodnych, intuicyjnych w obsłudze oraz drogich urządzeniach sprawia, że **mogą być w badanym świecie postrzegane jako symbol statusu**. O takim charakterze marki Apple pisano już wiele (Campbell i Pastina 2010; Krzysztofek 2011), my chcemy podkreślić, że o ile komputery z systemem macOS to około 13% globalnego rynku, to w USA jest to ponad 18%, a w Polsce niecałe 4% (StatCounter 2019). MacBooki występują w Polsce relatywnie rzadko i mogą być kojarzone z, jakkolwiek archaicznie brzmi to w 2019 roku, „Zachodem" czy „zagranicą". Zauważmy, że w Polsce zajmujący się data science pracownicy firm np. Pearson oraz iDash występując na wydarzeniach w rolach prowadzących/prelegentów korzystają z komputerów Apple {obserwacja 9, 26, 30, 38}, z takim laptopem występuje publicznie także np. Piotr Migdał, zaś z grona globalnych celebrytów badanego świata np. Hadley Wickham.

Część badanego świata może preferować pracę na komputerach personalnych pod kontrolą różnych dystrybucji GNU Linuxa, uznając maszyny z macOS za zbędny wydatek lub wyłącznie symbol statusu, zaś Windowsa za system niewygodny i nieefektywny. Takie osoby mogą korzystać z tzw. laptopów gamingowych lub serii ThinkPad firmy Lenovo lub innych tzw. biznesowych linii, ponieważ takie urządzenia mają opinię wydajnych i niezawodnych (szczególnie ThinkPady). Nie zdecydowaliśmy się opisywać tych różnych pozycji dotyczących preferowanego systemu operacyjnego ani marki urządzenia jako areny badanego świata, ponieważ nie zachodzi tutaj spór. Każdy system operacyjny czy urządzenie będzie akceptowalne, o ile sprzyja efektywności i wygodzie w wykonywaniu działania podstawowego i działań wspomagających. Zauważmy, że niemal wszystkie opisywane w części 5.3. technologie cyfrowe badanego świata działają na każdej z opisywanych platform: Windowsie, macOS i Linuxach. Twierdzimy natomiast, że **doświadczenie w korzystaniu wyłącznie z systemu operacyjnego Windows jest dla uczestników świata DS jednoznacznym wskaźnikiem bycia co najwyżej początkującym uczestnikiem**. Osoba, która nie korzystała nigdy z systemu Unix/Linux bez wątpienia nie

pracowała nigdy na serwerze ani maszynie w tzw. chmurze, a jak wskazywaliśmy w części 5.3.3. takie maszyny są powszechnie wykorzystywane, szczególnie w komercyjnym data science, pracują pod kontrolą różnych dystrybucji Linuxa i interakcja człowieka z nimi przebiega poprzez interfejs tekstowy, czyli terminal. Umiejętność korzystania z terminala i chociażby minimalna znajomość systemów operacyjnych Unix/Linux są zatem niemal tożsame i traktujemy je razem jako element wiedzy milczącej badanego świata. Umiejętność korzystania z terminala i różnych systemów operacyjnych nie jest jednak sama w sobie tym, co świadczy o byciu uczestnikiem badanego świata. Wskazuje wyłącznie na bycie zaawansowanym użytkownikiem komputera, ewentualnie na związek z szeroko pojętą informatyką.

W repertuarze wiedzy milczącej zaawansowanego użytkownika komputera jest także m.in. znajomość podstaw technologii *front end* (dosł. przedni koniec), – jak język znaczników HTML, CSS (kaskadowe arkusze stylów) oraz język JavaScript – czyli takich, które odpowiadają za wygląd i działanie stron www i innych aplikacji internetowych. Taka wiedza będzie przydatna szczególnie dla data scientistów zajmujących się tworzeniem interaktywnych wizualizacji danych dla klientów, ma także znaczenie przy zadaniach webscrapingu. Będzie to także wiedza o tym, w jaki sposób komputer zapisuje i wyświetla obraz, czyli znajomość pojęć np. kanałów RGB i przypisanej każdemu z nich wartości od 0 do 255 – to niezbędne każdemu, kto używa metod uczenia maszynowego dla danych w formie obrazów. Bez wątpienia będzie to również (silnie związany z technologiami *front end*) zestaw wiedzy i umiejętności dotyczących podstaw działania internetu, przeglądarek internetowych, wyszukiwarek internetowych, pojęć jak URL, domena, hosting itp., szczególnie do celów efektywnego wyszukiwania informacji np. na Stack Overflow, ale także przy wykonywaniu webscrapingu. Zaawansowany użytkownik posługuje się także np. pojęciem tzw. wagi plików/aplikacji, czyli ich rozmiarów wyrażonych w bajtach, rozumie różnicę rzędu wielkości między 10 KB a 10 TB. Niezwykle przydatną na etapie przygotowania danych będzie znajomość wyrażeń regularnych (*regular expressions*, regex) czyli wzorców znaków, pozwalających wyszukiwać np. słowa (a właściwie dowolne ciągi znaków) np. na podstawie tego, jak się one zaczynają, kończą, czy zawierają w sobie konkretny wzór. Przykłady można by mnożyć – jak wspominaliśmy także znajomość SQL-a może być uznawana za coś oczywistego, aż do poziomu przemilczenia.

Sądzimy, że dzieje się tak, ponieważ w naszym przekonaniu dominująca część uczestników świata DS to osoby, które już w wieku kilku-kilkunastu lat interesowały się

komputerami. Niektórzy mogą identyfikować się jako nerdzi lub geekowie, inni jako gamerzy/e-sportowcy, część jako entuzjaści nowoczesnych technologii, istnieje szerokie i nie budzące popkulturowych skojarzeń określenie *tech-savvy*, które należy rozumieć jako biegłość technologiczną[226]. Autodefiniowanie się jako osoby od dziecka zainteresowanej technologią lub matematyką opisujemy jako praktykę indywidualnej legitymizacji działania uczestników badanego świata (por. rozdział 6.). Podkreślmy także, że istnieje w badanym świecie część uczestników akcentująca swój brak zainteresowania technologiami w dzieciństwie (ale nie spotkaliśmy się z wyrażaniem braku zainteresowania lub niechęci do matematyki), którzy nabrali tejże biegłości w toku kolejnych doświadczeń naukowych lub zawodowych, związanych z badaniami ilościowymi czy pracą z danymi w ogóle. Niemniej już **jako uczestnik świata DS „trzeba" być osobą biegłą technicznie i mówiąc kolokwialnie – nie bać się technologii**. Brak strachu przez zepsuciem czy zniszczeniem systemów technologicznych, połączony z rozeznaniem w opisywanym wyżej repertuarze technologicznej wiedzy milczącej, w połączeniu ze znajomością języka angielskiego i umiejętnością korzystania z najbardziej aktualnych źródeł referencyjnych, w połączeniu z samodzielnością, poczuciem kontroli[227] nad systemami technologicznymi, a także przekonaniem, że technologie są dla człowieka zasadniczo proste składa się na to, co wyżej nazwaliśmy ową „konieczną" biegłością technologiczną uczestnika świata DS.

Podkreślamy namacalność (*palpable*) (Strauss 1978:121) pracy na komputerze, którą wykonują uczestnicy świata DS. Mamy na myśli korzystanie z klawiatury komputerowej, wciskanie klawiszy palcami. Skoro działanie podstawowe jest wykonywane niemal wyłącznie za pomocą pisania kodu, a nierzadko uczestnicy to osoby piszące z dużą prędkością czy nawet bezwzrokowo, to w ogóle **odrywanie dłoni od klawiatury postrzegać mogą one jako niewygodne i nieefektywne**, zbędny i nieergonomiczny ruch rąk. Jest to zatem kolejny argument za posługiwaniem się terminalem, ale także za posługiwaniem się **skrótami klawiaturowymi**. Nie należy nazywać tego wiedzą milczącą, ale twierdzimy, że do elementarza umiejętności uczestników świata DS należy co najmniej zestaw skrótów klawiaturowych dla systemu operacyjnego, którego się używa, oraz zestaw skrótów dla

226 Badacze społeczni od około roku 2000 używali określeń „tech-savvy", „GenTech" a także słynnego „digital natives" (tłumaczonego jako „cyfrowi tubylcy") jako nazwy pokoleń ludzi korzystających z komputerów i internetu od najmłodszych lat, nie znających świata bez tych technologii (Jain 2018; Mallan, Singh, i Giardina 2010; Pavlicek, Sudzina, i Malinova 2017). Podkreślmy, że choć niemal wszyscy uczestnicy badanego świata DS mogą być nazwani cyfrowymi tubylcami przynajmniej ze względu na swój wiek, to dla nich poziom wiedzy i umiejętności komputerowych tzw. cyfrowych tubylców jest poziomem laickim.

227 Kontrola nad maszyną była w różnych badaniach nad hakerami określana jako motywacja indywidualna hakerów (Zaród 2018:29-30, 36-37), zatem uznajemy to poczucie za immanentny składnik opisywanej przez nas biegłości technicznej, nie jako coś charakterystycznego dla data science.

najczęściej używanego IDE (np. Jupyter Notebook, PyCharm, RStudio). Klikanie myszką, by zmienić okno aktywnej aplikacji lub uruchomić linię kodu[228] będzie dla uczestników badanego świata jednoznacznym wskaźnikiem komputerowego laika, nawet do granicy kwestionowania możliwości uznania takiej osoby za uczestnika świata DS. Jako **element wiedzy milczącej traktujemy także samo korzystanie z klawiatury**, tj. szybkie na niej pisanie czy znajomość lokalizacji znaków niebiurowych, np. ~ _ ^ | `.

5.3.6. Stos technologiczny data science

Podsumowując, przyjrzyjmy się jednej z wypowiedzi osoby osadzonej w badanym świecie, która opowiada o własnych narzędziach pracy:

> *B: To chciałbym, w inną stronę teraz pójść i zapytać cię jakie masz narzędzia pracy?*
> R: Moje narzędzia pracy? Yyy ja pracuję na MacBooku Pro, większość większość mojego zespołu pracuje na MacBookach. Ee, mój język programowania wyboru to jest Python, mimo że zaczynałam w eR, przesunęłam się w stronę Pythona. Yyy pracuję w Jupyter Notebooku, zazwyczaj, bo moim zdaniem jest, jest łatwo, jest szybko, nie potrzeba dodatkowego oprogramowania. Eee co jeszcze? I zazwyczaj rozwiązania w chmurze Amazon Web Services i tak dalej. Ostatnio też zaczęłam nowe rozwiązania testować, takie na przykład jak Data bricks. Mmm, nie wiem co jeszcze, co jeszcze, jakie są jeszcze moje narzędzia pracy (…) Do komunikacji Slack i Google Hangouts. Eee [pauza] no i oczywiście najważniejsza chyba jest konsola, eee w której wszystko, wszystko robię. Jeszcze przez jakiś czas robiłam wizualizacje w Tableau, na szczęście już nie muszę [z uśmiechem].
> *B: Aha. Nie podobało ci się Tableau?*
> R: Nie jestem fanką [z uśmiechem].
> *B: Aha.*
> R: Co jeszcze? Jeszcze używam, używam programu od Cisco, do VPNów kiedy jestem, do używania VPNa[229] kiedy jestem poza biurem. Myślę że to są wszystkie moje codzienne narzędzia. Czasami muszę poszukać czegoś nowego, ale to jest taki codzienny standard, co otwieram każdego dnia i sprawdzam. (…) [Databricks] To jest [krótka pauza] rozwiązanie które, to jest środowisko które łączy Sparka z Amazon Web Services iii można bezpośrednio w nim uruchamiać kod, zarządzać swoim klastrem. Jest po prostu powiedzmy, przyjaznym miksem wielu narzędzi. Jest takim centrum dowodzenia.
> *B: Mhm, mhm, ok. A, mówiłaś o konsoli, to chodzi o wiersz polecenia, tak, w komputerze?*
> R: Tak. Tak, tak, chodzi o terminal.
> *B: O terminal? A dlaczego, bo bo, dobrze dobrze zapamiętałem że powiedziałaś, że to jest najważniejsze?*
> R: Dla mnie tak. Dla mnie to jest ważne, ponieważ yyy kiedy się łączy na przykład ze zdalną maszyną, z jakimś, jakimś komputerem w chmurze, tooo mogę to robić wyłącznie przez terminal i na przykład jak chcę odpalić na nim kod to wszystko muszę robię przez terminal, czy ściąganie danych czy na przykład wrzucanie danych do chmury, też wszystko robię przez terminal, czy jak uruchamiam jakieś skrypty to też staram się to robić przez terminal, co jest sporym ułatwieniem, eee i dlatego jest dla mnie najważniejszy. Zaraz po Pythonie chyba, ale to

228 Zmiana aktywnego okna to skrót Alt + Tab / Cmd + Tab, uruchomienie linii kodu (a także np. wysłanie emaila w aplikacjach poczty) to Ctrl + Enter.

229 VPN (*virtual private network*) czyli wirtualna sieć prywatna to technologia zwana tunelowaniem, która pozwala na bezpieczne połączenie przez internet pomiędzy dwoma użytkownikami (np. firmową bazą danych a pracownikiem pracującym zdalnie) w taki sposób, jakby połączenie odbywało się w sieci prywatnej, nie publicznej (czyli w internecie).

> [krótka pauza] jest oczywiste. (...)
> *B: Mhm, mhm, ok. Yyy aaa, a takie rzeczy jak kontrola wersji, też ci się przydaje?*
> R: Tak, używamy w firmie Gita, mmm myślę że Git jest standardem w środowisku, e iii staramy się trzymać **dobrych** praktyk używając Gita, takich typowych dla software engineering. Yee to mi się przydaje ale zazwyczaj [krótka pauza] albo dotychczas pracowałam w małych projektach, więc łatwo było, łatwo było wszystko ogarnąć. {wywiad 20}

Zwróćmy uwagę na wcześniej nie omawianą technologię Databricks, którą Rozmówczyni określa jako środowisko będące „przyjaznym miksem wielu narzędzi. Jest takim centrum dowodzenia" {wywiad 20}. Databricks to komercyjne narzędzie oferowane przez firmę Microsoft, „szybka i łatwa w obsłudze usługa analityczna do pracy zespołowej oparta na platformie Apache Spark":

> Uzyskuj szczegółowe informacje ze wszystkich swoich danych i twórz rozwiązania sztucznej inteligencji w usłudze Azure Databricks. Skonfiguruj środowisko Apache Spark™ w kilka minut, skaluj je automatycznie i pracuj nad wspólnymi projektami w interaktywnym obszarze roboczym. Usługa Azure Databricks obsługuje języki Python, Scala, R, Java i SQL, a także struktury i biblioteki nauki o danych takie jak TensorFlow, PyTorch i scikit-learn (Microsoft Azure 2019).

W badanym świecie praktyka stosowania tego rodzaju „przyjaznych miksów wielu narzędzi" jest rzadka. Należą do nich Databricks i wcześniej wspomniane DVC (narzędzie do kontroli wersji modeli i danych). Rozmówczynie mówiły o **testowaniu** tych rozwiązań, zatem to nie są rozwiązania wdrożone do systematycznej pracy polskich data scientistów. Można odnaleźć materiały porównujące zalety tego rodzaju platform, wymieniające np. Azure Databricks, Alteryx, IBM SPSS, KNIME, H2O.ai, RapidMiner, IBM Watson Studio, Dataiku Data Science Studio oraz Anaconda (Gualtieri i in. 2018; IT Central Station 2019; Muenchen 2019), jednak w świecie DS w Polsce poza opisywaną już przez nas Anacondą – która stanowi inną kategorię narzędzi niż pozostałe z wymienionych – żadne z tych narzędzi nie jest popularne, a część raczej nie byłaby uznana za narzędzia do data science[230].

Naszym zdaniem zestaw technologii używanych przez uczestników badanego świata, tak pracujących komercyjnie jak i naukowo oraz hobbystycznie, zdecydowanie bardziej niż np. uporządkowane instrumentarium chirurga przypomina **warsztat majsterkowicza**. Znajdują się w tym warsztacie narzędzia różnych firm, narzędzia stare jak i najnowsze,

230 Jak wskazywaliśmy w części 5.2 zbliżone rozwiązania to np. Orange, Caffe, WEKA, Rattle, które pozwalają wykonywać część analiz za pomocą GUI, można wskazań także na narzędzia SAP i SAS Institute. W polskim świecie data science część z nich jest postrzegana jako narzędzia analityki biznesowej tzw. BI, lub naukowej (te narzędzia to np. IBM SPSS, SAS, SAP, aplikacje Excel / Power BI firmy Microsoft), inne są używane marginalnie i nie traktuje się ich stosowania jako właściwego sposobu wykonania działania podstawowego, gdyż całość analiz nie jest lub nie musi być zapisywana w kodzie. Narzędzia takie jak Alteryx są nawet reklamowane jako narzędzia dla osób nie potrafiący programować.

niestabilne wersje testowe, nierzadko narzędzia chałupniczo przerabiane, wykorzystywane niezgodnie z pierwotnym przeznaczeniem, narzędzia jak gdyby posklejane w całość za pomocą srebrnej taśmy klejącej.

Cały zestaw technologii używany na różnych etapach i różnych poziomach projektu nazywany jest **stosem technologicznym** (*stack / technological stack* lub *full-stack*). Określenie to, używane w inżynierii oprogramowania bywa stosowane także w DS. Jak pokażemy w części 6.1.2., **stos technologiczny jest podstawowym obszarem odniesienia w praktykach zaświadczania o autentyczności** w świecie społecznym DS. W odniesieniu do używanego przy realizacji projektów stosu technologicznego oceniana jest przynależność osoby do badanego świata, bycie prawdziwym uczestnikiem świata DS, oraz oferty pracy/współpracy w różnego rodzaju firmach/zespołach/projektach data science. Technologie są także podstawą w wyznaczaniu granic społecznego świata DS oraz jego subświatów, a zatem rozeznanie w technologiach jest niezbędne do zrozumienia rozmaitych procesów zachodzących w społecznym świecie – głównie segmentacji i profesjonalizacji.

Na rysunku 36. przedstawiamy szeroki stos technologiczny data science. Omawiane w części 5.3. narzędzia pokazujemy w relacji do działania podstawowego świata społecznego DS (por. rys. 24.), nie zamieszczając wszystkich z omówionych wyżej technologii na jednym rysunku, a raczej wskazując na przypadki typowe.

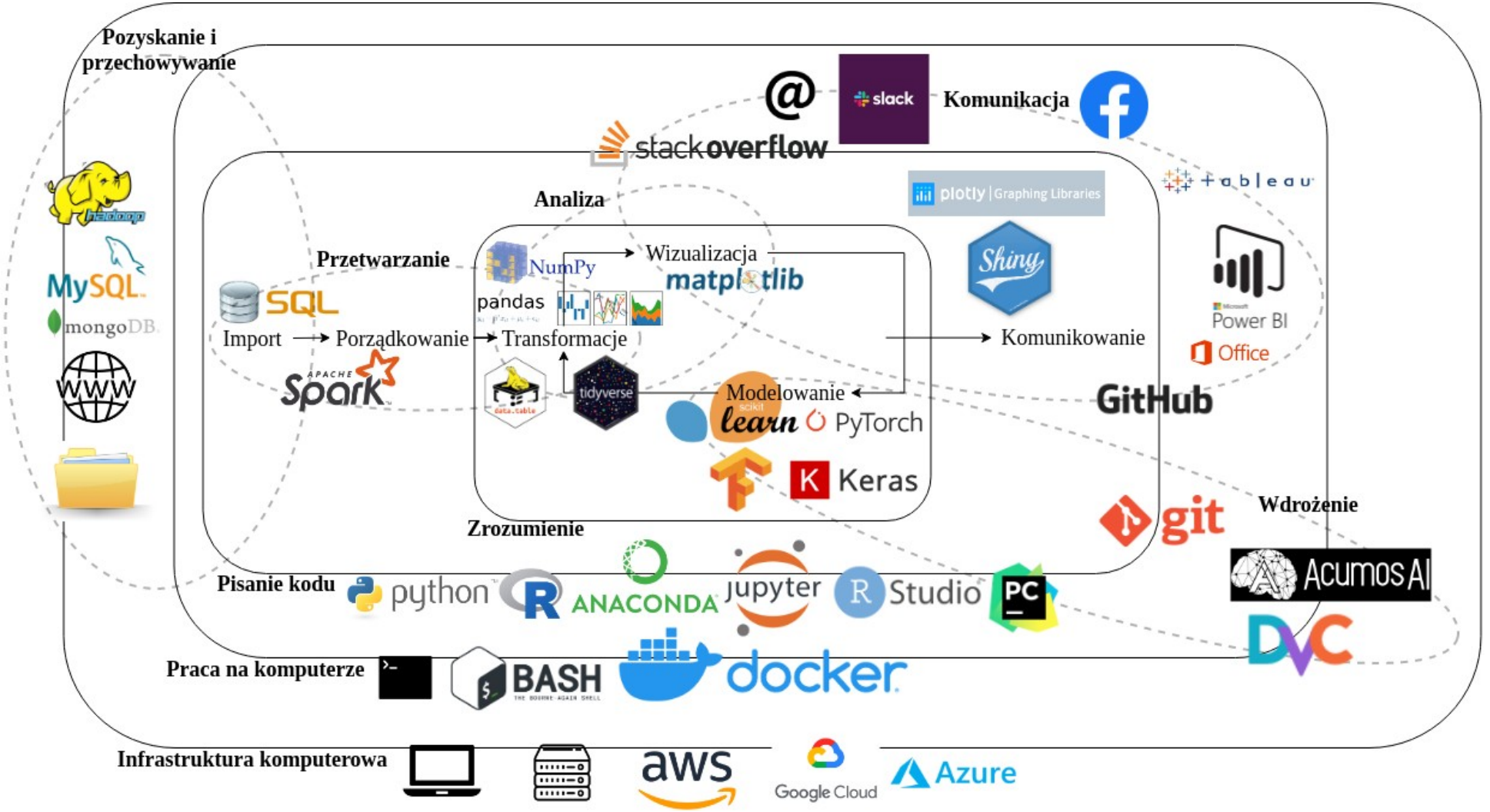
Podkreślmy, że różni uczestnicy mogą korzystać z różnych wycinków tego stosu technologicznego, szczególnie kiedy dotyczy to zakresu poza działaniem podstawowym badanego świata, czyli przechowywania i pozyskania danych, wdrożeń produkcyjnych oraz komunikacji.

Przykładowo:

- uczestnik pracujący naukowo i hobbystycznie może działając wyłącznie na własnym laptopie importować dane z plików (danych z własnych badań lub publicznych repozytoriów) oraz ściągać je z internetu metodami webscrapingowymi, używać języka R i IDE RStudio, przetwarzać, analizować i modelować dane za pomocą wybranych pakietów R (poza modelowaniem korzystając głównie z ekosystemu tidyverse), komunikować wyniki w artykułach naukowych i na meetupach, kontrolować wersje kodu przez Git i publikować je na GitHubie, korzystać ze źródeł wiedzy na Stack Overflow i komunikować się z innymi uczestnikami na Facebooku Data Science PL;

- data scientist w zespole data science w dużej firmie zajmującej się ubezpieczeniami może pracować głównie na laptopie, a czasami przy bardziej zasobochłonnych obliczeniach łączyć się z firmowym serwerem analitycznym, używać głównie języka Python w dystrybucji Anaconda, pisać kod w notatniku Jupyter lub za pomocą IDE PyCharm, importować dane z firmowych serwerów – baz danych OLAP za pomocą języka SQL i odpowiednich pakietów pozwalających wykonać to w środowisku Pythona, przetwarzać, analizować i modelować dane głównie za pomocą pakietów „NumPy", „pandas", „matplotlib" i „scikit-learn", konteneryzować konkretne części projektów za pomocą Dockera, kontrolę wersji kodu prowadzić w SVN, komunikować wyniki za pomocą narzędzi pakietu MS Office (prezentacje PowerPoint, pliki Excela) oraz budować interaktywne raporty w MS Power BI, komunikować się w zespole na Slacku, z klientami wewnątrz firmy mailowo, nie brać udziału we wdrożeniach ponieważ w tej firmie zespół data science ogranicza się do przekazywania plików .h5 z zapisem wytrenowanego modelu do działu IT, zaś po pracy zaliczać kolejne kursy MOOC i wykonywać hobbystycznie projekty z uczenia głębokiego za pomocą biblioteki TF przy użyciu chmury Google i za pomocą Gita umieszczać kod na prywatnym GitHubie;
- data scientist z kilkuletnim doświadczeniem w małej firmie/kliku firmach data science, które pracują dla różnych klientów w różnych branżach i krajach (tzw. start-upy lub firmy konsultingowe) może umieć posługiwać się każdym z narzędzi w prezentowanym stosie technologicznym (rys. 36.) i innymi technologiami, używając różnych narzędzi w różnych projektach, w zależności od specyfiki konkretnego zlecenia.

W tabeli 2. prezentujemy opis dla każdego z wyróżnionych elementów stosu w podziale na czterdzieści trzy technologie i ich role.

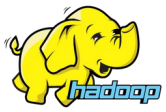


Rysunek 36: Stos technologiczny a działanie podstawowe DS

Źródło: opracowanie własne za pomocą draw.io, logotypy pobrane z oficjalnych stron www dostawców, inne rysunki na licencjach zezwalających na użycie
Uwaga: porównaj z rys. 24. oraz tab. 2.

Tabela 2: Stos technologiczny DS w podziale na technologie i ich role

Symbol	Nazwa	Opis	Rola w stosie technologicznym DS
	Komputer personalny	Laptop lub komputer stacjonarny pracujący pod kontrolą dowolnego systemu operacyjnego; własność firmy lub osoby fizycznej	Infrastruktura komputerowa
	Serwer wewnętrzny	Komputer w sieci lokalnej, bez połączenia przez internet; przeznaczony do wykonywania usług lub udostępniania zasobów dla innych komputerów w sieci, tzw. klientów; o większej niż komputer personalny mocy obliczeniowej i zasobach sprzętowych, przystosowanym do nieprzerwanej pracy; raczej własność firmy	Infrastruktura komputerowa
	Amazon Web Services	„Chmura obliczeniowa”: duże serwery zewnętrzne z dostępem przez internet; wiele klastrów komputerowych, których zasoby i moc obliczeniową lub usługi na tych klastrach wypożycza się na żądanie od właścicieli – firmy Amazon	Infrastruktura komputerowa
	Google Cloud Platform	J.w. firmy Google	Infrastruktura komputerowa
	Microsoft Azure	J.w. firmy Microsoft	Infrastruktura komputerowa
	Terminal	Program komputerowy służący korzystaniu z innych programów i systemu operacyjnego za pomocą interfejsu tekstowego (nie graficznego – GUI – jak np. w pracy biurowej); stosowane zamienne są nazwy „powłoka” (*shell*), „linia poleceń” (*command line*), „konsola” (*console*), bash	Praca na komputerze
	Bash	Domyślna powłoka (*The Bourne-Again Shell*) systemów Linux/Unix, a także język programowania skryptowego, którego się w niej używa; open source	Praca na komputerze
	Docker	Technologia konteneryzacji umożliwiająca zamrożenie całego środowiska pracy wraz z używanymi w danej chwili wersjami narzędzi do ponownego wykorzystania; bezpłatna wersja open source i płatne wersje rozszerzone rozwija firma Docker Inc	Praca na komputerze
	System plików	Ogólnie metoda przechowywania plików na nośniku (np. dysku twardym, płytach CD/DVD) bez użycia systemu bazy danych; pliki takie mogą być źródłem danych dla zastosowań DS	Pozyskiwanie i przechowywanie danych
	Internet / dane publiczne	Różne zasoby internetowe mogą być pozyskiwane do zastosowań DS za pomocą metod webscrapingowych, lub przez API; również dostęp do danych publicznych czy open source z użyciem internetu	Pozyskiwanie i przechowywanie danych

Symbol	Nazwa	Opis	Rola w stosie technologicznym DS
	MongoDB	System zarządzania nierelacyjną bazą danych w architekturze dokumentów; open source rozwijany przez firmę MongoDB, Inc. oferującej także usługi komercyjne m.in. „chmurę” obliczeniową	Pozyskiwanie i przechowywanie danych
	MySQL	System zarządzania relacyjną bazą danych do przechowywania danych tabelarycznych, także do dokumentów; bezpłatna wersja open source i płatne wersje rozszerzone rozwija firma Oracle	Pozyskiwanie i przechowywanie danych
	Hadoop	System rozpraszania składowania danych i operacji na danych na klaster komputerowy, umożliwia wykonywanie wielu operacji równolegle; także cały ekosystem technologii (np. Hadoop MapReduce + Apache Pig + Apache Hive + Apache Spark); open source rozwijane przez Apache Software Foundation	Pozyskiwanie i przechowywanie danych
	Python	Język programowania skryptowego ogólnego zastosowania; najpopularniejszy język do zapisywania działania podstawowego świata społecznego DS; open source rozwijany przez organizację non-profit Python Software Foundation	Pisanie kodu; **działanie podstawowe**
	R	Język programowania skryptowego przeznaczony do pracy z danymi (domenowy); drugi najpopularniejszy język do zapisywania działania podstawowego świata DS; open source rozwijany przez organizację non-profit The R Foundation	Pisanie kodu; **działanie podstawowe**
	Anaconda	Dystrybucja języków Python i R, menadżer pakietów i środowisk Conda, kolekcja ponad 1 500 pakietów dla DS, interfejs tekstowy i graficzny umożliwia m.in. łatwe korzystanie z notatnika Jupyter (dystrybucja głównie używana dla języka Python); bezpłatna wersja open source i płatne wersje rozszerzone rozwija firma Anaconda Inc. sponsorująca organizację non-profit NumFOCUS	Pisanie kodu; **działanie podstawowe**
	Jupyter	Interaktywne środowisko obliczeniowe w formie notatnika w przeglądarce internetowej; pozwala na edytowanie i uruchamianie dokumentów analitycznych, które są przystosowane do czytania przez ludzi; może być uruchamiany lokalnie lub w „chmurze”; wspiera około 40 języków programowania (notatnik głównie używany dla języka Python, najpopularniejszy sposób pisania kodu w tym języku w DS); open source	Pisanie kodu; **działanie podstawowe**
	RStudio	IDE (*Integrated development environment*) posiadające szereg funkcji wspomagających powtarzalne czynność; IDE używane prawie wyłącznie do języka R (użycie Pythona jest możliwe), zdecydowanie najpopularniejszy sposób pisania kodu w R (nie tylko w DS); bezpłatną wersję open source i płatne wersje rozszerzone rozwija	Pisanie kodu; **działanie podstawowe**

Symbol	Nazwa	Opis	Rola w stosie technologicznym DS
		firma RStudio sponsorująca organizacje non-profit NumFOCUS i Linux Foundation	
	PyCharm	IDE dla języka Python używane powszechnie także poza DS; bezpłatną wersję open source i płatną wersję rozszerzoną rozwija firma JetBrains	Pisanie kodu; **działanie podstawowe**
	SQL	Język programowania (*Structure Query Language*, strukturalny język zapytań), w DS używany jest komponent DML (*Data Manipulation Language*, język operacji na danych), pozwala na import danych z bazy do środowiska pracy; możliwe jest także wprowadzanie, aktualizowanie i usuwanie danych; jest przeznaczony wyłącznie do pracy z bazami danych i na zbiorach danych, nie ma żadnego innego zastosowania; różne dystrybucje open source i komercyjne	Pisanie kodu / Pozyskiwanie i przechowywanie / Przetwarzanie danych; **działanie podstawowe**
	Apache Spark	Zunifikowana technologia do rozproszonego przetwarzania danych; może wykonywać wiele rodzajów zadań np. obsługa zapytań zbliżonych do SQL-owych, uczenia maszynowego i przetwarzania danych w reprezentacji grafowej; open source rozwija Apache Software Foundation	Pozyskiwanie i przechowywanie / Przetwarzanie danych; **działanie podstawowe**
	NumPy	Pakiet Pythona do pracy z danymi w formie tabelek oraz obliczeń naukowych i inżynieryjnych; open source wspierane przez NumFOCUS	Przetwarzanie / analiza danych; **działanie podstawowe**
	pandas	Pakiet Pythona dostarczający struktur danych i narzędzi do ich analizy; open source wspierane m.in. przez NumFOCUS, RStudio, Anaconda	Przetwarzanie / analiza danych/ komunikacja ; **działanie podstawowe**
	data.table	Pakiet R o unikalnej konwencji zapisu, pozwala na wykonywanie szybkich manipulacji danymi do około 100GB; open source	Przetwarzanie / analiza danych; **działanie podstawowe**
	tidyverse	Ekosystem pakietów R o spójnej konwencji zapisu; open source wspierane przez RStudio	Przetwarzanie / analiza danych/ komunikacja ; **działanie podstawowe**
	matplotlib	Pakiet do wizualizacji danych w języku Python; open source wspierane przez NumFOCUS	Analiza danych/ komunikacja ; **działanie podstawowe**
	scikit-learn	Pakiet metod tzw. tradycyjnego uczenia maszynowego w języku Python; open source wspierane przez liczne uczelnie i firmy	Modelowanie; **działanie podstawowe**

Symbol	Nazwa	Opis	Rola w stosie technologicznym DS
	TensorFlow	Pakiet metod uczenia maszynowego w tym uczenia głębokiego głównie w języku Python; open source wspierane przez Google	Modelowanie; **działanie podstawowe**
	Keras	Nakładka wyższego poziomu dostarczająca zestaw narzędzi do budowania modeli uczenia maszynowego i uczenia głębokiego w prostszy sposób; głównie jako zaplecze Kerasa (*backend*) używany jest TensorFlow; open source wspierane przez Google	Modelowanie; **działanie podstawowe**
	PyTorch	Pakiet metod uczenia maszynowego w tym uczenia głębokiego głównie w języku Python (zapewnia korzystanie z pakietu Torch dla języka Lua); open source wspierane przez Facebook	Modelowanie; **działanie podstawowe**
	plotly	Pakiet dla Pythona i R do interaktywnych wizualizacji danych; open source i komercyjne rozwiązania firmy Plotly	Komunikacja; **działanie podstawowe**
	Shiny	Pakiet R do interaktywnych wizualizacji danych; open source i komercyjne rozwiązania firmy RStudio	Komunikacja; **działanie podstawowe**
	Tableau	Oprogramowanie zamknięte (*proprietary software*), aplikacja do interaktywnych wizualizacji danych w interfejsie graficznym; wersje darmowe i płatne firmy Tableau Software	Komunikacja
	MS Power BI	Oprogramowanie zamknięte, aplikacja do interaktywnych wizualizacji danych w interfejsie graficznym; możliwe używanie języków programowania m.in. SQL, Python, R; wersje darmowe i płatne firmy Microsoft	Komunikacja
	MS Office / MS 365	Oprogramowanie zamknięte, pakiet aplikacji do prac biurowych; do komunikacji świata DS z klientami głównie Excel (arkusze kalkulacyjne) i PowerPoint (prezentacje multimedialne); wersje darmowe i płatne firmy Microsoft	Komunikacja
	email	Poczta elektroniczna w dowolnej domenie stosowana głównie do komunikacji świata DS z klientami lub współpracownikami	Komunikacja
	Facebook	Popularny serwis społecznościowy ogólnego przeznaczenia; w DS używany m.in. do komunikacji w grupach uczestników świata DS, także przez osoby chcące wejść do świata	Komunikacja
	Slack	Aplikacja do pracy grupowej, platforma dla wysyłania wiadomości, plików i połączenia z innymi narzędziami; oprogramowanie zamknięte; wersje darmowe i płatne firmy Slack Technologies	Komunikacja
	Stack	Publiczne forum internetowe przeznaczone do	Komunikacja

Symbol	Nazwa	Opis	Rola w stosie technologicznym DS
	Overflow	zadawania pytań i udzielania odpowiedzi związanych z pisaniem kodu w różnych językach programowania (używane szeroko także poza DS)	
GitHub	GitHub	Internetowe repozytorium kodu połączone z systemem kontroli wersji Git; funkcje uznawania autorstwa i pracy grupowej, funkcje społecznościowe (np. obserwowanie, wyrażanie uznania); funkcje publikacyjne (np. licencje, opisy, strony www, prezentacje multimedialne); publiczne repozytoria darmowe, prywatne płatne; należy do firmy Microsoft	Komunikacja / pisanie kodu
git	Git	Zdecentralizowany system kontroli wersji kodu; służy do zarządzania i śledzenia ewolucji całego zbioru plików cyfrowych, które składają się na projekt; używany szeroko w tworzeniu oprogramowania (*software development*); obsługa przez interfejs tekstowy, w innych narzędziach (np. RStudio) także interfejs graficzny; open source wspierane przez non-profit Software Freedom Conservancy	Komunikacja / pisanie kodu / wdrożenie
Acumos AI	Acumos	Platforma do konteneryzacji (w połączeniu z Docker) określonych wersji modeli uczenia maszynowego wraz ze środowiskiem obliczeniowym; open source wspierane przez Linux Foundation	Wdrożenie
DVC	DVC	Platforma do kontroli wersji zbiorów danych i modeli uczenia maszynowego oraz śledzenia projektu (*Data Science Version Control*); open source wspierane przez firmę Iterative.ai	Komunikacja / wdrożenie

Źródło: opracowanie własne

Uwaga: porównaj z rys. 36.

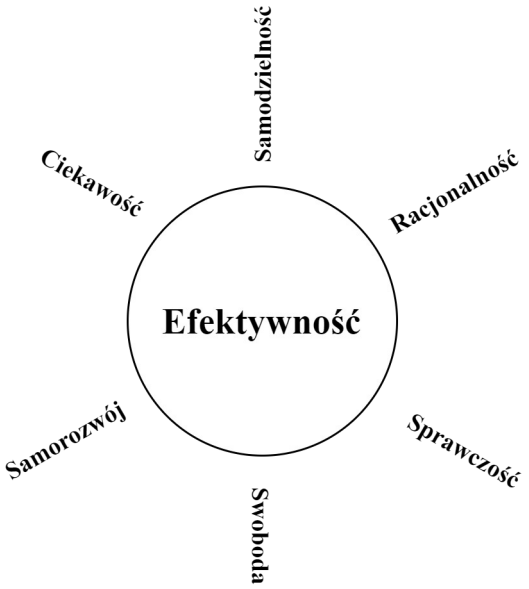
5.4. Wartości społecznego świata data science

Wartości w koncepcji społecznych światów rozumiane są inaczej niż w innych obecnych w naukach społecznych podejściach. Nie chodzi o deklarowane przez uczestników świata społecznego wartości ani o ustalanie hierarchii wartości, ale o

> wskazywane w narracjach i dyskursach społecznego świata **orientacyjne punkty pozwalające na legitymizację danego działania, a jednocześnie na dokonanie oszacowania jego autentyczności i poprawności** z punktu widzenia „interesów" danego społecznego świata lub jego segmentu. (…). **Wartości ujawniają się w działaniu**, gdy ludzie podejmują decyzje, co do swego postępowania. Ujawniają się w sposobie odgrywania ról, w konkretnych czynnościach [podkr. RŻ] (Kacperczyk 2016:44, 47).

Opisujemy siedem **wartości społecznego świata DS: efektywność, samodzielność, racjonalność, sprawczość, swobodę, samorozwój, ciekawość**. Działania w badanym

świecie są oceniane, legitymizowane, uznawane za autentyczne i poprawne zawsze w odniesieniu do centralnej kategorii efektywności. Schematycznie prezentujemy[231] to na rysunku 37.



Rysunek 37: Siedem wartości społecznego świata data science
Źródło: opracowanie własne za pomocą draw.io

Wyróżnione wartości łączą się ze sobą, są wzajemnie powiązane, a także zmieszane na rozmaite sposoby. Poniższe części mają zatem jedynie organizować nasz wywód, nie oznaczają rozłączności siedmiu opisywanych wartości, a pewne kategorie i kody będą powtarzały się, ujęte z różnej perspektywy. Pojawiające się rzadko odniesienia do literatury z dziedziny nauk społecznych są wynikiem przeglądu, który wykonaliśmy po opracowaniu opisywanych siedmiu wartości świata DS na podstawie własnych badań terenowych.

5.4.1. Efektywność

Peter Norvig, nawiązując do określenia DS jako najbardziej seksowny zawód świata wskazał, że

> [data scientist] Musi połączyć wiedzę z danego obszaru ze specjalistyczną wiedzą statystyczną i wdrożyć ją, korzystając z najnowszych konceptów informatycznych. **Ostatecznie seksowność sprowadza się do efektywności** (*In the end, sexiness comes down to being effective*) [podkr. RŻ]. W przedsiębiorstwie, które ma wiele skomplikowanych części ruchomych – wchodzących w interakcje z klientami, dostawcami, surowcami, produkcją i wszystkim innym – analityk jest w stanie poprawić efektywność o 10% lub więcej tylko poprzez manipulowanie bitami, nigdy nie dotykając atomów (Gutierrez 2014:viii–ix).

231 Inspiracją był rysunek matrycy sytuacyjnej Clarke (2005:73).

Efektywność jest naszym zdaniem centralną wartością badanego świata, w obu odcieniach znaczeniowych słowa „efektywność". W świecie DS ceniona jest zarówno efektywność (*efficiency*) jako dążenie do zwiększania wyniku (*output*) przy tym samym wkładzie na jego osiągnięcie (*input*) lub zmniejszania wkładu przy tym samym rezultacie, jak i efektywność (*effectiveness*) rozumiana jako relacja wkładu (*input*) lub wyniku (*output*) do rezultatu (*outcome*) (Mandl, Dierx, i Ilzkovitz 2008:3–4). Cenione jest takie działanie, które przynosi jak największy czy najlepszy wyniki na jednostkę wkładu, traktowanego jako koszt (to w rozumieniu *efficiency* wyrażonej równaniem: efektywność = wynik / wkład) oraz takie, którego wyniki lub koszt jest uznany za odpowiedni w relacji do założonego rezultatu działania (w rozumieniu *effectiveness*). Nie trzymamy się ściśle tego subtelnego rozróżnienia i nie spotkaliśmy się w badanym świecie z jego stosowaniem. Szczególnie w kontekście technicznych metryk oceny działania modeli uczenia maszynowego zamiast słowa efektywność może być używane określenie wydajność (*performance*), niemniej dalej piszemy po prostu o efektywności jako wartości.

Efektywność związana jest nierozerwalnie z racjonalnym podejściem do działania i nawet jeśli nie jest ona precyzyjnie skwantyfikowana, to uczestnicy badanego świata chętnie posługują się w jej określaniu pojęciami ilościowymi, jak „około trzy razy szybciej", „zrealizowaliśmy dzięki temu ponad 80% projektu". Używane są rozmaite ilościowe wskaźniki, np. w wizualizacji danych za Edwardem Tufte mówi się o relacji ilości zaprezentowanych danych do ilości użytej grafiki – im więcej tym lepiej (*data-ink ratio*, dosł. stosunek danych do tuszu) (por. Inbar, Tractinsky, i Meyer 2007; Tufte 1983) lub o współczynniku kłamstwa (*lie factor*) – dla poprawnej wizualizacji wynosi on 1, czyli stosunku informacji prezentowanej graficznie do liczb, które grafika ma przedstawiać (Biecek 2018e). Istnieje także wiele wskaźników służących do oceny modeli ML, o czym nieco dalej.

Efektywność może być mierzona i oceniona wyłącznie wtedy, kiedy uzyskano jakiś wyniki i zakładano jakiś rezultat, zatem **w badanym świecie cenione jest przede wszystkim działanie zrealizowane oraz działanie celowe**. Przykładem tego nastawienia na rezultat jest traktowane przez nas jako kod in vivo popularne w badanym świecie powiedzenie „na koniec dnia" (będące kalką z angielskiego *at the end of the day*). Tym samym efektywność jest nierozerwalnie związana ze sprawczością oraz racjonalnością. Przykładowo działanie, który postronny obserwator mógłby określić jako „bawienie się" nowymi narzędziami (np. pakietami dla Pythona), w świecie DS będzie legitymizowane jako „testowanie" nowych narzędzi, czyli jednocześnie bycie na bieżąco z pojawiającymi się nowinkami jak i ocenianie

tych nowinek w kategoriach przydatność do jakkolwiek rozumianej poprawy swojego działania.

Zauważmy, że **iteracyjne podejście** do wykonywania działania podstawowego sprzyja szybkiemu uzyskiwaniu wyników realizacji określonych jego elementów, co **postrzegane jest jako sprzyjające efektywności**, bowiem wyniki służą do podejmowania bieżących decyzji o kierunku dalszych działań. Tym samym jakoby nie traci się czasu na długotrwałe drążenie w niewłaściwym kierunku, bo już na wczesnym etapie po uzyskaniu nieobiecującego wyniku zmienia się kierunek działania. Przypomina to nastawienie na szybkie i częste porażki (*fail fast, fail often*), które jest nazywane mantrą firm technologicznych z Doliny Krzemowej. To nastawienie ma być elementem mitologii amerykańskiego ducha przedsiębiorczości, gdzie porażki poprzedzające ostateczny sukces są traktowane jako powód do dumy, jako wskazujące na ryzykowność przedsięwzięcia (czyli odwagę podmiotu podejmującego to przedsięwzięcie), jakoby nierozerwalnie związaną z wysoko cenioną innowacją. Ikonami tak pojętego sukcesu są np. Thomas Edison, Henry Ford czy Steve Jobs (Draper 2017). Co do zasady, to na wielokrotnym, iteracyjnym wykonywaniu operacji i korygowaniu błędów w celu optymalizacji wskazanego parametru bazują metody nadzorowanego uczenia maszynowego. Początkowo wydawało nam się zatem, że uczenie maszynowe jest implementacją tego amerykańskiego ducha przedsiębiorczości, a badany świat społeczny DS jest, zgodnie z owym duchem, nastawiony na osiąganie sukcesów w procesie odnoszenia szybkich i częstych porażek. Sądzimy jednak, że jest to wniosek przesadnie uproszczony. Oczywiście iteracyjna praca i poprawianie błędów to sposób działania uznawany w świecie DS za właściwy, ale ceniona jest efektywność i racjonalność, a nie ryzykowanie. **W świecie DS nie tylko nie ceni się ryzyka, ale też uczestnicy rzadko ryzykują utratę czegokolwiek innego niż czas**. Zatem choć sprawczość, szczególnie rozumiana jako wpływanie na coś realnego, praktycznego, np. koszty określonego procesu w organizacji, jest w badanym świecie wysoko ceniona, to znakomita część pracy osób zajmujących się data science polega na operowaniu na danych w cyfrowym laboratorium, gdzie ryzyka inne niż strata czasu są bardzo, bardzo małe. Takie eksperymenty mogą być kończone na różnych etapach zaawansowania, nie uzyskując żadnego wyniku praktycznego, tj. nie dochodząc do etapu wdrożenia. W laboratorium, w dużej mierze dzięki zapisywaniu operacji w kodzie, prawie wszystko jest powtarzalne, odwracalne, możliwe do zmiany. Z punktu widzenia osób zajmujących się DS i nie będących jednocześnie odpowiedzialnymi za budżet projektów (jak np. właściciele firm, kierownicy zespołów) nawet czas przeznaczony na tak określone porażki

projektów (czyli brak wdrożenia) może być postrzegany jako cenny z uwagi na indywidualny samorozwój, zdobyte doświadczenie itp., zatem nie jako koszt.

Pozostając przy równaniu „efektywność = wynik / wkład" twierdzimy, że w badanym świecie **zdecydowanie najczęściej w mianowniku jako wkład występuje czas**, przeznaczony na osiągnięcie wyniku. Nie spotkaliśmy się w badanym świecie z praktyką negatywnej oceny jakiegokolwiek działania jako wykonanego za szybko. Jeśli tylko rezultat jest oceniony jako odpowiedni, to mniej włożonego w realizację czasu może wyłącznie tę ocenę poprawić. Z uwagi na jednoznaczne przekonanie o szybkim rozwoju data science i generalnie technologii cyfrowych **istnieje bardzo duża presja do działania jak najbardziej efektywnie właśnie w kontekście poświęconego czasu** – bardzo trudno byłoby legitymizować uzyskanie dokładnie tego samego rezultatu wolniej niż to możliwe. Taka legitymizacja, by była skuteczna, mogłaby zostać sformułowana poprzez odniesienie do efektywności w konkretnej sytuacji praktycznej np. wykonałem coś wolniej za pomocą funkcji x z pakietu A, wiem o istnieniu pakietu B i że w nim inaczej zaimplementowana funkcja x działa o rząd wielkości szybciej, ale nie znam pakietu B tak dobrze jak A, zatem czas poświęcony na użycie B przeze mnie przewyższyłby czas uzyskany dzięki szybszemu działaniu funkcji x w B niż A. Czyli podsumowując – wybrałem rozwiązanie bardziej efektywne w danym kontekście.

Do **efektywności, a często bezpośrednio do szybkości, odwołuje się wartościowanie właściwie wszystkich opisywanych w części 5.3. technologii świata DS**. Dotyczy to także programowania jako takiego:

> Współcześnie programowanie umożliwia wykonywanie trylionów (10^{18}) operacji arytmetycznych na sekundę. Pozwala to na znaczne zwiększenie wydajności dostępnych rozwiązań, otwiera możliwość praktycznego wykorzystania istniejących idei, lub też tworzenia nowych pomysłów (Nowosad 2019:12).

Tak samo dotyczy to systemów bazujących na analizie danych czy uczeniu maszynowy, Już powstałe w latach 50. pierwsze komercyjne systemy oceny zdolności kredytowej klienta (*credit scoring*) – Przanowski uważa, że są one „prekursorami" współczesnych modeli w dziedzinie data science (2014: 13-15) – były budowane pod hasłem efektywności, a właściwie optymalizacji wybranych aspektów procesu oceny: „szybciej, taniej i obiektywniej" (Przanowski 2014:18).

Areny dotyczące sporów pomiędzy narzędziami konkurencyjnymi, jak przede wszystkim języki Python i R, ale także dowolnymi technologiami software'owymi czy

hardware'owymi, są zawsze – choć nie tylko – sporami o efektywność. Uczestnicy świata DS, nie tylko w Polsce, opracowują niezliczoną ilość tzw. benchmarków nastawionych na pomiar szybkości działania narzędzi (np. BrodieG 2018; Carneiro i in. 2018; Fierro, Salvaris, i Wu 2017; Gorecki i Yuan 2019; J. White 2012), istnieją także pakiety do mierzenia czasu działania kodu (Chang, Luraschi, i Mastny 2019; np. Wickham 2018c) i naszym zdaniem niezmiernie trudno byłoby znaleźć materiały, w których jest wykonana ocena narzędzia bez oceny szybkości jego działania. Jak już wskazywaliśmy na przykładzie Pythona i R w sporach o to, które narzędzie jest lepsze, uczestnicy mogą różnie definiować efektywność, zatem mogą podawać argumenty dotyczące nie tylko porównywania szybkości działania kodu dla tego samego zadania, mierzonej w sekundach, ale także np. ilości linii kodu (czas pisania), dostępności pakietów / funkcji (także czas pisania, bo szybciej skorzystać z gotowej funkcji niż pisać ją samodzielnie), czytelności kodu (łatwość i a zatem czas poprawiania błędów w kodzie, wdrożenia nowej osoby do projektu, łatwość ponownego wykorzystania kodu) czy nierzadko poczucia wygody lub łatwości pracy (pracuje mi się wygodniej w tym np. języku / pakiecie / IDE / systemie operacyjnym, zatem pracuję efektywniej). W całej części 5.3. pokazaliśmy liczne przykłady narzędzi, które poprzez ułatwienia mają czynić pracę wygodniejszą, a w konsekwencji bardziej efektywną, np. dystrybucja Anaconda, IDE RStudio, pakiet-nakładka na Tensor Flow czyli Keras. Korzystanie z terminala to przykład pozbycia się nakładki – czyli intefejsu graficznego – w celu poprawy efektywności oraz zwiększenia sprawczości i samodzielności człowieka na komputerze.

Samo uznanie, że jedynym prawidłowym sposobem wykonywania działania podstawowego jest zapisywanie operacji na danych w kodzie, jest wyrazem traktowania efektywności jako wartości w świecie DS. **Data science po prostu nie da się robić, klikając** (*you can't do data science in a GUI*) (Wickham 2018f). Znany z ilościowych badań naukowych postulat powtarzalności wyniku (*reproducibility*) realizowany jest właśnie za pomocą kodu – można go przecież uruchomić dowolną ilość razy i uzyskać ponownie wynik zapisanych czynności. Kod jest również możliwy do odczytania (*readable*), zatem przeglądając czyjś kod nawet bez uruchamiania możemy zrozumieć tok myślenia i operacji. Dzięki zapisowi w kodzie można także śledzić zmiany i porównywać kolejne wersje zapisu (*diffable*), zobaczyć proces ewolucji projektu, o czym mówiliśmy omawiając Git. Wickham wskazuje także, że klikanie w ogóle zwiększa ryzyko pomyłek w procesie operacji na danych, szczególnie pomyłek niezauważonych, a on sam używając Excela zwyczajnie boi się kliknąć złą opcję, szczególnie że niezwykle trudno potem odtworzyć to, co dokładnie się stało więc

nie wiadomo, co naprawiać (ibidem). Przy pisaniu kodu taki problem nie występuje – proces jest zapisany, człowiek widzi całą historię operacji, a nawet jeśli dochodzi do pomyłki, skasowania części kodu, nadpisania zmiennej czy zagubienia się w skrypcie, wystarczy uruchomić odpowiednie linie kodu do powtórzenia ostatniego etapu wykonanego zgodnie z zamierzeniem i sprawa rozwiązana. Zatem pisanie kodu w porównaniu do klikania sprzyja efektywności na wiele sposobów. Ogólnie rzecz ujmując, **klikanie się nie skaluje** (Mason 2010), zaś pisanie kodu tak. Co to oznacza? Jeżeli potrzebujemy np. pobrać jeden plik ze strony internetowej, prawdopodobnie wykonanie kilku kliknięć będzie szybsze niż napisanie skryptu do ściągania pliku. Jeżeli jednak chcemy pobrać sto plików i będziemy robić to codziennie przez miesiąc, klikanie będzie powtarzalnym, czasochłonnym wysiłkiem. Kiedy sytuacja zmieni się i będzie trzeba pobierać tysiąc plików pięć razy dziennie, w skrypcie wystarczy zmiana odpowiednich parametrów. Zatem dla rosnącej ilości zadań czy ilości danych kod będzie „skalować się", tzn. działać równie efektywnie (przy ograniczeniach technicznych) i nie wymagać dodatkowej pracy ludzkiej, zaś dla klikania zawsze więcej zadań będzie oznaczało więcej pracy (por. rys. 31.), a także zwiększać ryzyko ludzkich pomyłek, frustrację klikającego, jego zmęczenie itd.

Zauważmy, że **postulat powtarzalności wyniku** (*reproducibility*) **został przez świat data science zaczerpnięty z nauk przyrodniczych i przedefiniowany**. Powtarzalność nie ma zapewniać – jak w nauce – gromadzenia dowodów na wsparcie albo odrzucenie hipotezy, by wyeliminować fałszywe roszczenia do nowych odkryć i strzec dyscypliny działalności naukowej (Peng 2011b:1126). Powtarzalność w data science sprzyja efektywności (Bryan i in. 2019; Henderson i in. 2017; Wickham i Grolemund 2017). To przedefiniowanie to przykład „odgradzania się" od innych światów (Kacperczyk 2016:48), niemniej data science odgradza się od świata nauki tylko częściowo, częściowo powołuje się na naukowe korzenie odgradzając się od zwykłej działalności inżynierskiej, jaką przez uczestników świata DS przedstawiane bywa np. tworzenie oprogramowania (*software development*). Stąd już wcześnie mówiliśmy o DS jako działalności paranaukowej, czyli takiej, gdzie stosowane są metody naukowe lub zbliżone do naukowych, ale z nastawieniem na użyteczność i efektywność, a jak dalej pokażemy także sprawczość, nie aspekty poznawcze i prawdziwość. Zdaniem badającego rosyjskich data scientistów I. Lowrie, w inżynierii dominuje kryterium efektywność, zaś w nauce kryterium prawdy. Twierdzi on, że choć DS posługuje się tylko kryterium efektywności, szczególnie w kontekście oceny działania modeli uczenia maszynowego, to zmieniło ono technologiczne kryterium efektywności w standard

epistemologiczny (Lowrie 2017:3–9). Te rozważania Lowriego są słuszne w odniesieniu do odróżnienia data science od nauk przyrodniczych, co widać chociażby na opisanym wyżej rozróżnieniu w rozumieniu celu osiągania powtarzalnych wyników. Łatwo byłoby tutaj ocenić data science jako przynależne do kultury modelowania algorytmicznego (w której chodzi o wynik), a nauki przyrodnicze do kultury modelowania generatywnego (w której chodzi o wyjaśnianie) (Breiman 2001; Donoho 2015). Niemniej granica jest rozmyta, a w samym data science mogą być preferowane metody wyjaśnialne (np. regresja logistyczna, drzewa decyzyjne, różne metody modelowania statystycznego, metody optymalizacyjne znane z badań operacyjnych), w przypadku określonych problemów lub specyfiki klienta (np. branża finansowa, choć niekoniecznie działanie w niej będzie uznawane za DS). Istnieje także wiele metod służących wyjaśnianiu działania złożonych modeli ML[232], także w celu przeciwdziałania możliwym zarzutom prawnym lub etycznym o nieprzejrzyste i dyskryminujące decyzje podejmowane z ich użyciem. Niewątpliwie jednak dominującą pozycją w świecie DS jest wartościowanie efektywności (a także, jak zaraz wskażemy, sprawczości) działań w opozycji do nauki akademickiej – tak nauk przyrodniczych jak i ścisłych lub technicznych, tzn. akademickiej matematyki i informatyki. DS bywa w tym kontekście przedstawiane jako działania, które nie są sztuką dla sztuki, a rozwiązywaniem rzeczywistych problemów, dostarczaniem rzeczywistej wartości biznesowej, pracą na prawdziwy, brudnych zbiorach danych, robieniem czegoś, co naprawdę działa.

Efektywność jako centralną wartość widzimy wyraźnie w czterech używanych w badanym świecie pojęciach: **MVP, POC, „działa / nie działa" i tzw. „dowożeniu"**. Są to kody *in vivo* (Charmaz 2009:76–79).

MVP (*minimum viable product*) to produkt minimalnie gotowy do wprowadzenia na rynek, a w przypadku data science gotowy do wdrożenia produkcyjnego. Produkt zrealizowany minimalnym wkładem, ale działający, gotowy dla pierwszego klienta. Bywa używane słowo prototyp.

> Nowo powstały zespół DS za poleceniem managera zdecydował się użyć bazy NoSQL[233], pomimo że nikt w zespole nie miał doświadczenia z taką technologią. Zbudowali model, potem podjęli próbę skalowania swojej aplikacji. Jednak ponieważ próbowali skalować produkt za pomocą nieodpowiedniej do tego celu technologii [NoSQL], nigdy nie dostarczyli (*deliver*) go klientom. Kierownictwo firmy nigdy nie ujrzało zwrotów z tytułu inwestycji w ten projekt i doszło do wniosku, że inwestowanie w produktu bazujące na danych było zbyt ryzykowne i

232 Ten koncept to tzw. XAI (*eXplainable AI*), aktywny w polu jest znany w Polsce data scientist i naukowiec (matematyk) Przemysław Biecek, por. część 2.1.

233 Na razie pomijamy zjawisko fetyszyzacji modnych narzędzi – w tym przykładzie bazy NoSQL – więcej o zjawisku w części 6.

> nieprzewidywalne. Gdyby zespół DS zaczął od MVP, nie tylko mogliby zdiagnozować problemy z samym modelem, ale także przejść na tańszą, bardziej odpowiednią technologię [bazodanową] i zaoszczędzić pieniądze (Prendki 2018).

Takie MVP mogą być testowane już nie w laboratorium data science, czyli niemal zawsze na danych historycznych, ale w środowisku będącym wycinkiem środowiska produkcyjnego, tzw. piaskownicy (*sandbox*).

> R: (…) na razie jesteśmy na tym etapie właśnie testowania, rozwiązania w takiej piaskownicy analitycznej, i potem jeżeli się okaże że to działa i przynosi pożądane efekty no to będziemy to wprowadzać już tak na stałe że wszystkie sprawy będą już szły takim procesem.
> *B: OK, muszę dopytać co to znaczy że jest testowane w piaskownicy?*
> R: To znaczy że wybieramy tylko fragment yyy [krótka pauza] jakby naszych klientów, ee nie jest tak że, jeżeli tam coś się po prostu popsuje no to to nie odczujemy tego tak, to jest tylko wyrywek dostępnych naszych klientów na których możemy sobie pozwolić na pewne eksperymenty, eee i to jest tak, wtedy można powiedzieć że ewentualne ryzyko jest wpisane w koszta, czyli to nie jest coś co zaburzy no funkcjonowanie firmy i sprowadzi nas na dno jakieś finansowe, tylko to jest takie coś, że jeżeli to nie wyjdzie to co sobie policzyliśmy, albo popełniliśmy jakiś błąd, no to wtedy jakby odczujemy tylko małego pstryczka w nos, a nie że zawali się cała firma, nie? {wywiad 15}

Testy i ich kolejne iteracje wskazują na potencjalne kierunki zmian w produkcie – może dotyczyć to jakości danych, ścieżki przepływu danych (*pipeline*), przechowywania danych, przygotowania danych, etykietowania danych, mocy obliczeniowej sprzętu komputerowego, modelowania czy wdrożenia (ibidem) – a zatem testowanie MVP prowadzić ma do bardziej efektywnego wykorzystania zasobów (np. pieniędzy, ludzi, sprzętu, czasu) na poziomie zespołu data science czy całej firmy, bowiem pozwala przy relatywnie małym koszcie sprawdzić produkcyjną użyteczność wypracowywanego rozwiązania. Rozmówca {wywiad 15} akcentował to, że testy w piaskownicy to sposób na zmniejszanie konsekwencji ewentualnej porażki produktu, co jest właściwie zmniejszaniem mianownika w równaniu efektywności.

Podkreślmy, że MVP to bardzo zaawansowane stadium projektu DS. Zespół zajmujący się data science w większej organizacji może odpowiadać za monitorowanie testów MVP w piaskownicy i w porozumieniu z klientem wewnętrznym oraz ewentualnie zespołem IT ulepszać produkt w celu wykonania wdrożenia produkcyjnego. Mała firma data science, tzw. startup lub firma konsultingowa, może w zależności od konkretnego projektu w ogóle nie brać udziału w takich działaniach, zaś osoby zajmujące się DS głównie naukowo czy hobbystycznie nie mają z nimi styczności. Przypomnijmy także, że „w data science wypróbowujesz 10 pomysłów, jeden działa bardzo dobrze, a reszta – nie działa wcale (Vazquez 2017). DS to przede wszystkim eksperymenty, eksploracja, uczestniczenie w

projekcie dochodzącym do etapu MVP czy już wdrożenia nie jest w żadnym razie warunkiem bycia uznanym za uczestnika badanego świata (co widać jasno po zakresie działania podstawowego, por. 5.2.), choć udane wdrożenia w portfolio będą zdecydowanie podnosiły prestiż uczestnika świata i jego wartość na rynku pracy. W rozdziale 6. powiemy więcej o tworzących się w zespołach data science rolach Machine Learning Engineer, MLOps lub DataOps, które odpowiadają właśnie za wdrożenia i utrzymanie modeli.

POC (*Proof of Concept*) to pojęcie także czasami używane zamiennie z prototypem lub MVP, jednak prototyp i MVP są raczej rozumiane jako rozwiązania nastawione na testowanie produktu przez klienta, zaś POC to rozwiązanie przeznaczone do testów technicznych, wewnętrznych, raczej bez udziału klienta. Tym samym POC może być wcześniejszym niż MVP etapem projektu DS, który to etap można by zaklasyfikować jako komunikowanie wyników (por. rys. 24.) w celu uzyskania akceptacji osób decyzyjnych odnośnie dalszych kroków w kierunku gotowego produktu, wdrożenia produkcyjnego.

> *B: Okey, i to jest koniec projektu? W tym, na ten moment?*
> R: Nie, nie to, to miała być taka pierwsza, pierwsza faza, pierwszy POC żeby tam też chodzi o to, generalnie chodzi o to żeby te wytyczne sprzedać zespół, że jesteśmy świeżym zespołem dopiero co zbudowanym, i też chodzi o to żeby się pokazać.
> *B: Jasne.*
> R: I to chyba się udało. Z drugiej strony to był POC który robiliśmy na niewielkim obszarze w [nazwa kraju], gdyż kolejnym krokiem ma być już rozszerzenie całej analizy do całego obszaru [nazwa kraju].
> *B: Mhm, OK. A mm a czy jest jakiś plan, co ma być takim mm ostatecznym zakończeniem tego projektu? Takim ostatecznym, ostatecznym produktem?*
> R: Mhm, takim ostatecznym produktem który jest oczywiście poza nami no to jest tak naprawdę plan rozwoju sieci, czyli w jakim miejscu i jakie te maszty, jakie anteny poustawiać. {wywiad 18}
>
> R: POC - poof of concept Czyli nie zakładasz że rozwiązanie będzie produkcyjne, tylko robisz taką próbę aby sprawdzić czy się da, czy wyniki są obiecujące {konsultacja na LinkedIn z R po wywiadzie – wywiad 18}

Pojęcia MVP, POC i prototyp są zaczerpnięte z branży programistycznej (*software development*) i pomimo różnic pomiędzy nimi traktujemy je jako wskazujące na praktyki, dzięki którym projekty stają się bardziej elastyczne, co prowadzić ma do unikania sytuacji brnięcia w rozwiązanie nie przekładające się na zwrot inwestycji, czyli są to praktyki nastawione na efektywność – zmniejszanie kosztów i zwiększanie zysków.

Działa / nie działa odnoszone jest zawsze do kodu – kod musi działać. Może nie działać dokładnie w sposób zamierzony, jednak musi działać tym sensie, że zwraca jakiś wynik, daje rezultat. Poza efektywnością, która w tym wypadku oznacza generalnie

nastawienie na działanie celowe (na efekty), omawiana kategoria jest silnie związania z wartościami sprawczości i samodzielności, co rozwijamy dalej. Kod, który nie działa, tzn. nie zwraca wyniku, a tylko komunikat o błędzie, musi zostać naprawiony. Jak już wspominaliśmy, **nie ma znaczenia perfekcja**. Ma znaczenie efektywność. Także w tym kontekście uczestnicy świata DS chętnie powołują się np. na zasadę 80/20. Jedną z jej odmian jest przekonanie, że 20% czasu poświęconego na najważniejsze elementy projektu przekłada się na 80% wartości uzyskanej w projekcie, zaś pozostałe 80% czasu daje tylko 20% wartości.

> R: (…) Trzeba zacząć wtedy od tego, żeby jak najmniejszym nakładem pracy, pokazać mu jak najlepsze efekty, wtedy jest taki efekt wow, w sensie nie za, przynajmniej to jest, moje jakby, z małego doświadczenia moja obserwacja, że no najlepiej jest zaskoczyć nawet jeśli to nie jest w ogóle żadne science tak? Zrobić na początek takie rzeczy, jeśli to nie jest odgórnie zdefiniowany problem, tylko zrób coś, zrobić takie rzeczy które dają największy zysk, najmniejszym nakładem pracy, to zawsze takie się znajdują tak? żeby to był ten efekt wow, są jakieś oszczędności, pojawia się budżet, być może, być może tak, być może tak być może nie. No ale no to jest chyba lepsza droga jakby od razu tracić rok na budowanie nie wiadomo jakiego narzędzia, a potem się okaże że ono i tak jest niewiele lepsze od tego co było.
> *B: Mhm, mhm, OK*
> R: Zresztą to jest chyba taka zasada, to jest chyba zasada która działa wszędzie, nawet jak masz na przykład analizę danych, nie wiem jak ona się nazywa, ale coś na zasadzie 80% a 20. Czyli jeśli masz analizę danych to 80% czasu to jest czas, 80% czasu to jest przygotowanie danych tak, a 20% to jest ta, yyy już ta pełna analizę. Jeśli masz jakieś zadanie, to wystarczy 20% czasu żeby zrobić 80% pracy, apotem 80% czasu żeby jakby to dopieścić, tak, resztę, więc prawdopodobnie w każdym projekcie jest tak, że bardzo szybko można właśnie osiągnąć te 80% tego, a potem trzeba się zastanowić, czy warto dalej robić tak, czy tam już dalej ten koszt za bardzo nie rośnie. Wydaje mi się że to jakaś powszechna reguła jest, przynajmniej tak z doświadczenia i ze słyszenia. {wywiad 13}

To prowadzi do praktyk akceptowania niedoskonałości, legitymizowanych jako optymalizacja wkładu do rezultatu działania – jeżeli zespół DS przygotował model, który ma dopasowanie (*accuracy*) np. 90%, to decyzja o uznaniu tego modelu za wystarczająco dobry i przejście do budowania prototypu może być uznana za optymalną w tym sensie, że bardzo dużo czasu należałoby poświęcić na uzyskanie dokładności wyższej o kilka punktów procentowych, zaś w tym samym czasie można by np. przedstawić klientowi pierwszą wersję prototypu, uzyskać informację zwrotną i wykonać poprawki (które mogą dotyczyć czegoś zupełnie innego niż dokładności działania modelu, a ważnego z punktu widzenia klienta).

> Nie trzeba być super data cruncher żeby komuś pokazać jakieś rekomendacje które zmienią jego myślenie. Świat czeka na szybkie odpowiedzi, nie perfekcyjne ale pozwalające działać i zbliżyć się do celu (...).Czasem konsultujemy firmy które mają dział data science całkowicie odcięty od reszty firmy i nie daje żadnej wartości. {notatka terenowa, obserwacja 26}

Zwróćmy uwagę, że w tym kontekście część uczestników badanego świata uważa za bezwartościowe branie udziału w konkursach modelowania na platformie Kaggle. W tych

konkursach optymalizowana jest wyłącznie techniczna miara dokładności modelu, a o wygranej decydują minimalne różnice, uznawane za nieznaczące w komercyjnym DS. Osoby biorące udział w konkursach cyzelują model, poprawiając jego wyniki o ułamki punktów procentowych, czego właściwie nie praktykuje się w komercyjnym DS. Jednocześnie konkursy te polegają niemal wyłącznie na modelowaniu, bowiem po pierwsze, dane do konkursu są przygotowane[234] (zatem uznawana za żmudną i nudą, acz bardzo ważną faza pozyskania i wyczyszczenia danych jest ograniczona do minimum), po drugie, osoby budujące model, nawet jeżeli wygrają i uzyskają nagrodę pieniężną, to nie uzyskują żadnej sprawczości, bowiem oddają organizatorowi zbudowany model i nie mają żadnego związku z jego ewentualnym wdrożeniem.

> *B: Mhm, (...) a co myślisz o takich rzeczach, które się pisze, typu najbardziej seksowny zawód świata, najlepsza praca XXI wieku, ee no to, no to powiedz mi jak to jest mieć tą najbardziej seksowną pracę? Ze wszystkich.*
> R: No więc wcale nie tak seksownie jak się wszystkim wydaje. Yyy i przede wszystkim nie tak seksownie jak się to wydaje ludziom kształcącym się w tym kierunku na studiach. Co jakby widzę po yyy po juniorach przychodzących do pracy, ee jak, ee jak szybko zderzają się z rzeczywistością że te rzeczy które wydawało się że, że się robi cały czas, czyli przede wszystkim eee budowanie i uczenie modeli machine learningowych, to jest taki konik wielu osób które kończy studia. Eee podejrzewam że mogłeś słyszeć o czymś takim jak Kaggle ee czyli taka platforma na której
> *B: Tak! Na której są takie konkursy.*
> R: na której są konkursy z modelowania. Mmm to tak, to, to zauważyłem że wielu ludzi wyobraża sobie, że to jest właśnie praca data scientista, aa a to jest w najlepszym razie 20% całości, bo, bo dużo więcej czasu spędza się na takich mało seksownych zajęciach jak eeem poprawianie jakości danych, torturowanie osób z biznesu, żeby przyznały się co tak naprawdę mają na myśli oczywiście to nie to że oni nie chcą tego zrobić, tylko mmm skomunikowaniem kogoś kto myśli ymm liczbami z kimś kto nie myśli liczbami nie jest takim łatwym zadaniem. Mmm tak i te wszystkie rzeczy eee które ee są często bardzo żmudne i ee i nie polegają na tym że robi się kilka bardzo błyskotliwych rzeczy i nagle i nagle wszystkie metryki zaczynają wyglądać wspaniale, i, i można się tylko cieszyć, to, to, to, tak to jest żmudne i nie jest seksowne. Ale jednocześnie mm jakby patrząc z perspektywy tego mm kto w jakim stopniu wydaje mi się że już w znacznym stopniu i pewnie w coraz większym stopniu pewnie będzie kształtował rzeczywistość dookoła, no to się zgadzam że data science i, yyy, i sztuczna inteligencja to, to są właśnie te, te kierunki, które, które będą zmieniają i będą zmieniać świat. [pauza] Także no, w tym sensie można powiedzieć że to, to jest najbardziej seksowny. {wywiad 16}

Konkursy Kaggle mogą być postrzegane także jako coś nieciekawego, bo łatwego, znacznie bardziej oczywistego niż rozwiązywanie realnego problemu biznesowego. Z wymienionych powodów tzw. granie w Kaggle może być uznawane za stratę czasu, szczególnie przez silnie osadzonych uczestników świata DS (Brzeziński 2018).

234 Etap w projekcie DS, kiedy dane są dobrze przygotowane i można zacząć „bawić się" modelowaniem nazywany jest czasem „Kaggle moment" {obserwacja 46}.

To, że nie ma znaczenia perfekcja widzimy także na mikropoziomie działania, czyli w sposobie pisania kodu. Uczestnicy świata DS nie oceniają negatywnie praktyki interaktywnego uruchamiania kodu, poprawiania, sprawdzania dokumentacji do konkretnego pakietu czy sprawdzania komunikatu o błędzie na Stack Overflow, o ile odbywa się to szybko i skutkuje naprawieniem kodu. Mając działający kawałek kodu, ale nie do końca działający zgodnie z intencją uczestnika[235], wypada zadać pytanie na Stack Overflow, ewentualnie skonsultować się z koleżanką / kolegą. Wypada zwrócić się o pomoc z niedziałającym kodem tylko w trakcie warsztatów z kodowaniem na żywo, gdyż zgłoszenie problemów osobom prowadzącym postrzegane jest wówczas jako bardziej efektywne z punktu interesu uczestnika warsztatu niż samodzielne szukanie odpowiedzi. Niemniej jeżeli będzie to pytanie o czynność uznawaną za elementarną, a nie będącą treścią warsztatów, może być ono ocenione przez innych jako brak samodzielności a także niepotrzebnie zajmujące czas prowadzących.

W kategoriach działania /nie działania mówi się także o modelach uczenia maszynowego i innych metodach analitycznych i statystycznych. I znów muszą one w ogóle zadziałać – to jest tożsame z działaniem kodu, za pomocą którego je zapisano – ale też powinny działać dobrze. Sądzimy, że dobre działanie jest w tym przypadku tożsame z uzyskiwaniem wyniku bliskiego zaplanowanemu rezultatowi. Istnieją jednak różne podejścia do definiowania tego, co jest rezultatem. W kontekście modeli ML istnieje szereg ilościowych metryk ewaluacji modeli – w nadzorowanym uczeniu maszynowym najprostszą będzie dopasowanie odpowiedzi modelu do wartości zmiennej zależnej (*accuracy*), wyrażone jako stosunek wszystkich odpowiedzi do odpowiedzi poprawnych (czyli zgodnych z wartościami zmiennej zależnej), np. 0,9 lub 90% – i część uczestników badanego świata będzie koncentrować się właśnie na nich, uznając za pożądany rezultat np. accuracy wyższe od ustalonej wartości. W zadaniach klasyfikacji stosowane są także bardziej szczegółowe miary, pozwalające ocenić różnego rodzaju niezgodności odpowiedzi modelu z wartościami zmiennej zależnej[236]. Cześć uczestników będzie nazywała takie podejście najprostszym, a jednocześnie nieużytecznym i będzie oceniała dobroć działania modelu także w kategoriach użyteczności odpowiedzi modelu dla rezultatu w kategoriach osiągnięcia celu o charakterze

235 To może także wskazywać na wykrycie błędu w cudzym kodzie, najczęściej będzie to autor pakietu. Wówczas, gdy osoba upewni się co do powtarzalności błędu i przygotuje działający przykład, wypada zgłosić błąd np. jako issue na GitHubie. Takie działanie może być elementem budowania swojej reputacji jako osoby potrafiącej programować.

236 Typowe miary dla klasyfikacji binarnej to specyficzność (*specificity*) precyzja (*precission*) oraz czułość (*sensitivity* lub *recall*), dla klasyfikacji binarnej i wieloklasowej stosuje się także metrykę F-1 (w podstawowej postaci jest ona średnią harmoniczną precyzji i czułości). Dla różnych modeli ML stosowana jest także krzywa ROC (*Receiver Operating Characteristics*) i obliczana na jej podstawie powierzchnia pod krzywą (*area under the curve*, AUC) (por. Powers 2007).

biznesowym (np. zmniejszenia czasu odpowiedzi na reklamację klienta, zwiększenia ilości płatnych subskrybcji wśród użytkowników w serwisie streamingowym), oraz w kontekście całego projektu. Te dwie skrajne strategie definiowania rezultatów działania modeli to przeciwstawne pozycje na arenie, dotyczącej tego, czy uczestnicy realizują działanie podstawowe żeby rozwiązać problem, czy zrobić model? W części 6.1. przyglądamy się zagadnieniu bliżej.

Dobrze działająca wizualizacja lub inna metoda analizy danych to taka, która pozwala na wgląd w dane, pokazuje zależności pomiędzy zmiennymi, ewentualnie komunikuje coś w sposób zrozumiały, pozwala na skonsumowanie wyniku przez odbiorcę. Część uczestników świata DS zwraca dużą uwagę na fizjologiczne i psychologiczne mechanizmy percepcji informacji podanych w formie graficznej, część stosuje różne, także zaawansowane metody statystycznego testowania hipotez, wielu jest na bieżąco z najnowszymi publikacjami (nierzadko preprintami) w dziedzinie uczenia maszynowego. Dobrze działające są zatem metody nie tylko szybkie lub niekoniecznie szybkie, ale odpowiednio dopasowane do problemu, w sensie biznesowym i metodologicznym.

Podkreślmy raz jeszcze, że w badanym świecie nie ma znaczenia perfekcja. Popełnianie i naprawianie błędów, szczególnie technicznych[237], ale także metodologicznych, jest uznawane za nieusuwalny element wykonywania działania podstawowego. Iteracyjna praca metodą prób i błędów będzie uznawana za normalną dopiero od poziomu umiejętności uznawanych za elementarne – wielokrotne mylenie się przy napisaniu jednej linii kodu przy podstawowych funkcjach z pakietów np. NumPy, pandas, czy dplyr będzie jasnym sygnałem posiadania niskich kwalifikacji, podobnie kłopoty z wykonaniem modelu regresji logistycznej czy drzewa decyzyjnego. Zauważmy przy tym, że w badaniach psychologicznych i pedagogicznych popełnianie błędów, doświadczanie tzw. konstruktywnych porażek (p*roductive failure*) wskazywano jako to, co sprzyja nauce konceptów matematycznych i rozwiązywania problemów matematycznych (Kapur i Rummel 2012). Naszym zdaniem można to odnieść nie tylko do efektywności, ale także do innych wartości świata DS – samodzielności, sprawczości i swobody. W badaniach wskazywano, iż poczucie lęku (*anxiety*) przed matematyką u nastolatków jest związane z poczuciem bezsilności wobec rozwiązania problemu matematycznego (OECD 2015), zaś to poczucie lęku może być

237 Przykładowo pojawiają się żarty, w których według szablonów charakterystycznych okładek znanego w data science wydawnictwa O'Relly przygotowano fikcyjne tytuły techniczny podręczników, m.in.: „kopiowanie i wklejanie ze Stack Overflow", „googlowanie komunikatów o błędach", „próbowanie rzeczy, aż zadziałają", „pisanie kodu, którego nikt nie będzie mógł przeczytać", „zmienianie rzeczy i patrzenie, co się stanie".

powodowane m.in. przez nadmierny nacisk na prawidłową odpowiedź i zmuszanie do stosowania określonych procedur rozwiązywania zadań przez nauczycieli (Harper i Daane 1998). Także w świecie open source przyzwolenie na popełnianie i poprawianie błędów jest traktowane jako to, co sprzyja poczuciu sprawczości, a w konsekwencji osiąganiu efektów. Twórca jądra (*kernel*) Linuxa, Linus Torvalds wspominając wydarzenia sprzed 25 lat wskazywał, że

> Ze złymi decyzjami technicznymi jest tak, że zawsze można je cofnąć. Może to być bardzo frustrujące i oczywiście jest cały zmarnowany czas i wysiłek, ale jednocześnie to nie jest tylko marnowanie: był jakiś powód, dla którego źle skręciłeś i zdałeś sobie sprawę, że to było złe i nauczyło cię czegoś. Nie twierdzę, że to naprawdę dobra rzecz - oczywiście lepiej zawsze podejmować właściwą decyzję za każdym razem - ale jednocześnie nie martwię się szczególnie, kiedy dokonuję wyboru. Wolę podjąć decyzję, która okaże się później błędna niż zbyt długo ględzić w sprawie możliwych alternatyw (Cass 2016).

Zatem uczestnik, którego działanie będzie pozytywnie oceniane przez innych uczestników badanego świata w omawianych tutaj aspektach, podejmuje samodzielnie próby rozwiązywania napotykanych problemów, działa bez strachu przed popełnieniem błędu i z przekonaniem o własnym sprawstwie czyli tym, że jest w stanie poradzić sobie z rozwiązaniem problemu, samodzielnie w swobodny sposób wybiera sposoby podejścia do problemów oraz działa efektywnie, optymalizując różne aspekty wkładu pracy w porównaniu do jej efektu. Wydaje nam się zatem, że w badanym świecie ceniony jest taki sposób działania, który może być charakterystyczny dla osób określanych jako posiadające zdolności matematyczne lub ogólnie umysły ścisłe. Później pokażemy (w rozdziale 6.) jedną ze strategii legitymizowania własnego działania w świecie DS – deklarowanie tego, że jest się zainteresowanym lub uzdolnionym w dziedzinach ścisłych, matematycznych, technicznych już od dzieciństwa.

Nie jest więc tak, że data scientiści zwracają uwagę wyłącznie na szybkość:

> Istnieje jednak kilka reguł, z którymi zgadza się większość programistów. Pierwsza z nich mówi, że **wolny działający kod jest lepszy niż szybki niedziałający kod** {podkr. RŻ} (Nowosad 2019:15).

Uczestnicy badanego świata raczej **postrzegają swoje działanie w kontekście jak najbardziej efektywnej optymalizacji** jakości i szybkości pracy oraz **maksymalizowania wybranego aspektu**.

W świecie DS tak ważna jest efektywność, ale nikt nie uważa, że im więcej zrobi modeli dziennie czy im więcej linijek napisze kodu, tym lepszym jest data scientistem. także

to nie jest działanie „na ilość" w takim sensie. Szybkość jest bardzo ważna, ale nikt nie liczy efektowności czasu pracy tak, jak dla pracownika zbierającego truskawki. Data scientiści postrzegają swoją pracę jako twórczą, techniczną i naukową, zatem chodzi o minimalizowanie czasu, ale nie ilość tylko jakość rozwiązanych problemów. Sądzimy, że uczestnicy badanego świata spotykając się z zadaniem nastawionym na ilość problemów dążyliby do zautomatyzowania ich obsługi. Wspominane wyżej optymalizowanie widzimy jako dążenie, by **rozwiązać problem najlepiej jak to możliwe w jak najkrótszym czasie**. Zatem pomimo nastawienia świata DS na szybką, iteracyjną, nie doskonałą, efektywną pracę, jakość pracy – szczególnie w sensie metodologicznym – także jest dla uczestników kategorią odniesienia normatywnego. Raczej nie identyfikują się oni z tym, co nazywano kulturą firm technologicznych wyrażającą się w hasłach: „wdrażaj i poprawiaj później" (*launch and iterate*), oraz „działaj szybko i psuj" (*move fast and break things*) (Crawford et al. 2018:12). Bardziej może z akcentującym iteracyjne optymalizowanie hasłem „wydawaj wcześnie, wydawaj często" (*release early, release often*), pochodzącym z ruchu open source (Raymond 2001). W naszej opinii świat DS ceni racjonalność, scjentystyczne podejście do realizowanych działań, a także ciekawość realizowanych projektów, raczej nie dąży do szybkich wdrożeń kosztem tych wartości. Tym samym wartości nie stanowi szybkość sama w sobie. Stanowi ją efektywność, rozumiana jako optymalizacja.

W kontekście optymalizacji uczestnicy badanego świata DS chętnie powołują się na *no free lunch theorem* (dosł. twierdzenie o tym, że **darmowe obiady nie istnieją**). Ogólnie rzecz ujmując, to twierdzenie wskazuje, że nie ma zysku bez kosztu, zatem maksymalizowanie jednego wskaźnika odbywa się przez obniżanie innych wskaźników i nie ma rozwiązań po prostu najlepszych (Wolpert i Macready 1997) – obiad nigdy nie jest za darmo, ktoś / gdzieś / kiedyś za niego zapłacił. W odniesieniu do tego twierdzenia wskazuje się, nawet w materiałach do nauki podstaw uczenia maszynowego, że

> każdy algorytm klasyfikacji zawiera integralne założenia i żaden z modeli klasyfikacji nie przeważa nad innymi, jeżeli nie opracujemy założeń dotyczących danego zadania. W praktyce niezbędne okazuje się porównywanie przynajmniej kilku różnych algorytmów w celu wytrenowania i doboru najbardziej skutecznego modelu. Zanim jednak będziemy w stanie porównać różne modele, musimy najpierw ustalić metrykę służącą do pomiaru wydajności. Jedną z najbardziej popularnych metryk jest dokładność klasyfikacji, którą definiujemy jako stosunek poprawnie sklasyfikowanych wystąpień do nieprawidłowo określonych instancji (Raschka 2018:35).

Tego rodzaju racjonalne, a najchętniej także kwantyfikowalne podejście do optymalizacji uczestnicy badanego świata DS stosują do wielu aspektów swojego działania –

od oceny modeli ML, przez różne kwestie dotyczące używanych w projekcie narzędzi, przez kod i pisanie kodu, po np. skład zespołu projektowego do ogólnego rezultatu projektu. W rozmowach nieformalnych z uczestnikami świata DS spotykaliśmy się z stosowaniem kategorii optymalizacji, relacji wkładu do wyniku, ogólnie efektywności także w odniesieniu do przeróżnych sfer życia nie związanych z działaniem podstawowym ani w ogóle z data science, m.in. w kontekście wyboru środka transportu, zmiany pracy, przeprowadzki, wyboru studiów. W stosunku do pisania kodu za kolejny przykład nastawienia na efektywność może służyć postulat przygotowywania kodu możliwego do powtórnego użycia (*reusable*) – kod ma być modularny (*modular*, składać się z niewielkich, niezależnych części, co realizowane jest głównie za pomocą funkcji), poprawny (*correct*, robić to, co piszący ma na myśli), czytelny (*readable*, łatwo innej osobie czytającej zrozumieć, co robi kod), mieć styl (*stylish*, być pisany zgodnie z jednym stylem, np. PEP 8 dla Pythona), wszechstronny (*versatile*, rozwiązywać problem, który występuje więcej niż raz i przewidywać zmienność danych wejściowych) oraz kreatywny (*creative*, rozwiązywać problem dotychczas nierozwiązany lub znacząco poprawiać istniejące rozwiązanie) (Tatman 2019). Na efektywność, choć także na wartości humanistyczne, mogą powoływać się uczestnicy zaangażowani w działania na rzecz tzw. różnorodności (*diversity*) w data science. W polskim świecie społecznym DS jest to głównie sprawa równouprawnienia płci, o czym więcej w części 6.2.2.

Określenie **dowozić** jest traktowane jako nieformalne, stosowane raczej w rozmowach pomiędzy uczestnikami badanego świata i ich współpracownikami, natomiast w rozmowach z klientami czy w wystąpieniach publicznych używane jest słowo **dostarczyć**, będące, tak samo jak dowożenie, kalką z angielskiego *deliver* (por. Ameisen 2018). Oznacza ono zrealizowanie określonego zadania w umówionym terminie, który może być określany jako *deadline*. Wskazywano nam, że nastawienie na dostarczanie odróżnia data science od pracy naukowej, gdzie co prawda występują określone terminy realizacji np. grantów, ale zdaniem Rozmówcy nie jest to presja czasowa, znana mu z pracy w komercyjnym DS:

> *B: OK, a czy rzeczywiście jest takie rosnące zapotrzebowanie na ludzi, którzy będą data scientistami, i to jest w ogóle taka **cudowna** ścieżka kariery że się teraz trzeba na nią rzucać? (...) Nie wiem czy to ma jakiś sens czy to jest, czy to tylko tak się ładnie pisze?*
> R: Rzeczywiście tak jest. Rzeczywiście przechodzi teraz przez cały świat [pauza] rewolucja, taka autentyczna rewolucja, nie boję się tego słowa, porównywalna z właśnie z wynalezieniem elektryczności, a mało sobie osób zdaje sprawę jak jak mocna to będzie rewolucja, bardzo sobie mało osób zdaje sprawę że ona się wydarzy w ciągu jednego pokolenia, bo to jest ta różnica. Bo wcześniej to było [pauza] tak rozwodnione na pokolenia, a teraz to się wydarzy bardzo szybko, iiiii, osobami, które wykonują tą rewolucję są właśnie data scientiści.
> *B: Mhm*
> R: I, mmm, jasne jest że to jest bardzo dobra [pauza] praca, tak? Tylko dosyć niedużo osób się

> do niej nadaje. Czyli trzeba mieć dosyć specyficzny jak to się mówi skillset czyli tak jak mówię trzeba mieć całkiem nieźle programować, trzeba całkiem nieźle znać się na statystyce i czuć matematykę, trzeba też jednocześnie rozumieć procesy biznesowe klienta w domenie, w której klient pracuje i czwartą rzecz trzeba **dowozić**. Jest taka, na przykład, mmmm, sztuczna inteligencja na uczelniach [pauza] ale ona zazwyczaj nie jest skoncentrowana na tym żeby w miarę szybko dostarczyć klientowi, eeee, efekt. Ktoś bierze sobie grant, ma dwa lata na wykonanie jakiejś jakiś badań, no i sobie wykonuje tak? Tutaj w części biznesowej, komercyjnej, jednak dosyć duży nacisk jest na tempo. (...) nawet u nas to jest tak, że bardzo ciężko znaleźć osobę, która ma wszystkie kompetencje więc trzeba budować zespoły, ten jest trochę mocniejszy w tym, ten w tamtym, więc jak się ich sparuje to będzie to, co trzeba. I to, to, tak jak z piłkarzami. No przecież każdy może kopać piłkę i się potem okazuje, że jednak nie każdy, iiii, no to, to tak jakby się pan zapytał czy warto jest być dobrym piłkarzem? **Tak, warto jest być dobrym piłkarzem.** [z uśmiechem]
> *B: OK*
> R: Powodzenia. [z uśmiechem] {wywiad 2}

Zauważmy, że dla Rozmówcy dowożenie jest czwartym, silnie zaakcentowanym przez niego elementem składającym się na tę rzadko występującą hybrydę, jaką jest osoba nadająca się do pracy jako data scientist. Pierwsze trzy z wymienionych elementów bliskie są słynnej definicji data scientista zaproponowanej w 2010 r. przez Drew Conwaya (por. rys. 2.) – umiejętność programowania (u Conwaya *hacking skills*, czyli w naszym rozumieniu programowanie i biegłość technologiczna), znajomość matematyki i statystyki (literalnie *math and statistics knowledge*) oraz rozumienie procesów biznesowych klienta (*substantive expertise*, co oznacza bycie ekspertem w konkretnej dziedzinie i różni się od umiejętności zrozumienia biznesu kolejnych klientów na potrzeby kolejnych projektów, a taka jest właśnie specyfika pracy cytowanego Rozmówcy). Słowo „dowozić" pojawia się w Polsce w opisywanym znaczeniu także w innych środowiskach, gdzie praca ma charakter projektowy – np. w żargonie firm IT i żargonie tzw. korporacji.

W badanym świecie silnie cenione jest **dostarczanie czegoś, co działa**. Wysoko ocenione będzie wdrożenie produkcyjne modelu, być może nieco niżej zakończenie projektu na etapie POC lub MVP. Niemniej oznacza to, że wtedy, kiedy rezultatem projektu data science jest raport, składający się z np. z podsumowań danych (tabel czy wykresów), wyników testów statystycznych, wyników eksperymentów z modelami uczenia maszynowego, jest to uznawane za projekt nieciekawy, mało ważny, mało użyteczny. Jeżeli celem pracy jest samo generowanie podsumowań danych w celu sprawdzenia np. w którym sklepie odnotowano największy wzrost rok do roku w sprzedaży określonego produktu, to w świecie DS przeważnie nie będzie to w ogóle uznane za data science, a za analitykę danych (na pewno przy użyciu „tradycyjnego zestawu analityka" czyli arkusza kalkulacyjnego i zapytań w SQL-u, raczej przy zapisaniu całości podsumowań w R / Pythonie). Zagadnienie

opiszemy szerzej w rozdziale 6., ale już teraz twierdzimy, że z perspektywy uczestników świata DS to element modelowania za pomocą metod uczenia maszynowego jest tym, co odróżnia data science od analizy danych. Zatem jeżeli w projekcie przeprowadzono eksperymenty z modelami ML, to raczej nie ma wątpliwości co do uznania tego za data science. Jeżeli jednak rezultat projektu sprowadza się do prezentacji takich eksperymentów w raporcie, to również jest to nieciekawy projekt, mało użyteczny projekt.

> *B: A czy w każdym takim yyy, może inaczej spytam i co jest tym produktem który wy dostarczacie, co to jest że to dajecie i projekt jest zakończony?*
> R: OK, no to odpowiedź na to pytanie wypracowywała się przez kilka lat, więc dziś już jest w miaręęęęę w miarę **jasna** bo już to przetrenowaliśmy ale faktycznie, to jest duży problem w firmach które zaczynają, i w ogóle z konceptem data science przez długo był taki właśnie problem gdzie jest ta wartość dodana. Którą niby wszyscy gdzieś tam widzieli, ale rzadko na kontach nie? No to poza firmami typu Google, Facebook które no w oczywisty sposób, wiadomo że na tym zarabia. Więc tak pierwsze takim, ee niektóre są wyniki które **wprost** przywozi yy przynosimy, a inne są które przy okazji. Jeśli chodzi o te wprost. Po pierwsze to jest identyfikacja, tak? To jest raport który opisuje że coś się wydarza lub nie wydarza w takich ilościach z takich to a takich powodów, tak? To jest to jest pierwszy typ wyników, który przy który dostarczamy. Tak? Dlaczego ten proces się nie dzieje tak jakby się czyli wydawało na papierze że powinien działać. Dlatego dlatego dlatego, tutaj są problemy tu są problemy takie, tu opisujemy i tak dalej, tak? To jest taki raport. Do tego nam bardzo często powstaje też wizualizacja. Która mówi dobrze, to więc my stworzyliśmy wizualizację i pewne metryki pewne wyliczenia, które **śledzą** jak ten proces wygląda, od zeszłego roku do dziś, i jak teraz wprowadzimy jakieś zmiany, to będziemy co miesiąc widzieć widzieli, czy te zmiany wprowadzane przynoszą efekt biznesowy. To jest coś co uważam że warto robić i to jest coś czego brakuje w wielu projektach, tak? Czyli stworzenie to jest jedno, ale wsparcie w realizacji benefitów. To jest bardzo ważne dla klienta, tak? Kiedy on wydaje 100 000 na stwierdzenie że coś jest źle, to to bardzo **fajnie** znaczy, łatwo można to zrobić, tanio, dać mu narzędzie które pozwoli mu co miesiąc śledzić, czy jak on zmienił to co my mu powiedzieliśmy żeby on zmienił, to czy on **zarabia** te swoje 100 000 z powrotem, i jeszcze więcej nie? To jest pierwsze. Drugi to jest nieoczywisty efekt jest taki, żeby dojść do tego wyniku my przeprowadziliśmy cały proces. Identyfikacji wiedzy, ludzi odpowiedzialnych za konkretne elementy tego procesu, **dane** które są w **opłakanym** stanie, to jest ogromna też jest ogromny temat. (...) Kolejnym, oprócz tych wizualizacji takim bardzo pośrednim często wynikiem jest powstanie czegoś na wzór małej aplikacji lub dużej aplikacji która nie ludziom zarządzającym ale ludziom którzy na co dzień coś robią, w jakiś sposób ułatwia pracę. Lub życie, lub automatyzuje,
> *B: Czyli na przykład to rozpoznawanie mowy o którym wspominałeś*
> R: Dokładnie tak.
> *B: Dla konsultanta.*
> R: Dokładnie tak. No i teraz tych konsultantów siedzą tysiące, tak? I oni na co dzień korzystają tak? Czyli to jest produkt danowy, który trafia z powrotem [cmoknięcie] bo ja zauważyłem też często jest taki problem że bardzo wiele projektów analitycznych jest robionych na poziomie wysokim znaczy to jest dobrze, tak? Jest jakiś prezes zarządu i on tam będzie mógł, chciałby się dowiedzieć co u niego w firmie nie gra. Natomiast spory problem polega się na tym że jak się odkryje co nie gra, co z tego, co teraz nie? Więc on jest jednym konsumentem więc ja zawsze mówię że jeszcze jest ten drugi poziom, żeby stworzyć coś używając tych danych które oni tam mają, już teraz oczyszczonych i lepszych, żeby na co dzień mieć wpływ na pracę tych tysięcy ludzi. Bo to też mega się przekłada, nie? Jeżeli każdy tam tylko dolara dziennie zaoszczędzi czy tylko pięć dolarów, no to przez skalę, przez codzienność ma dużo większy wpływ niż to że prezes wie że tu ma na czerwono tu ma na zielono. To jest z tych dodatkowych, więc powiedzmy takie aplikacje danowe my na to mówimy data products, na poziomie operacyjnym

> yyy codziennych obowiązków tak? (...) czy takie podsumowanie statystyczne że coś jest takie lub inne już jest produktem? **Może** być produktem, natomiast ja jestem yy nie jestem szczęśliwy jeśli mój zespół proponuje tylko taki produkt, ponieważ uważam, że yy jego problemem jest słaba użyteczność. Powiedzenie komuś że jest źle [pauza] no jest pierwszym krokiem tak? Natomiast ja bardzo często chciałbym ussłyszeć co jak mu wskazać co w ten [cmoknięcie] tą informację w taki sposób żeby on miał szansę coś zrobić. Czyli po angielsku na to się mówi actionable. Czyli jeżeli to co robimy pomaga człowiekowi w biznesie wykonać konkretne akcje żeby coś tam poprawić, to ja jestem z tym OK. I wtedy najczęściej jest model. Niekoniecznie predykcyjny, tak jak mówię, czasami jest to model statystyczny który coś tam albo nie wiem, jeśli z dużej ilości danych jesteś w stanie mu wydłubać konkretne linijki którymi on się musi zająć, i te linijki mu wskażesz nie wiem, w języku mu zrozumiałym, to też już jest coś eee co on może skonsumować. {wywiad 6}

Rozmówca {wywiad 6} wskazuje, że raport jest typowym rezultatem projektów, ale jeżeli to jest jedyny rezultat, to projekt nie jest użyteczny (jak wizyta lekarska, która polega wyłącznie na postawieniu diagnozy). Nieco bardziej użytecznym rezultatem projektu – rezultatem, który w badanym świecie jest już uznawany za coś działającego – jest aplikacja do wizualizacji danych. Takie wizualizacje, tzw. dashboardy, przygotowywane np. jako strony internetowe / intranetowe przy użyciu pakietów „Shiny" lub „plotly", będą prezentować użytkownikowi podsumowania kluczowych danych, informacje w postaci różnych wykresów i wskaźników tzw. KPI – czasami za KPI stoi proste działanie matematyczne, czasami zaawansowany model statystyczny lub ML – a także pozwalać użytkownikowi na interakcje, jak samodzielne filtrowanie zakresu danych w różnych wymiarach. Tego rodzaju produkt jest uznawany za tym bardziej działający, im mniej ingerencji ludzkiej potrzeba do jego funkcjonowania (np. odświeżania danych), ale także działanie produktu może być oceniane w kontekście realnego wpływu w dokonywaniu zmian. Zdecydowanie najbardziej działający jest produkt, który bazuje na produkcyjnym wdrożeniu modelu, dającym automatyczną odpowiedź bardzo wiele razy dla bardzo wielu osób – „jeżeli każdy tam tylko dolara dziennie zaoszczędzi czy tylko pięć dolarów, no to przez skalę, przez codzienność ma dużo większy wpływ niż to że prezes wie że tu ma na czerwono tu ma na zielono" {wywiad 6}.

Na powyższym przykładzie wyraźnie widzimy, jak ważna jest nie tylko efektywność, ale jak duża jest wartość sprawczości w badanym świecie DS. Zobaczmy, jak negatywnie Rozmówca ocenia sytuacje, w której model, w sensie biznesowym, byłby użyteczny tylko teoretycznie, w raporcie, „na papierze", a nie praktycznie „po wdrożeniu":

> R: Powiedzmy ktoś celowo robi [pauza] ktoś **celowo** [pauza] jak to powiedzieć, chyba trzeba powiedzieć że to był sfingowany wypadek na tej zasadzie że ludzie się umawiają, walą tanim samochodem w drogi samochód i z ubezpieczenia taniego samochodu biorą pieniądze na naprawę drogiego samochodu i jeszcze tak pięć razy.
> *B: Aha.*
> R: Aaa, i teraz chodzi o to, żeby w momencie zgłoszenia szkody stwierdzić, **tak**, mamy do

> czynienia z wyłudzeniem, trzeba się temu przyjrzeć bliżej. I były dwie trudności w tym projekcie. Jedna trudność to nie wiadomo było czy **da się** to zrobić i trudno było w jakiś systematyczny sposób stwierdzić tak, na pewno da się to zrobić, trzeba było po prostu przeanalizować projekt, przeanalizować **proces** i korzystając z doświadczenia stwierdzić [pauza] **tak, da się** albo **nie, nie da się**. Druga sprawa było bardzo dużo zmiennych które potencjalnie były użyteczne, zmiennych [pauza] tych zmiennych były **setki**, iiii, trzeba było uniknąć problemów związanych z tym, że danych się nie rozumie. Aby upewnić się, że dane są przygotowane w dobry sposób i ta analiza **też** nas zaskoczyła mimo sporego doświadczenia, yyy, w takim sensie że jeśli byśmy się nie przyjrzeli zmiennym, praktycznie jedna po drugiej, to wtedy byśmy zbudowali model który na papierze może i działa, ale po wdrożeniu działa dużo, dużo **gorzej**. (...) Eee, no i **to** było drugie zaskoczenie, natomiast **efekt** był **też fajny**, znaczy on był zaskakujący w zasadzie był do przewidzenia, w ten sposób, żeee tam było **kilka** modeli w zależności od modelu było plus 10, plus 20 plus 30% do przodu, licząc w **pieniądzach**.
> *B: Mhm, zaoszczędzonych.*
> R: Potem, finalnie w udaremnionych wyłudzeniach. {wywiad 5}

5.4.2. Sprawczość

Sprawczość rozumiemy jako **sprawianie, że coś się dzieje**. Ten sens trafnie oddaje dobrze znane w socjologii pojęcie *agency.* To „dzianie się" jest tożsame ze zmianą, rozumianą jako przejście czegokolwiek z jednego stanu w drugi. W badanym świecie zdecydowanie najbardziej ceniona jest sprawczość w tzw. realnym świecie, rozpatrywana jest ona także w kategoriach skali – im większa skala wpływu, tym większa sprawczość. W powyższym przykładzie {wywiad 5} widać, że zmiana dokonywana jest z użyciem modelu uczenia maszynowego. Model, wdrożony do pracy operacyjnej w firmie ubezpieczeniowej przyczynił się do redukcji kosztów wyłudzeń ubezpieczeń o około 10 – 30% {wywiad 5}. Choć z punktu widzenia naukowców z dziedzin humanistycznych i społecznych mógłby być to rewelacyjny przyczynek do rozważań nad sprawczością aktorów nie-ludzkich, to nie podejmujemy ich tutaj, ponieważ taka perspektywa nie pojawia się w badanym świecie społecznym DS. Twierdzimy, że w badanym świecie data science **sprawczość postrzegana jest wyłącznie jako sprawczość ludzka**. Mówi się o rozmaitych narzędziach jako potężnych, wydajnych, ale to człowiek opanowuje używanie tych narzędzi, i używa ich do wykonywania zmiany. Narzędziami nazywamy zarówno technologie (hardware i software) jak i metody używane w DS.

Wydaje się nam to przekonanie o panowaniu uczestnika badanego świata nad narzędziami szczególnie ważne właśnie w DS, a konkretnie w modelowaniu metodami uczenia maszynowego. Metody uczenia maszynowego (ML) są używane do tworzenia systemów, w których decyzje podejmowanie są automatycznie bez udziału człowieka, na podstawie odpowiedzi modelu na dane wejściowe. W ML nie ma zaprogramowanych reguł, a

reguły te są wyestymowane z danych. W dodatku część metod pozwala na stosowanie danych wejściowych w postaci przystosowanej do percepcji ludzkiej, jak tekstów, obrazów czy dźwięków. Połączone jest to wysokim poziomem skomplikowania matematyczno-informatycznego ML. To wszystko, podsycane marketingiem, składa się na obecne w różnych dyskursach – popkulturze, mediach, naukach (przyrodniczych, humanistycznych, społecznych), biznesie – kategorie określania modeli uczenia maszynowego jako magii, inteligencji, władzy, statusu boskiego czy demonicznego algorytmów (por. Thomas, Nafus, i Sherman 2018). Z punktu widzenia uczestników świata DS tego rodzaju kategoriami, podkreślającymi sprawczość tzw. algorytmów, posługują się osoby spoza świata DS, które nie mają pojęcia o szczegółach technicznych i metodologicznych, ewentualnie osoby kierujące komunikaty marketingowe (mogą być to uczestnicy świata DS lub ich rzecznicy) do takich osób spoza świata DS. Rozwiniemy ten problem w części 6.2., niemniej już teraz zaznaczmy, że z punktu widzenia uczestników świata DS modelowanie metodami uczenia maszynowego to nie magia, a majsterkowanie, zaś tzw. sztuczna inteligencja to raczej hasło na slajdach w PowerPointcie niż coś działającego, zapisanego w kodzie. **W badanym świecie DS podkreślana jest w ten sposób silna podmiotowość człowieka, biegłego w posługiwaniu się narzędziami** (technologicznymi i metodologicznymi). W popkulturowych obrazach sztucznej inteligencji Bogiem są albo „konstruktorzy myślących maszyn", albo same te maszyny (Knapik 2018:13). Pierwszy typ obrazów jest zdecydowanie bliższy temu, gdzie sprawczość lokują uczestnicy badanego świata DS.

W komercyjnym DS akcentowana jest sprawczość w sensie **wpływu działania na wyniki finansowe firm-klientów**. Stosowane jest pojęcie dostarczania wartości (*deliver value*), ewentualnie dostarczania wartości biznesowej. Zauważmy, że pojęcie wartości (*value*) stosowane było już w definicjach big data. Tzw. definicje 5 Vs wskazywały jednoznacznie na potencjalną wartość finansową danych cyfrowych, zbieranych i przechowywanych przez różne podmioty (Gandomi i Haider 2015). W toku badań zauważyliśmy, że w komunikacji ze sceny, tak na konferencjach jak i meetupach, często pada określenie „wartość" bez ujętego wprost odniesienia do finansów. Mowa jest raczej o dostarczaniu wartości, korzyściach, zmienianiu czegoś, poprawianiu i ulepszaniu. Do akcentowania wyników finansowych działań data scientistów dotarliśmy głównie dzięki wywiadom indywidualnym. Szczególnie Rozmówczynie / Rozmówcy z najdłuższym stażem zawodowym – rozmawialiśmy z pojedynczymi osobami pracującymi więcej niż 15 lub 20 lat, co niekoniecznie oznacza tyle lat w DS – podkreślali, że **data science musi się opłacać**:

> *B: Yhm. No i co to, co to właściwie znaczy [data scientist], jakbyś mógł mi to wytłumaczyć jeżelibym słyszał o tym pierwszy raz w życiu?*
> R: Data science to jest, znaczy najpierw rzucę takim **banałem** że data science to nauka o danych prawda, a data scientist tylko tutaj właściwie bym rozgraniczył na jak z matematyką, jest matematyka teoretyczna i matematyka stosowana, to tak samo ja widzę data scientista jako kogoś kto **stosuje** tą naukę do danych dla całkiem realnych problemów żeby osiągnąć całkiem eeyy znaczy realne, w tym sensie realne, że yyy przydane nie wiem w jakiejś dziedzinie życia, w biznesie, w technice, yyy niekoniecznie także w jakichś zupełnie nie wiem, zupełnie innych dziedzinach w których się nie da zarobić, ale chodzi o sam fakt że w czymś co człowiek może jakoś użyć, żeby coś zrozumieć, zauważyć. Także to jak matematyka jest teoretyczna i stosowana, no to data scientist [krótka pauza] tak jak ja go widzę i siebie rozumiem, to jest właśnie ten ktoś kto się zajmuje matematyką stosowaną, tooo używa tej nauki o danych dlatego żeby żeby wspierać jakieś procesy, tu akurat w moim przypadku, w mojej organizacji, no ale, dlatego właśnie siebie widzę jako data scientista.
> *B: Yhm, no właśnie. A czym ty dokładnie* się zajmujesz?
> R: Yyy ja tutaj yy właściwe jakby znaczy, zajmuje się tutaj dwoma rzeczami. Raz to jestem ekspertem od dziedziny którą się zajmuje ta moja organizacja, a dwa że zajmuje się też już data science takim węziej pojętym, że używam pewnych technik technik data science do tego żeby yyyy z danych które ta moja organizacja produkuje, albo jej klienci produkują, żeby wyciągnąć jakieś ciekawe informacje które, które będą wartością dla mojej firmy, albo dla klientów, czyli de facto dla mojej firmy, bo to o wszystko chodzi żeby moja firma miała z tego pieniądze. (…) w ostatecznym rozrachunku chodzi o to żeby to data science się po prostu **opłacało**, a nie może być po prostu yyy w cudzysłowie cool działem, który jest takim dodatkowym kwiatkiem przy kożuchu, że [zmienionym głosem] **każda duża firma w branży IT musi mieć dział data science bo to dobrze wygląda**. Znaczy może i tak jest [śmiech], może to dobrze wgląda [z uśmiechem] i może warto się tym tak chwalić, marketing też ważną sprawą jest, nie mniej w ostatecznym rozrachunku chodzi o to żeby to na bieżąco wspierało procesy dużej organizacji i żeby przynosiło zyski jakieś. {wywiad 14}

Rozmówca uważa, że DS jako takie oznacza osiąganie użytecznych, mających praktyczne zastosowanie efektów. Dalej podkreśla, że w jego miejscu pracy, czyli zespole DS w dużej firmie, co oznacza w tym wypadku pracę dla tzw. klienta wewnętrznego, firma jest zorientowana na przełożenie pracy data scientistów na własne wyniki finansowe. Rozmówca z rozbawieniem mówi o praktykach prowadzenia działów DS tylko ze względów wizerunkowych / marketingowych. Sądzimy, że przez część uczestników badanego świata takie praktyki oceniane są negatywnie, jako nieracjonalne i nieefektywne w sensie marnowania wkładu, nie mającego przełożenia na wyniki oraz braku wyników, zatem i braku sprawczości.

Na poziomie indywidualnym wskazywano nam na praktyki pracy pozornej, pozornego i przedłużającego się w czasie eksperymentowania z modelami ML. Rozmówca {wywiad 23} określił data scientistów wykonujących taką pracę „bullshit artists" (w wolnym tłumaczeniu: mistrz bzdur). Takie działanie jest raczej negatywnie oceniane w badanym świecie:

> R: [rozmowa dotyczy tego, kto powinien być w zespole zajmującym się ML] (...) wiadomo, że potrzebny jest też ktoś, wiadomo jakiś stakeholder z biznesu, bo tak naprawdę każdy model jest o tyle dobry, o ile przynosi koniec końców złotówki, to trzeba po prostu **policzyć**, czy to się

opłaca, czy nie. Koszty utrzymywania [pauza] machine learning to jest taki obszar, w którym jak ktoś jest takim bullshit artist, to może w tym jak **pączek w maśle** funkcjonować, nie? Bo **da się** wziąć jeden model i przez 6 miesięcy go cały czas poprawiać. I **cały czas** będzie coś do roboty. I tu spróbujemy tego, a tam tego, a tu zbudujemy, a tam tu a tam tamto i **6 miesięcy płaci się komuś pensję**, niemałą, i zysku z tego żadnego nie ma. Potrzebny jest właściciel produkty, ktoś ze strony biznesowej, ktoś, kto rozumie, po co to w ogóle się robi, ktoś, kto nakłada **twarde** ograniczenia, co jest warunkiem sine qua non, żeby w ogóle produkt wdrożyć, jakie są warunki brzegowe wdrożenia, ile to przynosi, kiedy ciąć? Bo to też jest ważna, super ważna decyzja. Ja w tym projekcie, w którym pracuję, regularnie mi się zdarza sytuacja, kiedy mówię, nie, to może jeszcze byśmy spróbowali, może jeszcze tego, tego nie spróbowaliśmy, tego nie spróbowaliśmy, i przychodzi chłopak, który jak gdyby [pauza] on podejmuje decyzję, na temat tego produktu i mówi: **nie**. 40h już na to spaliliśmy, mamy taki wynik, jaki mamy, teraz to już jest diminishing returns[238]. Nie ma sensu spalić kolejnych 10h, po to, żeby poprawić to o 2%. I to jest rzecz, której żadne naukowiec nigdy nie zrobi, to jest rzecz, której żaden data engineer nie zrobi, tu jest potrzebny ktoś, kto jest ze strony biznesu, który dobrze rozumie to środowisko biznesowe, w którym dane rozwiązanie funkcjonuje.
B: Też mówiono mi, mmmm że data scientiści mają tendencję do robienia modeli i do poprawiania accuracy, albo nie wiem do używania jakiejś metody, która jest z nowego paperu, natomiast, że biznesowi chodzi o coś innego, tak? Chodzi o zarabianie, tak?
R: Jest taka zasada, którą ja wbijam cały czas studentom do głowy, właśnie, jak im się podoba modelowanie. Modelowanie jest **fajne**, oczywiście jest fajnie, jak się pojawi jakiś algorytm, który **sam** coś przewiduje, albo generuje słowa, obrazy, itd., to jest sexy i w ogóle, ale jest takie proste zdanie: feature engineering beats model engineering always[239], które oznacza, że generalnie to zaaplikowanie algorytmu na sam koniec, tooo jest wisienka na torcie. Różnice między dobrymi algorytmami będą niezauważalne, kryterium użyteczności algorytmu zazwyczaj będzie leżało **zupełnie** gdzie indziej, to będzie właśnie albo KPI, albo wyjaśnialność, albo czas działania, albo odporność na błędy. To będą **ważne** rzeczy, a nie to czy on ma 94 czy 95%, plus właśnie potrzeba takiego **bardzo** silnego spojrzenia biznesowego zz, i cięcia ścieżek, nie? Jest czas, to jest takie exploration vs exploitation[240] i jest czas, kiedy poszerzamy przestrzeń rozwiązań i szukamy tak po **omacku** mniej więcej, gdzie można by iść, i jest moment, kiedy to **studzimy** i mówimy: dobra, koniec. Teraz już doszliśmy do takiego momentu, zmieniamy to w produkt, teraz idzie to do produktyzacji. Już nie ma czasu. Już mamy kolejną rzecz do zrobienia. Ale to jest [pauza] **trud** zarządzani projektem ML-owym. To jest trudniejsze niż zarządzanie takim tradycyjnym projektem informatycznym. {wywiad 23}

Na wartość sprawczości w badanym świecie wyraźnie wskazuje uznawanie pisania kodu za prawidłowy sposób wykonywania działania podstawowego. **Sprawczość człowieka posługującego się do operacji na danych kodem jest znacznie większa niż posługującego się GUI**. Ma to zastosowanie także do ogólnej pracy na komputerze. Dla osób będących doświadczonymi koderami[241] omawiana różnica jest tak oczywista, że nie werbalizowały jej w wywiadach. Wgląd w tę sprawczość osiąganą dzięki kodowaniu dała nam narracja Rozmówczyni, która wywodzi się z akademii i kodować zaczęła po studiach magisterskich:

238 Pochodzące z ekonomii pojęcie „prawa malejących przychodów".

239 Oznacza to, że przygotowanie danych jest ważniejsze niż strojenie modelu ML (por. wypowiedź Dominika Batorskiego, Gontarz 2019:19).

240 Dosł. eksploracja vs eksploatacja.

241 Mogą być to np. ludzie z wykształceniem informatycznym lub matematycznym; programiści, którzy przeszli do DS; osoby interesujące się programowaniem od dzieciństwa, także osoby młode.

B: Rozumiem, rozumiem, a w początkach doktoratu, czy tam jeszcze na etapie yee magisterki rozumiem że uczyłaś się sama, takich rzeczy jak Python, czy właśnie praca w konsoli?
R: Powiem ci że nawet na etapie pracy magisterskiej się tego nie uczyłam, bo wydawało mi się że to nie jest dla mnie, że to za trudne i że tym się zajmują yy ludzie gdzieś tam w jakichś najlepszych laboratoriach na świecie, na polibudzie, ale że to nie mnie dotyczy [śmiech]
B: Aha, no to to jak to było że zaczęłaś?
R: [śmiech] Więc takie było moje podejście na etapie pracy magisterskiej. A w zasadzie zaczęłam jak to bywa eee często, eee że konieczność czy potrzeba jest matką wynalazków, jak to mówią, i tutaj w zasadzie potrzeba mnie zmusiła do tego żeby po prostu spróbować niezależnie do tego co myślałam na ten temat. (...) Nie wchodząc w szczegóły tego problemu, z którym się spotkałam, eee szukałam jakiegoś narzędzia, **gotowego**, które mogłoby mi to ocenić [ze śmiechem]. No i się rozczarowałam, bo takiego narzędzia nie było, a wiedziałam że ten projekt chcę zrealizować właśnie z zachowaniem jakichś takich najwyższych standardów, czyli muszę obronić wszystkie potencjalne pytania, czy krytykę recenzentów, no to starałam sama wejść w ich skórę i skrytykować swój projekt. No i to mnie doprowadziło do tego, że stwierdziłam że skoro nie ma narzędzia to ja muszę w takim razie coś sama wymyślić. Ee no i udało mi się skonceptualizować eee jak musiałby działać taki prosty skrypt, żeby to miało sens. No i zaczęłam szukać, emm fragmentów kodu, które rozwiązywałby mi każ [westchnienie] każdy kolejny kawałeczek, tego problemu. Tutaj dużą pomocą też się okazał eee kolega, który już biegle znał Pythona więc był w stanie też skonsultować, czy podążam w dobrą [pauza]
B: Czy podążam, czy podążam w dobrą stronę.
R: Tak. Natomiast w momencie, w którym puściłam ten skrypt i on mi podzielił te bodźce, eee zobaczyłam jaka jest w ogóle **potęga** w użyciu tego eee narzędzia, tak? Jakim jest znajomość języka programowania. Że ja sobie wymyślam problem, jestem w stanie znaleźć rozwiązanie prawie że od tak, to po prostu do mnie dotarł [krótka pauza] miałam takie [śmiech] głupio mówić objawienie [ze śmiechem], ale to było po prostu tak **niesamowite poczucie mocy sprawczej**, tego co można zrobić, że ja stwierdziłam, że jak ja bym się tego dobrze nauczyła, to ja bym w ogóle mogła rozwiązywać wiele problemów które napotykam w swojej pracy badawczej i w ogóle [krótka pauza] **wznieść**, to co robię na na wyższy poziom, że to nie musiałoby być zawsze takie sprowadzone do tego co jest wykonalne manualne, manualnie, albo dostępne, zrobione już przez kogoś. I to był dla mnie taki moment, że stwierdziłam, że skoro mi się udało yyy zrobić to, to dlaczego miałabym nie próbować dalej. I że [krótka pauza] że może nie ma sensu się zastanawiać, czy ja się do tego nadaje, czy też nie, tylko po prostu się wziąć do roboty [ze śmiechem]. {wywiad 21}

Podobne jak cytowana Rozmówczyni wrażenie „potęgi" osiąganej dzięki kodowaniu miał piszący te słowa (RŻ), widząc wyniki działania pierwszych własnoręcznie napisanych linii kodu i porównując je z wkładem powtarzalnego klikania w GUI, potrzebnego do uzyskania takiego samego wyniku. Rozmówczyni wskazuje, iż umiejętności programistyczne postrzega jako dające większą swobodę pracy – „to nie musiałoby być zawsze takie sprowadzone do tego co jest wykonalne manualne, manualnie, albo dostępne, zrobione już przez kogoś" {wywiad 21} – czyli inaczej mówiąc w kodzie można zapisać niemal wszystko, co człowiek wymyśli (o ile pozwalają na to umiejętności), zaś klikając człowiek pozostaje ograniczony przez polecenia zdefiniowane przez twórców oprogramowania. Pisząc kod można raczej **skoncentrować się na tym, jak coś zrobić, a nie czy da się to zrobić**. Dzięki tej sprawczości i swobodzie Rozmówczyni odczuwała, że ma możliwość osiągnięcia wyższego poziomu pracy badawczej niż gdyby nie potrafiła pisać kodu.

O ile osoby z małym doświadczeniem w kodowaniu mogą być zachwycone sprawczością, jaką daje korzystanie z kodu w porównaniu do korzystania z GUI, to bardziej doświadczeni koderzy mogą mieć podobne odczucia porównując kodowanie (w sensie programowania czynności i reguł) z modelowaniem metodami uczenia maszynowego. Mówiliśmy już, że modelowanie to najbardziej lubiana część pracy osób zajmujących się DS. Na przykładzie cytowanego wcześniej w tym kontekście fragmentu {wywiad 2} zauważmy, jak Rozmówca wskazuje, że rezultat uzyskiwany z użyciem ML jest nie możliwy do uzyskania wyłącznie „narzędziami informatycznymi":

> R: Ja odczuwam niesamowitą radość kiedy model jest w stanie robić coś rozsądnego. Tak, czyli na przykład bierzemy sobie właśnie jakieś aaa, jakieś zagadnienie rozpoznawania czegoś na obrazkach, zaczynamy uczyć ten model i on zaczyna **rzeczywiście** wykazywać takie po wynikach widzimy że on coś rozpoznaje potem jeszcze robimy parę różnych takich tricków żeby zobaczyć co powoduje że on rozpoznaje dany obiekt i patrzymy i mówimy no kurcze to ja bym zrobił tak samo.
> *B: Mhm*
> R: Nagle się człowiek orientuje, że yyy, że to **coś** to już jeeest takie coś bardzo **mądrego** [powoli] co udało mu się stworzyć. I ja jeszcze mam ten background eee informatyczny i ja wiem na przykład że narzędziami informatycznymi bym tego nie zrobił. {wywiad 2}

Spotkaliśmy się wielokrotnie ze stwierdzeniami, że pewnych problemów nie dałoby się rozwiązań „normalnym" programowaniem (opisywaniem czynności i reguł), a da się je w sposób akceptowalny rozwiązywać dzięki uczeniu maszynowemu (czyli estymowaniu reguł z danych). **Uczenie maszynowe jest zatem postrzegane jako jeszcze wyższy niż pisanie kodu poziom sprawczości**, a także efektywności, osiąganej dzięki komputerom:

> R: (...) Poszedłem na mechanikę z tego powodu że, mmm w liceum pałałem, no nie powiem że nienawiścią, ale ogromna niechęcią do informatyki z tego powodu że wydawało mi się, taki zawód dla nerdów, tak? Jeśli jest informatyk to z pewnością nie prowadzi on żadnego ciekawego życia, co on ciekawego w życiu może robić. Iii, plus nie miałem jakichś wybitnych nauczycieli, nie trafiłem na wybitnych nauczycieli informatyki przynajmniej do liceum, znaczy nie no byli lepsi lub gorsi, ale no bez szału. Nikt nie zaraził mnie jakąś pasją w stosunku do informatyki. Iii, co się, co się zmieniło, pod koniec studiów inżynierskich właśnie mechaniki i budowy maszyn, w trakcie tych studiów miałem też zajęcia z programowania i muszę przyznać że, tam trafił mi się jeden nauczyciel, który w taki fajny sposób prowadził zajęcia że mi się wtedy to wszytko podobało. Strasznie się wtedy w to wkręciłem. Pamiętam że sobie myślałem, kurczę jaka szkoda że yyy dopiero teraz się tym zainteresowałem, pewnie gdybym się tym wcześniej zainteresował too, to bym to robił. I tak sobie myślałem że kurczę w sumie to jest już za późno, jak już zacząłem studiować tę mechanikę, no to pewnie będę się kręcił dookoła tego, w życiu. No ale potem jakoś tak, powoli, krok po kroku, krok po kroku zacząłem się tym coraz bardziej interesować. (...) ta książka którą przeczytałem [krótka pauza] angielski tytuł to jest „Second Machine Edge" (...) I ogólnie treść jest taka, że to co, w czym teraz żyjemy to jest taka druga rewolucja przemysłowa, tak? Pierwsza rewolucja przemysłowa opierała się na tym że weszły, że weszły maszyny do obiegu, no i one niesamowicie przyspieszyły rozwój ludzkości, teraz druga rewolucja przemysłowa polega na tym że do obiegu weszły komputery, tak? A potęga komputerów polega na tym, że załóżmy ja napiszę jakiś program komputerowy, jestem

> w stanie go zmultiplikować nie wiem, z miliard razy, dwa miliardy nie wiem, ile mi tylko pozwoli nie wiem pamięci komputer tak, to ja mogę sobie mnożyć, wysyłać, kopiować i tak dalej i tak dalej. Z maszynami tak nie jest [krótka pauza] mam tutaj na myśli nie wiem, ogólnie z produkcją czegokolwiek, tak? Jeśli wyprodukuję chleb to ten chleb jest dostępny tylko tu, w tym miejscu, jeśli będę chciał ten chleb nie wiem, załóżmy wyprodukować w Japonii, no to tam muszę mieć swoją fabrykę, ewentualnie muszę go wysłać przez pół świata. No to, widać tutaj ogromną różnicę i **potencjał**, maszyn, tak? I wtedy tak sobie pomyślałem, kurczę strasznie mi się spodobała ta koncepcja że, że maszyny mają taką **siłę** w sobie, znaczy maszyny, no dokładnie komputery, ale to co mi się najbardziej w tym wszystkim podobało to to że, yyy no komputer to jest tylko narzędzie tak, too komputer, póki co jeszcze nie myśli iii właściwe wszystko to co on zrobi, to to jest tutaj w głowie, tak? I strasznie mi się spodobała właśnie ta koncepcja, że wystarczy posiąść odpowiednią wiedzę, i możesz wydawać wiesz polecenia komputerom, które będą to wszystko robiły i które zwielokrotnią noo w nieograniczonym stopniu twoje możliwości, tak. I to właśnie bardzo bardzo mi się spodobało, wtedy właśnie jeszcze myślałem że będę robił jakieś takie strony i tak dalej, eee że w tę stronę to pójdzie i pojechałem po raz drugi, bo to był łączony taki wyjazd na studia do [nazwa kraju] i tam miałem pierwsze w swoim życiu zajęcia ze sztucznej inteligencji i wtedy no bardzo bardzo mi się spodobało i stwierdziłem że to jest to, że w to chcę pójść. {wywiad 11}

Osiąganie jeszcze większej niż dzięki tradycyjnym narzędziom IT, sprawczości i efektywności jest ogólną legitymizacją stosowania uczenia maszynowego, a nawet DS jako takiego. Wdrażanie modeli ML może zatem być przez uczestników badanego świata postrzegane jako kolejny, po tzw. cyfryzacji czyli w ogóle przejściu do pracy na komputerach, krok w rozwoju ludzkości[242]:

> R (...) No no myślę że żyjemy w czasach dużych przemian, iiiiii chyba jest ten kierunek że sobie w głowie widzimy Jetsonów albo Odyseję Kosmiczną i jak rozmawiamy z HALem i dużo jest automatów i roboty, cyborgi z Gwiezdnych Wojen i Terminatorze nas o otaczają, więc wydaje się że to jest pewien naturalny kierunek w którym zmierzamy, jako ludzie no i dziś potrzeba baaardzo dużo ludzi którzy by potrafili zaprząc yy te wszystkie eee systemy, sprząc ze sobą i wprowadzić uczenie maszynowe w jak najszerszym miejscu więc, no myślę sobie że zapotrzebowanie nie jest przesadzone. Tak? Jeśli faktycznie chcielibyśmy żyć w takim świecie który z nami rozmawia i ma świadomość, to faktycznie będzie potrzebnych nas baaardzo wielu, tak jak dziesięć lat temu było potrzebnych programistów. Jak mówiliśmy że wchodzimy w IT i będziemy teraz wszystko mieć komputerowe, to była potrzebna rzesza informatyków którzy to wykonają fizycznie, więc ja myślę że teraz potrzebujemy rzeszy data scientistów i modelarzy którzy kolejną warstwę yyy na to nałożą, nie? {wywiad 6}

Mogą się w badanym świecie pojawiać odniesienia do tego, że jest to jakoby naturalny kierunek rozwoju ludzkości, łącznie ze stosowaniem analogii rozwoju technologii do ewolucji istot żywych, jak np. komputery uczą się widzieć (o metodach analizowania obrazów np. z użyciem uczenia głębokiego).

To, że zdaniem uczestników badanego świata, data science nie jest – w odróżnieniu od nauki akademickiej – sztuką dla sztuki, a rozwiązywaniem rzeczywistych, praktycznych

242 Nie zauważyliśmy, by pojawiało się to w badanym świecie, ale w literaturze istnieją pojęcia trzeciej i czwartej rewolucji przemysłowej – trzecia to cyfryzacja, a czwarta polega na stosowaniu tzw. sztucznej inteligencji (por. Paprocki 2016).

problemów, widzimy jako przejście podobne do tego, co Ewa Domańska nazwała przesunięciem „od kontemplacji do zmiany" w humanistyce. Na gruncie nauk humanistycznych już ponad dekadę wstecz wskazała ona na zwrot performatywny / zwrot ku sprawczości, który polega na postrzeganiu badań przede wszystkim jako narzędzi rozumienia i analizy świata w celu „wpływania na zmiany, które w nim zachodzą oraz prowokowania tych zmian" (Domańska 2007:55–56). Uczestnicy świata data science odgradzają się (Kacperczyk 2016:48) od świata nauki akademickiej także wskazaniem na różnice w sprawczości – zarówno tym, że znacznie bardziej cenią sprawczość, jak i osiągają znacznie wyższą sprawczość. Może być to także praktyka legitymizowania działalności badanego świata DS, zarówno kierowana wobec klientów, jak i wobec początkujących uczestników lub wannabesów. W pierwszym przypadku argumentacja może przebiegać następująco: „potrzebujecie data science, czyli rozwiązania waszych problemów biznesowych za pomocą danych, a nie bezwartościowego teoretyzowania, nie zatrudniajcie akademików", w drugim „pracuj u nas jako data scientist, nie idź na doktorat, nie tylko więcej zarobisz ale będziesz robić coś, co naprawdę działa, a nie produkować teksty, których potem nikt nie czyta". Te przykłady są przejaskrawione, a podziały między DS a nauką akademicką bardziej skompilowane i rozmyte. Niemniej chcemy zaznaczyć, że badany świat społeczny jest silnie zwrócony ku sprawczości i bardzo wysoko ceni to, co Domańska nazwała mocnym podmiotem (2007:56). Mocny podmiot ma, jej zdaniem, charakter hybrydowy, współdziała z innymi aktorami ludzkimi i nie-ludzkimi, co w naszym przekonaniu wybitnie pasuje do badanego świata społecznego – wykorzystywanych w nim przeróżnych narzędzi technologicznych i metodologicznych, zaangażowania w ruch open source, organizacji meetupów czy hackatonów. Domańska wskazuje, że:

> W tle „zwrotu performatywnego" oraz „zwrotu ku sprawczości" (*the agentive turn*) stoi – jak wspomniałam wyżej – kategoria **zmiany jako wartości pozytywnej** we współczesnym świecie. **Podmiot, który ma siłę sprawczą, staje się w tym nastawionym na zmiany świecie najbardziej pożądany**. [podkr. RŻ] Dokonywać zmian, być ich sprawcą, a nie obiektem – oto pożądany model (…) (Domańska 2007:56).

Łatwo byłoby nam wywnioskować, jak zresztą zrobiliśmy to we wcześniejszej pracy, że generalnie taka działalność jak data science / tzw. big data czy AI to odpowiedź na wysoki poziom niepewności we współczesnym świecie, to

> pozornie trafna odpowiedź na wyzwania ponowoczesności – przeładowanie informacjami, szybkość zmian, pragnienie konsumpcji, tonięcie w morzu możliwości; szeroko rozumianą „mozaikowość" i „płynność" życia" (Żulicki 2017:202).

O tym samym mówiła Domańska wskazując, że w humanistyce przesunięcie od nastawienia na kontemplację do nastawienia na zmianę widzi jako wyraz „poczucia tracenia przez ludzi kontroli nad światem" (2007:55). Nie mamy żadnych podstaw do twierdzeń o globalnych procesach społecznych, ale wybrane praktyki legitymizacji (por. rozdział 6.) świata DS widzimy jako nastawione na obietnice zwiększenia kontroli nad jakoby coraz szybciej zmieniającym się i coraz bardziej skomplikowanym światem i dzięki temu uzyskaniu większej sprawczości oraz większej efektywności w wybranym obszarze.

Sprawczość osiągana dzięki technologii może być uwodzicielska dla ludzi zajmujących się data science – o tym mówił dla prasy słynny sygnalista Christopher Wylie (por. 2.2.3.4.), pracujący niegdyś w Cambridge Analytica (Levy 2019). Wylie wspominał swoje zafascynowanie możliwościami targetowania reklamy politycznej dzięki danym z Facebooka, i choć jako gej i obywatel Kanady nie miał wspólnych poglądów z amerykańską prawicą, to zaangażował się w projekt tak bardzo, że nie myślał o realnych konsekwencjach swoich działań. „Kiedy piszesz skrypt w Pythonie, to nie masz wrażenia, że robisz cokolwiek komukolwiek. Nie myślisz: 'jak to może faktycznie skrzywdzić ludzi?'" (ibidem). Wylie twierdził, że słyszał podobne do swojej historie od innych inżynierów, np. pracujących w Google czy Facebooku – nie myślą oni o społecznych konsekwencjach swojej pracy, tylko optymalizują wybrane metryki.

> Facebook i Google od dawna przywabiają inżynierów, obiecując, że podejmowane przez nich decyzje będą miały ogromny wpływ (*impact*), bowiem ich produkty docierają do miliardów ludzi. Przypadek Wylie'go pokazuje, że ci technologiczni czarodzieje powinni dokładnie pytać, co to za wpływ. Kiedy zajmują się interesującym problemem (...) ci ciekawi i zmotywowani badacze i koderzy zbyt łatwo mogą ślepo przeć naprzód (*charge ahead blindly*). Pokusa epickiego wpływu jest trudna do powstrzymania (*The lure of epic impact is hard to restrain*) (Levy 2019).

Prezentowany wyżej fragment to materiał prasowy, który podlega medialnym prawom ekonomii uwagi, a ponadto jego bohater Ch. Wylie promuje w nim swą świeżo wydaną książkę. Niemniej prezentujemy go, gdyż autor trafnie ujął **połączenie między sprawczością a ciekawością** – obie uznajemy za wartości badanego świata, więc zdecydowanie mogą być atrakcyjne dla osób zajmujących się data science.

Wskazywano nam też, że oczekiwania klientów wobec usług związanych z DS bywają wygórowane (o czym więcej w części 6.2.1., gdzie modelowanie traktujemy jako arenę), co świadczy o przecenianiu sprawczości, możliwej do osiągnięcia dzięki projektom data science. Także niektórzy uczestnicy badanego świata mogą przeceniać sprawczość własną lub

dziedziny DS jako takiej, o czym jeden z bardziej doświadczonych Rozmówców mówił następująco:

> *B: Nie, chciałem spytać o twoją opinię co sądzisz o tym, że to jest jakaś najlepsza praca świata,*
> R: [śmiech]
> *B: Najseksowniejszy zawód świata*
> R: Tak tak [ze śmiechem] słyszałem
> *B: I o tych wszystkich bardzo medialnych hasłach?*
> R: Yyy uch! Co ja o tym sądzę? No wiesz, to jest z jednej strony fajne [pauza] bo to pomaga pracować. Tak uważam, bo buduje pewną percepcję. Znaczy uważam że to co robimy jest trudne, jest nowe, więc to jest tak naprawdę stwarzanie baaardzo wielu rzeczy od nowa, to bardzo dynamicznie się rozwija więc ten cały szum medialny się w to wpisuje i pomaga, i dobrze dobry dobrze nas usytuławia na rynku, to na pewno. Ale z drugiej strony nie uważam żeby moja praca była trudniejsza niż nie wiem wystrzelenie ludzi na Marsa tak? Więc to nie jest tak że tu będzie jesteśmy teraz **herosami** i w ogóle złotym palcem dotykamy i rozwiązujemy problemy świata i lekarstwo na raka się nagle objawia nie? Więc też trzeba mieć w sobie pokorę i wiedzieć że to też jest pewna praca która no po prostu codziennie się ją wykonuje raz lepiej się udaje, raz gorzej, duży poziom procent projektów w ogóle bardzo szybko umiera, myślę że yyy [pauza] na cztery projekty myślę ze dwa może tak pół na pół będzie. Pół projektów połowa projektów w ogóle odpada bardzo szybko, ponieważ okazuje się że to nie jest zadanie dla tego typu takiej analizy danych i nie uda się tego rozwiązać, a z tej połowy jeszcze tam część odpada na dalszych etapach. Na przykład dostępności danych. Albo jakości tych danych, albo w ogóle ilości tych danych. Więc to też nie jest tak że to jest panaceum i wszystko można tym wyleczyć. Więc to też uczy pokory i trzebaaa [pauza] mówić nie. To nie jest proste, jak człowiekowi się wydaje że jest supermenem bo jest teraz data scientistem to może wszystko, [cmoknięcie] no **nie** może no. {wywiad 6}

W toku badań wielokrotnie spotykaliśmy się z historiami akcentującymi realny wpływ, zmianę rzeczywistości czy inaczej ujęte synonimy sprawczości, osiąganej dzięki realizacji projektów DS. Zarówno indywidualnie, jak i szczególnie w roli rzeczników swoich firm / instytucji uczestnicy badanego świata wypowiadają zdania w rodzaju: „my zmieniamy świat", „robimy rzeczy niemożliwe", „rozwiązujemy problemy, których nikt inny nie potrafi rozwiązać". Nie jest jednak tak, że uczestnikom badanego świata jest obojętne, w jakiej dziedzinie wpłyną na świat. Część uczestników świata DS zwraca uwagę nie tylko na siłę wpływu, ale na znaczenie tego wpływu.

> R (...) reklama to jest taka że jak wciska reklamy Viagry i tego gdzieś tam w spamie to jest ciężko sobie wiesz spojrzeć w lustro że robię coś co jest przydatne dla ludzkości, no bo, bo **nie**, prawda? Jakieś tam historyjki, półspamerskie rzeczy [niezrozumiale] żeby z tego mieć pieniądze, no to nie ma poczucia że robisz coś dla misji. W różnych firmach się różne rzeczy robi, ja akurat chcę mieć poczucie, nie jak to mówił Marks alienacji pracownika od tego, od [pauza] od owocu pracy, bo dla mnie to jest istotne. Dużo bardziej wolę być tym freelancerem, mieć to poczucie że ja pracuję nad odpowiednim projektem z odpowiednimi ludźmi. (...) **Mnie** to motywuje, nie wszystkich, nie wszystkich. (...) Na pewno mógłbym sobie siedzieć w Londynie w jakichś tam finansach i zarabiać dużo więcej pieniędzy, dużo więcej, około tam 200 000 może nie w pierwszym roku, po kilku latach około 200 000 funtów rocznie, tego rzędu. No ale, ani bym nie miał poczucia że robię coś wartościowego (...). Pieniądze nie są dla mnie celem, wyłącznie środkiem. {wywiad 26}

Naszym zdaniem marketing ma relatywnie najgorszą opinię w badanym świecie. Spotkaliśmy się w naszych badaniach z krytyką etyczną tej branży (Żulicki 2019). Raczej nie jest to specyficzne tylko dla polskiego świata DS – już w początkowym okresie kształtowania się branży DS w USA twórca słynnej wizualizacji prezentującej DS (por. rys. 2.) napisał, że „żaden młody hacker nie dorastał, marząc o sprzedawaniu reklam" (Conway 2014). Podobnie Jeff Hammerbacher, uznawany za jednego z twórców współczesnego znaczenia terminu „data science", były pracownik Facebooka i LinkedIn, powiedział publicznie wielokrotnie cytowane słowa, że „najlepsze umysły mojej generacji zastanawiają się, jak sprawić, by ludzie klikali na reklamy (…) to jest do dupy (*[t]he best minds of my generation are thinking about how to make people click ads […] that sucks*)" (Halford i Savage 2017:1140).

Rozmówczynie / Rozmówcy pracujący w obszarach, które określali jako służące wyłącznie zarabianiu pieniędzy lub nieciekawe (m.in. marketing, usługi finansowe) raczej wskazywali chęć zmiany branży na taką, która jest w ich opinii bardziej znacząca, wartościowa, ważna, służy ludzkości. Twierdzimy, że w badanym świecie za taką uważana jest przede wszystkim medycyna.

> *B: Mhm, a masz jakąś taką pracę marzeń?*
> R: [westchnie] pracę marzeń? Ojej, ciężko powiedzieć! Nie! Chyba nic takiego żeby powiedzieć że konkretnie to i, na pewno tak jak właśnie tutaj jak ci mówiłam bardzo podoba mi się branża medyczna. Jak, yy robiłam kursy w [nazwa kraju] to widziałam bardzo wiele zastosowań analizy danych takich które mają rzeczywisty **wpływ** yy na to żeby na przykład diagnozować choroby albo poprawić pracę lekarzy, no i to są takie zastosowania które **naprawdę** ta praca jeszcze dodatkowo, oprócz tego że sama praca jest ciekawa, to jeszcze jej efekty sprawiają przyjemność. To znaczy tak motywują, wiadomo że robi się to naprawdę w fajnym celu nie tylko, zazwyczaj celem w firmach takich sektora prywatnego, przedsiębiorstwach jest zysk. No to też jest dobry cel, jak najbardziej uzasadniony, ale to jest dodatkowa motywacja pracować z takimi danymi i i właśnie na przykład w branży medycznej, no. Więc myślę że to mogłoby być ciekawe, ale żeby powiedzieć o jakiejś jednej konkretnej firmie, to to nie, to nie. Ciężko mi teraz powiedzieć. {wywiad 4}

To w medycynie, czy ogólnie branży biomedycznej, wskazywano na możliwość wzięcia udziału w projekcie, który „zrewolucjonizuje coś w świecie" {wywiad 15} w takim sensie, że np. wydłuży ludzkie życie:

> R: (…) Tylko że mówię, akurat nie jest to, no nie zawsze do końca jest to, co chciałbym robić. Ja bym może chciał takiej biologii, albo nawet jakiejś takiej, branży, biomedycznej powiedzmy tam też jest silny ten nacisk na analizę danych, ale to chyba też się wiąże z tym że musiałbym doszkolić się w konkretnych metodach, no ale do gdzieś do tej około biologii bym chciał wrócić, no na pewno to jest fajniejsze niż finanse, i z tego bym chciał uciec, nie. Chociaż tu też jest największa kasa, no ale to, też nie chodzi o to [krótka pauza] żeby zarabiać nie wiadomo ile, tylko żeby mieć fun.
> *B: Ok. Ale to będzie też głupie pytanie, ale dlaczego biomedyczna branża jest fajniejsza niż finanse?*

> R: Kurde no [pauza] no nie wiem dla biologa [Rozmówca jest z wykształcenia biologiem – przyp RŻ] jest to na pewno ciekawsze, nie? [śmiech]. Jakby jest to bliżej tego co chciało się robić wcześniej nie, eee i [krótka pauza] tutaj też jakieś takie masz poczucie że jak rozwijasz nie wiem nowe, leki czy coś tam, no to rozwijasz nowe leki, które sprawiają że ludzie [krótka pauza] eee dłużej żyją czy coś nie, albo nie wiem jesteś w stanie wpłynąć, znaczy albo wziąć udział chociażby w jakimś takim dziwnym projekcie, który zrewolucjonizuje coś [krótka pauza] w w w, na świecie, a finanse to by było ciężko doszukiwać się rewolucji?, nie [pauza] przynajmniej z poziomu bycia pracownikiem banku, czy jakiejś innej instytucji finansowej. {wywiad 15}

Jednocześnie w tego rodzaju wypowiedziach wskazywano nastawienie na samorozwój i to, by praca była ciekawa:

> *B: No dobra, to [pauza] już częściowo, już częściowo o tym mówiłaś, co w tej pracy lubisz, co ci się w niej podoba, ale spytam cię wprost, co najbardziej lubisz w tej pracy, co ci się najbardziej w niej podoba. No bo mówisz że ją lubisz, tak?*
> R: Tak.
> *B: To co najbardziej w niej lubisz?*
> R: Em, **wyzwania** i rozwiązywanie problemów, które się pojawiają na drodze, iii [pauza] no tak, no głównie to. Znaczy to też mi się, to mi się zawsze bardzo podobało w programowaniu też, to rozwiązywanie problemów no i to myślenie algorytmiczne, jak coś zrobić najszybciej, najlepiej i najbardziej wydajnie. W machine learningu tam jest też **dużo** większe pole popisu na to, bo im algorytm dział szybciej to tym lepiej, tak? Coś się wykonuje 1000 razy więc to też się nakłada. Na przykład przy robieniu aplikacji webowych to nie ma aż takiego znaczenia, czy coś jest napisane optymalnie czy nie. Więc to jest najfajniejsze.
> *B: Ok.*
> R: Znaczy fajne jest też to, że tak naprawdę, czasami w niektórych projektach możesz faktycznie zmieniać życie ludzkie. Na przykład rekrutowałam się do takiej firmy [nazwa kraju i nazwa firmy], ona, oni to jest aplikacja ym aplikacja mobilna generalnie i oni robią, rozpoznają na podstawie zdjęć czy to jest rak skóry, czy nie. Robią zdjęcie znamion i oni to rozpoznają. I rozmawiałam z CTO tej firmy i on w trakcie w ogóle rozmowy rekrutacyjnej opowiedział mi parę historii i on miał łzy w oczach. Albo na przykład człowiek mi pisał że dzięki waszej aplikacji wykryłem że mogę mieć raka, poszedłem do lekarza i dzięki temu żyję. To jest, to jest fenomenalne.
> *B: Mhm Ok. A jaka była by taka twoja, wymarzona praca?*
> R: Oj [westchnienie] a nie wiem. To jest teraz bardzo drażliwy temat, bo jestem cały czas w tym procesie rekrutacyjnym.[śmiech]
> *B: No właśnie, no bo szukasz teraz pracy.*
> R: No. Znaczy ja szukam generalnie jako machine learning engineer, właśnie, że chcę programować jednak te rzeczy, te modele, i tak dalej. I czy, wymarzona praca dla mnie byłaby niekoniecznie w tym **co** robię tylko bardziej z **kim** robię. Chciałabym być w teamie ludzi, którzy się na tym znają, eem i uczę się od nich po prostu. To na tym się bardziej teraz skupiam niż na konkretnym co. {wywiad 8}

Takie nastawienie na ciekawą pracę pojawia się także w formie porad dla początkujących data scientistów:

> Moim przyjaciołom, kończącym właśnie studia, mówię: „Nie, nie idźcie robić w finansach, jeśli to nie jest coś, co naprawdę was ekscytuje. Jest tyle innych interesujących rzeczy." Za trzydzieści lat naprawdę nie będziesz zwracał uwagi na te lepsze zarobki [w branży finansowej niż w innych – przyp. RŻ]. To nie będzie mieć znaczenia, jeśli będziesz pracować nad problemami, które są dla ciebie interesujące i znaczące (*problems you find interesting and*

meaningful). (Gutierrez 2014:317)

Niemniej można spotkać się z opiniami, że charakterystyka pracy w komercyjnym DS jest zawsze taka, iż najważniejszy jest wyniki finansowy i jest on właściwie jedynym kryterium oceny jakości projektów DS:

> Model jest konstruowany w taki sposób, by osiągał najwyższą skuteczność oraz rentowność, nie zaś w kierunku 'sprawiedliwości' lub 'dobra zespołu'. Taka jest oczywiście natura kapitalizmu. Zyski są dla firm jak tlen - sprawiają, że przedsiębiorstwa żyją. Z ich perspektywy byłoby skrajnie głupie, wręcz nienaturalne, gdyby miały odrzucać potencjalne oszczędności (O'Neil 2017:108).

Pojęcie „działa / nie działa" także wskazuje na wartość sprawczości w badanym świecie. **Istnieje praktyka kwestionowania autentyczności działania, tzn. nie uznawania działalności za data science wtedy, kiedy projekt nie jest nastawiony na rezultat w postaci czegoś działającego**. Podkreślmy, że nie chodzi o sytuację, w której pomimo starań nie udaje się uzyskać zadowalających rezultatów i projekt zostaje zakończony bez rezultatów. Chodzi o intencjonalne nastawienie wyłącznie na rezultat w postaci deklaratywnej, czyli wykonania pewnej analizy danych i sporządzenia raportu czy prezentacji z wynikami lub rekomendacjami. Część uczestników badanego świata może wyśmiewać takie rezultaty, mówiąc, że to nie jest data science, a np. konsulting, nie uznając takich deklaratywnych rezultatów za coś, co rozwiązuje rzeczywisty problem. Wyraźnym przykładem omawianej praktyki było zdarzenie na jednym z meetupów zorientowanych na język R {obserwacja 25}. Pierwszy prelegent prezentował model optymalizacji cen dla klienta z obszaru handlu detalicznego, akcentując jak ważne w projekcie być zrozumienie potrzeb firmy-klienta oraz teorie ekonomiczne. Przyszedł czas na pytania z publiczności:

> „kiedy klient się dowiedział, czy wasz produkt działa, czy wyniki są dokładne? Czy dopiero jak zapłacił?" Śmiech.
> jeden człowiek pyta: „czyli co tak właściwie dostarczyliście? Slajdy?" śmieje się cała sala
> odpowiedź – ten projekt jest w toku, to nie są ostateczne rezultaty.
> inne pytanie: „Czy państwo używali R?" Mega śmiech na sali.
> Odpowiedź: R był ułamkiem stacku technologicznego, ale był w korze (od core[243]) naszego projektu, była też apka w Shiny {notatka terenowa, obserwacja 25}

Kolejny prelegent z tej samej, co pierwszy firmy, odniósł się do opisywanego zdarzenia w następujący sposób:

> „Mówi się że firmy konsultingowe produkują slajdy (sala śmiech) a my w analitics [czyli dziale analitycznym firmy konsultingowej – przyp. RŻ] chcemy dawać coś działającego, chociaż prototyp" {notatka terenowa, obserwacja 25}

243 Słowo *core* to dosł. rdzeń, oznacza centrum, najważniejszy element.

Przekonanie o możliwości rozwiązywania rzeczywistych problemów i zmieniania świata na lepsze za pomocą technologii nazywano technologicznym solucjonizmem (Carr 2013). Nie możemy jednak uznać, że w świecie społecznym DS uczestnicy cenią solucjonizm, skoro ma on oznaczać m.in. postrzeganie złożonych fenomenów kulturowych czy politycznych jako transparentnych, ewidentnych i dających się łatwo optymalizować (ibidem). Uczestnicy badanego świata cenią raczej scjentystycznie pojętą racjonalność (o czym więcej w części 5.4.6.) w połączeniu z opisaną wcześniej efektywnością, co wyraża się np. w akceptowaniu niemożliwości lub nieopłacalności rozwiązania pewnych problemów za pomocą aparatury data science oraz w akcentowaniu wagi doprecyzowania problemu i jego metryk, konsultacji z ekspertami dziedzinowymi itp. W badanym świecie solucjonizm jako ślepe przekonanie o tym, że technologia (a dokładniej ML) jest dobrym i skutecznym rozwiązaniem dla niemal każdego problemu byłoby ocenione jako mało racjonalne, świadczące o braku doświadczenia w wykonywaniu działania podstawowego, więc charakterystyczne dla wannabesów lub co najwyżej początkujących uczestników świata DS. Nawet w sondażach KDnuggets zauważono, że przekonanie o pozytywnym wpływie „automatyzacji, AI i ML" na społeczeństwo globalne spada wraz ze wzrostem liczby lat doświadczenia w rozwijaniu tych technologii. Szczególnie w podziale na grupy od 0 do 4 lat doświadczenia i powyżej 4 lat zaobserwowano niższy odsetek respondentów wskazujących na pozytywny wpływ (ponad 60% w grupie 0-4 przy niecałych 40% dla powyżej 4) i wyższy wskazujących wpływ negatywny (około 10% dla 0-4 i 30% powyżej 4) (Piatetsky 2017a).

Autorzy analizujący kodeksy etyczne tzw. AI zauważyli zapisane w nich jednoznaczne przekonanie o ogromnym potencjale sztucznej inteligencji – zarówno pozytywny, jak i negatywny wpływ AI jest w takich kodeksach powszechnym, uniwersalnym zmartwieniem (*the positive and negative impacts of AI are a matter of universal concern*) (Greene i in. 2019). Sądzimy, że w badanym świecie DS jednoznacznie panuje taki właśnie pogląd: narzędzia, jakimi dysponuje data science (technologiczne i metodologiczne) mają ogromny wpływ na dalsze losy ludzkości, Ziemi i świata. Rozwój DS, a w szczególności uczenia maszynowego, postrzega się jako nieunikniony, jest on imperatywem, dotknie każdej dziedziny życia. Wskazuje to na wysoką wartość sprawczości w świecie DS. Jednocześnie narzędzia, w tym AI czy uczenie maszynowe, postrzegane są jako neutralne w ludzkich rękach. Sądzimy, że według uczestników świata DS o pozytywnym / negatywnym wpływie wolno mówić dopiero na przykładzie konkretnych projektów, a najlepiej na zrealizowanych wdrożeniach. Może to akcentować ludzką odpowiedzialność za cel używania tych

uznawanych za potężne narzędzi, co widzimy jako kolejny argument za przywiązaniem badanego świata do mocnej, sprawczej podmiotowości ludzkiej. „Technologia jest **niewinna** (*innocent*, podkr. org.). To data scientist ustawia dane wejściowe (*input*) i zadaje modelowi zmienną celu" (Casas 2018). W badanym świecie DS oraz na przecięciu ze światem biznesu {obserwacja 22} pojawiają się porównania do noża / młotka, którymi można przygotować posiłek / wbić gwóźdź lub zabić człowieka[244]. Data science czy konkretnie ML widziane jest jako narzędzie ogólnego zastosowania, które samo w sobie nie podlega ocenie moralnej. Prowadzi to do tego, że w badanym świecie nie kwestionuje się etyczności data science jako takiego – nikt nie powie, że coś takiego jak analiza danych, statystyka, programowanie i uczenie maszynowe to generalnie złe i szkodliwe pomysły. Niemniej sądzimy, że **artykułowana neutralność narzędzi jest równa temu, że są one domyślnie dobre**, oceniając tę dobroć w odniesieniu do wartości świata DS. Narzędzia te mają przecież zwiększać efektywność i sprawczość.

To, by dzięki rozumieniu rzeczywistości poprzez dane „mieć wpływ" na świat jeden z Rozmówców wskazywał jako ulubioną część pracy DS, zaś problemy z tym „ulepszaniem" świata jako coś powodującego niezadowolenie z pracy:

> *B: Mhm okej. A czy, czy lubisz tą pracę? Tak w ogóle?*
> R: Zdecydowanie.
> *B: A co ci się w niej podoba?*
> R: Emm no chyba właśnie to, że mogę mieć wpływ, że i że mogę rozumieć. Tak bym to ujął. Eee bo to jest też coś co spowodowało że poszedłem na studia na fizykę. Ee i konkretnie fizykę teoretyczną, żeby bo, znajduję przyjemność w, wytłumaczaniu świata. [pauza] W budowaniu jakiś modeli, w sensie, modeli rzeczywistości w sensie koncepcyjnym. Eem i, i na tej podstawie jakby odkryw, odkrywaniu nowych rzeczy w ten sposób. No tak, i w zasadzie, to samo, dzieje się, dzieje się teraz w mojej pracy tylko na trochę innym poziomie, już nie zajmuję się ee kwarkami, czy atomami, tylko klientami.
> *B: Mhm, OK. A czy jest coś czego nie lubisz w tej pracy, coś co ci się nie podoba?*
> R: Mm, [pauza] czy jest coś co mi się nie podoba? [pauza]
> *B: Może nie ma czegoś takiego?*
> R: [pauza] Mmm, właśnie próbuje się zastanowić czy [pauza] czy jakieś aspekty, amm to znaczy może ujmę to w ten sposób, mmm [pauza] czasem ta praca bywa frustrująca [pauza] ee przez eee przez to yy jak duże problemy napotyka się po drodze, mmm przede wszystkim mm ze względu na jakąś taką niemrawość, czy oporność organizacji na zmianę, na, na ulepszanie i to można powiedzieć że to często staje na przeszkodzie temu żeby, yy żeby osiągać to co się chciało osiągnąć, żeby yyy [pauza] to przeciwdziała temu zmienianiu świata na lepsze, o którym mówiłem wcześniej. Na lepsze nie w sensie żeby nie było głodu tylko żeby ten mały fragment rzeczywistości lepiej funkcjonował.
> *B: OK, rozumiem.*

244 Przykłady można by mnożyć, np. tak w wywiadzie dla portalu sztucznainteligencja.org.pl wypowiedział się Przemysław Biecek: „przecież nie mówimy, że młotek jest etyczny albo nieetyczny, tylko że osoba, która go użyła, podjęła bardziej lub mniej etyczne decyzje. SI [sztuczna inteligencja] jest narzędziem i w kwestiach etycznych nie zdejmowałbym odpowiedzialności z człowieka, który to narzędzie stworzył, lub który z niego korzysta" (Chojnowski 2020).

> R: Oczywiście. [pauza] Ee także tak, czy t jest coś czego nie lubię w mojej pracy to, tak bym tego nie ujął, ale to jest powodem że, tego że mmm czasem jestem niezadowolony mocno z dnia pracy. {wywiad 16}

5.4.3. Samorozwój i samodzielność

Opisujemy łącznie samorozwój i samodzielność, ponieważ te wartości są w badanym świecie powiązane ze sobą tak, że samorozwój człowieka obywać się ma niemal wyłącznie samodzielnie. W naszych wywiadach pojawiały się rzecz jasna historie o znaczących innych, np. w postaci nauczycieli, mentorów, partnerów, niemniej sądzimy, że **w świecie DS szczególnie ceni się samodzielny wysiłek oraz panuje przekonanie, że rozwój człowieka polega głównie na samodzielnej pracy**. Takie postrzeganie samorozwoju związane jest z charakterem wykonywania działania podstawowego DS – pisanie kodu odbywa się niemal wyłącznie samodzielne przed ekranem komputera. Fizyczna obecność innych osób pracujących na swoich komputerach nie ma naszym zdaniem dużego znaczenia, a praktyki programowania w parach (*pair programming*) (Williams i in. 2000) są niezwykle rzadkie. Zaród taki styl pracy w kolektywach hakerskich nazwał „wspólnotą indywidualnej koncentracji" (Zaród 2018:220) i ten styl obowiązuje w świecie DS w sytuacji fizycznego dzielenia przestrzeni przez dwóch i więcej uczestników badanego świata.

Samodzielność nie oznacza samotności. W badanym świecie praktykowanych jest wiele form współpracy grupowej i wzajemnej pomocy, co omawialiśmy w kontekście narzędzi (5.3.4.). Zatem o ile działanie podstawowe jest realizowane samodzielnie i zazwyczaj za dany projekt DS jest odpowiedzialna jedna samodzielna osoba, to występuje praktyka konsultowania się i tzw. dzielenia wiedzą:

> *B: Mhm okey. A mógłbyś opowiedzieć mi o swoim zespole i o tym jak wygląda praca w tym zespole, tak na co dzień?*
> R: Mhm, zespół mamy niewielki tak rzędu 5-7 osób. Amm w ostatnim czasie można powiedzieć że to trochę fluktuuje, więc eee więc pozostańmy przy tym zakresie 5-7. [długa pauza] Głównie, to znaczy <u>każdy analityk czy tam data scientist yy u mnie w zespole yy samodzielnie pracuje nad swoimi projektami, sobie zleconymi eeee rzadko zdarzają się projekty które jednocześnie tak jakby na full time zaangażowanych jest dwóch, dwóch analityków z mojego zespołu eeee głównie właśnie każdy ma swoje projekty i ewentualnie wspieramy siebie nawzajem dyskusją poradami czy wiedzą z konkretnych obszarów biznesu. [pauza] Co jeszcze można powiedzieć. No tak to w zasadzie jeżeli chodzi o ten element zespołowy to jest przede wszystkim wymiana wiedzy i idei a już konkretne zadania to już każdy sam realizuje</u>. {wywiad 16}

Podkreślmy, że praktyki dzielenia się wiedzą, kodem i wynikami pracy własnej oraz udzielania pomocy innym osobom są w świecie DS powszechne i cenione, tak w formie

zapośredniczonej przez internet (np. na GitHub, LinkedIn, blogi, Stack Overflow, Slack) jak i w kontaktach twarzą w twarz (codzienna praca, meetupy, konferencje, warsztaty, hackatony). Każda z tych praktyk jest także nastawiona na samorozwój oraz autoprezentację – budowanie swojego wizerunku zarówno w badanym świecie jak i wobec ludzi spoza świata, np. rekruterów, pracodawców. Autor podręcznika dla początkujących koderów J. Nowosad wskazuje na popkulturowy mit samotnego programisty:

> Mit programisty mężczyzny jest też powiązany z wymienionym kilka akapitów niżej mitem samotnego programisty. (…) W popkulturze osoba, która potrafi programować spędza czas samotnie, gwałtownie wpisując kolejne linie kodu do komputera w ciemnym pokoju. W rzeczywistości jednak większość profesjonalnych programistów pracuje w zespołach, których członkowie pracują nad różnymi aspektami tego samego problemu. Pisanie programów często wymaga współpracy różnych osób, dlatego też umiejętność pracy w grupie jest coraz istotniejsza. Warto dodać, że współpraca nad pisaniem programów nie musi odbywać się w jednym pokoju czy budynku. Ze względu na charakter takiej pracy i możliwości technologiczne, wiele formalnych i nieformalnych grup pracuje zdalnie nad projektami. Wiele przykładów takich zachowań można znaleźć przyglądając się otwartemu oprogramowaniu (ang. *open-source software*) na platformie GitHub (np. https://github.com/trending/r) (Nowosad 2019:14–15).

Samorozwój jest w badanym świecie raczej postrzegany jako działanie bezustanne, polegające na pięciu aspektach. Po pierwsze, byciu na bieżąco z nowymi technologiami i narzędziami dla DS. Po drugie, poznawaniu domeny lub kolejnych domen, z których pochodzą dane, z którymi się pracuje. Po trzecie, poznawaniu pracy innych osób zajmujących się DS, także w innych domenach, stosujących inne metody i inne technologie. Po czwarte, zgłębianie lub przynajmniej odświeżanie wiedzy teoretycznej z zakresu matematyki, uznawanej za niezmienne podstawy DS (np. algebra liniowa, statystyka matematyczna). Po piąte, podejmowanie wyzwań, może pojawiać się określenie w rodzaju „wychodzenia ze swojej strefy komfortu", czyli np. zmiana pracy na inną domenę danych, udział w hackatonie, podjęcie się organizacji konferencji, podjęcie projektu niekomercyjnego itp. Cześć uczestników świata DS podkreśla, że ważne jest także rozwijanie kompetencji biznesowych i miękkich, jak np. umiejętności doprecyzowania celów biznesowych projektów, rozumienia działa różnych rodzajów firm-klientów, zdolności komunikacji z osobami nie-technicznymi. Mówiliśmy wyżej o przynajmniej odświeżaniu wiedzy matematycznej – podkreślmy, że w świecie DS nie uważa się, że raz osiągnięty, wysoki poziom kompetencji jest już dany raz na zawsze, a raczej, że nie używając określonych narzędzi i metod zapomina się je i trzeba uczyć się powtórnie.

Zwracamy szczególną uwagę na pojęcie **bycia na bieżąco**. W badanym świecie dominuje przekonanie o tym, że data science jako dziedzina zmienia się i rozwija się bardzo

szybko, oraz że podobne, szybkie tempo przemian w różnych aspektach charakteryzuje współczesny świat jako taki. Zatem nie dotyczy to tylko rozwoju narzędzi charakterystycznych dla DS, choć w praktyce duża część działań uczestników to śledzenie nowo pojawiających się narzędzi (konkretnie np. repozytoriów kodu, pakietów, artykułów naukowych prezentujących metody ML). Część uczestników może mówić o eksponencjalnym tempie zmian, przyspieszeniu technologicznym (co charakterystyczne np. dla Raya Kurzweila, por. 1.1.2.), bywa to także praktyką legitymizowania data science i praktyką marketingową – upraszczając trzeba wdrażać AI, bo firmę przegoni konkurencja – panuje w tej sferze „wyścig zbrojeń" {wywiad 19}. W kontekście samorozwoju jako wartości świata DS to bycie na bieżąco oznacza, że należy nieustannie uczyć się i podejmować działania nierutynowe co najmniej w pięciu wyżej wyróżnionych aspektach. Jako kod in vivo traktujemy używane w świecie DS pojęcie **state-of-the-art**, które oznacza najbardziej zaawansowaną technologię lub metodę w danym momencie. W badanym świecie DS należy zatem znać state-of-the-art i wiedzieć, że za miesiąc czy dwa coś innego może być owym najbardziej zaawansowanym na daną chwilę rozwiązaniem. Należy znać najbardziej aktualne dokonania i rozumieć ich tymczasowość.

To przekonanie o konieczności bycia na bieżąco i znajomości state-of-the-art występuje równocześnie z opisywanym nieco wcześniej przekonaniem o nieuchronności dalszego rozwoju DS. Rozmówca {wywiad 19} wskazywał, że będzie się zmieniać struktura ról w DS (mniej stanowisk analitycznych, więcej związanych z różnymi aspektami ML oraz inżynierii danych, o czym więcej w części 6.), niemniej nie spodziewa się spadku wysokiego popytu na pracę osób zajmujących się DS, ponieważ „to nie jest tak że jedna firma odpuści żeby być w tyle" {wywiad 19}. Poproszony przez nas o spekulacje na temat tego, co mogłoby zatrzymać lub spowodować regres w branży DS wskazał nieco żartobliwie na wybuch na Słońcu niszczący całą ziemską elektronikę, ewentualnie silne regulacje prawne – choć te drugie raczej zmieniłyby charakter branży niż doprowadziły do zapaści:

> *B: Ok, ok. Aaa jak myślisz jak się będzie ta branża, data science, dalej rozwijała?*
> R: [pauza] Mmm, znaczy wydaje mi się że będzie, że może się **zmniejszać** liczba takich stanowisk powiedzmy analitycznych, i tak dalej. Bo [pauza] może w końcu uda się firmom zrobić narzędzia na tyle proste, że ludzie bezpośrednio powiedzmy **decyzyjni** będą mogli sobie w nich wyklikać to co chcą, albo ustawić takie narzędzia po prostu gdzie ktoś może sobie **wpisać** pytanie, tekstem, że niby to ma mu wyświetlić na przykład odpowiednie tam wykresy, i tak dalej, z danych których potrzebuje. Więc w takich obszarach powiedzmy, od tej strony może się trochę zmniejszać, liczba stanowisk, yyy ale pewnie będzie rosła liczba, liczba stanowisk w takich obszarach, mmm jak właśnie ten machine learning, gdzie powiedzmy badasz też, też taki masz, masz taki typowy problem, bierzesz typowy algorytm, patrzysz czy działa i wypluwasz na produkcję, i na takich gdzie, w których wymagają robienia częściowo researchu, gdzie

> bierzesz **paper**, gdzie już jakieś rzeczy porobione były, tą metodę, powiedzmy implementujesz ją i w jakimś stopniu **badasz** co w niej nie działa, co można poprawić i ulepszasz. Ten, ten masz do zrobienia, to mi się wydaje będzie rosło, bo to jest coś czego jednak maszyna już nie może, nie będzie mogła zrobić, w ten sposób zautomatyzować, czego się nie da zautomatyzować, eee i [pauza] **też** myślę, jednak w obszarze data engineeringu też mogą być, wzrosty zapotrzebowania, mmm, bo to **jednak** są takie różne rzeczy z data engineeringu, to są one, w tych aspektach, gdzie jednak jest ważne zrozumienie co się w danych dzieje [krótka pauza] mmm, dlatego że jednak danych będzie generowanych coraz więcej, mmm, i będzie potrzeba coraz więcej ludzi którzy będą w jakimś stopniu to umieli [pauza] powiedzmy przeczyścić. To już się nie będzie nazywało po prostu stanowisko data engineer, a może nie wiem, data clearance engineer, coś takiego.
> *B: Mhm, mhm, ok. Mmm, a, yyy [krótka pauza] jak myślisz, na zasadzie takiego eksperymentu myślowego, czy, co by się musiało stać żeby nastąpiła, yyyy jakaś taka zapaść właśnie w tej branży data science, żeby, no nie wiem, żeby biznes przestał się interesować w ogóle jakimiś rzeczami typu data driven? [krótka pauza] Co by takiego, co by takiego musiało się stać, na zasadzie eksperymentu, jak myślisz?*
> R: Najbardziej? No to myślę, odpowiedni rozbłysk na słońcu, który by wyemitował fale EMP[245] które by zniszczyły elektronikę [z uśmiechem].
> *B: [śmiech] Ok.*
> R: I wtedy byłoby zawieszenie do momentu konstrukcji nowej elektroniki nie? bo, a to po prostu byłoby dlatego żeby człowiek się cofnął, no myślę do XIX wieku, nie? co najmniej. Nie? Z takich rzeczy powiedzmy, nie mamy klęski na świecie, jakiejś wirusowej, hmm [pauza] co by mogło być? [długa pauza] Mmm, no jedyne co mi przychodzi do głowy, co by mogło być **realnie**, no to jakieś **dziwne regulacje prawne**, które by uniemożliwiały od strony praktycznej [krótka pauza] wyciąganie wniosków z danych które generują ludzie. [krótka pauza] Ale wtedy myślę też bardziej by się to, tak naprawdę **zmieniło** w jakimś stopniu, zakres pracy data scientista, czy zakres takich kompetencji, dotyczący szerokiej wiedzy prawnej, co wolno robić, a co, jak można ominąć.
> *B: Mhm, no i to by chyba trochę sparaliżowało w ogóle, wszystkie biznesy, prawda, jeżeli te regulacje by poszły jakoś tak super daleko, nie?*
> R: Znaczy no tak mi się wydaje nie? ale mówię takie rzeczy, po prostu które by **mogły** być. Ee, nie wydaje mi się żeby coś tak specjalnie mogło, yee zmniejszyć no bo to jest w pewnym stopniu jednak, **wyścig zbrojeń** [krótka pauza] yyy w firmach. I to nie jest tak żee [krótka pauza] [cmoknięcie] jedna firma odpuści żeby być w tyle, nie? Ja to widzę w ten sposób. Tak. {wywiad 19}

W podobnym duchu wskazywano nam, że **branża DS będzie się zmieniać w kierunku automatyzowania czynności wykonywanych obecnie przez ludzi, ale nie ma takiej możliwości, by przestała się rozwijać** czy wygasła:

> *B: Też na zasadzie eksperymentu myślowego yyy co by się mogło takiego stać, żeby data science przestało się rozwijać? Żeby w ogóle jakoś to wygasło? Żeby pozwalniali tych data scientistów wszystkich lub prawie wszystkich? Żeby firmy tego nie chciały? Żeby w ogóle temat umarł?*
> R: Nie wiem o jakich ty abstrakcjach mówisz [śmiech] Nie, żartuję sobie. Wiesz co, **nie**, no tak nie będzie.
> *B: Rozumiem, że to jest niemożliwe?*
> R: No nie wydaje mi się. **Oczywiście** te role będą ewoluowały, czyli to, że te algorytmy coraz bardziej będą się automatyzować, to jest fakt. Teraz rozmawiałam z moim kolegą, oni już pracują nad rozwiązaniem automatyzacji tego, żeby wybierały się same najlepsze modele czyli to będzie coraz bardziej się automatyzowało, ale nadal będą potrzebne osoby, które będą potrafiły prezentować przed biznesem konkretne informacje, które będą rozumiały, na jakiej

245 EMP to *electromagnetic pulse* czyli impuls elektromagnetyczny.

> zasadzie to jest tworzone. Więc te role będą ewoluowały. W co co, jeszcze możemy sobie gdybać. Nie mam **pojęcia**, ja nie wiem co będzie. Kiedyś lubiłam sobie mówić tak no, za pięć lat, za dziesięć lat. Na dzień dzisiejszy to idzie tak szybko, że ciężko mi planować, 3 lata temu to ja w ogóle jeszcze data science nie zajmowałam się, ten termin jakbyś się mnie zapytał, to bym zrobiła takie [R szeroko otwiera oczy] Więc na pewno będzie to coraz bardziej się szerzyło i coraz bardziej też **powszedniało** tak na dobrą sprawę, w pewnym momencie to stanie się coraz bardziej powszednie, ale będzie się też wiązało z tym, że rzeczywiście te automatyzacje zmuszą niektórych do gdzieś tam zmiany swojej profesji, dokształcenia się. {wywiad 22}

Rozmówczyni wyraźnie odnosi się także do szybkiego tempa zmian i nieprzewidywalności tego, co będzie się działo za kilka lat.

W powyższym fragmencie mowa także o „spowszednieniu" {wywiad 22}. Część uczestników badanego świata odnosi się w tym kontekście wprost do kategorii mody, szumu medialnego (*hype*) wokół data science czy AI sądząc, że **nie tyle branża przestanie się rozwijać, co opadnie wobec niej entuzjazm**. Jest to także związane z przekonaniem o konieczności bycia na bieżąco i z rozumieniem aktualności oraz tymczasowości panujących mód:

> *B: Ok. To dążąc już ku końcowi, yyy chciałbym spytać o twoje plany zawodowe na później?*
> R: Więc [pauza] zdaje sobie sprawę że nic nie jest dane raz na zawsze, już byłem, już trochę, no już dosyć długo pracuję, tak teraz będę miał 20 lat pracy no nie, także to moja pierwsza praca po studiach związana z telekomunikacją i wielokrotnie już byłem świadkiem takich [pauza] **mód**, ta moda tych mód, które po prostu, to był **niezwykły hype** w ogóle na coś, o jej, były takie, takie znaczy nawet, podejrzewam że nawet jakbym ci powiedział słowo to **Parlay**, yyy to nawet byś nie wiedział
> *B: Nic mi to nie mówi.*
> R: A po prostu to, ile było na ten temat konferencji, ile na ten temat. Ogólnie była idea taka żeby yyy sieci radiowe, to jest przykład nie masz się z tym przyzwyczajać chodzi o idee, żeby sieci operatora wystawiały pełen interfejs i żeby można było aplikację internetową i różne usługi oferować za pomocą, za pomocą tych aplikacji w internecie, stricte już bazując na na sieciach komórkowych. Konsorcja powstawały wielkie, yyy to były wielkie konferencje jedna za drugimi, były biuletyny tam te największe raporty Gardnera i tak dalej, które każdy kosztował to po tysiąc dolarów, a tam można, tam było na ten temat produkowane, a teraz nawet już **strony** nie ma po tym. I ja byłem wielokrotnie świadkiem takich właśnie wielkich, takich hype'ów, takich mód, który po prostu miały w ogóle zmienić oblicze telekomunikacji, **świata**, w ogóle i całego biznesu, iii mówię o tej swojej działce, którą znam. I pamiętam że ja też się podpalałem raz za razem i pamiętam jak jeszcze pierwszy **raz** coś takiego było, to po prostu taki stary inżynier mówi że on już to widział wiele razy takie, takie, takie emocje i potem to nic nie dało, myślę ooo co ten gość tam, co ten gość gada no nie, i teraz, i teraz **sam jestem takim gościem**, który też czasami wręcz studzi emocje i i i mówi że coś, że yy jak jest jakiś hiper optymistyczny pomysł, że użyjmy data science i zobaczymy jakie będą wyniki, to mówię, spokojnie. I i i [pauza] nie traktuję tego data science jako tego czego będę, o jaaa! teraz złapałem Pana Boga za nogi i już będę tym żył do końca życia. Nie, no to po prostu jest kolejna rzecz którą się dobrze bawię, no tak jak zmieniałem też zupełnie rzeczy którymi się zajmowałem będąc jeszcze w takim dziale telekomunikacji, **bardzo** różne, to teraz też w tym data science się dobrze bawię i przez najbliższe lata, czuję że to jest coś co będzie mi sprawiało radość, ale nie przywiązuję się do tego, jestem, **teraz** to już wiem na pewno [z uśmiechem], że za parę lat, może za pięć, może za sześć, nie wiem, nie chcę tutaj strzelać, bo nie wiem aż tak bardzo, to będzie jedna z wielu mód, teraz jak teraz ktoś jest na przykład nie wiem, specjalistą

od Java Enterprise Edition, co kiedyś w ogóle było, też **mega** hype był na to, no to teraz ok, nadal ma pracę, no ale już nikt nawet no jakośśśś, już się jakoś tak o tym nie mówi w ogóle. Uważa się to za technologię ciężką i w ogóle, gdzieś tam jeszcze w korporacjach też to jest wspierane i tak dalej, ale no szału nie ma, chociaż nadal dobrze zarabia i jest fajnie, ale już, ale już chyba nikt o tym nawet nie mówi, nie myśli że to może być fajne, po prostu ktoś na tym dobrze zarabia i jeszcze pewnie będzie przez najbliższe dziesięć lat, albo. Java staje się takim nowym COBOLem i i tak też będzie jeszcze przez jakiś czas z data science, ale ale powstaje coś nowego, co być może przykuje uwagę ludzi którzy chcą się uczyć rzeczy nowych, albo no będzie takim nowym, takim, taką super rzeczą. {wywiad 14}

Bycie samodzielnymi i rozwijanie się jest przez część uczestników badanego świata DS widziane jako wynik ciekawości, a dokładniej bycia zainteresowanym data science. O ciekawości jako wartości świata DS piszemy w kolejnej części 5.4.4. W poniższym fragmencie zwracamy szczególną uwagę na to, że zdaniem Rozmówczyni bycie zainteresowanym DS sprawia, że nie „trzeba" (w sensie przymusu) a „można" (w sensie możliwości) wciąż się rozwijać. To, czy będzie mieć poczucie ciągłego samorozwoju jest dla Rozmówczyni kryterium, decydującym o kontynuacji albo zmianie pracy:

B: OK, a co tak w ogóle najbardziej w twojej pracy ci się podoba?
R: Najbardziej podoba mi się to, że jest różnorodna. Że wymaga myślenia. Że projekt za projektem to nie są powtarzalne rzeczy. Jakieś tam części się powtarzają jasne, ale jakby za **każdym razem** jest to trochę inny problem do którego trzeba [pauza] wymyślić jak jakieś podejście. Trzeba przepracować, wypróbować kilka różnych możliwości, czasami wymyślić coś zupełnie nowego, skorzystać poszukać nowych metod jakie, teraz yyy powstają, iiiii no właściwe za każdym razem można jakoś poprawić te swoje metody algorytmy, booo to jest **bardzo** prężnie rozwijająca się dziedzina, mam wrażenie że jak mało która teraz, iiii właściwie cały czas śledząc rozwój yyy rozwój tej dziedziny, czytając różne artykuły patrząc jakie się pakiety pojawiają chociażby tam właśnie w eRze, można się dowiadywać zupełnie nowych rzeczy i robić w zupełnie inny sposób to co się jeszcze robiło jaką konkretną metodą, yyy nie wiem **pół roku** temu chociażby. Więc to jest tak że można **dużo** yyy, trzeba dużo **myśleć**, dużo **szukać** różnych źródeł i dużo yyy być **cały czas na bieżąco**. Z tym jak wygląda ten, ten yy data science. To jest chyba najciekawsze, no i tak, no i te problemy są różnorodne, jakby jeszcze teraz w ramach jednej pracy, może jestem cały czas w tej branży [usługi finansowe] i tak spotkam różne modele, a wiem żee w pewnym momencie mogę stwierdzić że, chcę zmienić pracę i wtedy znaleźć sobie zupełnie inną dziedzinę, w której moja wiedza i moje umiejętności również znajdą zastosowanie. I jakby to, to jest właśnie super. Że mogę zacząć nową, pracować [pauza] w branży medycznej, chociażby. Nie ma, nie ma ograniczeń, właściwie, yyyy, no, **żadnych**. Jest po prostu ciekawa praca. (...)
B: Mhm, OK. Jeszcze wróćmy do tego co mówiłaś przed chwilą, że jest mnóstwo nowych pakietów i tak dalej, że trzeba być na bieżąco, ja też właśnie takie wrażenie odniosłem na konferencji że to jest taka branża w której się trzeba non stop uczyć!
R: Tak.
B: Że jakby to co umiesz dzisiaj to za chwilę będzie, no może nie że przestarzałe ale za chwilę będą nowe rzeczy które też trzeba znać. Czy ty też masz takie wrażenie albo czy to rzeczywiście tak jest z twojego doświadczenia?
R: Tak, też tak uważam, to znaczy ja tego zupełnie nie widzę jak taką nie wiem, wadę czy że trzeba, tylko właśnie dużą zaletę że **można** cały czas, poznawać nowe rzeczy i cały czas się uczyć i że jest na to miejsce i że jest też wszyscy w tym środowisku widzą taką potrzebę i że dzięki temu to się rozwija i i naprawdę można z tego mieć duże **zyski**. Że **zna się** takie wszystkie nowości, jakby to, to też mi się podoba w tej pracy, że można się cały czas rozwijać.

(...)
B: A jakie masz, jakie masz dalsze plany zawodowe? Co byś jeszcze chciała robić?
R: Dalsze plany zawodowe? To na razie właśnie pracuję, pracuję w tej firmie i wszystko wygląda całkiem fajnie cały czas się rozwijam, i póki póki się rozwijam i póki widzę takie możliwości eee dla siebie że uczę się nowych rzeczy to na pewno będę chciała w tej firmie jeszcze pracować, nie wiem ile czasu to będzie. Po prostu, w momencie kiedy zobaczę że przestaję się rozwijać, że gdzieś tam staję w miejscu to będę chciała sobie szukać nowej pracy (...) także nie mam nic sprecyzowanego a mam bardzo dużo możliwości więc po prostu życie życie pokaże [pauza] co, co będzie mi w aktualnym czasie odpowiadać, ale na pewno zawsze się chcę rozwijać, to będzie główne kryterium czy czy chcę zostać w danej pracy czy zmienić czy szukać czegoś nowego. (...)
B: A czy myślisz że to ma jakiś sens, artykuły które się pojawiają że data science to jest w ogóle najlepsza branża na świecie, najbardziej seksowny zawód świata to już chyba z milion razy to przeczytałem w różnych miejscach, aaa, co o tym myślisz? Czy to ma jakiś sens czy to tylko dziennikarze sobie tak lubią pisać?
R: Na pewno takie tytuły są chwytliwe, też takie artykuły czytałam, a czy ma sens, noo właściwie to pewnie podobne rzeczy są co jakiś czas pisane o innej branży, więc to co rok co dwa lata jakaś inna branża zaczyna być modna, tam, nie wiem co tam było wcześniej, aleee myślę że to jest praca która może przynosić dużo satysfakcji i i jeśli ktoś tak **dobrze** pokieruje swoją karierą no i nie wiem, tak właśnie jest rzeczywiście zainteresowany tym, to myślę że naprawdę to może być taka dobra praca do której się przyjemnie przychodzi, właśnie przez to że nie jest to monotonna praca, że można się cały czas rozwijać, że wymaga takiego **myślenia** i no różnych zdolności, i i jest też dosyć dobrze płatna więc to też na pewno jest plus, eeee i to tak myślę że rzeczywiście to może być jeden z takich fajniejszych, ciekawszych zawodów na świecie. A czy najlepszy to już nie mnie oceniać no bo nie mam doświadczenia w innych zawodach więc nie mogę powiedzieć.
{wywiad 4}

Stykając się w wywiadach wyłącznie z wypowiedziami łączącymi ciekawość, samorozwój i samodzielność w sposób wyżej przedstawiony, przez wiele miesięcy sądziliśmy, że badany świat DS składa się tylko z uczestników „rzeczywiście zainteresowanych" {wywiad 4}, dla których praca jest pasją lub co najmniej hobby. Twierdzimy jednak, że choć ciekawość, samorozwój i samodzielność są wartościami świata DS, to są uczestnicy nie będący tak pojętymi „zainteresowanymi" DS. Wskazywano nam, że istnieje kilka poziomów takiego bycia zainteresowanym, z których najwyższy to pasjonat, i że do takich właśnie osób mamy w naszych badaniach etnograficznych dostęp, bywając na meetupach i konferencjach. Podkreślmy jednak, że nawet w przypadku uczestników, dla których stabilna praca jest ważniejsza od samorozwoju, „nie jest tak że oni w ogóle nie inwestują w rozwój, ale no jakby widać dysproporcje [w porównaniu z pasjonatami]" {wywiad 15}. Zatem **uczestnik świata DS jest zobowiązany wobec tego świata do ciągłego samorozwoju**:

B: To nie, przepraszam wiem, to jeszcze mam, jeszcze mam krótkie pytanie. Czy mógłbyś mi poradzić z kim jeszcze powinienem porozmawiać, czy może gdzie powinienem się wybrać, żeby to środowisko data science lepiej poznawać?
R: Mmm no nie wiem, a czy w waszym mieście jest może jakaś grupa taka meetupowa, która

się spotyka?
B: Jest, jest, jest. Teraz jest taka grupa która się nazywa [nazwa meetupu].
R: Ok. No to to, to wydaje mi się że to są najfajniejsze miejsca, bo one [krótka pauza] w jakiś sposób, no to też nie jest w sumie dobre, bo to nie jest reprezentatywna próbka tego co się dzieje na rynku tak, tylko to właśnie zrzesza tylko takich pasjonatów, którzy kurde jeszcze w czasie wolnym przyjadą ci na spotkanie, żeby gadać o [krótka pauza] tabelka i modelach. Eee [krótka pauza] bardziej mnie zastanawia teraz jak tu dotrzeć do ludzi którzy przychodzą i traktują to jak robotę, tak jak są programiści pasjonaci, a są tacy którzy przyjdą, poklepią iii, trochę kodu i wracają do domu, nie myślą o tym programowaniu. Zastanawiam się jak dotrzeć do takiej grupy ludzi, nie? Bo też mamy takich w zespole, znaczy, może nie jest tak że oni w ogóle nie inwestują w rozwój, ale no jakby widać dysproporcje, nie? Są tacy którzy chodzą na te meetupy, a są tacy dla których ważniejsze jest, żeby wybudować dom iii spłodzić potomstwo, a nie siedzieć na spotkaniu, z bandą innych przeważnie facetów, pić piwo i rozmawiać o modelach, nie? Emm to jak do takich dotrzeć, to nam się trzeba zastanowić, a to może być trudne.
B: Rozumiem. Dobrze, no to ja też będę pytał w takim razie, o takich ludzi [z uśmiechem].
R: Kurde no, teraz teraz zagwozdka no jakby tutaj, jak się do takich dobrać. No bo w pewnym sensie jeżeli na przykład wybierasz ludzi [krótka pauza] zzz konferencji czy innych spotkań, no to grupa z punktu widzenia statystycznego jest zaburzona, nie? Tak samo jak badania ankietowe, to są zawsze ludzie którzy są skłonni wypełnić twoją ankietę, no to
B: No tak masz rację.
R: Więc jak robisz [krótka pauza] yy ankietę wśród miłośników psów, no to wiadomo że to jest już zaburzona grupa tak, i i i badasz, badasz to co ludzie myślą o prawach zwierząt na przykład i rozwieszasz to na forum psiarzy, no to wiadomo że będziesz miał zaburzone te. I podobnie może być tutaj kurde no, i teraz jak się dobrać do ludzi którzy się niechętnie pokazują w środowisku [krótka pauza] a pracują w tym, no nie wiem.
B: A mogę cię namówić na taką spekulację, żebyś ocenił jak to jest, w populacji data scientistów ilu jest zaangażowanych, a ilu jest mało zaangażowanych?
R: No to czekaj, powiedzmy że tam spojrzę na nasz zespół
B: Ok.
R: Iii, no ciężko, bo my mamy teraz takich dwóch studentów [krótka pauza] Albo nie, mamy 20% studentów
B i R: [śmiech]
R: Eee no to, to oni są jeszcze tacy zaangażowani, ciekawi świata [krótka pauza] eeem mamy 30% osób które już właśnie idą w tą stronę, stabilnej pracy i nie szukają jakby, specjalnie rozwoju nie wiadomo jakiego, tylko jest im dobrze u nas w firmie.
B: To zostaje nam 50 prorozwojowych?
R: Mmm, nie, nie. Też myślę że mógłbym, wyróżnić [krótka pauza] myślę że też można, yyy rozbić na [krótka pauza] dwie grupy. I jedna to by była taka właśnie która cały czas szuka, szuka nowości, chodzi na spotkania, angażuje się w ogóle społecznie do tego stopnia, że myśli nad jakimiś projektami, które można by robić czy to w ramach stowarzyszenia, czy właśnie w ramach firmy, ale jakieś takie właśnie przełomowe rzeczy, jakby gdzie się angażuje zewnętrze firmy do tego i tak dalej. A są tacy że są pomiędzy właśnie, tym [krótka pauza] ustatkowywaniem się, a jeszcze potrzebą rozwoju i szukania nowości, nie? [pauza] Także tutaj no.
B: Ok. A siebie do którego klastra zaliczył?
R: No ja jeszcze jestem z tych takich wariatów, którzy nie myślą o stabilizacji iii, jeszcze myślę przez kolejne tak, nie wiem 3 do 5 lat to jeszcze będę, będę tym który chodzi ii, chodzi na spotkania, interesuje się i w sumie stabilizacją jeszcze nie interesuje. (...) No, ale też gdzieś się tam udzielam, chodzę na te spotkania, rozmawiam z ludźmi, którzy są głębiej w temacie, żeby się gdzieś tam uczyć cały czas. Więc ja jestem jeszcze tym pasjonatem [krótka pauza] zobaczymy co to, na jak długo mi jeszcze starczy takiej energii, żeby to robić. {wywiad 15}

W powyższym fragmencie Rozmówca uzasadnia swoją chęć samorozwoju tym, że „nie myśli o stabilizacji" {wywiad 15}, niemniej zobowiązanie do samorozwoju dotyczy w ten sam sposób uczestników bardziej nastawionych na stabilizację np. w sensie życia rodzinnego:

> *B: Mhm, iii a jakie masz dalej plany zawodowe?*
> R: [pauza] Ja to jestem w takiej sytuacji że mam rodzinę i dwójkę małych dzieci także dla mnie najważniejsza jest stabilizacja, rozwijanie oczywiście swoich kompetencji bo myślę że ta ścieżka kariery wiąże się z tym że **człowiek musi się uczyć** po pierwsze chce po drugie musi albo odwrotnie musi a potem chce, ee, i nie wyobrażam sobie żeby ktoś, w tej branży nagle stwierdził OK już umiem wszystko i już nie uczę się. Także na pewno rozwój, eeee, nowe narzędzia które gdzieś tam przyspieszają ulepszają, możliwości i właśnie tego typu **konferencje** [rozmawiamy w trakcie konferencji] gdzie można poznać **zainspirować** się tym jak to robią inni [pauza] tak. Na razie nie myślę o zmianie ale nie, nie wykluczam znaczy, teraz mam taką sytuację że mam **bardzo** blisko do pracy [śmiech] mam dość **fajne** warunki, ale nie ukrywam że jeżeli ktoś zaproponowałby mi jakieś lepsze **wyzwanie** i, yyy, lepsze warunki finansowe to, no, myślę że w przypadku kiedy firma miałaby potrzebę restrukturyzacji to nie miałaby wyrzutów sumienia żeby [pauza] się mnie pozbyć także [pauza] {wywiad 3}

To przekonanie o konieczności ciągłego rozwijania się może być wzmacniane w badanym świecie przez to, że uczestnicy świata DS znają lub sami wykorzystują możliwości automatyzacji pracy ludzkiej, co może wywoływać poczucie bycia zagrożonym tzw. bezrobociem technologicznym.

> B: *Okey, aa czy teraz ee też musisz się jeszcze uczyć różnych rzeczy? Czy już wszystko umiesz?*
> R: Nie [śmiech] nie, nie. Jakbym uznała, że wszystko umiem to by był koniec. {wywiad 17}

Przemysław Biecek podczas wystąpienia[246] na warszawskim Data Science Summit 2019 stwierdził, że najważniejszymi obecnie trendami w DS jest właśnie automatyzacja modelowania (autoML) oraz wyjaśnialność modeli (XAI) {obserwacja 46}.

Zgodnie z centralną kategorią efektywności **uczenie się rzeczy nie związanych z właśnie wykonywanym projektem może być uznawane za bezsensowny wysiłek**, bowiem nie niesie żadnej wartości praktycznej, zajmuje czas, który można by przeznaczyć na osiągniecie konkretnego rezultatu, a w dodatku ta nie użyta od razu w praktyce wiedza najpewniej zostanie szybko zapomniana i nawet jeśli byłaby potrzebna np. za rok, to i tak trzeba będzie się uczyć tego samego ponownie. W ten sposób niektórzy uczestnicy badanego świata mogą kwestionować sensowność takich działań, jak np. czytanie podręcznika, udział w kursie MOOC, udział w warsztatach na konferencji, kontynuowanie studiów drugiego stopnia lub podjęcie studiów podyplomowych.

246 Prezentacja dostępna publicznie na https://www.slideshare.net/PrzemekBiecek/xai-or-die-at-data-science-summit-2019-149824427, dostęp 19.05.2020 r.

Twierdzimy, że osiągnięcie pewnego poziomu kompetencji technicznych i metodologicznych połączone z przekonaniem o wartości samorozwoju oraz połączone z przyzwyczajeniem do podejmowania samodzielnego wysiłku może prowadzić część uczestników badanego świata do **przekonania o możliwości skutecznego nauczenia się niemal wszystkiego**. Jest to odmiana poczucia sprawczości, która polega na przekonaniu podmiotu o byciu zdolnym do zrozumienia lub zrobienia, co zechce. Sądzimy, że sama owocna nauka pisania kodu może sprzyjać kształtowaniu się takiego przekonania u osób uczących się, a jednocześnie nastawienie na samodzielność i poczucie własnej sprawności może sprzyjać osiąganiu sukcesów w pisaniu kodu.

Kodowanie może być postrzegane przez laików jako umiejętność trudno dostępna, wymagająca elitarnych kompetencji umysłowych:

> Kolejny mitem jest mit programisty geniusza. W tym micie programują tylko osoby, która ma nadludzką pamięć oraz wyróżniającą wiedzę matematyczną. Oczywiście, takie cechy przydają się w programowaniu, ale nie są do niego wymagane. W programowaniu częściej od dobrej pamięci przydaje się umiejętność szybkiego znalezienia rozwiązania czy odpowiedzi na problem w internecie. Programista nie musi znać na pamięć setek różnych poleceń i funkcji, ważne że umie je zidentyfikować. Natomiast zamiast głębokiej wiedzy matematycznej do większości zadań programistycznych wystarczy podstawowa znajomość algebry. Z tym mitem wiąże się też inna kwestia - założenia że ten programista geniusz posiadł całą wiedzę programistyczną. Podobnie jak nauka języka obcego, nauka języka programowania wymaga dużo pracy i czasu. Dodatkowo, języki programowania czy techniki programowania zmieniają się znacznie częściej niż języki naturalne, dlatego też częścią programowania jest ciągłe uczenie się (Nowosad 2019:15).

Wyraźnie widzimy, że Nowosad jako antidotum na ten mit proponuje samodzielne wyszukiwanie informacji w internecie i ciągłe uczenie się. Jedna z Rozmówczyń wskazywała, że programowanie wydawało jej się czymś tajemniczym, magicznym, a środowisko ludzi potrafiących to robić odbierała jako hermetyczne:

> *B: Mhm, mhm, ok. Ale to, mmm [pauza] może tak zapytam, a dlaczego myślałaś że się nie nadajesz do programowania? [krótka pauza] Czy że to jest nie dla ciebie?*
> R: Wiesz co bo nie miałam nigdy wcześniej możliwości żeby spróbować, faktycznie, eee w ramach w ogóle całej, całego procesu edukacji, nie miałam takiego momentu żeby ktoś mi po prostu pokazał, czy w jakiś sposób mnie zachęcił, żebym miała z tym styczność na tyle żeby, żeby się przekonać. Jedyny obraz, który gdzieś się pojawiał, no to najczęściej kolegów właśnie z, czy na wcześniejszych etapach edukacji tak, z nie wiem, z klas o profilu mat-fiz, którzy jakieś tajemnicze rzeczy pisali w tym kodzie, i to było takie bardzo hermetyczne środowisko, i z tego względu mmm nie mając żadnego doświadczenia, nie znając w zasadzie nikogo bliżej ktooo [krótka pauza] kto by się tym zajmował, ale też kto chciałby się podzielić tą wiedzą w przystępny sposób, eee bo to też chyba jest ważne, na ile eee ktoś chce mmm pokazać programowanie jako eem no taką przydatną umiejętność, a nie, a nie tylko [śmiech] wiesz, robienia do tego takiej aury, aury tajemniczych magicznych sztuczek, bo trochę też też też czasem tak bywa. No to myślę że ta kwestia dostępności przede wszystkim, bo potem jak, bo tutaj zaczęłam mówić o tym że zaczęłam wykorzystywać Pythona w zasadzie żeby trochę

> ułatwić sobie pracę, ale to jeszcze nie jest analiza. I potem mnie zaczęło zastanawiać, no to w takim razie jakby to wyglądało eee [krótka pauza] eem gdyby nie korzystać z tego softu z którego korzystam, na przykład z SPSSa (...) gdybym nie musiała ręcznie tylu rzeczy wyszukiwać w SPSSie to jakby to miało wyglądać. I eee napisałam do takiego profesora na, który na [nazwa wydziału matematycznego] prowadził analizę danych w eR, to tam miało inną nazwę, mniejsza z tym, czy mogłabym po prostu pochodzić jak wolna słuchaczka na te jego zajęcia i zobaczyć z czym to się je. No i oczywiście, eee on jako pedagog z powołania [z uśmiechem], bo naprawdę naprawdę zajęcia prowadził niesamowite, zgodził się yyy i chodziłam na jego wykłady i laboratoria. I on w ramach tych swoich zajęć, eee wprowadzał bardzo wiele metod z tego, z tego co nazywamy tak, uczeniem maszynowym, więc bardzo szybko przy okazji tego podłapywania eRa, podłapywałam ten sposób myślenia który się wiąże z budowaniem tych modeli, z testowaniem modeli i tak dalej. No i generalnie po takim intensywnym semestrze chodzenia na te zajęcia, czytania na ten temat, to po prostu wsiąkłam, zwłaszcza że, on miał taki sposób prowadzenia tych zajęć, taki narracyjny, to profesor [nazwisko profesora], eee [śmiech] on opowiadał o tych analizach w taki sposób że że człowiek zaczynał wierzyć że w zasadzie nie ma nic piękniejszego co można by robić w życiu tylko właśnie analizować te dane [śmiech] eee no w każdym razie, to tak mocno rozbudziło moje zainteresowanie i już potem zaczęłam, no dużo sama czytać na ten temat, eee dłubać sobie kolejne, znaczy robić tak kolejne kursy online i przymierzać się do tego jak, w jaki sposób mogłabym [krótka pauza] za pomocą kodu przejść te kolejne etapy analiz. {wywiad 21}

W powyższym fragmencie widzimy wyraźną rolę znaczącego innego – profesora, który prowadził analizę danych w R i przyjął Rozmówczynię jako wolną słuchaczkę. Niemniej nauka pisania kodu odbywała się z własnej inicjatywy Rozmówczyni, a zajęcia rozbudziły zainteresowanie i ukierunkowały dalsze samodzielne wysiłki.

W świecie DS ceniona jest samodzielność także jako cecha opisywana w psychologii jako **wewnątrzsterowność** lub wewnętrzne umiejscowienie kontroli (*internal locus of control,* LoC). Sądzimy, że jest to związane, po pierwsze, z tym, że świat społeczny DS jest młodą i niezinstytucjonalizowaną całością społeczną, a po drugie, ze specyfiką działania uczestników świata DS. W pierwszym przypadku chodzi o to, że nie ma jednoznacznej instytucjonalnej ścieżki zostania data scientistem, nie ma uznanego, formalnego sposobu zaświadczania o kwalifikacjach, a osoby, które obecnie określają się jako data scientists czy zajmujące się data science mogą mieć różne wykształcenie i doświadczenie zawodowe (w których niekoniecznie pada nazwa „data science"). Tym samym w badanym świecie zdecydowana większość uczestników to samoucy, a dopiero nieliczne osoby rozpoczynające karierę w 2018 – 2019 roku mogą być absolwentami studiów o profilu data science:

> R: Jest atmosfera takiej burzy mózgów myślę bardzo, niekończącej się no wynikającej też z charakteru data science, a że tam są sami data scientiści, no to nikt się jakby nie ukrywa, nikt się nie kryje z tym że nie potrafi przewidzieć co zrobi za dwa dni i po prostu jak jest przyblokowany to siadamy tu i teraz i rozkminiamy ten temat i każdy się podłącza, a przynajmniej część. No i to są osoby które są z różnych dziedzin wiedzy jakieś, nie wiem fizyka, jakaś astronomia i tak dalej i tak dalej, a część osób jest też po moim kierunku [ekonomicznym – przyp. RŻ], a ktoś tam jest po jakimś biochemie i tak żeby, dużo około

> ścisłych rzeczy, ale nikt stricte z data science, dopiero jest jeden chłopak który to studiował, więc jest **farciarzem**, bo studiował rzeczy które my sami musieliśmy **zdobywać** z internetu. I to też jest fajne, bo jest bardzo taki, znaczy to jest takie środowisko bardzo zdywersyfikowane przez to, więc fajnie bo każdy może coś ze swojego doświadczenia wyciągnąć. {wywiad 10}

Samodzielne uczenie się narzędzi technologicznych jest szczególnie ważne w przypadku osób wywodzących się z matematyki lub nauk przyrodniczych, nie mających wykształcenia informatycznego. Typowym scenariuszem, właściwie nie tylko dla takich osób, jest uczenie się nowych narzędzi na potrzebę konkretnego projektu czy problemu:

> *B: Dobra yyy a mmm jestem ciekaw jak się nauczyłeś programowania jako biolog? Bo rozumiem że to jest bardzo potrzebne w pracy data scientista?*
> R: Jest. Znaczy powiedzmy od razu że yy bardzo częstooo to są mylone koncepty, my nie jesteśmy programistami.
> *B: OK*
> R: To jest moje zdanie. Ja siebie nie uważam za rasowego programistę. Bo programowanie jest mega szerszym konceptem. **My** tak naprawdę używamy języków programowania, do bardzo specyficznego celu jakim jest analiza danych. Więc [pauza] Python czy Java yy mogą zbudować ogromne aplikacje z interfejsem użytkowników, a my naprawdę używamy tych języków do w bardzo wąskim celu. Więc czy ja bym powiedział że jestem programistą? No coś tam w życiu zaprogramowałem, natomiast nie są tooo, pewnie gdyby to zobaczył rasowy informatyk programista to by się zaśmiał że to jest mega podstawowe bo kod ma kilkadziesiąt linijek, tak naprawdę, i bardzo specyficznieeee działa, yyyy więc to po pierwsze po drugie nauczyłem się sam poprzez zapotrzebowanie. To znaczy tak ja w ogóle jestem fanem technologii, **zawsze** byłem, byłem graczem i byłem koderem i się zawsze tam **bawiłem** technologiami, komputery zawsze mnie tam interesowały, więc miałem jakąś tam naturalną [pauza] nie miałem tej bariery żeby powiedzieć o niee, to jest trudne. No i jak się jak się uczyłem analizy danych, to chciałem robić to coraz mocniej ciekawsze rzeczy dla mnie wymagały ode mnie coraz mocniejszych maszyn, które musiałem sobie sam zaprogramować postawić, stworzyć bo na przykład instytut mówi słuchaj no my **nie mamy** klastrów, mamy tu jakieś komputery, jak chcesz to spróbuj. No po prostu musiałem się wtedy nauczyć też i trochę informatyki, pod konkretne zapotrzebowanie tak?
> *B: No tak.*
> R: No i to w ten sposób zacząłem się uczyć eRa, zacząłem się uczyć Pythona, zacząłem tam o Julii, później się okazało że wiele danych siedzi w bazach danych, tak? SQLowych. To żebym ja te dane mógł zanalizować to mam dwa wyjścia mogę łazić do IT i wiecznie się prosić o dane, a mogę też **sam** się tego nauczyć i oni mówią sobie tu masz bazę weź tu sobie korzystaj, tak? Nie zawracaj nam głowy. Co też chętnie uczyniłem [z uśmiechem]
> *B: [śmiech]*
> R: I nauczyłem się SQLa, tak? Na potrzeby też analizy danych. Więc są to, **na pewno** to trzeba powiedzieć, że pewne zacięcie pewne rozumienie jak działają systemy IT, jak działają sieci, jak działają komputery, jak działają bazy danych te wszystkie paradygmaty te wszystkie założenia trzeba znać, faktycznie trzeba to w tym się **poruszać**, bo faktycznie to co my robimy jest 100% zależne od IT. Tak technologicznie. Tego się na kartce nie zrobi. Więc, eee natomiast yyy **ja** jestem samoukiem, i uważam żeeee do robienia data science ten poziom wystarcz jest wystarczający, tak? {wywiad 6}

Po drugie, specyfika działania badanego świata DS sprawia, że wewnątrzsterowność i wewnętrzne LoC jest w tym świecie pożądane – pozytywnie będzie oceniana w DS osoba, która samodzielnie osiąga rezultaty, samodzielnie definiuje cel lub cele szczegółowe, nie

czekając na polecenia, z własnej inicjatywy zadaje pytania, szuka odpowiedzi, jest gotowa podejmować decyzje o nowym kierunku eksperymentów, a przede wszystkim ogólnie wychodzi z inicjatywą, nie czeka na zestaw gotowych instrukcji.

5.4.4. Ciekawość

W języku polskim słowo „ciekawość" oznacza po pierwsze, **ciekawość jako cechę człowieka –** bycie ciekawy czegoś, zainteresowanym czymś, także bycie tzw. ciekawskim lub dociekliwym. Po drugie, oznacza **ciekawość jako cechę przedmiotu –** coś jest ciekawe lub interesujące. Twierdzimy, że w świecie DS ciekawość ceniona jest we wszystkich tych znaczeniach.

Ciekawość jako cecha człowieka jest tym, co wskazywano nam jako charakterystykę łączącą wszystkich uczestników świata DS. Rozmówca uważa, że DS to nie tylko zawód ,ale typ myślenia (co było dla nas markerem do szerszego niż kategoria zawodu czy pracy skonceptualizowania badanego wycinka rzeczywistości społecznej, a ostatecznie przyjęcia ramy pojęciowej społecznych światów):

> R: Ale znaczy jak na razie to pan się pyta o moją pracę, ale to, czy to ma wpływ na to jakim my jesteśmy plemieniem?
> *B: No pewnie! Mmm, ja myślę, że tak, ja myślę, że tak, bo, mmm, mnie się wydaje, że jakby to jest plemię takie jakby to powiedzieć pod sztandarem zawodowym, które się schodzi, tak? Bo nie wiem czy jest coś innego co was łączy, czy to jest tak, że praca was łączy, nie? Że data scientist to jest zawód, tak bym to nazwał, tak?*
> R: Ale też typ myślenia na przykład. To znaczy myślę, że taką istotną cechą, to znaczy każdy ma ją w różnym zakresie jest ciekawość. U nas w zasadzie nie ma ludzi nie-ciekawych, każdego tam coś ciekawi [pauza] no mamy swój nerdowski humor [pauza] mamy pewne specyfi [R wskazuje na swój T-shirt] no sam pan widzi jak ja wyglądam, aaa, w garnitury się ubieramy na śluby i na pogrzeby, tak? No i tego typu, tak? Mmmmm [pauza] duża część myślę z nas jednak gdzieś tam błądzi myślami tak mocno i nawet podczas rozmów kwalifikacyjnych wychodziło. Rekrutowałem też też dosyć sporą ilość informatyków [pauza] no to mam wrażenie że jednak ta grupa data scientistów jest **jeszcze** bardziej specyficzna niż informatyków. {wywiad 2}

Bycie ciekawym jest rozumiane na co najmniej dwa sposoby, które zasygnalizowaliśmy wyżej:

- ciekawość czegoś, zainteresowanie czymś – tym czymś jest DS, ewentualnie dziedziny jej bliskie jak ML, programowanie, matematyka. Tak pojęta ciekawość może być postrzegana w badanym świecie jako coś, co sprzyja samorozwojowi w tym sensie, że osoba „rzeczywiście zainteresowana" DS (por. część 5.4.3.) rozwija się i uczy nowych rzeczy wiedziona przede wszystkim ciekawością, czyli własną motywacją wewnętrzną. W konsekwencji to **bycie zainteresowanym jest w badanym świecie**

strategią potwierdzania autentyczności, czyli legitymizacji siebie jako prawdziwego uczestnika świata DS. Negatywnie są w świecie DS oceniane motywacje własnych działań w rodzaju „uczę się metody X i dzięki temu zatrudnię się w firmie A, gdzie zarobię dwa razy tyle co obecnie". Raczej wypada wskazać na motywacje typu „słyszałem o metodzie X i tak mnie zaciekawiła, że wczoraj pół nocy nie spałem i czytałem o tym, a dzisiaj w pracy znalazłem nowy pakiet właśnie do tego i już nie mogę się doczekać, aż w domu to uruchomię".

- ciekawskość lub dociekliwość – rozumiane jako niezadowalanie się łatwymi, powierzchownymi odpowiedziami i wyjaśnieniami, dążenie do poznania szczegółów, doprecyzowywanie i uściślanie, a jak sądzimy także chęć do zrozumienia, eksperymentowania, szukania, zgłębiania, rozpracowywania problemów. Podkreślmy, że nie tyczy się to każdego i zawsze działania w badanym świecie, bowiem w odniesieniu do centralnej kategorii efektywności w pewnych sytuacjach dociekanie szczegółów będzie oceniane negatywnie jako strata czasu lub energii umysłowej. Zdaniem niektórych uczestników świata DS tak rozumiana ciekawskość czy dociekliwość jest cechą indywidualną, bez której nie można być uczestnikiem badanego świata. **Osoba zajmująca się DS musi zatem legitymizować się zarówno chęcią jak i zdolnością do zadawania pytań**. Część uczestników badanego świata może uważać, że DS wymaga także pomysłowości, kreatywności, twórczego czy ogólnie nieszablonowego podejścia do działania. Początkowo wydawało nam się to niezwiązane z ciekawością, jednak sądzimy, że wskazuje to zarówno na wartości ciekawości jak i samodzielności oraz sprawczości w badanym świecie w tym sensie, że pozytywnie ocenia się tworzenie nowych rozwiązań (technicznych i metodologicznych), podejmowanie nierozwiązanych czy po prostu słabo zdefiniowanych problemów, zaś negatywnie rozwiązania szablonowe i wtórne, dobrze rozpoznane – takie traktowane są jako nieciekawe (co dalej rozwiniemy).

Przyjrzyjmy się, co o zadawaniu pytań napisał Wickham w podręczniku dla początkujących data scientistów używających języka R:

> Kluczem do zadawania dobrych pytań doprecyzowujących (*follow-up questions*) będzie poleganie zarówno na swojej ciekawości (*curiosity*) – o czym chcesz dowiedzieć się więcej? – oraz na sceptycyzmie (*skepticism*) – jak może to być mylące? (Wickham i Grolemund 2017).

Jedna z Rozmówczyń, osoba na stanowisku menedżerskim, uznała ciekawość – którą rozumie zarówno jako sprzyjające samorozwojowi i samodzielności bycie zainteresowanym

DS, jak i jako chęć do zadawania pytań i zrozumienia problemów biznesowych – za element ważniejszy niż wiedza i doświadczenie, przy zatrudnianiu na najniższe (tzw. junior) stanowiska data scientist:

> *B: (...) jakby musiał wyglądać taki kandydat na junior data scientista twoim zdaniem, żeby on miał jakiś sens?*
> R: Wiesz co, zacznę tak naprawdę od tego, że na pewno musi być to osoba ciekawa. Znaczy w sensie ciekawa rozwiązywania problemów. Musi mieć ciekawość, o.
> *B: Cechuje ją ciekawość?*
> R: Tak, cechuje ją **bardzo** wysoka ciekawość. Umiejętności, no wiadomo, podstawy programistyczne musi mieć, ale taki junior to wiesz, to wszystko jest do ukształtowania. Ja w ogóle uważam, że lepiej wziąć osobę, która może ma mniejszą wiedzę, ale większy zapał niż większą wiedzę, ale tak wiesz, trochę mi się nie chce już bardziej rozwijać. No i powinien mieć według mnie data scientist chęć zrozumienia problemu biznesowego, czyli rzeczywiście wychodzenie od hipotez biznesowych, i to jest taki najlepszy kandydat. No fajnie, jeśli ma już jakieś doświadczenie, ma ileś kursów przerobionych, wiesz, na juniora nie musi być wymagane te wdrożenia produkcyjne, bo to jest junior, ale na pewno znajomość algorytmów, czyli te wszystkie podstawy, jeśli chodzi o teorię, no to musi znać. Musi wiedzieć, gdzie, jak ich używać, do czego. Tylko pamiętajmy, to jest cały czas junior, więc ważne, żeby miał podstawy, chęci i zapał do zrozumienia biznesu, a nie tylko do klepania kodu, że tak powiem, no bo wtedy to niech idzie na data inżyniera i to też jest super. Ale przy data scientiście musi mieć ten aspekt biznesowy. {wywiad 22}

Znaczące jest w powyższej wypowiedzi to, że **DS to nie osoba „tylko do klepania kodu”** {wywiad 22}. Sądzimy, że ciekawość jako cecha indywidualna uczestników oraz jako cecha przedmiotu (DS czy ML jako ciekawe oraz tzw. „ciekawe projekty”, o czym dalej) jest w badanym świecie chętnie wskazywana jako to, co odróżnia DS od innych, raczej inżynieryjnych dziedzin IT – tam tylko klepie się kod. Uczestnicy świata DS postrzegają się jako bliższych nauce niż zwykłej, raczej nieciekawej z ich perspektywy inżynierii, może pojawiać się określenie R&D (*research and development*, dosł. badania i rozwój). Z tej perspektywy programowanie jako dziedzina – czyli tworzenie oprogramowania (*software development / engineering*) może być uznawana w świecie DS właśnie za coś zwykłego, nudnego, nieciekawego, nie zapewniającego samorozwoju, polegającego na szablonowym rozwiązywaniu dobrze zdefiniowanych problemów. DS przeciwnie, jest ciekawe i wymaga bycia ciekawym, żeby je robić:

> [rozmawiamy o tym, jakie osoby aplikują do zespołu DS, który prowadzi Rozmówca] (...) Zdarzają się programiści, ale rzadko udaje się nam dojść do porozumienia. Przy całym moim szacunku i sympatii do programistów i sam jestem człowiekiem IT, chyba podejście, myślenie jest troszeczkę inne. Yyyyy, mam wrażenie że programiści bardzo by chcieli i przyzwyczajeni są do takiego trybu, że dostają gotowe wymagania, od klienta, że ta aplikacja to ma robić to to to to to, i tamto. Tu ma być zielona i tu ma mieć guzik. I oni lubią pracować według tych wytycznych. Tak? Natomiast w **naszych** projektach problem jest często bardziej abstrakcyjny, i [cmoknięcie] z góry nie ma gotowej definicji tego problemu. I często widzę że też programiści borykają się z tym. Znaczy mam tutaj kilku też ludzi którzy są po IT, z wykształcenia, i widzę

> żeee, to jest po prostu borykają się z tym innym podejściem, mindset[247] nazwijmy to. (...) Ja jestem taką mam osobowość że raczej kontaktową, lubię ludzi lubię pracować z ludźmi dla mnie ludzie są taką główną częścią tego co robię, a drugie też jestem ciekawską osobą bardzo z natury, więc dobrze się czuję w sprawach, które nie są już wyjaśnione. Tak? Czyli **bardzo** lubię, jestem rybą mętnej wody. Się śmieją tutaj koledzy ze mnie w pracy że ja najczęściej biorę projekty które albo się nie udały, przynajmniej z dwa razy, albo nikt nie wie jak się za nie zabrać, albo w ogóle problem jest tak rozmytyyy, że pffffff, to są sprawy ale ja to dlatego też zostałem naukowcem przez chwilę byłem, tak? Z tej naturalnej ciekawości. Więc akurat te rzeczy które są u nas w zawodzie dla wielu osób wyzwaniem, to mnie cieszą, bo mam taki charakter. {wywiad 6}

Mówiąc o ciekawości jako cesze przedmiotu musimy zacząć od zrozumienia, **co uczestnicy świata DS uznają za ciekawe**. Już w powyższym fragmencie {wywiad 6} widzimy, że ciekawe są sprawy nie mające oczywistego rozwiązania: trudne do zdefiniowania, trudne w realizacji, niewyjaśnione. Zauważmy, że omawiana wypowiedź wyraźnie wskazuje także na wartość sprawczości – „ja najczęściej biorę projekty które albo się nie udały, przynajmniej z dwa razy, albo nikt nie wie jak się za nie zabrać" {wywiad 6}, co znaczy: inni nie dają rady realizować tak trudnych projektów, jak data scientist. Uczestnicy badanego świata chętnie postrzegają „coś ciekawego" jako wymagającą ich wysiłku intelektualnego zagadkę. Ciekawy może być przede wszystkim problem, w drugiej kolejności dziedzina (wymieniane wcześniej finanse i marketing będą w badanym świecie raczej uznawane za mało ciekawe), ewentualnie metody (szczególnie uczenie głębokie lub metody niestandardowe), ale raczej nie technologie. „Ciekawy projekt" oznacza jednoznacznie „taki, którym warto się zająć" i uczestnikom uzasadniającym w ten sposób np. zmianę pracodawcy inni uczestnicy raczej nie zadają kolejnych pytań o motywy decyzji. Ciekawy projekt to także taki, w którym konkretny uczestnik zgodnie ze zobowiązaniem do samorozwoju będzie się dalej rozwijać, zatem pewne przedmioty są uznawane za ciekawe/nieciekawe na poziomie indywidualnym – zarówno w odniesieniu do bycia zainteresowanym jakimś przedmiotem, jak i relacji własnych kompetencji do poziomu trudności przedmiotu. W tym sensie dla określonego uczestnika ciekawa może być praca w agencji marketingowej czy w banku. Niemniej twierdzimy, że w badanym świecie DS istnieją także przedmioty powszechnie uznawane za ciekawe. Jeżeli chodzi o domeny, to naszym zdaniem są to medycyna, pojazdy autonomiczne i bezpieczeństwo (wykrywanie wyłudzień/oszustw, bezpieczeństwo cyfrowe). Co do metod, to przynajmniej część uczestników za ciekawe uznaje uczenie głębokie (*deep learning*), zaś raczej powszechnie za takie uznane będą metody niestandardowe, czyli np. połączenie typowych modeli ML w jeden bardziej złożony asamblaż (*model ensemble*),

247 *Mindset* dosł. sposób / typ myślenia

wykorzystanie typowego modelu ML lub statystycznego w sposób inny niż typowy, a także użycie mało znanych metod statystycznych pochodzących z wąskiej dziedziny badań ilościowych (jak ekonometria, badania operacyjne, bioinformatyka, meteorologia) i zaadaptowanie ich do innej domeny. Za ciekawe mogą być uznawane dane nieustrukturyzowane (teksty, obrazy, dźwięk), które standardowo modelowane są za pomocą uczenia głębokiego. Nie spotkaliśmy się z określaniem technologii jako ciekawych, ewentualnie określano niektóre z nich jako fajne. Niemniej „fajne" jest mało znaczące poza tym, że wskazuje na pozytywną ocenę – jako fajne określano nie tylko technologie (np. klastry, GPU) ale i konkretnych pracodawców, ciekawe problemy, rozwiązywanie zagadek, korzystanie z technologii open source, możliwość pracy zdalnej, zatem rozmaite sprawy wskazujące na różne wartości badanego świata.

Jeden z Rozmówców, określający się jako data scientist i właściciel firmy DS, wskazywał, że najbardziej w swojej pracy lubi rozwiązywanie problemów, moment „kombinowania" {wywiad 5} nad rozwiązaniem zagadki (co nazywa przetłumaczeniem problemu na analitykę) oraz to, że każdy kolejny projekt jest niestandardowy. Jednocześnie ta niestandardowość i nierutynowość projektów sprawia, że trudno mu zwiększyć skalę działania firmy, bowiem nie może zatrudnić niedoświadczonych osób do wykonania nowych projektów według ustalonego schematu:

> *B: OK, chciałem spytać, aaa, co pan najbardziej lubi w tej pracy data scientista?*
> R: Hm! To jest trudne pytanie. [pauza] Chyba najbardziej lubię rozwiązywać problemy [pauza] nie takie że komuś nie działa komputer i trzeba mu załatwić tylko takie problemy w sensie problemu analitycznego czy biznesowego, i znowu to tłumaczenie, gdzie zdarza się, że problem wydaje się niewinny, a my siedzimy we **czterech** gdzie każdy ma doświadczenie powiedzmy od 10 do 40 lat [pauza] i próbujemy to przetłumaczyć na analitykę i nam nijak nie wychodzi. I to zawsze jest **fajne** samo to kombinowanie, jak się wchodzi w taki powiedzmy rodzaj flow, że rzuca się tymi pomysłami zastanawia się co dobrze a co źle to chyba to najbardziej lubię, mmm, ten moment, i **czasem** lubię to że w każdym projekcie robi się coś niestandardowego przynajmniej w tych projektach które my mamy, z drugiej strony tego też nie lubię, na tej zasadzie że ta moja część która jest powiedzmy [pauza] byłym naukowcem no to to to lubi, to jest fajne, [pauza] z drugiej strony jak chciałbym żeby firma bardziej rosła to chcę mieć takie projekty że się dogaduję z klientem i pach, się ten projekt się robi zawsze tak samo. A to się praktycznie nigdy nie zdarzy jeżeli się chce zrobić projekt na poziomie, to za **każdym** razem robi się coś inaczej, na tej zasadzie jakieś uwarunkowania biznesowe są inne, funkcja celu tak mówiąc technicznie będzie inna, jakieś tego typu rzeczy. I za każdym razem wygląda to trochę inaczej. No i właśnie, to jest fajne bo jest **przygoda** intelektualna a z drugiej strony jest niefajne bo, **utrudnia** to wzrost firmy że chciałoby się przekazać, ustrukturyzować sporo projektów i powiedzieć projekty takie to my robimy zgodnie z tym schematem, zatrudniamy sobie trzech świeżaczków i wy to róbcie.
> *B: Mhm, OK. Czyli czyli bardzo trudno a w zasadzie nie da się tak pracować schematycznie.*
> R: Są takie projekty, że się da, są takie że się nie da, zależy też czy się chce projekt zrobić dobrze. Bo tak generalnie jak chce się projekt zrobić dobrze [pauza] noo to trzeba pokombinować. Zazwyczaj. Jak chce się zrobić projekt **jakoś** [pauza] to sprowadza się w

> standardowy sposób do problemu analitycznego, jest to budowa modelu klasyfikacyjnego buduje się model rozmawia się o tym z klientem być może się koryguje być może nie i tyle. {wywiad 5}

Podkreślmy, że „jak się chce projekt zrobić dobrze no to trzeba [zazwyczaj] pokombinować" {wywiad 5}. Schematyczne podejście jest możliwe, ale Rozmówca, jak i wielu uczestników badanego świata ocenia to negatywnie. Zarówno jeżeli mowa o podejściu preferowanym przez firmę czy instytucję oraz pojedynczych uczestników, to **działanie schematyczne, rutynowe, może być oceniane przez innych uczestników jako coś, co w ogóle nie jest DS, a „zwykłą" inżynierią lub „zwykłą" analityką**. Wskazuje to nie tylko na wartość ciekawości w badanym świecie DS, ale także omawianego już samorozwoju (5.4.3.), oraz swobody (5.4.5).

> R: (...) Póki co to całe data science to jest **sztuka** w tym sensie, że są jakieś ogólne wytyczne jak to się robi, ale [pauza] nic nie jest na sztywno ustalone, jedynym w zasadnie kryterium czy czy dobrze robisz czy źle jest sprawdzić jak twój model się zachowuje więc duża część to jest takie trochę, takiego na chybił trafił, trochę słuchanie własnej intuicji [pauza] zupełnie nie jak programowanie. [pauza]
> *B: Mhm*
> R: Więc to jest fajne.
> *B: A czego pan nie lubi w tej pracy? [długa pauza]*
> R: Jak powiem klientów to to będzie źle odebrane [z uśmiechem]
> *B: Nie, ja się nie obrażę [z uśmiechem] i ani pana klienci też się ode mnie o tym nie dowiedzą*
> R: Nie nie klientów może nie. Ale to czego nie lubię to nie, nieokreśloności tego czym mamy się zająć. Czyli że jest sytuacja gdzie ja w zasadzie nie wiem co jest moim celem, tak? Dostaję informację zwrotną od klienta pod tytułem wiesz co, to jeszcze nie to tak? To myślę że to jest mniej więcej podobny stopnień, ee, wkurzenia jak ma grafik który dostaje informację że no ładne ale trochę jeszcze nie za ładne i zrób z tym coś. No to to jest mnie więcej ten poziom, jeżeli byśmy mieli **dobrze** opisane [pauza] co ma być aa co ma być zrobione i w jaki sposób no to było fajnie z drugiej strony gdyby nasi klienci wiedzieli co mają do zrobienia i w jaki sposób to by do nas nie przychodzili. Więc nie wiem czy to jest jakaś odpowiedź. {wywiad 2}

Tym samym samo istnienie niedokreślnych, trudnych czyli ciekawych problemów, które można zaadresować za pomocą danych cyfrowych i metod ich przetwarzania / analizy / modelowania może być w badanym świecie uznawane za legitymizujące istnienie świata DS jako takiego.

W kilku różnych badaniach twórców oprogramowania open source oraz hakerów wskazywano na ciekawość jako motywację do podejmowania i podtrzymywania ich działań. Zaród uważa, że w tym wymiarze hakerzy podobni są do naukowców według tzw. klasycznego etosu naukowca, opisanego przez Roberta K. Mertona (Zaród 2018:24–25, 288–89). Ciekawość jako motywacja do działania w świecie data science jest, jeśli nie dominującą

(czego nie badaliśmy), to **obowiązującą w tym sensie, że uczestnikom wypada powoływać się na ciekawość jako własną motywację** do uczestnictwa w świecie DS.

Wyraźnie widzimy związek wartości świata DS – ciekawości, samodzielności, samorozwoju, efektywności – z postmodernistycznym rozumieniem pracy (Haratyk, Biały, i Gońda 2017:154–56). Choć mówimy o świecie społecznym, w którym działanie podstawowe może, ale nie musi być wykonywane komercyjnie i może, ale nie musi być źródłem dochodu, to rozumienie „pracy" (czy raczej działania podstawowego) jako rozwinięcia / spełnienia cech życia jednostki (ibidem) oraz interpretowanie „pracy" jako kwestii wyboru oraz rozwoju osobistego lub rozwoju życiowego (zgodnie z postmodernistycznym imperatywem samorealizacji, opisanym m.in. przez Małgorzatę Jacyno (ibidem)) są zgodne z poglądami uczestników świata DS i znajdują wyraz w wymienionych wyżej wartościach.

5.4.5. Swoboda

Uczestnicy świata DS traktują swobodny wybór własnych narzędzi pracy raczej jako coś oczywistego. Negatywnie odnoszono się w naszych badaniach do sytuacji, w której narzędzia są odgórnie narzucane, np. przez szefa zespołu, firmę / instytucję zatrudniającą czy klienta. Podkreślmy, że wyjątkowo negatywnie uczestnicy badanego świata będą oceniali odgórne narzucanie narzędzi w postaci oprogramowania zamkniętego (*proprietart software*). Szczególnie tyczy się to podstawowego środowiska pracy DS, gdzie swoboda wyboru jest zawężona do dwóch dominujących języków programowania skryptowego – Pythona i R. Narzędzia typu SAS czy SPSS raczej nie są używane przez uczestników badanego świata z własnego wyboru. Zaznaczymy, że **swoboda i elastyczność są w badanym świecie uznawane za zalety zarówno R jak i Pythona jako takich**. Z jednej strony te zalety płyną z tego, że są to języki programowania, a nie oprogramowanie z klikanym interfejsem graficznym (GUI), zatem użytkownik jest znacznie mniej ograniczony w dostępnych opcjach. Z drugiej strony są to narzędzia open source, zatem istnieje swoboda w stosowaniu ich do rozmaitych celów, swoboda modyfikacji i budowania rozszerzeń (przede wszystkim w postaci pakietów), a te rozszerzenia rzecz jasna zwiększają elastyczność, a także możliwości (zatem i sprawczość).

Spotkaliśmy się z praktyką podkreślania tego, że używany język programowania jest językiem wyboru (to kalka z angielskiego *language of choice*) – osoba używa go dlatego, że chce, a nie musi. Jest to strategia nastawiona na wygodę i łatwość pracy w tym sensie, że człowiekowi pracuje się wygodnie za pomocą narzędzia, które mu odpowiada. Istnieje w

badanym świecie omawiany już przez nas postulat dobierania właściwego narzędzia do problemu, nie kłóci się to zupełnie ze swobodą wyboru narzędzia, bowiem biegłość i sympatia np. co do języka Python u konkretnej osoby może być traktowana jako jedno z kryterium dopasowania narzędzia. Ostatecznie ta swoboda wyboru narzędzi ma rzecz jasna służyć efektywności. Zatem choć to, jakimi narzędziami posługują się inni uczestnicy ma dla indywidualnego uczestnika znaczenie przy wyborze narzędzia (co opisywaliśmy m.in. jako wyraz efektu sieciowego por. 5.3.1.), to zdecydowanie pozytywnie oceniana jest w badanym świecie sytuacja, kiedy uczestnik może wejść do nowego projektu / zespołu / instytucji z wybranymi przez siebie narzędziami:

> *B: Mhm, OK. A teraz w nowej pracy będziesz czego używał?*
> R: No wszystko wskazuje na to że i eRa i Pythona. Znaczy, dostałem informację że, większość ludzi tam używa Pythona, ale że eRa też, znaczy są otwarci na to żeby każdy używał tego czego potrzebuje akurat w danej chwili, eee no ale no siłą rzeczy jest też trochę projektów w którym trzeba współpracować z innymi więc no jednak ten kod musi być w jakimś języku. Powiedzieli mi że akurat ostatnio dołączył też do firmy jakiś gość, który też od wielu lat w eRze pisze, więc no to jest akurat spoko, że na pewno ten eR jest tam reprezentowany i ktoś go używa. No, na przykład na rozmowie kwalifikacyjnej rozmawiałem z gościem który będzie na, no jest na takiej samej pozycji na jakiej ja będę, w sensie jako analityk w tym zespole, no i on mówi że wszystko robi w Pythonie, tak? Więc myślę że i z tego i z tego będę korzystał akurat. {wywiad 9}

Jako coś oczywistego w badanym świecie traktowana jest także swobodna komunikacja i ubiór. Nie jest to swoboda w sensie dowolności, a w sensie nieformalności. Zwracanie się per „Ty" przez dwudziestoletnich uczestników świata DS do dwukrotnie starszych uczestników z tytułami profesorskimi lub na stanowiskach kierowniczych jest traktowane jako norma, jak sądzimy – zgodnie ze zwyczajami angloamerykańskimi. Taka forma komunikacji obowiązuje zarówno na wydarzeniach świata DS, jak i generalnie pomiędzy uczestnikami świata DS.

> *B: (...) Rozumiem że wszyscy [w zespole DS] jesteście na ty?*
> R: Tak.
> *B: Nie ma czegoś takiego że proszę, proszę pana?*
> R: Nie, nie, nie.
> *B: Aaa do szefów jak się zwracacie?*
> R: Też na ty.
> *B: Macie jakiegoś prezesa?*
> R: Znaczy mamy [imię A], który jest [krótka pauza] CEO, tak? Ale no, na **ty**, normalnie.
> *B: Rozumiem że też jest na ty.*
> R: Tak.
> *B: Rozumiem żee, yyy [imię A] co ty pierdolisz, nie będę tego robić, to też jest normalne? [z uśmiechem]*
> R: [śmiech] Może nie co ty pierdolisz, ale można, można się z nim nie zgodzić, można się postawić bez żadnych problemów. To jest takie start-upowe środowisko jednak, więc jest raczej płaska struktura, dość luźna atmosfera, więc to jest bardzo fajne. Nie trzeba przychodzić w

garniturze, ani nic takiego.
B: A jak się trzeba ubierać?
R: No tak żeby ci było wygodnie [z uśmiechem]
B: Ok. A na spotkania z klientami? [pauza] Znaczy masz w ogóle spotkania z klientami?
R: Eee no teraz nie mam, ale no na spotkania z klientami tak, no trzeba jednak jakąś marynarkę może założyć czy coś takiego. Najczęściej z klientami są calle[248], bo mam najczęściej klientów z zagranicy, czy ze Stanów, więc, to są calle z nimi. Wtedy to w sumie nie trzeba się ubierać jakoś, marynarka tylko [ze śmiechem].
B: Spodnie mogą być od dresu?
R: Tak, tak. To jak na rozmowach kwalifikacyjnych jestem [które odbywają się przez komunikatory wideo, np. Skype], marynarka na górze, ale spodnie mam dresowe [ze śmiechem] {wywiad 8}

W powyższym fragmencie wyraźnie widzimy, że **swobodna komunikacja dla uczestników badanego świata oznacza, poza nieformalnością, także bezpośredniość i możliwość wyrażania własnego zdania**, np. niezgody w kontakcie z przełożonym w pracy. Swoboda jest w tym wymiarze bliska samodzielności, dotyczy bowiem m.in. podejmowania decyzji o kolejnych czynnościach we własnej pracy, organizacji czasu pracy, a także omawianego już doboru narzędzi. Nie każdy uczestnik badanego świata ma pełną decyzyjność w omawianych obszarach, ale tak pojęta swoboda jest traktowana jako coś cennego, szczególnie w kontekście zatrudnienia. Dalej powiemy o dość powszechnych praktykach elastyczności czasu i miejsca pracy.

Wartość swobody w relacji z przełożonym widzimy w badanym świecie także w negatywnym ocenianiu przez uczestników praktyk tzw. mikromanagementu. Mikromanagement ogólnie oznacza sposób zarządzania ludźmi bazujący na szczegółowej kontroli nawet drobnych aspektów pracy. Uczestnicy cenią raczej przeciwny styl zarządzania jak i współpracy – taki, gdzie każdy ma dużą dozę samodzielności i swobody, oraz odpowiedzialności za własne działanie, a także gdzie panuje raczej płaska struktura:

B: OK, a co ci się najważniej podoba w tym co robisz teraz?
R: [pauza] Najbardziej podobają mi się ludzie. Aaa, **w końcu** pracuję z ludźmi, z własnego wyboru, **nie chciałem** już **prowadzić** organizacji, szczególnie że było to w całkiem innej branży, ee, natomiast miałem tam cały czas do czynienia z ludźmi którzy **oczekiwali** że to ja im powiem co oni mają robić, gdzie to ja chciałem korzystać z ich potencjału z ich kompetencji żeby to **oni** mi powiedzieli cooo powinniśmy zmienić, szczególnie że to oni mieli na przykład na co dzień do czynienia z klientem. I to jest dość męczące kiedy okazuje się że ta cała odpowiedzialność spada, no właśnie, na kogo? Na kim powinno to spoczywać czy na wszystkich? Czy na jakimś wybranym fragmencie. Natomiast yyy w przypadku takiego zespołu gdzie jest nas kilkanaście osób codziennie mamy spotkania pracujemy yyy w pięciu miastach w czterech krajach także spotykamy się **tylko i wyłącznie** na Skype [pauza] i pracujemy. Nikt nikogo nie potrzebuje, nikt nie ma potrzeby żeby nikogo pilnować, stać na nim z **batem** kontrolować godzin przyjść i wyjść z pracy bo każdy wie co ma robić. Każdy chce to robić każdy chce to ulepszyć to co robimy, iiii jest ten taki **ferment** polegający na yyy ścieraniu się

248 Od słowa *call*, oznacza rozmowy telefoniczne, także przez internetowe komunikatory głosowe lub wideo.

> różnych idei pomysłów, yyyy, po prostu bycia **lepszym** w tym co się robi. To jest, i to jest coś, co [pauza] co jest **najlepsze** w tym wszystkim. W tej pracy tak. {wywiad 3}

Wiele treści zawiera w sobie końcówka powyższej wypowiedzi – widzimy w niej odniesienia nie tylko do omawianej wartości swobody, ale także do efektywności czy dokładniej do optymalizacji („każdy chce to ulepszyć to co robimy"), ciekawości i samodzielności („każdy wie co ma robić. Każdy chce to robić") oraz samorozwoju („jest ten taki ferment polegający na ścieraniu się różnych idei pomysłów, po prostu bycia lepszym w tym co się robi" {wywiad 3}).

Co do ubioru to zdecydowanie obowiązuje casual, osoba w formalnej marynarce / żakiecie mocno wyróżnia się na wydarzeniach w badanym świecie, garnitur / kostium wskazuje raczej na osobę spoza świata DS, dodatki formalne jak krawat, koszula ze spinkami, teczka czy buty ze skóry lakierowanej widzieliśmy wyłącznie na konferencjach biznesowych {obserwacja 22, 37}. Jak cytowaliśmy wyżej, w świecie DS należy ubierać się „tak żeby ci było wygodnie" {wywiad 8}. Niemniej w toku badań nie spotkaliśmy się z praktyką noszenia dresów. Jednocześnie w komercyjnym DS nie spotkaliśmy się ze sformalizowanym dresscodem. Nie istnieje także nieformalny dresscode badanego świata, świat ten jest dość różnorodny, choć można wskazać tendencje.

Po pierwsze, plecaki są znacznie bardziej popularne niż torby / teczki. Spotkaliśmy się głównie z używaniem plecaków o pojemności około 20 litrów, nierzadko modeli producentów dobrej jakości sprzętu turystycznego (np. Deuter, Vaude, Thule). Wielu uczestników badanego świata nosi na co dzień laptop przy sobie, stąd używanie plecaków i to takich z odpowiednią przestrzenią do bezpiecznego i wygodnego przenoszenia komputera. Plecaki noszone są tak do t-shirtów i jeansów, jak i do marynarek, sukienek. W rozmowach nieformalnych określano kobiecy zestaw składający się z casualowej sukienki i plecaka na laptop jako „stylówka z politechniki".

Po drugie, obowiązują dwa nierozłączne style ubierania się, dla uproszczenia nazywamy je „techniczny" i „schludny". Techniczny polega przede wszystkim na noszeniu T-shirtów z logotypami technologii lub wydarzeń o charakterze technicznym[249]. Utrwalonym w popkulturze elementem tego stylu ubierania się jest bluza z kapturem, jednak jest ona w świecie DS mniej popularna niż opisywane t-shirty, ponieważ jednoznacznie kojarzona z IT. Niemniej niektórzy uczestnicy badanego świata noszą takie bluzy, także z logotypami

249 Inną praktyką jest oklejanie pokrywy laptopa naklejkami o zbliżonej tematyce – robią to prawie wszyscy uczestnicy badanego świata DS, o czym więcej w rozdziale 6.

technologii, zdarza się też, że są to bluzy służbowe – z logotypem firmy, której uczestnik jest pracownikiem (szczególnie w roli oficjalnej reprezentacji firmy, np. prelegenta na wydarzeniu). Styl schludny składa się z raczej stonowanych elementów ubioru casualowego – gładkie t-shirty, koszule, bluzki, swetry, jeansy, sukienki, buty sportowe, buty skórzane – w porównaniu ze stylem technicznym jest bardziej dopracowany i zadbany w sensie połączenia kolorów czy dopasowania rozmiaru do sylwetki i jak nam się wydaje jakości odzieży. Połączenia w rodzaju: casualowa marynarka i t-shirt z zeszłorocznej konferencji są także w pełni akceptowane. Latem w pełni akceptowane są osoby noszące krótkie spodnie, sandały lub klapki. Nie spotkaliśmy się jednak z ubiorem eksponującym ciało ani w ogóle urodę w sposób seksowny, jak głębokie dekolty / rozpięte głęboko guziki, ubrania mocno przylegające do ciała (u mężczyzn i kobiet), wyraźny makijaż u kobiet. Nie ma praktyk noszenia wyraźnych czy wskazujących na wysoki status materialny ozdób, biżuterii, zegarków, jedynym dość popularnym dodatkiem do ubioru jest smartwatch / opaska fitness (poza oczywistym plecakiem z laptopem w środku). Techniczny styl ubierania może być w badanym świecie łączony z kategorią nerda i naszym zdaniem raczej z mężczyznami niż kobietami:

> *B: (...) czy kogoś kto się zajmuje data science jesteś w stanie rozpoznać po wyglądzie?*
> R: Niee, no w ramach populacji, no może inaczej. Jako chodzi ogólnie o typ osoby technicznej to jest często typ osoby którą się rozpoznaje z daleka.
> *B: Na czym on polega?*
> R: No to by trzeba zobaczyć, to są bardzo często jakieś koszulki z konferencji, z technologii, ogólnie taki t-shirt i tak dalej. A takie coś też takie że często to są takie z Aspergerem, z małym kontaktem z ciałem, tak chodzą tak tak właśnie [R przygarbia się i unosi barki] no tak dosyć sztywnie i tak widać **myśli** są są w siebie w myślach, no to może być programowanie dowolne czy tam nauka. Często to jest też styl wypowiedzi, ostatnio powiedziałem do jednej dziewczyny czy jest matematyczką może. Takich wyrażeń może, jednak **nie**, jest lingwistką ale to jest blisko taki styl wypowiedzi, styl wszystkiego. To często jest chodzenie, jak pojechaliśmy na konferencję data science z koleżanką to zgadywaliśmy sobie kto może mieć zespół Aspergera po tym jak mówi na scenie. I to była kwestia taka że osoby które [pauza] takie jak ktoś ma takie sztywne ruch albo taką mowę ciała albo przez całe nie zmienia pozycji tylko wiesz, tak samo, tak samo. Ale wiesz co ciekawe osoby które tak dynamicznie chodzą, to widać że to jest takie rytmiczne jakieś wyuczone, to jest tak często że osoby wiedzą o tym [pauza] i tak chcą **na siłę** to aplikować. (…) To nie jest że data science ma zawsze takie coś czy że tak wygląda, ale cechy data scientista to jest ciężko rozpoznać bo to jest w wielu miejscach podobne. Bo to też są różne teamy, różne osoby, bardziej matematyczne, bardziej naukowe, bardziej projektowe, bardziej startupy bardziej cośtam [pauza] data science jest bardzo ogólne. Są osoby jak artyści, są jak programiści, jak biznesowi, więc to jest takie, wiesz o co chodzi? Też są różni artyści, ale czasami widać że ktoś coś kreuje, ten styl jest taki przemyślany, niektóre osoby są takie po prostu schludnie ubrane, są osoby które przychodzą w ogóle wiesz w bluzie wyciągniętej, a to jest znany jakiś jeden z najlepszych specjalistów data science na świecie. {wywiad 26}

Po trzecie, z uwagi na ubiór nie byliśmy w stanie rozróżnić uczestników silniej i słabiej osadzonych w badanym świecie. Podobnie rzecz się miała ze stanowiskiem, branżą czy w ogóle obszarem działania uczestnika – wszyscy ubierają się raczej w zbliżonym, nieformalnym stylu. Sądzimy, że jest to kolejnym wyrazem traktowania swobody jako wartości w badanym świecie w tym sensie, że nie podkreśla się (przynajmniej w sferze wizualnej) hierarchii pomiędzy uczestnikami świata DS.

Jak mówiliśmy w części 5.3.4., osoby zajmujące się DS często pracują bez ściśle określonych godzin, mogą pracować zdalnie, a współpraca z osobami z innych miast czy międzynarodowa nie należy do rzadkości. Mówimy także o osobach pracujących komercyjnie, w tym zatrudnionych na etatach. Nawet wśród takich osób nie spotkaliśmy się ze sztywnym systemem godzin pracy, np. 8 – 16. Jeżeli takie osoby mają nałożony przez pracodawcę obowiązek ośmiogodzinnego dnia pracy, to raczej samodzielnie decydują o godzinie jej rozpoczęcia, np. między godziną 7 a 10. Zdecydowanie bardziej elastyczny lub nienormowany czas pracy, a właściwie podejmowania działania podstawowego i działań wspomagających, mają osoby zajmujące się DS akademicko oraz hobbyści. Uczestnicy badanego świata cenią szczególnie swobodę czasu pracy, zaś swobodę miejsca mogą traktować jako coś oczywistego, bowiem i tak niemal całość zdań wykonują w środowisku cyfrowym, za pomocą laptopa z dostępem do internetu:

> *B: Aha, ok. Yy, yy, czyli jak często się spotykacie w biurze? Jak często widujesz innych ludzi z zespołu?*
> R: [z uśmiechem] Teraz, jako że zmieniam trochę podzespół będziemy się widywać myślę 3, 4 dni w tygodniu, eee wcześniej byłam w biurze może [krótka pauza] 1 dzień, 2 dni w tygodniu, ale ja jeszcze, jeszcze często jestem za granicą, więc bywały miesiące że w ogóle nie przychodziłam do biura iii wszystko się odbywało przez Google Hangoutsy. (…)
> *B: Za granicą byłaś też w ramach swoich obowiązków, czy to jakiś inny był powód tych wyjazdów?*
> R: Eee ja często wyjeżdżam prywatnie za granicę i po prostu można powiedzieć że takie workcation[250], że trochę wakacji ale też pracuję normalnie 8 godzin dziennie.
> *B: Aha, aha, ok. Mmm to może mogłabyś nam opowiedzieć jak taki twój dzień pracy wygląda? No pewnie wygląda inaczej jak jesteś zdalnie, a inaczej jak jesteś w biurze, yyy ale, no jakbyś mogła opowiedzieć o takich paru, paru przypadkach twojego dnia pracy.*
> R: Mmm, ja jestem trochę w wyjątkowej sytuacji, bo ja jeszcze muszę chodzić na studia, więc czasami yy poza pracą jeszcze muszę pojechać na uczelnię która się znajduje niemal 30 minut, 40 minut od biura, ale generalnie pracę zaczynam, jak pracuję z domu, koło 5:40 rano, koło 6:00, jak pracuję z biura to o godzinie 6:12, więc jestem zawsze pierwsza. Kończę pracę,

250 Słowo *worcation* jest połączeniem *work* (praca) i wakacje (*vacation*). Oznacza pracę zdalną wykonywaną spoza miejsca zamieszkania, połączoną z turystyką i rekreacją w miejscu pobytu. Osoby, dla których worcation jest głównym stylem pracy mogą być określane terminem *digital nomads* (dosł. cyfrowi nomadzi). Nie spotkaliśmy się z takimi osobami w toku badań, ale niektórzy Rozmówcy / Rozmówczynie wskazywały, że styl pracy digital nomads jest dla nich atrakcyjny i chcą tak pracować w przyszłości. Z praktyką worcation spotkaliśmy się wśród uczestników świata DS wielokrotnie. Zauważmy, że nie są traktowane jako digital nomads osoby mieszkające w Polsce, a pracujące zdalnie na kontraktach np. z USA.

> zazwyczaj koło 14:00 lub 15:00. Większość mojego dnia pracy yy to siedzę przed komputerem, robię robię rzeczy które które sobie zaplanowałam na dany dzień. Spotkania zajmują mi może godzinę dziennie, dwie godziny dziennie, generalnie są czasami dni bez spotkań, a czasami są dni z długimi spotkaniami. Mmm, tak tak tak wygląda mój eee, mój dzień pracy. Jak pracuję zdalnie to, jest praktycznie dokładnie tak samo. Jak są dni że muszę pojechać do szkoły, too zazwyczaj pracuję zdalnie, żeby jakoś sobie ułatwić podróżowanie tego dnia. {wywiad 20}

Organizacja pracy Rozmówczyni {wywiad 20} byłaby oceniona jednoznacznie jako dobra przez uczestników świata DS – osoba ta ma zapewnioną swobodę i dużą dozę samodzielności ze strony pracodawcy, przy tym jest zdyscyplinowana i stara się optymalizować swoje czynności w czasie.

Przypomnijmy, że choć centralną wartością badanego świata jest naszym zdaniem efektywność, to jest ona postrzegana w kategoriach optymalizacji. Pozwala to zrozumieć, jak wartość efektywności i swobody przekłada się na powszechną w badanym świecie niechęć do pracy nastawionej wyłącznie na jak najszybsze wykonywanie jak największej ilości zadań. Kiedy nie ma komponentu jakości, nie ma tak cenionego przez uczestników czasu na eksperymentowanie i popełnianie błędów, na testowanie nowych narzędzi, na konsultacje z innymi data scientistami, na samorozwój. Nie spotkaliśmy się w komercyjnym DS z praktyką wyznaczania celów o charakterze sprzedażowym, co zdaniem jednego z Rozmówców przekłada się na dobrą atmosferę w zespole DS:

> *B: OK. A yyy Trochę idąc już, już w inną stronę chciałem zapytać jaka jest w ogóle atmosfera u ciebie w pracy, w twoim zespole?*
> R: Bardzo dobra. Bardzo taka, mm przyjacielska i zrelaksowana, [pauza] am też myślę że ze względu na to że nie jesteśmy, emm zespołem operacyjnym który ma konkretne yyy konkretne cele do zrealizowania. Ayy także na przykład miesiąc trzeba zamknąć z jakimś wynikiem, aa to, to no wiadomo że są u mnie w firmie zespoły, które, którym zdarzają się bardzo ciężkie okresy, tak w moim zespole nie ma takiego problemu, nie, nie ma takiej presji która by powodowała jakieś, jakieś stresy czy napięcia, więc amm to bardzo, bardzo dobrze wpływa na komfort pracy.
> *B: Mhm, czyli nie macie czegoś takiego jak cel sprzedażowy tak? Czy coś tego typu?*
> R: Nie mamy, tak.
> *B: Mhm, OK. A też*
> R: Oczywiście poszczególne zadania mają swoje deadline'y ale eee ale to są mmm to nie są deadline'y na których, na których że tak powiem wisi przyszłość całej firmy. [pauza] Tylko jakieś rozsądne dla wszystkich stron danego zagadnienia ustalony termin który ma być pewnym odnośnikiem dla tego jak projekt jest realizowany. {wywiad 16}

Stąd w badanym świecie raczej **mało cenione są firmy, w których istnieje duża odgórna presja na ilość wykonywanych projektów w krótkim czasie, przekładająca się na praktyki przymusowego wyrabiania nadgodzin czy pracy w dni wolne, pracy na urlopie wypoczynkowym** itp. Sądzimy, że data scientiści postrzegają swoją pracę jako w

dużej mierze twórczą i polegającą na eksperymentach, zatem od pracodawców i przełożonych spoza świata DS oczekują swobody co do organizacji pracy w czasie.

> *B: A chciałbym jeszcze spytać o twoje miejsce pracy, yyy jak w ogóle w pracy panuje atmosfera?*
> R: [pauza] Hmmm, raczej jest luźno, my też, kurczę nie wiem czy to jest coś co my sobie sami ogarniamy, chociaż ostatnio była napinka że jest aż za luźno, bo już cały czas ktoś od nas siedzi w kuchni nie, no, i trochę walczymy z tym śmiejąc się pod nosem, booo stwierdzamy że kurde, to nie jest taśma produkcyjna, i nie jest tak że usiądziesz i klepiąc po prostu kolejne linie kodu to coś osiągniesz tak, bo to trzeba mieć pomysł na rozwiązanie, czasami trzeba właśnie usiąść, pogadać w ogóle o dupie marynie i dać pracować podświadomości no [krótka pauza] ale ludzie tego nie rozumieją za bardzo no i i musimy trochę o to walczyć, jakby nie przejmujemy się tym, co mówią i jak krzyczą na nas, jak ktoś idzie pograć w piłkarzyki, no to, to jest udowodnione że tak trzeba no. Także atmosfera jest raczej luźna, no my tam nie mamy jakichś targetów stricte tak, nikt nam nie mówi że musimy osiągnąć jakiś wynik finansowy, albo oddamy zadanie albo nie i jak jest pomysł na to żeby wprowadzać jakieś miary produktywności analityków, również, bo tam gdzieś firma się [krótka pauza] prze trochę przeee przerabia na tego leana[251] całego, zarządzanie takie [krótka pauza] leanowe. Tak więc teraz Toyota jest super i ja to wykorzystuję w rozmowach z kierownictwem kiedy mówię, a słuchajcie w Toyocie wprowadzili 6-godzinny dzień pracy iii, na razie jest to w formie żartu, mam nadzieję że kiedyś to się przebije do świadomości i będziemy pracować 6 godzin. Eeem [krótka pauza] no, i ii i zaczyna wprowadzać się takie miary produktywności, ale [krótka pauza] **póki co** to nie jest tak że ktoś dostaje [krótka pauza] jakieś reprymendy, dostajemy reprymendy za to że na razie nie logujemy sobie czasu i jakby sami nie tworzymy sobie bazy danych do wyciągania wniosków na przyszłość, jeśli chodzi o zarządzanie naszą pracą, ale nie jest tak że jesteśmy z tego rozliczani. O, w ten sposób, więc też nie rzutuje to na [krótka pauza] swobodę pracy iii nikt nie mówi do ciebie sorry, ale teraz nie mogę, bo mam wiesz 30 sekund na oddanie zadania, na przykład nie? {wywiad 15}

Model działania – może być on określany jako „orka" lub „praca na ciągłym ASAPie[252]" – jest dla uczestników świata DS po prostu nieoptymalny, nie tak pojmują oni efektywność.

Przy tym **niektórzy uczestnicy tłumaczą, że czasem lubią lub chcą pracować kilkanaście godzin dziennie, ale nie z przymusu**. Pojawiają się w tym kontekście odniesienia do rozwiązywania ciekawego problemu, przeżywania przygody intelektualnej, doświadczania stanu przepływu[253] (*flow*). Jednocześnie nie spotkaliśmy się z praktykami nadużywania swojego zdrowia fizycznego dla pasjonackiego poświęcenia się pracy – raczej osoby mówiące o kilkunastogodzinnej czy całonocnej pracy zaznaczały, że kolejnego dnia pracują krótko lub biorą wolne. Nie spotkaliśmy się w badanym świecie z praktyką przyjmowania innych środków pobudzających niż kawa i inne napoje kofeinowe.

251 Od *lean management*, dosł. szczupłe zarządzanie, to koncepcja zakładająca m.in. dużą elastyczność działania i eliminowanie strat, wywodząca się z firmy Toyota.

252 ASAP to akronim *as soon as possible* (dosł. tak szybko jak to możliwe).

253 Termin *flow* stworzył psycholog Mihaly Csikszentmihalyi, oznacza ono „przepływ" jako stan doświadczenia optymalnego, polegającego na skoncentrowanym wysiłku wykonania czegoś trudnego i wartościowego. Badania wskazują, że przepływu doświadczają tak np. wspinacze (Kacperczyk 2016:481-483) jak i hakerzy (Zaród 2018:220-221).

Na wartość swobody w badanym świecie wskazuje też **praktyka oceniania się i wzajemnego oceniania uczestników w kategoriach swobody w zmianie pracy**. Do problemu wrócimy jeszcze omawiając dokładniej procesy zaświadczania o autentyczności, niemniej już teraz przypomnijmy, że termin data scientist to przede wszystkim nazwa stanowiska pracy lub zakresu obowiązków zawodowych. Zatem jest bez wątpliwości stosowany wobec uczestników realizujących komercyjnie działanie podstawowe świata DS (por. 5.2.1.), ale nie każdy z taką nazwą stanowiska pracy będzie uznawany przez innych uczestników za prawdziwego data scientista. Jednym z kryteriów bycia prawdziwym data scientistem jest to, czy ma się swobodę w zmianie pracy. Osoba związana tylko z jedną firmą, a szczególnie tylko z jedną domeną zastosowań data science może być postrzegana raczej jako ekspert dziedzinowy niż data scientist – ekspert od stosowania danych. Może być też kwestionowany samorozwój takiej osoby. Przy tym uczestnicy mają łatwo obserwowalne wskaźniki potencjalnej swobody w zmianie pracy, czyli po pierwsze, faktyczne zmiany miejsca zatrudnienia, po drugie, bycie nagabywanym przez rekruterów z innych firm.

5.4.6. Racjonalność

Racjonalność w sensie wartości społecznego świata DS rozumiemy jako posługiwanie się rozumem, a nie emocjami czy intuicją. Są od tego nastawienia na rozum odstępstwa, ale generalnie w badanym świecie ceniona jest racjonalność, na którą składają się ścisłość myślenia i kwantyfikacja.

Ścisłość myślenia polega na logice i precyzji. Uczestnicy badanego świata oceniają ją głównie na podstawie wypowiedzi, tak ustnych jak i pisemnych. Najłatwiej było nam zauważyć to na meetupach. Piszący te słowa (RŻ) był na wczesnym etapie badań zaskoczony dociekliwością i krytycznością pytań, zadawanych prelegentom przez uczestników. Typowe są pytania w rodzaju: „co dokładnie oznacza...", „zdefiniuj...", „jaki był cel tego…", „dlaczego użyliście metody / technologii...", „po co zrobiliście...", „jaka była dokłada wartość metryki...". Bez wątpienia takie interakcje są przykładem prób sił – testowania autentyczności jednych uczestników przez drugich, poprzez wskaźniki np. znajomości narzędzi, wiedzy matematycznej, bycia na bieżąco, efektywności, bycia prawdziwie zainteresowanym czy właśnie ścisłości myślenia. Występują one także w innych sytuacjach, np. w codziennych interakcjach członków zespołów DS:

> *B: (...) Yyy o tym się też trochę trudno może opowiada, ale mmm zazwyczaj w takich, w takich właśnie niewielkich zespołach właśnie powstaje taki wewnętrzny humor, ludzie się z czegoś*

śmieją, mają jakieś swoje żarty, yyy ewentualnie mają jakieś rzeczy na które zazwyczaj narzekają albo coś w tym rodzaju. Czy, czy, czy byłbyś tak miły i o czymś takim mi opowiedział ze swojego zespołu?
R: Mmm, o kurczę
B: Ja wiem że to jest trochę niewdzięczne jak się o tym rozmawia teraz przez Skype'a, tak? A nie jest się tu i teraz.
R: Mało tego że niewdzięcznie, nawet ciężko mi takie rzeczy zidentyfikować mimo że one istnieją, [pauza] to, to jakby jest lekko podświadome i nawet, nawet nie uświadamiam sobie co i jak bardzo jest insiderskie[254]. [pauza] Mmm, no jeśli chodzi o narzekania to oczywiście wszyscy zawsze narzekają na yyy jakość danych, to jakie śmieci można znaleźć w różnych systemach. Eee no i na yyy na biznes jako taki, że [pauza] że oczekuje gruszek na wierzbie. [pauza] Ee i ee że przez to współpraca jest ciężka i się trzeba użerać, ale [pauza] myślę, że w taki przyjacielski sposób do tego podchodzimy. Eee aa a jeśli chodzi o humor, mmm [pauza] oczywiście dużo jest docinek związanych ze precyzją wypowiedzi i, eeee i tym co kto miał na myśli, co wynika jakby ze, [pauza] z tego że większość ludzi jest po matematyce, albo ja właśnie po fizyce matematycznej, gdzie, gdzie em [pauza] to czym człowieka tłuką od samego początku to jest definicje, precyzja i tak dalej. I to później siłą rzeczy przenosi się eee przenosi się na całe życie. [pauza] Amm to myśle że to taka wyróżniająca cecha jeżeli chodzi o humor w moim zespole, bo po za tym to już są bardziej kwestie personalne po prostu. Relacje między nami. Czyli, czyli się śmiejemy, z tego jacy jesteśmy, kto jaki jest, a nie eeee a nie opowiadamy żartów o, nie wiem regresji logistycznej czy sieciach neuronowych. {wywiad 16}

Tego rodzaju zabiegi, jak sądzimy w łagodnej formie, stosowali Rozmówcy / Rozmówczynie wobec badacza (RŻ) podczas wywiadów oraz w rozmowach przed i po nich. Pytano o metodologię, np. w jaki sposób są dobierane osoby do badania, czy wyniki będą generalizowane oraz o badanie czy naszą pracę doktorską jako takie, np. cel badań, hipotezy, a także delikatnie krytykowano i proponowano różne strategie. Również w trakcie wywiadów często pojawiały się prośby o doprecyzowanie, zdefiniowanie lub doprecyzowywano wypowiedzi:

B: Mhm. Czyli ty jesteś nerdem? Uważasz siebie za nerda?
R: Eee, zdefiniujmy nerda.
B: No, ja ja nie wiem. Może ty mi wytłumaczysz?
R: Nie wiem, ja chyba nie lubię tego określenia. Tak samo jak
B: Czyli nie lubisz tego określenia?
R: Nie, nie lubię takich określeń, które **nic** nie znaczą. Tak na przykład, tak samo jak hipster. To nic nie znaczy. Każdy może być hipsterem i tak samo każdy może być nerdem, no albo programuje i ma swoje żarciki, albo uwielbia anime oglądać, albo czytać komiksy, tak? Każdy z tych osób może być nerdem na inny sposób. {wywiad 8}

B: Wiesz co to przy okazji, przy okazji muszę zapytać eee o taką rzecz, bo jak, jak słucham niektórych ludzi mam wrażenie, że mm jest jakaś wojna pomiędzy no nie wiem jak to powiedzieć środowiskami, czy pomiędzy językami, pomiędzy eRem a Pythonem yyy
R: Na pewno bardziej środowiskami niż językami, bo języki nie walczą. {wywiad 16}

254 Od słowa *insider*, czyli ktoś wewnątrz grupy

Podobnie w odpowiedziach Rozmówczynie / Rozmówcy dążyli do wypowiedzi logicznych i precyzyjnych, kierując się wytycznymi znanymi z ilościowych dziedzin nauki, np.:

- nie używali tzw. wielkich kwantyfikatorów (jak: zawsze, wszyscy, nigdy, nikt) relacjonując własne obserwacje, dodawali „moim zdaniem" lub „chyba / raczej";
- ostrożnie wyrażali się o związkach przyczynowo-skutkowych, akcentując, że nie świadczy o nich ani korelacja, ani następstwo czasowe;
- wskazywali na brak informacji o czymś, czasami z zaznaczeniem, że brak dowodu nie jest dowodem braku.

Co do logiki, to mamy wrażenie, że uczestnicy świata DS przeważnie stosują w komunikacji werbalnej słowa „i", „lub", „albo" tak, jak operatory logiczne koniunkcji, alternatywy i alternatywy rozłącznej. Generalnie cenione są raczej wypowiedzi ścisłe, spójne, logiczne – sądzimy, że uczestnicy świata DS mogliby zaakceptować postulat dążenia do maksymalizowania współczynnika: przekazanej treści do ilości wypowiedzianych słów.

Kwantyfikację rozumiemy jako tendencję do ilościowego postrzegania świata, posługiwania się ilościowymi metrykami i danymi ilościowymi. Uczestnicy badanego świata stosują w codziennej (i niekoniecznie dotyczącej data science) komunikacji takie kategorie, jak procenty (np. 20% czasu), ilorazy (np. 3 razy szybciej, stosunek A do B), pojęcia dużo/mało, często/rzadko itp. Jak wspominaliśmy w różnych kontekstach pojawiają się odniesienia do tzw. zasady Pareto czy inaczej 80/20. Istnieje praktyka kwantyfikacji działania podstawowego, np. w kategoriach czasu pisania kodu, ilości linii kodu potrzebnych do wykonania operacji, czasu działania kodu, a przede wszytkim używanych jest wiele ilościowych metryk w analizie i modelowaniu danych – od statystycznego testowania hipotez, przez metryki dobroci dopasowania modeli, do metryk wydajności czy kosztu produkcyjnego działania modelu. Truizmem będzie stwierdzenie, że same dane traktowane są zawsze jako ilościowe, tak ustrukturyzowane jak i nieustrukturyzowane.

Uczestnicy badanego świata chętnie korzystają z rozmaitych źródeł informacji o charakterze ilościowym. Podkreślmy, że nie spotkaliśmy się z najmniejszymi sygnałami tego, by osoby zajmujące się DS uważały, że dane mówią same za siebie, a eksperci domenowi stracili na ważności (por. części 2.2.1. i 2.2.2.) – przeciwnie, twierdzimy, że powszechne jest korzystanie z dokonań ekspertów w postaci artykułów naukowych (nie tylko z zakresu metod i technologii stosowanych w DS, ale dotyczących domeny, np. ekonomii przy projekcie dla

branży FMCG), statystyk publicznych, a nawet źródeł popularno naukowych. Nie będzie ceniony jednak ekspert, który wygłasza kategoryczne sądy bez odwołania do wyników badań empirycznych. Cześć uczestników badanego świata może cenić także informacje o charakterze jakościowym, w tym te nienaukowe i potoczne, jak materiały z mediów, wywiady, polemiki czy opinie np. doświadczonego pracownika firmy-klienta. Zaznaczmy, że omawiane korzystanie z informacji ilościowych w badanym świecie zdecydowanie wykracza poza sferę zawodową, poza działanie podstawowe, a nawet poza uczestnictwo w badanym świecie. W rozmowie prywatnej jedna z osób osadzonych w świecie DS mówiła, że rozważa przeprowadzkę do pewnego miasta m.in. dlatego, że według wieloletnich danych pogodowych jest to wyjątkowo ciepłe miasto jak na dany kraj – padło określenie w rodzaju „to wyraźny outlier na wykresie". Popularne w badanym świecie jest także prowadzenie rozmaitych badań i analiz na potrzeby własne. Mogą być to benchmarki metod i technologii, ale powszechnie wykorzystywane są także sondaże (np. związane z wydarzeniem lub platformą internetową, por. 5.1.). Zasadniczo cenione jest argumentowanie na bazie danych lub wyników badań, co dwukrotnie pojawia się w poniższym fragmencie:

> *B: Ok, rozumiem. Wiem że może, może się trochę ciężko rozmawia o takich rzeczach, tak wynika z mojego doświadczenia, ale no w takich zespołach małych jak się pracuje, często są jakieś żarty, czy jakieś tam, mmm noo takie rzeczy wewnętrzne z których się ludzie śmieją, czy czy może na które narzekają. Yyy czy jest coś takiego w twoim zespole, czy mógłbyś o czymś takim opowiedzieć?*
> R: No jest, oczywiście że jest, jest bardzo mocna kultura żartu, u mnie, w zespole. Też jesteśmy stosunkowo wulgarni, ale niedawno czytaliśmy nową książkę, lubimy popierać nasze tezy badaniami, więc jest taka książka „Bluzgaj zdrowo" gdzie pani analizuje właśnie [krótka pauza] **koherentność**, jest takie słowo po polsku, yyy koherentność, teraz powiedzmy że ta koherentność grupy gdzie jakby predyktorem jest liczba wulgaryzmów, no to generalnie [krótka pauza] grupy które przeklinają bardziej się trzymają ze sobą i się lepiej wśród nich pracuje. No jest to zauważalne. (...) No, ale jest dużo, dużo żartów, dużo czarnego humoru, narzekanie też jest oczywiście nie? Tam nie, przedmiotem narzekania jest nie tylko płaca, nie, ale tak to raczej [krótka pauza] to co, no ewentualnie to jak się czepiają że siedzimy za dużo w kuchni, kiedy tak naprawdę jakby [krótka pauza] nikt tego nie zweryfikował, czy rzeczywiście siedzimy za dużo, czy nie, czy to jest jedna osoba, czy cały zespół, nikt tego nie sprawdzał i nie liczył, nie? {wywiad 15}

Choć w badanym świecie cenione są dane i informacje o charterze ilościowym, to nie zgadzamy się na określenie, że uczestników cechuje dataizm jako ideologia zasadzająca się na wierze w obiektywność kwantyfikacji i zaufaniu do agentów zbierających dane (Dijck van 2014:198). Sądzimy, że w badanym świecie występuje wiara w kwantyfikację w tym sensie, że ceni się podejście ilościowe bardziej niż inne podejścia i porządki wyjaśniania świata, jako podejście racjonalne i umożliwiające optymalizacje. Świat jako taki raczej postrzega się w kategoriach obliczeniowych; pojawiło się określenie, że data scientist „myśli liczbami"

{wywiad 16}. Na pewno nie jest to jednak ślepa wiara w dane. W świecie DS powszechna jest troska o jakość danych i o rozpoznanie obciążeń (*bias*) i szumów występujących w zbiorach danych. Pojawia się także troska o rozpoznawanie własnych „biasów", rozumianych jako ukryte założenia, przekonania:

> *B: Chciałem zapytać o to pojęcie „data driven", które się tutaj pojawiło kilka razy. Co to właściwie znaczy?*
> R: Ja w ogóle jestem ogromną fanką Cassie Kozyrkov[255], żeby nie powiedzieć psychofanką, bo ona bardzo fajnie tłumaczy. Jest coś takiego jak data inspired i data driven. Data inspired to jest rzeczywiście takie tak, że widzisz te dane, ale gdzieś tam z tyłu głowy już masz założenia. Data driven to jest rzeczywiście bazowanie, podejmowanie decyzji, patrząc na dane, ale też walcząc trochę ze swoimi biasami. Bo wydaje mi się, że jedno, co jest bardzo niedopowiedziane na głos w całym trendzie data science i artificial intelligence to jest to, że no my powinniśmy się zacząć bardzo mocno edukować, jeśli chodzi o nasze biasy, czyli pewnie znasz Kahnemana „Pułapki umysłu"[256]. Ja teraz przechodzę, ee strasznie fajna książka „Seeking wisdom"[257], gdzie Warren Buffett, Charlie Munger opowiadają z punktu widzenia ekonomicznego, jak sobie radzić właśnie przy decyzjach z konkretnymi przykładami, co może wpłynąć na nasze błędne decyzje. Bo prawda jest taka, że algorytmy algorytmami, ee musimy też wiedzieć, że wszystkie dane, które są wykorzystywane przez algorytmy też są biasowane przez nas, czyli przez użytkowników, czy tam przez osoby tworzące te dane. I nawet podejmując decyzję opartą na danych, która jest lepsza, musimy mieć też po pierwsze taką świadomość tego, jakie mogą być nasze ograniczenia, jakie mogą być ograniczenia danych, które były użyte do analizy. To jest druga sprawa. I trzecia sprawa – ja nadal uważam, że najlepsze decyzje to nie są decyzje podejmowane super tylko na danych, bo jednak nie przeskoczymy też tej takiej wiedzy branżowej. Uważam, że jednak najlepsze decyzje to są takie, które podejmuje osoba, która zna się na branży, ma pewne takie pojęcie. (...) Więc ja myślę, że no **ekstremalnie** ważna w dzisiejszych czasach jest edukacja o tym, jak podejmować inteligentne decyzje.
> *B: OK. Ale z tego, co mówisz, rozumiem, że yyy same dane nie wystarczą? Że muszą być dane plus człowiek i jego rozum, tak? Jego znajomość branży. Być może to było już parę lat temu, być może ta moda już się skończyła, ale było trochę takich tekstów, mówiących o tym, że dane zastąpią wszystkich. Naukowcy nie będą potrzebni. (...) Ale to rozumiem, że nie podpisałabyś się pod czymś takim?*
> R: **Definitywnie** nie. Definitywnie bym się nie podpisała w takim znaczeniu, że, znaczy tak nasza rola się zmieni. My musimy być dużo bardziej analityczni. Tak jak mówię, koniec końców to, jakie będą zawody, to zawody będą się opierały na podejmowaniu decyzji. Druga sprawa jest też sprawa odpowiedzialności. No bo kto to będzie odpowiedzialny za decyzję, którą podejmie algorytm? To już są takie bardzo etyczne sprawy, które myślę, że też są bardzo mocno jeszcze nieadresowane. Są rzeczy, które można automatyzować i rzeczywiście nie jest tam potrzebna ingerencja człowieka, ale powiedziałabym, że to są po prostu bardziej rzeczy takie, które nie wymagają dodatkowej wiedzy branżowej i odpowiedzialności. {wywiad 22}

Nie jest to abstrakcyjna, epistemologiczna krytyczność wobec danych jako takich, która reprezentowana jest przez płynącą z nauk społecznych i humanistycznych krytykę tzw. big data czy pojęcia danych surowych (por. 2.2.1.), bowiem uczestnicy świata DS wierzą, że kwantyfikacja może być cennym – przede wszystkim w kategorii efektywności – sposobem

255 Chief Decision Scientist w Google .

256 Chodzi o pracę: Daniel Kahneman (2012) „Pułapki myślenia. O myśleniu szybkim i wolnym", Poznań: Media Rodzina.

257 Praca: Peter Bevelin (2007) „Seeking Wisdom: From Darwin To Munger", Malmö: Post Scriptum AB.

wyjaśniania świata, a teoretycznie także obiektywnym. Przy tym wierzą (a jak naszym zdaniem sami by to określili: wiedzą), już na poziomie metodologicznym czy wręcz operacyjnym, że kwantyfikacja jako np. metody zbierania czy przechowywania danych jest w praktyce zawsze niedoskonała. Widać to wyraźnie w pojęciach sygnału i szumu, z którego zawsze składają się tzw. realne, brudne zbiory danych. Zatem uczestników świata DS cechuje niskie zaufanie do agentów zbierających dane, a także pojawia się ograniczone zaufanie do zdolności intelektualnych ludzi, także siebie samej/samego. Wyraźnie widzimy jednak silne przywiązanie do weberowskiej „intelektualnej racjonalizacji", która oznacza

> wiedzę o tym, albo wiarę w to, że gdyby człowiek tylko chciał, to mógłby w każdej chwili przekonać się, że nie ma żadnych tajemniczych, nieobliczalnych mocy, które by w naszym życiu odgrywały jakąś rolę, ale wszystkie rzeczy można, w zasadzie, opanować przez kalkulację. Oznacza to jednak odczarowanie świata. Nie jesteśmy już jak dzikusy, którzy w takie moce wierzyli i sięgali do magicznych środków, by duchy opanować lub przebłagać – dziś rolę tę spełniają techniczne środki i kalkulacje. To właśnie przede wszystkim oznacza intelektualizacja jako taka (Weber 1989:121–22).

Niemniej w świecie DS lub wśród biznesowych rzeczników świata DS pojawiają się odwołania do magii, próby zaczarowania DS, przynajmniej w oczach klientów (por. 6.2.1.2).

Podsumowując, **w dataizm jako pułapki ślepego entuzjazmu metodologicznego wobec kwantyfikacji nie wpadają osoby silnie osadzone w świecie DS. To one lub ich rzecznicy zastawiają owe pułapki na tych, którzy mogą być zainteresowani płaceniem za usługi DS**. W biznesie istnieje pojęcie *data driven* (dosł. napędzanie danymi), które widzimy jako klarowną deklarację dataizmu i być może wyraz skuteczności działania wspomagającego świat społeczny DS, czyli marketingowego zastawiania pułapek myślowych związanych z kwantyfikacją (por. Żulicki 2019). Przy tym może być tak, że nie tylko marketing, ale i zakup jest mistyfikacją. Wskazywano nam w tym kontekście, że niektórzy klienci czyli tzw. biznes deklarują chęć bycia data-driven, ale raczej używają wyników analiz do legitymizowania określonych działań niż działają na podstawie wyników, a ponadto z powodu presji czasowej nie poświęcają uwagi wynikom analiz (ibidem). Zauważmy, że odpowiedzią świata DS na takie podejście może być praktyka odchodzenia od analiz i raportów w kierunku produkcyjnego wdrażania modeli ML – mają one przecież z jednej strony bazować na danych, a z drugiej być czymś działającym, co nie ma wymagać dodatkowej uwagi i angażowania zasobów klienta, a poprawiać efektywność jego biznesu w sposób zautomatyzowany.

DS jako działalność paranaukowa jest pozornie bliska pozytywizmowi, czyli orientacji racjonalnej i zdecydowanie modernistycznej. Andrzej Szahaj (2004:16–17) takie podejście nazywa „scjentyzmem" i uznaje je za ideologię, która zakłada zdobywanie wartościowej wiedzy wyłącznie za pomocą metod naukowych zmatematyzowanego przyrodoznawstwa. Dalej pisze o scjentyzmie jako „religii nauki", która zasadza się na oczekiwaniu, że

> poznanie naukowe zbliży nas ostatecznie do poznania jakiejś absolutnej Prawdy o świecie i naszym w nim miejscu. Kult prawdy ściśle łączy się z optymizmem poznawczym: poznanie nie zna żadnych niemożliwych do pokonania barier i przeszkód, prowadzi ono prostą drogą do ostatecznego celu – poznania świata takim, jakim on jest (Szahaj 2004:17).

Być może ów optymizm poznawczy jest charakterystyczny dla badanego świata. Powtórzy jednak, że **zobowiązaniem świata DS nie jest dostarczenie prawdy absolutnej o świecie. Jest nim zwiększanie efektywności, rozumianej jako optymalizacja**. W tym sensie zgadzamy się z Lowrie (2017:3–9), mówiącym o algorytmicznej racjonalności osób zajmujących się DS w Rosji.

Odstępstwa od racjonalności w omówionych wyżej aspektach pojawiają się w badanym świecie w formie powoływania się na intuicję i doświadczenie w wykonywaniu działania podstawowego. Jak już mówiliśmy, DS jest postrzegane w badanym świecie jako dziedzina wymagająca podejścia nieschematycznego, twórczego, a nie odtwórczego, rozwiązywania rozmytych i trudnych do zdefiniowania problemów, eksperymentowania. Doświadczenie i nabyta wraz z nim intuicja w wykonywaniu poszczególnych elementów działania podstawowego są traktowane jako cenne atrybuty osoby zajmującej się DS, jak sądzimy szczególnie dlatego, że doświadczenia nabyć można wyłącznie przez czas i kolejne, realizowane projekty – nie ma żadnego skrótu. Doświadczeni uczestnicy dysponują zatem niezastępowalną niczym intuicją czy ogólnie repertuarem milczącej wiedzy i umiejętności, który jest niedostępny i niemożliwy do przekazania ogromnej rzeszy uczestników mało doświadczonych i początkujących. To doświadczenie ma zatem pozwalać na intuicyjne decyzje dotyczące np. kolejnego kroku po serii nieudanych eksperymentów z różnymi modelami uczenia maszynowego na etapie przygotowywania POC. To doświadczenie ma pomagać w decyzjach wtedy, kiedy nie ma możliwości odwołania się do żadnej skwantyfikowanej metryki. Doświadczenie ma pomagać również w mikroaspektach wykonywania działania podstawowego, jak interpretacji wykresów na etapie eksploracyjnej analizy danych:

> Odpowiednio wyszkolony statystyk może odczytać kaprysy wykresu Q-Q, tak jak szaman może

> odczytać wnętrzności kurczaka, przy podobnym zastosowaniu zasad naukowych. Interpretowanie wykresów Q-Q jest bardziej trzewnym (*visceral*) niż intelektualnym ćwiczeniem. Niewtajemniczeni często są omamieni (*mystified*) przez ten proces. Kluczem jest doświadczenie. (Zeileis i in. 2016:cytat 105.) {obserwacja 9}

Przypomnijmy, że w badanym świecie pojawia się przekonanie, że wiedza i umiejętności rosną wykładniczo wraz z doświadczeniem w wykonywaniu działania podstawowego (rys. 23.), zatem ta intuicja doświadczonych uczestników jest także jakoby racjonalnie wyjaśniana.

Opisując dalej modelowanie jako arenę (por. 6.2.1.) wskażemy, że świat DS lub jego rzecznicy zastawiają nie tylko pułapki entuzjazmu metodologicznego wobec kwantyfikacji, ale także pułapkę DS, a nawet bardziej modeli ML lub tzw. sztucznej inteligencji, jako magii. Ponieważ laik wierzy w AI „na słowo" (Knapik 2018), to komunikaty marketingowe płynące ze świata DS (lub w imieniu tego świata) mogą odwoływać się do kategorii magii i bazować na obietnicach magicznego, zautomatyzowanego i napędzanego danymi usprawnienia biznesu. Są to pozornie racjonalne komunikaty, zaczarowujące DS w oczach części potencjalnych klientów, i nie spotykają się one z powszechną akceptacją uczestników świata DS. Odwołania zarówno do magii jak i entuzjazm wobec kwantyfikacji widzimy także w marketingu usług związanych z DS, ale nie polegających na wykonywaniu działania podstawowego i świadczonych czasem nawet przez osoby nie będące uczestnikami świata DS – mamy na myśli szeroką gamę usług edukacyjnych np. studiów, kursów i szkoleń skierowanych przynajmniej częściowo do tzw. wannabesów, czyli osób pragnących zostać data scientistami i do początkujących uczestników świata DS.

5.4.7. Podsumowanie wartości społecznego świata data science

Opisywane wyżej wartości, ze szczególnym naciskiem na sprawczość i efektywność oraz nie ujętą przez nas osobno, ale bez wątpienia obecną merytokrację, można opisać jako typowe dla neoliberalizmu (Gershon 2011; Lorenz 2012; Wrenn 2015). Węziej widzimy je jako wywodzące się z jednej strony z z tzw. światopoglądu technokratycznego, z drugiej zaś z etosu naukowca akademickiego wg. Mertona.

W odniesieniu do tzw. światopoglądu technokratycznego widzimy bliskość wartości świata DS do tego, co J. Kurczewska nazwała konstytuującym ów światopogląd „terrorem przyszłości" (1997:18–35). Terror ten polega, po pierwsze, na tym, że technokratyczna wizja, odwołująca się do nauki i techniki jest wizją jedyną, doskonałą, prawdziwą i pewną, bowiem

daną przez przyrodę – nie tworem technokratycznych myślicieli, a dziełem samej przyrody zniewalającej całkowicie przebieg i kształt życia społeczeństw ludzkich. Po drugie, co Kurczewska określa jako pogląd merytokratyczny, uznać tę wizję za konieczną i obowiązującą mogą tylko ci posiadający odpowiednio wysokie kompetencje w scjentystycznie czy pozytywistycznie postrzeganej nauce. Po trzecie, ci posiadający odpowiednio wysokie kompetencje w podanym zakresie to tylko technokraci, zatem tylko technokraci są predystynowani do roli autorytetu w kwestii przyszłej rzeczywistości, autorytetu od którego nie ma odwołania.

> Takie przeświadczenia musiały prowadzić technokratę do przekonania, że jego wizja przyszłości ma ze wszech miar charakter obligatoryjny, a on sam jest najlepszym, niezastąpionym i „wybranym" czy „zesłanym" przez Prawdę autorytetem intelektualnym, jedynym znającym prawdziwą przyszłość (Kurczewska 1997:19).

Uczestnicy świata DS nie są tak opisanymi, modernistycznymi technokratami, niemniej cechuje ich podobnie jednoznaczne, choć niezupełnie bezalternatywne postrzeganie przyszłości. Nie wywodzą oni jednak swojej wizji przyszłości tylko z przekonania o jej zgodności z determinującą ludzkie życie przyrodą – choć jak pokazaliśmy wcześniej, pojawiają się takie legitymizacje, jak np. ML będący kolejnym stadium naturalnego rozwoju ludzkości – ale **z przekonania o efektywności jako ogólnoludzkiej wartości najwyższej**. Zatem może i dałoby się wskazać alternatywne wizje przyszłości, w których ludzie zrezygnują ze zbierania danych cyfrowych i ich analizy oraz modelowania, ale są to wizje kompletnie nieprawdopodobne, właściwie bezsensowne z punktu widzenia uczestników świata DS, bowiem nieefektywne, nieoptymalne i nieopłacalne w sensie ekonomicznym. Podkreślmy po raz kolejny, że zgadzamy się z poglądem Lowriego (2017), że **świat DS zamienił scjentystyczne kryterium prawdy na kryterium efektywności, zatem wizja przyszłości nie jest wizją świata od początku do końca odczarowanego, wyjaśnionego dzięki kwantyfikacji** (choć w świecie DS część uczestników dąży np. do osiągnięcia wyjaśnialności modeli ML, opracowując metody otwierania czarnych skrzynek), **a wizją świata efektywnego i optymalnego w najlepszy możliwy sposób.**

Wymieniane przez Mertona elementy etosu naukowca: bezinteresowność, swoboda poszukiwań naukowych, merytokracja, dostęp do wyników pracy naukowej i dochodzenie do prawdy w wyniku współpracy wskazywał zarówno M. Castells, jako wartości przyświecające twórcom internetu (Juza 2016:200–201), jak i Zaród, jako wartości podobne do etosu hakerów (2018:33, 40, 90).

Z wymienionych wyżej elementów etosu naukowca jedynie bezinteresowność nie ma żadnego zastosowania do wartości świata DS – interesowność w sensie dbania jednostki o siebie: o własny interes finansowy, a szczególnie samorozwój jest nie tylko całkowicie akceptowana, ale i ceniona tak dalece, że (w przypadku samorozwoju) staje się zobowiązaniem. Niemniej pamiętajmy, że uczestnicy zdecydowanie wyżej cenią uzasadnianie indywidualnych wyborów jako „bycie prawdziwie zainteresowanym" tą czy inną domeną, metodą, technologią niż uzasadnianie tylko korzyści finansowych. Pokazuje to, iż w badanym świecie silna jest kultura indywidualizmu samorealizacji, gdzie sukces osobisty postrzegany jest jako egorealizacja (Jacyno 2007:242–43). Zatem indywidualne wybory uczestnika świata DS uzasadniane tylko względami finansowymi będą w badanym świecie postrzegane jako prowadzące do „toksycznego sukcesu" (ibidem) – takiego, w którym życie jednostki jest niepełne, a ona sama zaniedbana, gdyż zrezygnowała z siebie. Wybory uzasadniane głównie ciekawością, a także pragnieniem samorozwoju będą postrzegane jako samorealizacja, czyli coś prowadzącego do sukcesu zdrowego, zgodnego ze sobą.

Kolejna różnica jest taka, że w świecie DS ceniona jest współpraca, szczególnie na poziomie indywidualnym zaś mniej na poziomie organizacji, niemniej jej celem nie jest dochodzenie do prawdy, a zwiększanie efektywności (zaś celami pośrednimi współpracy jest ponownie samorozwój, zwiększanie własnej efektywności, budowanie marki osobistej, budowanie sieci kontaktów itp. indywidualne korzyści poboczne).

Badany świat DS jest bardzo merytokratyczny i jesteśmy przekonani, że równanie zaproponowane w 1958 r. przez Michaela D. Younga: *IQ* + *effort* = *merit* (Jackson 2001; Yair 2007) zdobyłoby w tym świecie akceptację (nie tylko z uwagi na elegancką formę matematyczną). Widzimy w badanym świecie wielkie przywiązanie do wartości indywidualnego wysiłku (*effort*): indywidualnej nauki, samodzielnego rozwiązywania napotykanych problemów technicznych i metodologicznych, doświadczenia, czasu poświęconego na wykonywanie kolejnych projektów. Pojawiają się jednak także odwołania do *IQ*, a właściwie szeroko pojętych dyspozycji indywidualnych, nawet wrodzonych – zdolności matematycznych czy intelektualnych w ogóle, „ciekawskości" jako posiadanego typu myślenia, bycia zainteresowanym matematyką lub programowaniem od dziecka, zdolności do szybkiego uczenia się. Może to prowadzić do merytokracji w sensie elitarnej klasy społecznej (Yair 2007). Mamy na myśli praktykowane przez część uczestników samopostrzeganie badanego świata jako wysoce obdarzonej różnego rodzaju kapitałem sprawczej elity, nagrodzonej ową elitarną pozycją przez społeczeństwo za swoje unikalne

zdolności i wyjątkowy (ale i wyjątkowo dobrze zoptymalizowany) wysiłek. Także samopostrzegania się indywidualnego, szczególnie dobrze osadzonego w świecie DS uczestnika jako elitarnego dobra rzadkiego w kontekście rynku pracy i w odniesieniu do światów, na których przecięciu DS powstało (matematyka akademicka, ilościowe akademickie badania empiryczne, informatyka) oraz świata, z którym bezustannie się przecina (biznes).

Swoboda poszukiwań naukowych ceniona jest w zmienionej postaci: jako swoboda w zajmowaniu się takimi projektami, jakie ciekawią indywidualnych uczestników, swoboda w eksperymentowaniu i popełnianiu błędów w trakcie realizacji projektów i swoboda w korzystaniu z takich narzędzi, jakie uczestnik uznaje za najbardziej mu odpowiadające i dopasowane do rozwiązywanego problemu, zatem w konkretnej sytuacji najbardziej efektywne. Istnieje napięcie pomiędzy światem DS a nauką akademicką w tym sensie, że niektórzy uczestnicy świata DS (np. doktoranci lub byli akademicy) mogą poddawać w wątpliwość to, czy DS czy akademia dają jednostce większe możliwości tak pojętej swobody. Więcej o tym powiemy pokazując procesy stawania się uczestnikiem świata DS.

Dostęp do wyników pracy naukowej i nie tylko naukowej jest powszechnie akceptowany w świecie DS. Otwartość kodu (open source!) i metod, powszechne praktyki dzielenia się wiedzą, konsultowania się uczestników między sobą są postrzegane jako jednoznacznie wartościowe, bo sprzyjające efektywności, swobodzie, samodzielności, samorozwojowi oraz zwiększające sprawczość. Udostępnianie wyników własnych działań, np. w postaci pakietów, należy do typowych działań wspomagających w badanym świecie i traktowane jest jako element strategii budowania własnego wizerunku i prestiżu, marki osobistej, upubliczniania swojej pracy w celu zaświadczenia o swoim doświadczeniu, wiedzy i umiejętnościach i zaświadczenia o własnej autentyczności jako uczestnika świata DS. Są to także elementy strategii budowania wizerunku organizacji – firm, zespołów DS czy jednostek akademickich.

Wyjątkowo dobrze pracuje tutaj koncepcja późnej nowoczesności Anthony'ego Giddensa, Scotta Lasha i Ulricha Becka, którzy, jak zwięźle ujął to Piotr Sztompka, uznają, iż „nowoczesność bynajmniej nie przeminęła, lecz przeciwnie, występuje obecnie w najbardziej wyrazistych formach" (Sztompka 2007:576). Opisywane przez nas wartości świata DS można by przecież wywodzić nie tylko z mertonowskiego etosu naukowca, ale nawet z cech nowoczesności zaproponowanych przez Augusta Comte'a, m.in. pozytywizmu jako sposobu myślenia bazującego na metodologii nauk przyrodniczych i ścisłych, organizacji pracy

nastawionej na zysk i efektywność, stosowania nauki i technologii w procesach produkcyjnych; choć nie tyczy się to bezpośrednio wartości to nawet comte'owska koncentracja siły roboczej w centrach miejskich (por. część 5.1.1.) jest zgodna z charakterystyką badanego świata DS. Zgadzamy się z P. Sztompką, że: „dostrzeżone przez Comte'a rysy nowoczesności znajdą się jak dotąd (…) w każdym podobnym katalogu" (2007:559).

Patrząc na systemy budowane z udziałem zespołów DS (por. 5.2.4.) jako na wdrożenia socjotechniczne, tzw. maszyny społeczne, zgadzamy się także z K. Krzysztofkiem, iż

> maszyny społeczne nie mogą funkcjonować w pustce aksjonormatywnej jako twory nastawione tylko na maksymalizację produktywności. Jeśli mają być osadzone tylko w logice społeczeństwa kierowanego, to będzie to konstruktywizm z określoną narracją ideologiczną. A może one same wymuszą *autopoiesis*, samoorganizację, samoregulację, agregowanie większych całości *bottom up*, czyli mielibyśmy wizję kontynuacji społeczeństwa liberalnego z indywiduum jako jego znakiem rozpoznawczym (Krzysztofek 2011:139–40).

Jako szeroką ramę aksjonormatywną, nie będącą tylko ramą dla samego świata społecznego DS (bo w nim najważniejszą wartością jest efektywność), ale ramą umożliwiającą i podtrzymującą istnienie świata DS, wybieramy raczej drugą z wymienionych przez K. Krzysztofka – liberalizm / neoliberalizm. Przynajmniej w komercyjnej sferze działalności DS i wtedy, gdy dane dotyczą ludzi, projekty DS nastawione są głównie na personalizowanie. Jest to owoc liberalnego kultu wolnej, racjonalnej jednostki, w którym wyborca wie lepiej, klient ma zawsze rację itd. (por. Harari 2018:281).

Rozdział 6. Dynamika społecznego świata

Pozostając w uniwersum podstawowych pojęć przyjętej ramy teoretycznej (por. część 3.1. i rys. 3.), opisujemy w tym rozdziale zachodzące w badanym społecznym świecie procesy (6.1.) oraz areny (6.2.) tego świata. Nazywamy to ogólnie dynamiką badanego przez nas społecznego świata DS.

Nie oznacza to, aby elementy badanego świata opisane w części 5. były czymś statycznym, trwałym. Wręcz przeciwnie. Są one w koncepcji społecznych światów/aren traktowane nie tylko jako zmieniające się w czasie, ale także niestabilne, rozmyte i zamazane w sensie ontologicznym – tak, jak i społeczny świat/arena w ogólności (Clarke 1991; Kacperczyk 2016). W rozdziale 6. mówimy więc o dynamice, czyli o procesach i arenach społecznego świata, które oczywiście także cechuje wspomniane rozmycie. Są one jednak czymś zasadniczo innym, niż elementy społecznego świata w tym sensie, że są dynamiczne jako takie: proces polega na zmianie pomiędzy różnymi stanami czy etapami, zaś arena polega na sporze.

Podstawą empiryczną rozdziału są głównie wyniki badań własnych, co podkreślamy raz jeszcze.

6.1. Procesy zachodzące w społecznym świecie data science

Wcześniej koncentrowaliśmy się na technologiach świata DS (5.3.), które pomagają zrozumieć jak odbywa się działanie podstawowe oraz to, że są elementem ściśle związanym z zachodzącymi w świecie procesami. W tej części proponujemy bardziej syntetyczne i abstrakcyjne ujęcie procesów: wyznaczania granic społecznego świata, legitymizacji i zaświadczania o autentyczności, i segmentacji świata społecznego (w tym profesjonalizacji jako jednej z konsekwencji segmentacji). Zatem w częściach 5.3. oraz 6.1. realizujemy drugi z celów szczegółowych rozprawy.

6.1.1. Wyznaczanie granic społecznego świata

Właściwie każdy ze zidentyfikowanych w naszych badaniach świata DS procesów moglibyśmy opisywać w kategoriach wyznaczania granic społecznego świata. Tak samo rzecz ma się jeżeli chodzi o areny. Spójrzmy najpierw na legitymizację i zaświadczanie o autentyczności. Praktyki legitymizacji podejmowanych w społecznym świecie działań i

zaświadczania o autentyczności jego uczestników są przecież bez wątpienia nastawione – choć nie wyłącznie, ale także – na obronę granic społecznego świata przed „podobnymi innymi" (Kacperczyk 2016:46) oraz na określaniu innych i samookreślaniu się jako uczestników konkretnego świata społecznego.

> W istocie jedną z najbardziej uderzających cech wielu społecznych światów jest nieustanna wewnętrzna dysputa oraz podejmowanie decyzji odnośnie do wyobrażeń o ich własnych granicach. Ujawnia się to w licznych pytaniach zadawanych przez członków i dyskusjach dotyczących tego, czy dana czynność, osoba czy produkt jest „rzeczywiście" reprezentatywny dla nich i dla ich świata. Dzieje się tak, dlatego że nawet uczestnicy, członkowie społecznego świata, zazwyczaj nie znają ostatecznej odpowiedzi na pytanie o granice ich świata, a ich główna działalność przeplata się zazwyczaj z procesem permanentnego negocjowania własnego statusu, roli i granic własnych działań oraz nieustannego podejmowania wysiłku samookreślania się (Kacperczyk 2016:38).

Przechodząc do procesów segmentacji jak i aren, możemy je w tym miejscu potraktować zbiorczo jako powiązane ze sobą, dynamiczne składowe społecznego świata. Są one z jednej strony wyrazem tego, że uczestnicy świata pracują bezustannie nad definiowaniem jego granic (w sensie makro – co tenże świat odróżnia od innych i w sensie mikro – czy konkretna osoba, także ja, jest i w jakim stopniu jest uczestnikiem owego świata). Z drugiej strony wskutek dokonywania się segmentacji oraz wskutek zachodzącego na arenie sporu redefiniowane są nowe granice „starego" świata i „nowych" subświatów.

> Zapytaj o granice świata – a pojawi się arena! Zapytaj o arenę – a natychmiast wyłonią się granice subświatów! W każdym przypadku do głosu dojdą zasadnicze różnice w pojmowaniu istoty podstawowego działania (Kacperczyk 2016:43).

Zgadzamy się z A. Kacperczyk co do tego, że z samej koncepcji teoretycznej społecznych światów czy społecznych światów/aren wynika to, iż uczestnicy pracują wciąż i wciąż nad określaniem granic dlatego, że granice te są płynne i niejednoznaczne.

> Brak wyraźnie zdefiniowanych granic społecznego świata stanowi jego nieodłączną cechę strukturalną, a odpowiedzią na ten brak granic jest ich nieustanne poszukiwanie przez uczestników. Pomimo płynności, porowatości czy braku wyraźnych granic społecznego świata, dla działających aktorów problem granic nie znika. On narasta wraz z ich płynnością. Im trudniej określić z góry, kto należy, a kto nie do społecznego świata i jaki status w nim posiada, tym bardziej ożywiona staje się debata nad kryteriami lokalizowania siebie i innych w społecznym świecie, i tym szybszy proces różnicowania i „pączkowania" subświatów (Kacperczyk 2016:40).

W tej części zajmiemy zatem nie wszystkimi możliwymi składowymi wpływającymi na wyznaczanie granic społecznego świata, bowiem należałoby wówczas umieścić tutaj prawie całość naszych rozważań. Będzie to więc głównie przedstawienie tego, **jak odbywa się**

wchodzenie w granice społecznego świata data science w Polsce – przekraczanie bariery efektywnej komunikacji:

> Granice społecznego świata są płynne i porowate, a zakreśla je bariera efektywnej komunikacji. Oznacza to, że społeczny świat trwa dzięki wspólnocie znaczeń wytwarzanej i podtrzymywanej przez uczestników w toku działań komunikacyjnych, a elementem niezbędnym dla utrzymania i reprodukowania symboliki kolektywnej, uzyskiwania porozumienia umożliwiającego tworzenie wspólnych linii działania jest język, jakim posługują się uczestnicy. Fizyczny wymiar działania, zawierający się w wykonywaniu konkretnych czynności przy użyciu określonej technologii jest tak samo istotny, co jego wyrażona w słowach idea oraz językowe uzasadnienia. W społecznym świecie działanie i narracja na temat działania wzajemnie się uzupełniają. Przestrzeń działania i przestrzeń komunikacji stanowią jedność (Kacperczyk 2016:56-57).

Opisujemy pięć typowych (w sensie typu idealnego Maxa Webera), lecz nie całkowicie rozłącznych ścieżek wejścia do świata społecznego DS. Proponujemy następujące nazwy tych ścieżek:

- młodzi absolwenci matematyki / statystyki;
- z programowania do data science;
- robiliśmy data science zanim to się tak nazywało;
- eksperci domenowi przechodzą do data science;
- z doktoratu do data science.

Młodzi absolwenci matematyki / statystyki to osoby najczęściej rozpoczynające pracę na stanowisku data scientist lub zbliżonym jeszcze w trakcie studiów pierwszego stopnia (rzadziej drugiego) i kontynuujące ją po uzyskaniu dyplomu. Takie osoby mogą planować karierę w DS już od pierwszych lat studiów i od tego czasu brać udział w kursach MOOC, stażach czy praktykach, uczestniczyć w wydarzeniach świata DS i organizować je oraz w ramach studiów wybierać związane z DS przedmioty kierunkowe i specjalizacje. Osoby te uważają DS za wybór oczywisty, naturalny dla matematyków / statystyków i mogą sądzić, że posiadane wykształcenie stanowi o ich przewadze na rynku pracy data science. Przewaga ta jest oczywiście dyskutowana, kwestionowana przez inne pozycje w badanym świecie, które wyższą wartość przyznają np. umiejętnościom koderskim, doświadczeniu zawodowemu, rozumieniu problemów biznesowych, czy różnym miksom tych i innych umiejętności. Echa owych dyskusji, krążących wokół wartościowania wykształcenia czy znajomości matematyki / statystyki, a nawet tzw. zdolności matematycznych, opiszemy zarówno w kontekście legitymizacji i zaświadczania o byciu „prawdziwym data scientistem", jak i w kontekście aren skoncentrowanych na narzędziach badanego świata (np. sporze o język Python vs. język R).

Z naszych badań wynika, że omawiana ścieżka dotyczy głównie osób kończących kierunek „matematyka”. Takie osoby mogą uważać się za elitę w elicie (jaką jest świat DS) w tym sensie, że, po pierwsze, matematykę uznają za kierunek studiów trudny, wymagający wiele pracy i wysiłku umysłowego; po drugie, uznają matematykę za najtrudniejszy, kluczowy i – co jest silnie akcentowane – **niezmienny** fundament pracy z danymi: przetwarzania, analizy i modelowania dowolnymi metodami i z użyciem dowolnych narzędzi (które to w przeciwieństwie do matematycznych fundamentów zmieniają się szybko, ale za to względnie łatwo się ich nauczyć).

> R: Zainteresowanie to było takie, naturalne połączenie z tym że zawsze się interesowałam matematyką, eee, i no już w liceum i wcześniej, poszłam właśnie potem na studia matematyczne, eee więc myślę że to jest jedne z najfajniejszych początków do data science, właśnie wiedza ta taka ustrukturyzowana wiedza matematyczna, potem łatwo można wyjść rozumiejąc te algorytmy i metody, nauczyć się samych narzędzi już takich informatycznych już nie jest trudno. {wywiad 4}

Z drugiej strony poznaliśmy także odcinających się od takiej pozycji w dyskursie, a pracujących w DS absolwentów matematyki mówiących, iż „nie czują się matematykami”.

Zbliżona ścieżka może dotyczyć osób kończących kierunki z dziedzin obliczeniowych nauk empirycznych, jak ekonometria, bioinformatyka, geoinformacja, ale raczej nie występuje wśród nich przekonanie o charakterystycznej dla matematyków elitarności. Dodatkowo zarówno osoby kończące takie kierunki studiów jak i matematykę mają nierzadko epizody pracy jako programiści (lub w ogóle pracy w IT) lub epizody ze studiami doktoranckimi. Zatem opisywana ścieżka splata się z dwiema innymi – „z programowania do data science” oraz „z doktoratu do data science”.

Z programowania do data science w sensie ilościowym może być obecnie ścieżką najczęściej występującą (por. rys. 10.) – w ankietach środowiskowych wykształcenie informatyczne deklarowano najczęściej, od nieco ponad 28% odpowiedzi respondentów w badaniu Kaggle 2017, do około 54% w ankiecie Stack Overflow 2018. Wielokrotnie wskazywano nam także, że rosnąca w DS popularność języka Python (Nunns 2017; Piatetsky 2017b, 2018c; Strong 2018; Borowiecki i Mieczkowski 2019:36–37) oznacza napływ osób z wykształceniem informatycznym / programistów (to niekoniecznie oznacza to samo) do DS. Python jest bowiem popularnym i uznawanym za łatwy językiem programowania skryptowego do ogólnego stosowania (por. część 5.3.1.) i osoby z wykształceniem informatycznym albo już znają Pythona albo łatwo jest im zacząć się tym językiem posługiwać. Pochodzący ze środowiska statystyki akademickiej język R jest nierzadko przez

informatyków / programistów uznawany np. za nieintuicyjny czy ograniczony, szczególnie w przypadku wdrożeń produkcyjnych modeli uczenia maszynowego. Spór o zastosowania języków Python i R to najbardziej wyraźnie widoczna arena dotycząca technologii badanego świata.

Wśród osób wchodzących do świata społecznego DS tą ścieżką, widzimy charakterystyczne uzasadnienia własnych decyzji koncentrujące się wokół kategorii: ciekawości (którą opisaliśmy jako wartość świata DS w części 5.4.4.), samorozwoju (por. 5.4.3.) elitarności oraz przeciwstawiania sobie działalności badawczej i inżynieryjnej. Zatem w dużym uproszczeniu bycie programistą jest **nudne**, zaś data scientistem **ciekawe**; programowanie jest czymś relatywnie łatwiejszym, bardziej dostępnym, egalitarnym zaś DS jest trudne, niedostępne, elitarne; programista jest raczej „zwykłym" inżynierem wykonującym rutynowe prace zgodnie ze schematami, a data scientist robi coś twórczego, odkrywczego, niestandardowego.

Z drugiej strony osoby z dużym doświadczeniem programistycznym na wysokim poziomie mogą być szczególnie cenione w DS – tak symbolicznie jak i w pieniądzach – za ogromny wkład w dostarczanie (tzw. dowożenie) działających produktów (por. sprawczość jako wartość w części 5.4.2.), a wokół takiej koncentrującej się na wdrożeniach roli w zespołach / działach DS wyrasta obecnie jeden z najmłodszych subświatów profesjonalnego świata DS o nazwie machine learning engineering (inżynieria uczenia maszynowego, także DataOps). Spośród osób znanych publicznie w polskim świecie DS tę ścieżkę przeszedł m.in. Vladimir Alekseichenko.

Robiliśmy data science zanim to się tak nazywało to ścieżka, z wykorzystaniem której nie tyle ludzie wchodzą do DS, a raczej DS przychodzi do nich. Mowa o osobach zajmujących się od kilkunastu czy kilkudziesięciu lat przygotowaniem, analizą i modelowaniem danych, m.in. z użyciem języków programowania, w takich dziedzinach, jak np. bankowość, ubezpieczenia, telekomunikacja, energetyka. Mogą być to także ludzie, którzy równolegle bądź na pewnym etapie kariery zdobyli stopień doktora i pracowali naukowo w dziedzinie matematyki stosowanej / statystyki lub innych obliczeniowo zorientowanych dziedzinach. Takie osoby określające się niegdyś np. jako statystycy, specjaliści od data mining, specjaliści od modelowania ryzyka mogą obecnie uważać się za data scientistów. Cześć z nich faktycznie współtworzyła społeczny świata data science, część dołączyła do świata zauważając, że jest on bliski ich działaniu i wartościom. Pojawia się w tym kontekście zagadnienie tzw. **rebrandingu statystyki** czyli zorientowanego

marketingowo nazwania na nowo pewnej dobrze znanej już działalności w celu wywołania wrażenia nowości, wyjątkowości, otwarcia dotąd niedostępnych możliwości, a w konsekwencji zdobywania nowych rynków, klientów i zwiększanie zysków finansowych. Niemniej nie wszyscy robiący DS zanim się ono tak nazywało, czyli osoby o najdłuższym doświadczeniu zawodowym (por. rys. 15., 16., 17.) uzyskują najwyższe zarobki.

Statystycy (ogólnie mówiąc) zajmujący pozycję w dyskursie – „data science to tylko rebranding tego, co robiliśmy od 30 lat" – będą odcinać się od bycia uczestnikami świata data science i określać data science jako chwilową modę: tak, jak minęła moda na data mining, tak minie moda na DS, zaś statystyka pozostanie (co jest tożsame z omawianym wyżej przekonaniem o matematyce jako najważniejszym, niezmiennym fundamencie DS, ale prowadzi do innych wyborów). Widzimy to jako obronę granic społecznych światów przed „podobnymi innymi" (Kacperczyk 2016:46), czyli światem DS, który niewątpliwie wywodzi się także ze statystyki i matematyki stosowanej. Taka pozycja obronna czy przynajmniej osobna jest reprezentowana przez część środowisk akademickich oraz branże o konserwatywnym podejściu do pracy z danymi, której najbardziej wyraźnym przykładem jest branża farmaceutyczna. W branży farmaceutycznej centralną wartością jest dokładność analiz i głównie z uwagi na szybko zmieniające się środowisko obliczeniowe (np. wersje pakietów) nie zaadaptowała ona szeroko języków R ani Python. Pozostaje przy oprogramowaniu SAS (wraz z językiem SAS 4GL), zapewniającym powtarzalne działanie skryptów analitycznych napisanych nawet w latach osiemdziesiątych. Nie zaadaptowała także szeroko metod uczenia maszynowego ani głębokiego. Istnieją jednak firmy w branży medycznej identyfikowane z DS, jak mająca oddział w Poznaniu szwajcarska Roche.

Inni doświadczeni statystycy, szczególnie ci, którzy współtworzyli świat społeczny DS, mogą być w badanym świecie szczególnie cenieni właśnie ze względu na doświadczenie. Oczywiście osoby te nierzadko zbudowały wizerunek i uzyskały status autorytetów świata społecznego poprzez swoje wystąpienia publiczne, książki, kursy, działalność naukową i dydaktyczną, ale – jak pokażemy omawiając praktyki legitymizacji – **doświadczenie jako takie** w badanym świecie może być cenione szczególnie wysoko z uwagi na to, jak młody jest to świat i jak rzadkim dobrem są w nim osoby doświadczone. Jako doświadczone określa się osoby mające za sobą rozwiązanie wielu „realnych problemów" i znające szeroką gamę narzędzi matematycznych oraz informatycznych (nie tylko te narzędzia, które są na dzień dzisiejszy najbardziej modne). Ścieżka ta może splatać się z innymi spośród przez nas wyróżnionych, a szczególnie z „z doktoratu do data science", niemniej jest ona jakościowo

inna. Dotyczy ona wyłącznie osób wykonujących działanie zbliżone do działania podstawowego świata DS wcześniej, niż zaczął się ów świat kształtować (jako nazwane uniwersum dyskursu). Spośród osób znanych publicznie w polskim świecie DS tę ścieżkę przeszli m.in. Przemysław Biecek i Artur Suchwałko.

Eksperci domenowi przechodzą do data science to ścieżka, która pokazuje, że deklaracje w rodzaju „dane mówią same za siebie" (por. część 2.2.1.) są już przestarzałe. Oczywiście wciąż pojawiają się one w innych niż DS środowiskach, przez które przechodzi fala fascynacji koncepcją podejmowania decyzji bazując na danych, a nie na teorii / opiniach i większość omawianej w 2.2.1. krytyki epistemologiczno-metodologicznej pozostaje aktualna (Żulicki 2019). Niemniej dane mówiące jakoby same za siebie były strategią legitymizacji charakterystyczną dla wczesnego etapu popularyzacji tego, co wówczas nazywano big data (w żargonie nie-technicznym), ale obecnie w badanym świecie DS straciły znaczenie. Być może takie deklaracje były wyłącznie nastawione na zewnątrz i należy je traktować nie tylko jako legitymizację DS, ale i węziej – jako marketing nastawiony na wywoływanie wrażenia o pojawieniu się nowych, niesamowitych możliwości (jak wspominaliśmy wyżej mówiąc o rebrandingu statystyki). Obecnie co najmniej od kilku lat do świata społecznego DS i do zatrudnienia jako data scientiści wchodzą osoby specjalizujące się w konkretnej domenie, z której pochodzą dane. Jak wspominaliśmy wcześniej uczestnicy osadzeni w badanym świecie uznają, że współpraca z ekspertem domenowym jest konieczna do wykonania projektu DS:

> Bardzo ważnym jest, aby zaangażować eksperta domenowego. Tego mnie osobiście nauczyła praktyka: rezygnować z projektów, w których brakuje współpracy z ekspertem domenowym. Powinno się mieć przydzielonego eksperta na np. 4h w tygodniu do dyskusji i interpretacji wyników. W każdym trochę bardziej złożonym problemie jest to konieczne. Uwierz mi, sprawdziłem na własnej skórze :) (Alekseichenko 2019a).

Z tego względu data scientiści będący wcześniej znawcami domeny mogą być uznawani za wyjątkowo wartościowych członków zespołów DS, raczej w dużych firmach, gdzie zespół / dział DS pracuje dla klienta wewnętrznego – czyli w jednej / niewielu zbliżonych domenach. Niemniej osoby przechodzące omawianą ścieżkę wykonują ogrom pracy nad wizerunkiem i tożsamością, używając rozmaitych praktyk legitimizacyjnych i będąc poddawane rozmaitym testom na autentyczność. „Muszą" bowiem udowadniać bycie ekspertem od stosowania danych z użyciem programowania (czyli data scientistą), a odciąć się od bycia ekspertem dziedzinowym biegłym w pracy z danymi. Jeżeli pozostaje ona przy pracy w jednej domenie, kwestionowany może być samorozwój (por. 5.4.3. i 5.4.5.) takiej osoby.

Ścieżkę z doktoratu do data science przechodzą osoby mające stopień doktora, współpracujące/pracujące obecnie lub kiedyś na uczelniach wyższych, a także będące w trakcie doktoratu lub po porzuceniu studiów doktoranckich / nie obronieniu pracy doktorskiej. Warto odróżnić jednak osoby, które odeszły z akademii od tych, które równocześnie prowadzą swoje kariery akademickie i uczestniczą w świecie DS.

Te drugie to szerokie spektrum uczestników świata społecznego. Od tych, którzy „robili DS zanim to się tak nazywało", poprzez dość jednoznacznie uznawane za uczestników świata DS osoby jednocześnie piszące doktorat / zatrudnione na uczelniach i na stanowiskach data scientistów lub świadczące usługi DS dla różnych klientów w modelu freelance (dotyczy to raczej naukowców z dziedziny matematyki/statystyki lub informatyki. Charakterystyczne dla tej pozycji będzie np. realizowanie grantu naukowego w partnerstwie uczelni z firmą, publikowanie artykułów i występowanie na konferencjach naukowych oraz na wydarzeniach i w mediach badanego świata z podwójną lub wymienną afiliacją). Granice są rzecz jasna płynne i zmienne w czasie. Różne, nawet maleńkie subświaty spierają się o konkretne praktyki lub konkretne osoby, negocjując nieustannie granice tego, gdzie zaczyna się i kończy bycie uczestnikiem świata DS, „prawdziwym" data scientistem czy naukowcem. Spektrum zamykają osoby, których uczestnictwo w świecie DS jest kwestionowane, np. w przypadku naukowców tytułujących się jako „data scientiści", a nie mających żadnego doświadczenia w komercyjnych zastosowaniach uczenia maszynowego, będących wyłącznie dydaktykami (takie praktyki mogą być przez uczestników osadzonych w świecie DS określane jako podszywanie się pod świat DS w celach marketingowych, jak rekrutacja na płatne studia zaoczne lub podyplomowe, tzw. fake data science). Jeszcze inną kategorię stanowią naukowcy używający typowych dla DS narzędzi i metod w różnych dziedzinach obliczeniowych. Ci mogą odcinać się lub w ogóle nie interesować światem społecznym DS, lub wręcz przeciwnie – to właśnie spośród tej kategorii naukowców lub niedoszłych adeptów akademii rekrutują się osoby porzucające ją na rzecz pełnego i oczywiście zawodowego uczestnictwa w świecie DS.

Zatem wracając do pierwszych z wymienionych grup przechodzących omawianą ścieżkę – tych, którzy odeszli z akademii do DS – to w ich narracjach widzimy odwołania do kategorii **ciekawości, sprawczości, swobody i samorozwoju**. Każda z tych wartości przyciąga młodych, uznających się za sprawnych intelektualnie i obliczeniowo ludzi do rozpoczęcia kariery w świecie nauki (co jest właściwie równoznaczne z rozpoczęciem studiów doktoranckich). Po czasie, nierzadko krótszym niż potrzebny był na ukończenie

rozprawy doktorskiej, przychodzi **rozczarowanie wartościami akademii**. Osoby identyfikujące się z ciekawością, sprawczością, swobodną i samorozwojem z czasem nabierają przekonania, że świat akademii będący dla nich grupą odniesienia normatywnego kieruje się w swej codziennej działalności wartościami zgoła innymi, co może prowadzić do konstatacji iż uczelnie to „**feudalne instytucje w których (...) dominują rozgrywki polityczne** [podkr. RŻ] i żee to zżera bardzo wiele zasobów" {wywiad 21}. Nową grupą odniesienia normatywnego może stać się dla takich osób świat DS – szczególnie, że mówimy o osobach już przynajmniej częściowo posługujących się narzędziami i metodami świata DS i nierzadko będących już na obrzeżach świata DS, np. bywają one na wydarzeniach świata DS, przeszły kilka kursów MOOC i zaczęły wprowadzać poznane narzędzia i metody do swoich analiz naukowych, mogą uczestniczyć w hobbystycznych projektach DS lub brać udział w konkursach Kaggle. Nazywamy ich (oraz podobnych) w naszej pracy wannabesami, szczególnie jeżeli są to osoby chcące wejść w DS zawodowo.

Podkreślmy, że do DS przyciągają ich nie tylko te wartości, którymi rozczarowały się one w akademii, ale także inne wartości charakterystyczne dla DS, które nie są z akademią kojarzone. Chodzi przede wszystkim o **efektywność**, którą opisujemy jako centralną wartość badanego świata społecznego – zauważmy odwołanie do tej kategorii we fragmencie, w którym Rozmówczyni {wywiad 21} de facto wskazuje, że uczelniane rozgrywki polityczne i feudalna organizacja pracy jest nieefektywna, bo przeciwproduktywna wobec tego, co postrzega ona jako cel działania akademii (niżej obszerniejszy cytat z rozmowy). Nie jest to ani jednostkowe, ani charakterystyczne dla Polski. Już kilka lat temu powstawały artykuły poradnikowe, pokazujące jak przejść z akademii do data science, pojawiły się i do dziś istnieją komercyjne kursy DS przeznaczone dla osób z doktoratem – w Europie np. S^2DS (*Science to Data Science*), w USA i Kanadzie *Insight Data Science Fellows Program* (Migdał 2016). Przykłady można by mnożyć. Przypomnijmy jedynie, że wielokrotnie cytowana w tej rozprawie C. O'Neil przeszła z akademii do DS (by później stać się krytykiem konsekwencji społecznych stosowania modeli), ponieważ na uczelni dokuczało jej „ślimacze tempo prac", chciała być „częścią postępującego szybkim krokiem prawdziwego świata" oraz „pokazać ludziom, że mój intelekt może przekładać się na pieniądze" (O'Neil 2017a:64).

Prezentujemy dłuższy fragment cytowanej rozmowy, by pokazać więcej szczegółów omawianej ścieżki:

> *B: Mhm, rozumiem. A dlaczego teraz bardziej przerzuciłaś się na Pythona?*
> R: Tam [niezrozumiałe] jestem głównie ze względów użytkowych, mianowicie mmm w

większości firm z którymi miałam jakikolwiek kontakt, okazuje się że no produkcyjnie tak, czyli w tym ostatecznym wykorzystaniu eee modelu, eee w biznesie jednak dominuje Python, jako taki język bardziej yyy bardziej uniwersalny, no ze względu na to że w eRze, eRa możesz głównie wykorzystywać do celów analiz statystycznych, natomiast Python w ogóle w deweloperce jest wykorzystywany, więc to daje możliwość stworzenia na przykład jakiejś aplikacji, czy to aplikacji webowej, w której yyy będą te informacje, jakby rezultat tych analiz od razu może być implementowany, albo można stworzyć cały pipeline yyy w chmurze, co też jest coraz bardziej popularne nie? czyli jeżeli masz na przykład te serwisy w stylu Microsoftowego Azure'a, albo eee Amazonowego AWSa i tak dalej, chmura Googla i tak dalej, eee to za pomocą Pythona możesz sobie mmm jakby stworzyć taki między innymi Pythona, bo to nie jest jedyna opcja, ale to jest bardzo taki popularny język do tego żeby stworzyć sobie cały taki przepływ mmm tego co się dzieje eee z danymi nie, od tego jak one są pozyskiwane, potem jest jakiś preprocesing tych danych, one są czyszczone, dopiero potem jest analiza yyy iii pytanie potem [z uśmiechem] co się dzieje po etapie analizy nie, czy tam jest, jakieś dodatkowe jeszcze etapy do tego dochodzą. Nie mniej yyy no to daje takie szerokie spektrum zastosowań, no i nie ukrywajmy to że to jest wykonywane, wykorzystywane w praktyce w biznesie, no to tutaj było takim powiedzmy argumentem kluczowym no ze względu na to że że ja też osobiście chciałabym zmienić swoją ścieżkę kariery iii eee zacząć pracować po prostu w zastosowaniach takich hmm biznesowych, aaa ten aspekt naukowy zostawić dla siebie tak **hobbystycznie**, tak wracając do twojego pierwszego pytania, żeby kontynuować po prostu pracę badawczą eee już bardziej dla siebie, może [niezrozumiałe] dla świata, ale to w takim trochę innym sensie, nieinstytucjonalnym, nie?

B: Mhm, ok, czyli po doktoracie planujesz pracować w biznesie?

R: Eee, albo nawet przed ukończeniem, bo w zasadzie ja jestem teraz w takim momencie, w którym mam nabyte dane, mam zakończony cały program, cały cykl studiów doktoranckich i w zasadzie zostało mi w cudzysłowie tylko zrobienie analiz i napisanie eee doktoratu [śmiech], ale uznałam właśnie że chciałabym już zdobywać takie doświadczenie eee praktyczne, żeby zobaczyć właśnie jak się pracuje z tymi różnymi problemami biznesowo, a doktorat pisać sobie równolegle, także hmm no miejmy nadzieję że [ze śmiechem] że właśnie jeszcze przed zakończeniem coś się uda zrealizować.

B: Aha, a jakiś konkretny obszar ten komercyjny cię interesuje, czy czy może nie wiem, jakaś konkretna firma? Oczywiście jeżeli chcesz powiedzieć.

R: [śmiech] To mogę ci powiedzieć prywatnie, żeby za dużo nie gdybać [ze śmiechem], (...) Natomiast wracając że tak powiem do naszej, yyy naszej rozmowy tutaj, tooo chciała, dla mnie najważniejsze jest żeby to miejsce yyy pracy wnosiło jakąś wartość dodaną [ze śmiechem] do świata, nie wiem żeby to może głupio nie brzmiało, ale żeby żeby te analizy nie przyczyniały się **tylko i wyłącznie** do zwiększania tak, przepływu pieniędzy z jednego konta na drugie, ale żeby faktycznie to był jakiś taki produkt który komuś służy i jest, jest w jakimś sensie użyteczny, i to by mi sprawiało taką realną eee realną przyjemność, poza poza tym tak, w jaki sposób można by pracować, czy tym jaka jest nie wiem kultura organizacji, atmosfera w zespole i tak dalej. Że hmm rodzaj danych to jedno, ale ale sama też taka użyteczność wykorzystania danych eee to jest coś co jest dla mnie tutaj kluczowe.

B: Mhm, mhm, ok. A yyy czyy jakby to powiedzieć czyy praca na uczelni nie jest czymś takim, gdzie robi się właśnie coś ważnego dla świata, a nie tylko zarabia pieniądze?

R: [śmiech] Ok. I tutaj wchodzimy właśnie w te kwestie, co do których miałam wątpliwości na początku. W sensie co, dlatego cię pytałam czy będę anonimowa.

B: Jak nie chcesz mówić, to nie mówisz. Pamiętaj. [z uśmiechem]

R: (...) świat akademicki wymaga bardzoo wielu zmian, żeby faktycznie to było miejsce w którym yyy pracujesz dla dobra nauki, yyy boo yyy doświadczenia nie tylko moje, ale bardzo wielu doktorantów z którymi mam regularny kontakt, z różnych ośrodków wskazują na to, że dużo częściej to są po prostu takie feudalne instytucje w których [ze śmiechem] dominuje, dominują rozgrywki polityczne i żee to zżera bardzo wiele zasobów, y aaa niekoniecznie daje to pole [ze śmiechem], do rozwoju danej, danej dziedziny, że czasem mam wrażenie, że że ten prawdziwy rozwój nauki jest eee na dalekiej liście priorytetów [ze śmiechem] paradoksalnie

> tak, osób które które zajmują jakieś wyższe stanowiska, więc eee w tym sensie, nie, nie uważam że uczelnia jest takim miejscem i wydaje mi się że dużo, że współcześnie dużo więcej można zrobić przez eee projekty eee z takiej to nie jest jedna inicjatywa, tylko w zasadzie to jest taki nurt, open science, eee open grup, czy konsorcjów, czy grup osób rozproszonych na całym świecie, które współpracują w yym w pracy nad jakimś problemem, ale w taki bardziej niezależny sposób, no i dzięki temu są skupione po prostu naaa, na konkretnej pracy, a niee walczeniu z przeszkodami infrastrukturalno-personalnymi [ze śmiechem]. {wywiad 21}

Na marginesie zauważmy, jak wyraźnie w wyżej cytowanym wywiadzie widać słuszność myśli Clarke i Star piszących o infrastrukturze technologicznej jako „zamrożonych dyskursach" (*frozen discourses*) (2008:115). Pozornie mało abstrakcyjny i „bezpieczny" w rozmowie temat zmiany stosowanego narzędzia do wykonywania działania podstawowego pozwolił nam na przejście do głębokich, abstrakcyjnych problemów identyfikowania się z wartościami świata społecznego.

Nierzadko wywiady i rozmowy nieformalne z osobami przechodzącymi opisywaną ścieżkę zbaczały w kierunku szerokiej krytyki współczesnej akademii w Polsce i na świecie. W trakcie badań dowiedzieliśmy się m.in. o publicznej zbiórce pieniężnej zorganizowanej przez doktor nauk humanistycznych Katarzynę Dawidowicz na wydanie jej książki „Koszmar doktoratu". Książka zawiera wywiady z doktorantami i doktorami relacjonującymi rozmaite sytuacje poniżania i wykorzystywania ich, głównie przez osoby promotorów[258]. Nie jest naszym celem zajmowanie stanowiska w sprawie krytyki akademii, chcemy jedynie przekazać uwagi B. Stieglera, który pisze o przemianach zachodzących na uniwersytetach w obliczu rozwoju technologii informacyjnych i kapitalizmu. Upraszczając, według Stieglera, współczesne uniwersytety w dobie społeczeństwa hiperprzemysłowego (autor odrzuca „bajki o społeczeństwie poprzemysłowym") przyczyniają się do szerzenia głupoty i braku odpowiedzialności, bowiem racjonalizacja charakteryzująca społeczeństwa przemysłowe „prowadzi do regresu i nawrotu nierozumu". Myśl ta pochodzi z 1944 r. z pracy „Dialektyka Oświecenia". Stiegler stwierdza, że Oświecenia (*Aufklärung*) należy stale bronić przed nim samym, ponieważ dąży ono do „zwrócenia się ku samemu sobie jako wiedza, która stała się bezmyślnością". Dzieje się tak, bowiem Oświecenie staje się racjonalizacją, czyli urzeczowieniem (Stiegler 2017:50–63). W nurcie krytyki Steglera to, że doktoranci lub młodzi doktorzy odchodzą z akademii do komercyjnego DS można zatem określić jako utowarowienie ich wiedzy i umiejętności z pola nauki:

258 Zbiórka na pomagam.pl (https://pomagam.pl/KoszmarDoktoratu, dostęp 02.06.2020 r.) zakończyła się zebraniem 5 325 zł z planowanych 14 000 zł, a Dawidowicz wydała książkę samodzielnie i dystrybuuje ją przez portale olx.pl oraz allegro.pl (https://www.olx.pl/oferta/ksiazka-koszmar-doktoratu-katarzyna-dawidowicz-CID751-IDDfuZI.html, dostęp 02.06.2020 r.)

> Sprostytuowanie rozumu i teorii polega na zaciągnięciu ich na służbę racjonalizacji, pojmowanej nie tylko jako sekularyzacja społeczeństwa (w rozumieniu Maksa Webera), lecz także jako legitymizacja, to znaczy racjonalizacja w rozumieniu Ernesta Jonesa i Zygmunda Freuda, w której owo odwrócenie znaku (na skutek czego rozum prowadzi do nierozumu, a progres do regresu) jest **usprawiedliwiane i uzasadniane** [podkr. org.] pod maską samego rozumu (Stiegler 2017:121-22).

Stiegler (ibidem) uważa, że racjonalizacja polega tu na przyjęciu, że „nic nie da się zrobić", nie ma alternatywy – co pisze w języku angielskim (*there is no alternative*), z uwagi na znany akronim TINA.

W świecie społecznym DS nie pojawiają się tak negatywnie wartościujące głosy krytyczne wobec akademii i osób z niej pochodzących. W języku uczestników świata DS będzie raczej mowa o tym, że doktoranci / doktorzy **monetyzują swoje naukowe umiejętności** metodologiczne, analityczne, obliczeniowe aplikując je na gruncie DS. Osoby przechodzące omawianą ścieżkę – pamiętając, że ma ona kilka odmian – mogą być cenione w badanym świecie za rygor metodologiczny ale także ciekawość, tendencje do eksploracji. Niemniej nie może to być przesadnie zagłębianie się w eksplorację, gdyż najważniejszą wartością badanego świata społecznego jest efektywność. Obiegowo w świecie DS mówi się, że naukowcy pracują nieefektywnie, powoli i bez presji, cyzelując szczegóły; zaś „prawdziwe" DS to praca efektywna i „wystarczająco dobra", nastawiona na osiągnięcie optimum między zużytymi zasobami a osiąganym efektem, stanowiącym rozwiązanie tzw. realnego problemu. Tym samym osoby rekrutujące się z nauki akademickiej mogą być testowane przez uczestników osadzonych w badanym świecie pod kątem efektywności i sprawczości działania.

Spośród osób znanych publicznie w polskim świecie DS ścieżkę „z doktoratu do data science" przeszedł m.in. Piotr Migdał.

Każdą z wyróżnionych ścieżek można opisać w odniesieniu do różnic pomiędzy wartościami świata społecznego DS a wartościami w szerokich światach społecznych (dla uproszczenia będziemy je nazywać obszarami), **z którymi DS przecina się i przenika**, a w pewnym stopniu z nich się wywodzi. Chodzi o obszary: nauki akademickiej (matematyki i empirycznych nauk obliczeniowych), technologii informacyjnych (dalej IT; szczególnie tzw. programowania, czyli *software development* oraz baz danych) oraz biznesu (rozumianego jako różnego rodzaju działalność komercyjna w rozmaitych dziedzinach). Opisywane porównanie schematycznie prezentujemy na rysunku 38. Określamy wagę, jaką z perspektywy uczestników świata DS wyróżnione obszary (DS, nauka, IT, biznes) nadają

siedmiu wartościom badanego świata. Na potrzeby tego uproszczonego porównania przyjmujemy, że świat DS jako całość nadaje wysoką wagę (na rys. 38. oznaczoną jako „3") każdej z siedmiu wyróżnionych wartości, tj. efektywności, ciekawości, samorozwojowi, swobodzie, sprawczości, racjonalności i samodzielności.

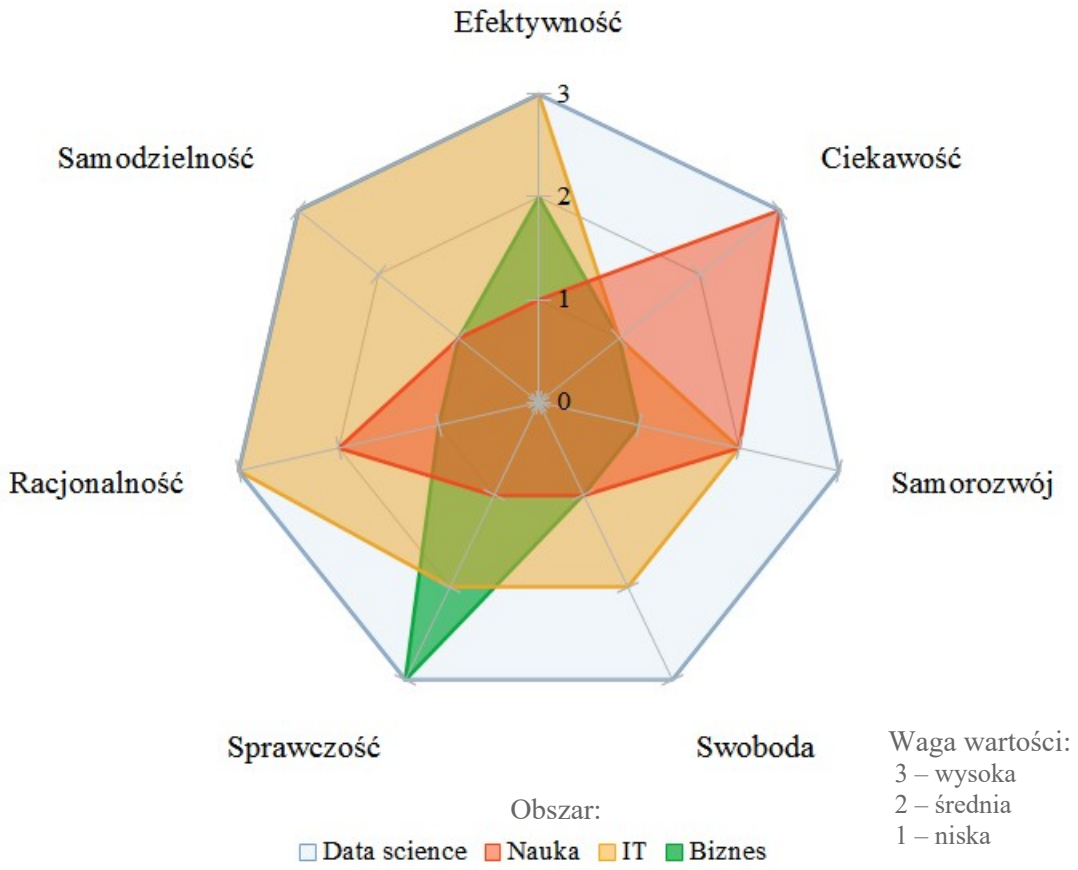
Obszar nauki akademickiej (na rys. 38. wyróżniony kolorem czerwonym) jest przez uczestników świata DS postrzegany jako wysoko ceniący ciekawość, średnio (waga „2") samorozwój i racjonalność, a nisko (waga „1") pozostałe z wyróżnionych wartości. Zatem dla uczestników świata DS głównie perspektywa zrealizowania „ciekawego projektu" może stanowić o atrakcyjności nauki akademickiej. Niska waga sprawczości spowodowana jest tym, iż nauka akademicka zajmuje się głównie jakoby wytwarzaniem artykułów / książek mających niewielki wpływ na obszary poza akademią – na pewnie nie zajmuje się budowaniem „czegoś działającego". Niskie wagi efektywności, samodzielności i swobody spowodowane są hierarchiczną, „feudalną" organizacją pracy na uczelniach i głównie publicznym jej finansowaniem.

Obszar IT jest w perspektywie wartości świata DS najbardziej z nim zbieżny spośród trzech wyróżnionych. IT jest przez uczestników świata DS postrzegane jako wysoko ceniące racjonalność, samodzielność i efektywność, zaś średnio samorozwój, swobodę i sprawczość. Niska jest jedynie waga ciekawości – jej powodem jest wielokrotnie przez nas opisywane postrzeganie informatyki czy konkretnie tworzenia oprogramowania (*software development*) jako zwykłej, nudnej pracy inżynieryjnej, w której rozwiązuje się problemy za pomocą względnie (tj. w porównaniu do DS) schematycznych metod i narzędzi, choć oczywiście buduje się „coś działającego". Szczególnie istotne wydaje się nam, iż uczestnicy świata DS tak samo jak programiści czy ogólnie informatycy, postrzegają siebie jako osoby techniczne. Oznacza to nie tylko osoby biegłe techniczne, ale także cechujące się tym, że wartościują rzeczywistość w kategoriach efektywności, racjonalności i samodzielności. Może za wartość charakterystyczną dla światów technicznych należałoby uznać także samorozwój i sprawczość. Niemniej uczestnicy świata DS postrzegają swoją dziedzinę za bardziej sprzyjającą samorozwojowi niż informatyka. Wskazują, iż DS jest na znacznie wcześniejszym niż informatyka stadium rozwoju, rozwija się znacznie bardziej dynamicznie i w pewnych zastosowaniach wypiera rozwiązania informatyczne bazujące na zapisywanie reguł w kodzie, stanowiąc zarówno nowe jak i lepsze rozwiązanie wielu problemów. Innym wątkiem jest większe ryzyko niepowodzenia w eksperymentalnych projektach DS niż w jakoby szablonowym tworzeniu oprogramowania. Podobnie uznają DS za bardziej od tejże

sprawcze – głównie z uwagi na stosowanie metod uczenia maszynowego, a nie „tylko" zapisywanie w kodzie reguł zaplanowanych przez ludzi.

> Dotychczasowe doświadczenia - realizowane przez kilkanaście zespołów firmy rozsianych po całym świecie - pokazują, że projekty realizowane z wykorzystaniem tego typu metod [ML] wymagają – w odróżnieniu od typowych projektów inżynierskich – specyficznych umiejętności, charakterystycznych dla przedsięwzięć o profilu badawczym. W szczególności istotne okazuje się być podejście obejmujące, m.in., stawianie hipotez, wykorzystywanie różnorodnych metod analizy danych, zdolność do przeprowadzania eksperymentów obliczeniowych z użyciem wielu metod, prezentacja wyników z zastosowaniem metod statystyki opisowej, lub znajdowanie istotnych statystycznie prawidłowości. W związku z tym, zdaniem panelisty, warto rozszerzyć dotychczasowy program magisterski o wskazane umiejętności. Prelegent podkreślił również, że obserwowany obecnie dynamiczny rozwój metod uczenia maszynowego owocuje szeregiem publikacji w czasopismach naukowych. Ich śledzenie wymaga nie tylko znajomości zaawansowanego aparatu formalnego, orientacji w danym obszarze badawczym, ale również czasu. Z perspektywy firm konkurujących na rynku śledzenie najnowszych trendów w obszarze uczenia maszynowego jest często kosztowne, by mogło być prowadzone na odpowiednio wysokim poziomie. Naturalnym rozwiązaniem wydaje się być w tej sytuacji współpraca pracowników nauki z przedstawicielami biznesu (Czarnowski i in. 2018:21).

Biznes jest przez uczestników świata DS postrzegany jako zbieżny z DS jedynie w wymiarze wysokiego wartościowania sprawczości. Oznacza to także, że w świecie DS za kluczowy czynnik sprawczy mogą być uznawane pieniądze – zazwyczaj to od budżetu biznesu-klienta zależy, jaki wpływ na rzeczywistość będzie miał realizowany projekt DS (np. jaki zespół ludzi i ich kompetencji można opłacić, jaka infrastruktura obliczeniowa będzie dostępna, jaka będzie skala i zakres wdrożenia produkcyjnego tworzonych modeli). Z punktu widzenia osób osadzonych w świecie DS biznes średnio ceni efektywność, zaś nisko pozostałe z siedmiu wartości badanego świata. Zauważyliśmy, że uczestnicy świata DS nisko oceniają także wartość racjonalności w biznesie, uznając, że biznes, co prawda, chętnie deklaruje bycie „napędzanym danymi" (*data-driven*), ale faktycznie decyzje zapadają z uwagi na przesłanki dalekie od scjentystycznie pojętej racjonalności – gry polityczne lub personalne, przeczucia i intuicja, stereotypy, kaprysy decydentów, choć celem nadrzędnym pozostaje niezmiennie zysk finansowy.



Rysunek 38: Porównanie obszarów w perspektywie wartości społecznego świata – DS a nauka, DS a IT, DS a biznes

Uwaga: oś liczbowa reprezentuje wagę wartości w obszarach z perspektywy uczestników świata DS

Źródło: opracowanie własne za pomocą Libre Office Calc

Podsumowując, za ważny element w sporze o granice społecznego świata DS uznajemy to, iż jego **uczestnicy postrzegają się w dużej mierze jako osoby techniczne**. Wskazuje na to fakt, że uczestnicy widzą dużą zbieżność w tym, jak ważne są wartości ich świata i obszaru IT. **Niemniej nie tylko techniczne** – w ocenie uczestników **świat DS jest jedyną w swoim rodzaju hybrydą** (por. rys. 2.), która w niespotykany dotąd i nieporównywalnie do niczego efektywny sposób pozwala na rozwiązywanie realnych problemów (w tym głównie biznesowych), nie tylko za pomocą najnowocześniejszych technologii, ale także najlepszej aparatury matematycznej i metodologii empirycznych nauk obliczeniowych. DS jako świat społeczny powstało i dynamicznie trwa na przecięciu trzech obszarów: nauki, informatyki i biznesu, postrzegając siebie jako elitę, która nie należy do żadnego z tych obszarów, a od każdego z nich jest „lepsza” (przyjmując za punkt odniesienia wartości świata DS). W kolejnej części przyglądamy się, jakie opowieści budują uczestnicy badanego świata by tę

„lepszość" – a mówiąc bardziej abstrakcyjnie: sens istnienia społecznego świata oraz jego normatywny ład (por. Kacperczyk 2016:56) – uzasadnić.

6.1.2. Legitymizacja i zaświadczanie o autentyczności

Światy społeczne uzasadniają racje swojego istnienia. Czynią to zarówno światy „młode", jak i te bardziej ustabilizowane, konstruując rozmaite opowieści dotyczące ich działalności oraz tożsamości. Owa praktyka ciągłego konstruowania uzasadnień wynika

> (...) z potrzeby uzyskania poparcia dla własnych działań i zawiera roszczenie zaakceptowania przez otoczenie sposobu funkcjonowania danego społecznego świata. Jest też jednocześnie sposobem dystansowania się wobec innych światów i ustanawianiem granic między nimi (Kacperczyk 2016:55).

Uzasadnienia mogą być wskazaniem na wyjątkowy sposób wykonywania działania podstawowego oraz na konkretną technologię, jaką dysponuje dany świat lub segment świata. W skali makro zachodzi proces „wyznaczania granic prawidłowej działalności" (gdzie negocjowane jest to, kogo wolno uznawać za uczestnika i „prawdziwego" uczestnika świata społecznego). W skali mikro budowany jest szereg objaśnień, uzasadnień, wymówek pozwalających pojedynczym aktorom kontynuować działanie podstawowe (Kacperczyk 2016:36). Legitymizacja i zaświadczanie o autentyczności realizowane są głównie poprzez zabiegi komunikacyjne o charakterze objaśniania, neutralizowania lub teoretyzowania.

6.1.2.1. Dużo danych i rozwój mocy obliczeniowej…

Najbardziej popularnym, relatywnie najstarszym i obecnie uznawanym za niezmiernie oczywistą prawdę o powstaniu świata społecznego DS (nie tylko polskiego, ale globalnego) jest następujące objaśnienie: **data science powstało, ponieważ z uwagi na gwałtownie rosnącą ilość dostępnych danych i rozwój mocy obliczeniowej komputerów przy jednoczesnym spadku ich kosztów** (tak danych jak i mocy) **możliwe, a zarazem opłacalne stało się rozwiązywanie realnych problemów z wykorzystaniem metod uczenia maszynowego.** Metody ML znane były naukowcom od lat 50. XX wieku, ale w badanym świecie objaśnia się, że dopiero w pierwszej dekadzie XXI w. można było je realnie wykorzystać – jakoby przekroczony został wówczas punkt krytyczny, poniżej którego dostępnych danych było za mało, ich przechowywanie za drogie i zbyt pracochłonne, moc obliczeniowa zbyt droga i zbyt powolna, poniżej którego wreszcie mało kto wierzył nie tylko w opłacalność, ale w ogóle możliwość efektywnego używania uczenia maszynowego w rozwiązywaniu tzw. realnych problemów.

To niezwykle wpływowe objaśnienie stosowano od czasów popularności terminu „data mining", i niemiłosiernie eksploatowano w długim okresie panowania nie-technicznego ujęcia terminu „big data". Wówczas szczególnie akcentowano szybko rosnącą ilość dostępnych danych, które zaczęto przedstawiać jako **zasób do monetyzowania**. Już w 2012 r. opisywany zabieg sportretował Scott Adams (rys. 39.), odnosząc się z humorem także do orwellowsko-panoptycznej narracji o wszechwiedzących danych, wyśmiewając metaforę „chmury" oraz trafnie pokazując wiarę w wielkie dane jako mesjasza, który za odpowiednio wysoką opłatą zbawi biznes od bankructwa.

Rysunek 39: Big data w komiksie "Dilbert" – legitymizacja powstania świata społecznego DS

[tekst komiksu]
Boss: „Konsultanci mówią, że codziennie powstają trzy kwintyliony bajtów danych. One przychodzą zewsząd. Wiedzą wszystko. Według księgi Wikipedii nazywa się to *Big Data*. Big Data żyje w chmurze. Wie, co robimy. W przeszłości nasza firma zrobiła wiele złego. Ale jeśli zaakceptujemy Big Data na naszych serwerach, będziemy uratowani przed bankructwem. Spłaćmy swe winy."
Alice: „Czy jest już za późno, by stanąć po stronie zła?"
Dilbert: „Ciiiiii! Słyszy Cię."

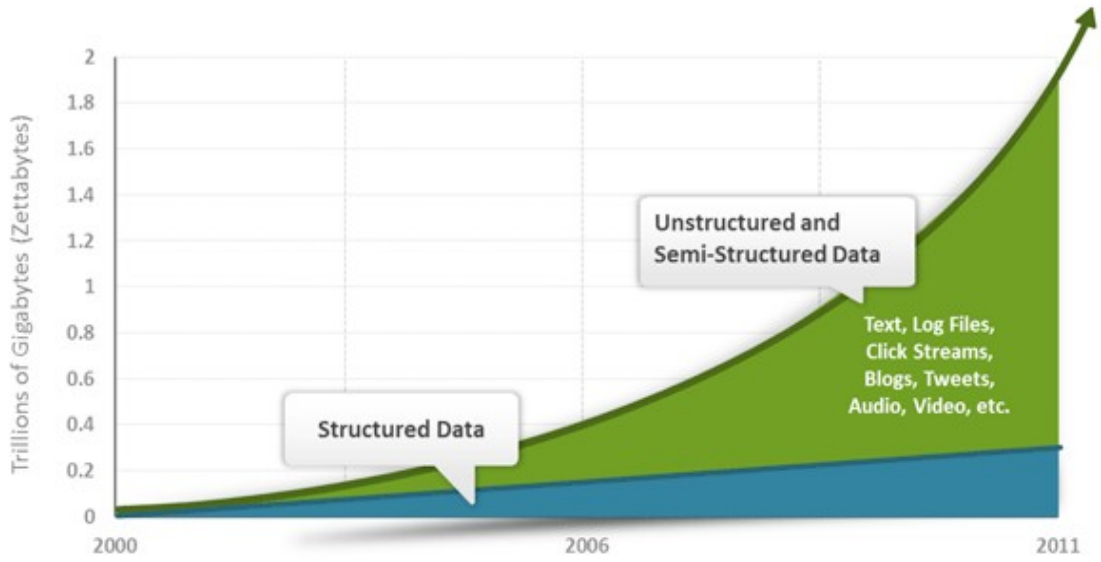
Źródło: http://dilbert.com/strip/2012-07-29, dostęp 12.10.2016 r.

Opisywane objaśnienie wykorzystano szeroko, np. w pełnej entuzjazmu książce *Big Data. Rewolucja...* (Cukier i Mayer-Schönberger 2014). Autorzy zaczęli kreślenie obrazu „epoki wielkich danych" od astronomii, gdzie dzięki rozpoczętemu w 2000 roku programowi badawczemu Sloan Digital Sky Survey zebrano w kilka tygodni więcej danych niż w dotychczasowej historii tej dyscypliny. Przez kolejne 10 lat trwania badań zebrano 140 terabajtów informacji. Ruszający w 2016 r. program Large Synoptic Survey Telescope miał gromadzić taką ilość danych co pięć dni. W naukach medycznych rozpoczęte w 2003 r.

zdekodowanie ludzkiego genomu poprzez sekwencjonowanie trzech miliardów podstawowych par DNA trwało kilka lat. Już w roku 2013 możliwe było odczytanie podobnej sekwencji w jeden dzień. Ogromną ilość danych generują jednak nie tylko badania naukowe. Niejako „przy okazji" różnego rodzaju czynności wielkie dane tworzą użytkownicy narzędzi dostarczanych przez technologicznych gigantów, np. Google i Facebook. Pierwsza z firm w 2012 r. przetwarzała dziennie 24 petabajty informacji. Druga co godzinę otrzymywała do publikacji około 10 milionów fotografii i mniej więcej trzy miliardy „polubień" oraz komentarzy w ciągu doby. Również w 2012 użytkownicy serwisu YouTube w liczbie 800 milionów osób dodają około godziny nowych filmów co sekundę. W tym samym roku dzienna liczba twittów wyniosła ponad 400 milionów (Cukier i Mayer-Schönberger 2014:21-22). W różnych publikacjach stosowano porównania do „zalewu" danymi czy „eksplozji informacyjnej", nierzadko argumentując na rzecz korzystania z owych cennych zasobów (Anderson 2008; Brynjolfsson i McAfee 2015:88–89; Cerf 2007; Cukier i Mayer-Schönberger 2014:22; Larose 2005:xi, 4; Paharia 2014:60). Zdaniem Cukiera i Mayer-Schönbergera:

> od nauki po ochronę zdrowia, od bankowości po Internet – różne branże, ta sama historia – ilość danych na świecie szybko rośnie, prześcigając nie tylko nasze maszyny, ale naszą wyobraźnię (2014:22).

W 2008 roku wielkość ogółu cyfrowych danych szacowano na pół zettabajta (ZB), czyli 500 milionów terabajtów. W 2011 roku było to około 2,5 ZB. Szacunki firmy Cisco wskazują, że w 2020 roku będzie to mniej więcej 35 ZB danych. Podkreślany jest fakt, że eksplozja danych towarzyszy dywersyfikacji rodzajów danych, szczególnie tych nieustrukturyzowanych, jak pliki tekstowe, muzyczne, wideo, itp. (Sumbul 2014). Na rysunku 40. prezentujemy przygotowaną przez M. Sumbula wizualizację przyrostu wolumenu i charakteru danych. O ile w jego ujęciu ilość danych ustrukturyzowanych rośnie liniowo (kolor niebieski), to dla danych nieustrukturyzowanych (kolor zielony) przyrost ma charakter wykładniczy.



Rysunek 40: Wykładniczy przyrost ilości danych w podziale na ich rodzaj – dane ustrukturyzowane i nieustrukturyzowane

Źródło: Figure 1: Growth of the data (Sumbul 2014)

Dane stały się zatem postrzegane jako zasób naturalny, stały się „nową ropą" dla wielu branż (Gogołek 2016; Puschmann i Burgess 2014; Sadowski 2019; Watson 2015), magiczną kopalnią diamentów nieustannie dostarczającą nowych minerałów, chociaż pierwotne złoża wyczerpały się (Cukier i Mayer-Schönberger 2014:140-141). Pochodną takiego postrzegania danych były silnie krytykowane przez nauki społeczne tezy (por. część 2.2.1.) o tym, że „surowe" dane stanowią obiektywną prawdę o świecie lub iż „mówią same za siebie", więc wiedza teoretyczna i ekspercka stała się bezużyteczna (badany świat społeczny raczej odcina się od owych tez) oraz także krytykowane zjawisko zwane danetyzacją.

Twierdzimy, że opisywane wyjaśnienie, a szczególnie jego część mówiąca o gwałtownie rosnącej ilości wartościowych danych, pełni obecnie nie tylko funkcję legitymizacji świata społecznego DS (a niegdyś tego, co nazywano data mining i później big data). Opowieść tę jako prawdziwą przyjęła nauka akademicka, biznes; obecnie przenika ona do tekstów popularnonaukowych, np. Y. N. Hararіego:

> Jeśli chcemy zapobiec skupieniu całego bogactwa i władzy w rękach niewielkiej elity, kluczową sprawą jest uregulowanie kwestii prawa własności danych. W starożytności największą wartość miała ziemia, polityka była walką o władzę nad ziemią, a jeśli zbyt wiele ziemi znalazło się w rękach zbyt niewielu ludzi – społeczeństwo dzieliło się na arystokrację i plebs. W epoce nowożytnej ważniejsze od ziemi stały się maszyny i fabryki, a walki polityczne skoncentrowały się na kontrolowaniu tych podstawowych środków produkcji. Jeśli zbyt wiele maszyn znalazło się w rękach zbyt niewielu ludzi – społeczeństwo dzieliło się na kapitalistów i proletariuszy. **W XXI wieku jednak zarówno ziemię, jak i maszyny przyćmią dane i to one staną się**

> **najważniejszym skarbem, a polityka będzie walką o władzę nad przepływem danych** [podkr. RŻ]. Jeśli dane znajdą się w rękach zbyt niewielu ludzi – ludzkość podzieli się na różne gatunki. Wyścig po dane już trwa, a prowadzą w nim tacy giganci, jak Google, Facebook, Baidu i Tencent. (...) W rzeczywistości nie chodzi im wcale o sprzedawanie reklam. Jest raczej tak, że przykuwając naszą uwagę, potrafią gromadzić ogromne ilości danych na nasz temat, które są warte wielokrotnie więcej niż dochody z reklam. Nie jesteśmy ich klientami – jesteśmy ich produktem (Harari 2018:111).

W badanym świecie DS opisywane wyjaśnienie uznawane jest za fakt – oczywistą prawdę. Autorzy książki poświęconej pracy z danymi tekstowymi z użyciem języka R rozpoczęli tekst słowami:

> Jeśli pracujesz w analityce danych lub w data science, tak jak my, znany jest Ci fakt, iż dane są generowane przez cały czas w coraz szybszym tempie. Możesz być nawet trochę zmęczony ludźmi, prawiącymi o tym fakcie kazania (*You may even be a little weary of people pontificating about this fact*) (Silge i Robinson 2018).

Podkreślmy, że ta legitymizacja może być używana także na poziomie mikro. Przyjrzyjmy się fragmentowi, w którym Rozmówca z wieloletnim doświadczeniem w IT uzasadnia swój wybór ścieżki kariery – zajął się „sztuczną inteligencją" w czasie, kiedy po raz pierwszy zaczęła ona „naprawdę działać" dzięki dostępowi do dużej ilości danych i wystarczającej mocy obliczeniowej:

> *B: Mhm, OK. Skąd w ogóle wziął się pan w data science? Może jak pan zaczynał? Jakie jakie jakie są pana początki w branży?*
> R: Ja dosyć dużo pracowałem z bazami danych, to też dziedzina pokrewna i jak ja wchodziłem w informatykę, to było prawie 30 lat temu, to wtedy sztuczna inteligencja istniała ale nie działała dobrze. Dopiero niedawno okazało się, że jak się ma dosyć dużo danych i jak się ma dosyć szybkie procesory i jeszcze parę innych rzeczy, metod wymyślonych to nagle to zaczyna działać. Więc ona się zaczęła tak powiedzmy parę lat temu naprawdę działać, no i jak się zorientowałem, że kurcze to działa no to, to **fajnie** to jest to, co chcę robić. {wywiad 2}

Opisywana legitymizacja, być może z uwagi na swoją popularność, straciła na znaczeniu na rzecz opowieści skoncentrowanej na terminie „sztuczna inteligencja", dotyczącym głównie „czegoś działającego" (szczególnie w zadaniach dotyczących pracy z tekstem, obrazem i dźwiękiem), a nie ilości danych czy mocy obliczeniowej. Pojęcie sztucznej inteligencji jest jednak raczej używane w komunikacji nie-technicznej, marketingowej, płynącej na „zewnątrz" ze świata DS. I choć w jego użyciu akcentuje się takie wartości badanego świata, jak sprawczość i efektywność osiąganą dzięki wdrażaniu produktów bazujących na AI, to widzimy owo pojęcie jako element areny (por. część 6.2.1), a nie strategię legitymizacji rozumianą przecież także jako uzasadnienie racji istnienia świata społecznego w komunikacji „wewnątrz" świata.

Można oczywiście traktować opisywane wyjaśnienie wyłącznie jako fakt, i być może w tej optyce moglibyśmy odpowiedzieć na pytanie: co uczyniło możliwym przecięcie, prowadzące do zaistnienia badanego świata społecznego w konkretnym czasie? Wydaje nam się, że koszt mocy obliczeniowej i koszt przechowywania danych spadł poniżej punktu nieopłacalności. Podobnie ekonomicznie możliwy stał się dostęp do dużej ilości danych z nieistniejących wcześniej źródeł. Brak było (i częściowo wciąż jest) regulacji prawnych dotyczących wykorzystania danych i ich analizy oraz modelowania, zatem raczej nie istniały przeszkody inne, niż techniczne. Usuwanie przeszkód technicznych pozwoliło na rozpoczęcie produkcyjnego stosowania osiągnięć akademickich, czyli metod uczenia maszynowego na coraz większą skalę. Wysokie budżety firm i ich atrakcyjna, bardzo „nieakademicka" kultura pozwoliły im pozyskać naukowców do pracy w DS. Sprzyjające były właśnie także warunki kulturowe – z jednej strony względnie powszechna zgoda na danetyzację, przejawiająca się np. w globalnie akceptowanych modelach biznesowych, gdzie użytkownik jest produktem generującym dane w zamian za bezpłatne usługi (jak Facebook i Google), z drugiej „późnonowoczesna" kultura firm technologicznych nastawionych na kwantyfikowalność, efektywność, sprawczość, akceptację niedoskonałości i iteracyjności pracy (adaptowana do innych podmiotów wraz z upowszechnianiem się DS).

Pozostając w sferze faktów musimy podkreślić, że dane cyfrowe są przechowywane i przetwarzane za pomocą materialnej infrastruktury komputerowej. Infrastruktura ta, w pewnych przypadkach nazywana „chmurą" obliczeniową, zużywa materialne zasoby środowiska naturalnego planety Ziemia (por. 5.3.3). Z uwagi na ciągły przyrost ilości danych stanowi to rosnące zagrożenie ekologiczne. Przy stałym wzroście ilości danych cyfrowych na poziomie 20% rocznie, za około 300 lat ilość energii niezbędnej do podtrzymania takiej produkcji danych przekroczy ilość energii, która jest obecnie konsumowana przez całą ludzkość (Vopson 2020).

6.1.2.2. Prawdziwy data scientist

DS w Polsce jest na tyle mało zinstytucjonalizowanym światem społecznym, że **nie istnieje formalna droga zaświadczania o autentyczności** uczestnika. Nie zauważyliśmy, aby jakiekolwiek formalne wykształcenie było w badanym świecie jednoznacznym wskaźnikiem tego, że ktoś jest uznawany za jego uczestnika. Obecnie nie zmienia tego pojawianie się na rynku pracy absolwentów studiów z „data science" w nazwie – ani studiów podyplomowych, ani magisterskich, ani żadnych innych. Podobnie nie widzimy w badanym

świecie, by istniał jednoznacznie uznawany certyfikat czy dyplom zaliczenia jakiegokolwiek kursu lub egzaminu. Rzecz jasna nie ma także czegoś takiego jak uprawnienia do pracy jako data scientist.

Także w przypadku DS jako pracy zarobkowej formalna nazwa stanowiska „data scientist" (np. na umowie o pracę) nie jest jednoznacznym wskaźnikiem bycia uczestnikiem badanego świata społecznego, a tym bardziej nie jest wskaźnikiem bycia prawdziwym data scientistem. Uczestnicy badanego świata wskazywali nam na praktyki podszywania się pod DS (zwane fake data science), które stosują np. rekruterzy, promując ofertę pracy będącą „faktycznie" – tzn. z punktu widzenia uczestników badanego świata – pracą w charakterze analityka danych nazwą stanowiska „data scientist", i przyciągając w ten sposób wannabesów lub początkujących uczestników świata społecznego DS. Najbardziej jaskrawym i jednocześnie łatwym do identyfikacji przez uczestników badanego świata przykładem takich fałszywych ofert jest nazywanie stanowiska „data scientist" przy wymaganiach ograniczonych do znajomości Excela (por. 6.2.3.), ewentualnie innych aplikacji biurowych, elementów języka SQL (por. 5.3.2.) lub narzędzi do wizualizacji danych (np. MS Power BI, Tableau). Z drugiej strony nazwy stanowisk pracy nie zawierające słów „data science" nie są w ocenianiu konkretnych osób traktowane jako wskaźnik bycia człowiekiem spoza badanego świata. Ogólnie formalne nazewnictwo w miejscach pracy uczestników świata DS jest uznawane za mało informatywne lub po prostu nie nadążające, zapóźnione wobec procesów zachodzących w tym społecznym świecie.

Powtórzmy, że aby być uznanym za uczestnika świata społecznego DS w Polsce należy realizować działanie podstawowe, polegające na pisaniu kodu w ramach wszystkich trzech wyróżnionych obszarów, tj. przetwarzania, analizy i modelowania danych (por. 5.2.) w jednym procesie, by osiągnąć postawiony cel (co określane jest w badanym świecie jako przeprowadzenie projektu od początku do końca). Poza tym należy mieć kontakt z innymi uczestnikami społecznego świata DS, uczestniczyć w wydarzeniach i uczestniczyć w komunikacji badanego świata, czyli brać udział w dyskursie o prawidłowym sposobie wykonywania działania podstawowego (por. Kacperczyk 2016:18), a więc w konsekwencji o granicach świata społecznego i o jego wartościach.

Na wysokim poziomie abstrakcji moglibyśmy powiedzieć, że autentyczny uczestnik świata społecznego to taki, który poza spełnianiem powyższych, podstawowych „kryteriów" uczestnictwa w świecie stanowi dla innych **uosobienie wartości** tego społecznego świata. Pamiętać przy tym należy, że żaden świat społeczny nie jest jednorodnym i niezmiennym

monolitem, zatem to, czyje działanie uosabia wartości danego świata społecznego jest nieostre, płynne, zmienne w czasie, negocjowane, będzie przedmiotem sporów pomiędzy subświatami. Niemniej w naszym przekonaniu opis etnograficzny społecznego świata polega przede wszystkim na podaniu konkretnych szczegółów, zaś w mniejszym stopniu na abstrakcyjnym teoretyzowaniu. Zatem przedstawiamy zidentyfikowane przez nas praktyki zaświadczania o autentyczności uczestnictwa w świecie DS w trzech kategoriach:

- powołania na „właściwe" technologie;
    - narzędzia,
    - metody.
- powołania na doświadczenie w „rozwiązywaniu realnych problemów";
- powołania na cechy osobiste.

Właściwe technologie – narzędzia w świecie DS to te, które umożliwiają realizowanie działania podstawowego w sposób zgodny z wartościami tego świata. Każda z technologii wyróżnionych w centralnej części rysunku 36. (a właściwie w części reprezentującej działanie podstawowe świata DS; zachęcamy do porównania z rysunkiem 24.) jest uznawana za narzędzie właściwe dla uczestnika świata DS. Na razie pomijamy spory dotyczące konkurencyjności i komplementarności narzędzi, co dalej w części 6.2.3. opisujemy jako arenę oraz spory o wartości świata społecznego.

Przykładem popularnego zestawu „właściwych" narzędzi (używane są nie do końca wymienne określenia „stos" i „ekosystem") uczestnika świata DS jest opisany już w części 5.3. sześcioelementowy zestaw technologii open source (Piatetsky 2018d). Składają się nań: **Python** – język programowania; **Anaconda** – dystrybucja Pythona do zastosowań DS (w tym notatnik Jupyter i repozytorium conda); **scikit-learn** – pakiet metod tzw. tradycyjnego uczenia maszynowego; **TensorFlow** (TF) – pakiet stosowany głównie do metod uczenia głębokiego; **Keras** – tzw. nakładka ułatwiająca korzystanie z TF; **Apache Spark** – zunifikowana technologia do rozproszonego przetwarzania danych. Podkreślmy, że ten sześcioelementowy ekosystem pomija narzędzia oczywiste dla uczestników świata DS, i te nie traktowane jako narzędzia charakterystyczne wyłącznie dla DS, czyli pakiety Pythona do zadań przetwarzania i analizy, w tym wizualizacji danych (np. **pandas, NumPy, matplotlib**). Rzecz jasna pominięto też narzędzia, które w naszej pracy opisujemy jako przemilczane / przezroczyste. Chodzi o język **SQL**, znajomość tzw. **terminala i języka bash** oraz wybitnie

oczywistą w kontekście przywołanego źródła samowiedzy[259] znajomość **języka angielskiego**. Wskazywano nam wielokrotnie, że np. oferty pracy na stanowisko data scientist zawierające w wymaganym zestawie narzędzi takie technologie, które nie są uznawane za „właściwe" oceniane są jako podszywanie się pod DS. Tak oferta zawierająca obowiązkową biegłość w zakresie Excela i Tableau / Power BI będzie „faktycznie" ofertą dla analityka danych, zaś jeżeli wymagany jest język programowania Java czy inne niskopoziomowe oraz technologie konteneryzacji będzie to „faktycznie" oferta dla programisty lub dla związanego z DS stanowiska inżynierskiego (data engineer, machine learning engineer).

Łatwą do zaobserwowania praktyką autoprezentacji i zaświadczania o autentyczności uczestników świata DS jest ozdabianie pokrywy personalnego laptopa naklejkami. Pamiętajmy, że laptop jest dla nich zazwyczaj urządzeniem nieodłącznym, noszonym w odpowiednim plecaku niemal zawsze przy sobie. Uczestnicy często widzą się nawzajem ze swoimi laptopami i rozpoznają je. Nieznajome osoby mogą rozpocząć interakcje pytaniem o naklejkę. W trakcie nieformalnej rozmowy podczas jednego z obserwowanych wydarzeń autor (RŻ) był świadkiem następującej sceny: w chwili wychodzenia z pubu zauważono pozostawiony plecak. Znalazcy nie rozpoznając plecaka natychmiast wyciągnęli laptop i oglądając naklejki oraz samo urządzenie zgadywali do kogo należy zguba.

Piszący te słowa miał to samo wrażenie co I. Lowrie – że fabryczny, czysty laptop etnografa jaskrawo wyróżnia się spośród oklejonych laptopów uczestników świata DS i pokrewnych środowisk technicznych (Lowrie 2016:26). Jako badacz (RŻ) zatem chętnie gromadził naklejki i umieszczał je na swojej pokrywie. Naklejki to często logo wykorzystywanych technologii oraz wydarzeń świata społecznego. Widzimy oklejanie laptopów jako praktykę zaświadczania o autentyczności poprzez powołanie na „właściwe" narzędzia i prezentację swojego uczestnictwa w dyskursie badanego świata, a także odwołanie do zdobytego doświadczenia. Pojawiają się także naklejki z logo firm, czasami uczelni, a także np. grup meetupowych lub organizacji organizujących meetupy. Spotyka się naklejki z logo technologii nie będących charakterystycznymi dla DS, a związanymi np. z ruchem open source, oraz rozmaite nalepki prezentujące światopogląd, gust czy zainteresowania:

259 Źródłem są wyniki ankiet (Piatetsky 2018d) anglojęzycznego portalu KDnuggets, z którego korzysta międzynarodowa społeczność DS.

Rysunek 41: Naklejki na pokrywie personalnego laptopa uczestnika świata DS

Źródło: fotografia wykonana przez Remigiusza Żulickiego za zgodą Piotra Migdała

Zgadzamy się z Lowiem (2016:26-28), że laptopy uczestników świata DS są przyozdabiane naklejkami w sposób dość chaotyczny, naklejki mogą nakładać się na siebie, być brudne lub podniszczone. Traktowanie personalnego laptopa w sposób nieco niedbały, nonszalancki, swobodny, jako rzecz nie wymagającą przesadnej troski – narzędzie do ciężkiej pracy i szybkiego zużycia – jest zgodne z podkreślaną w badanym świecie tymczasowością i szybkim dezaktualizowaniem się technologii, a zatem ze zobowiązaniem uczestników do ciągłego samorozwoju.

Posługiwanie się opisywanym wyżej zestawem narzędzi do wykonywania działania podstawowego jest w badanym świecie uznawane za wskaźnik bycia autentycznym uczestnikiem świata DS. Nie jest to wskaźnik jedyny ani jednoznacznie zaświadczający o autentyczności. Jest to raczej zestaw podstawowy (choć oczywiście świadczący o wysokim poziomie znajomości technologii badanego świata przez osobę go użytkującą, bowiem w naszym przekonaniu poziom „wysoki” jest najniższym z akceptowanych dla uczestników badanego świata, por. rys. 43). Przy czym dla większości uczestników badanego świata może

być zestawem najczęściej używanym – także przez uczestników silnie osadzonych w świecie DS. Widzimy jeszcze dwa inne „wskaźniki" odwołujące się do narzędzi – strategie zaświadczania nie tylko o autentyczności, ale i o maestrii uczestnika świata DS.

Po pierwsze, jest to względnie swobodne **posługiwanie się wieloma narzędziami konkurencyjnymi lub komplementarnymi do wykonywania działania podstawowego**. Takie działanie traktowane jako tzw. dobieranie narzędzia do problemu, a więc oceniane jako dobre w kategoriach sprawczości i efektywności. Stanowi ono wyraźny wskaźnik bycia autentycznym, silnie osadzonym uczestnikiem społecznego świata DS. Tyczy się to np. posługiwania się nie tylko najpopularniejszymi językami Python i R, ale językami słabiej w badanym świecie znanymi i jednocześnie posiadającymi inne zalety (jak szybkość) niż R i Python – np. Scala lub Julia. Podobnie rzecz ma się np. z używaniem pakietów do uczenia głębokiego. O maestrii nie świadczy używanie PyTorcha zawsze i do wszystkiego. Wskazuje na nią raczej używanie, w zależności od konkretnych wymagań danego projektu, najbardziej optymalnego pakietu – np. PyTorch, TF, Theano czy CNTK. Nawet w obrębie jednego języka programowania o maestrii świadczy używanie pakietów Tidverse i data.table (do przetwarzania danych), nie tylko jednego z nich do każdego, kolejnego problemu. Jako kod in vivo traktujemy określenie „**agnostyk technologiczny**", które oznacza człowieka nie przywiązanego silnie do konkretnego narzędzia, a zwracającego uwagę głównie na rozwiązywany problem i korzystającego z szerokiej gamy rozmaitych narzędzi. W tym kontekście raczej „wypada" mieć na laptopie naklejki kilku konkurencyjnych technologii, a nie tylko tych komplementarnych, stanowiących jeden ekosystem.

Powoływanie się na doświadczenie w **posługiwaniu się różnymi rodzajami danych z różnych źródeł** także stanowi strategię zaświadczania o autentyczności. Trudno powiedzieć, aby dane były narzędziem, bowiem są raczej materią pracy uczestników świata DS. Uczestnikom dość łatwo jednak ocenić rodzaje danych stosowanych przez konkretną osobę poprzez pryzmat technologii, za pomocą których dane były przechowywane i przetwarzane, a także poprzez pryzmat użytych metod do pracy z tymi danymi. Przykładowo jeżeli ktoś mówi, że w projekcie mają kłopoty z bazą w Mongo DB i planują jutro zacząć modelowanie od LDA, to mowa o dużym wolumenie dokumentów, czyli o danych tekstowych. Zatem za autentycznego uczestnika raczej będzie uznana osoba, która pracowała np. zarówno z danymi tabelarycznym z baz danych z dostępem przez SQL, z fotografiami zapisanymi w plikach oraz z danymi tekstowymi pozyskanymi za pomocą webscrapingu, a nie osoba pracująca wyłącznie z jednym rodzajem danych. Podobnie rzecz ma się z wolumenem danych, zatem o

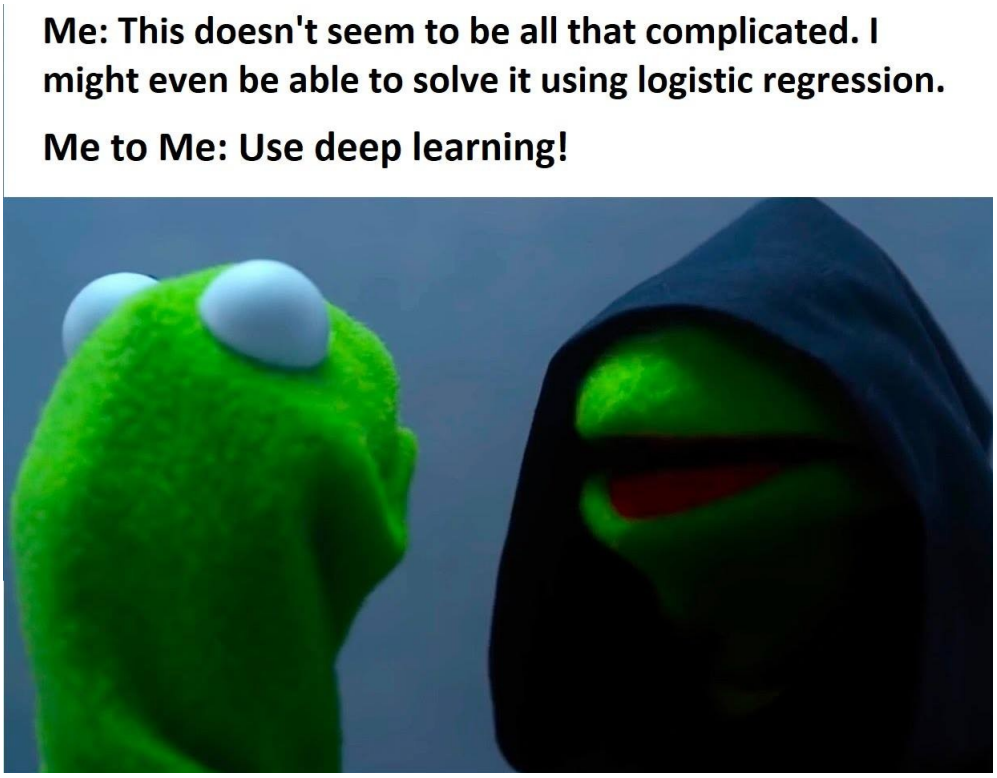
autentyczności będzie zaświadczać raczej doświadczenie zarówno z małą, jak i dużą ilością danych. Zarówno wolumen jak i rodzaj danych (i nierzadko zastosowane metody) uczestnicy potrafią ocenić także przez pryzmat zastosowanej infrastruktury komputerowej. I znowu za bardziej autentyczne mogą być uznawane osoby, które realizowały projekty zarówno na laptopach, jak i na maszynach wirtualnych AWS oraz lokalnych serwerach firmowych.

Po drugie, **samodzielne budowanie narzędzi**. Można być uznanym za autentycznego uczestnika świata DS korzystając jedynie z gotowych rozwiązań, przeważnie przyjmujących formę pakietów. O maestrii świadczyć będzie jednak pisanie własnych pakietów, początkowo raczej na własny użytek „chałupniczy". Może wymagać to użycia innych języków programowania niż Python / R, jak C++. W kolejnych etapach silniejszego osadzania się uczestnika w świecie występuje dość powszechna praktyka udostępniania coraz bardziej dopracowanych wersji budowanych narzędzi w otwartym dostępie, np. na GitHubie. Wraz z dalszym zdobywaniem uznania, łatwo obserwowalnego dla uczestników dzięki dostępowi do informacji m.in. o ilości pobrań pakietu czy tzw. gwiazdek (*star*) na GitHub, dalszym etapem może być publikacja autorskiego pakietu w repozytoriach uznawanych za oficjalne, jak CRAN dla języka R.

Właściwe technologie – metody w świecie DS – to najogólniej rzecz ujmując metody różnorodne. Rzecz jasna strategią zaświadczania o autentyczności jest powoływanie się na dobór właściwej metody do rozwiązywanego problemu praktycznego, wraz z uwzględnieniem specyfiki projektu. Autentyczny uczestnik to zatem taki, który względnie swobodnie korzysta z metod tradycyjnego uczenia maszynowego, uczenia głębokiego (*deep learning*), podstawowych i zaawansowanych metod statystycznych, metod eksploracyjnej analizy danych, a także rozmaitych metod statystycznych lub analitycznych adaptowanych z węższych specjalizacji obliczeniowych dziedzin akademickich, np. biologii, badań operacyjnych. Nie spotkaliśmy się z określeniem **agnostyk metodologiczny**, ale sądzimy, iż można by je zastosować w sposób analogiczny, jak kod in vivo „agnostyk technologiczny". Mówimy przede wszystkim o zadaniach modelowania danych, chcemy jednak wyraźnie podkreślić, że o ile **uczestnik świata społecznego DS musi legitymizować się znajomością i doświadczeniem w stosowaniu rozmaitych metod modelowania danych** (w tym bezwzględnie tradycyjnego uczenia maszynowego i głębokiego), **to nie musi zawsze stosować uczenia maszynowego ani głębokiego**. Wręcz przeciwnie – silnie osadzeni uczestnicy wskazywali nam, że za rozwiązanie rzeczywistego problemu za pomocą modelowania można uznać nawet automatyczne generowanie „tabelki" (w sensie zestawienia

częstości czy statystyk opisowych w podziale na grupy), o ile rozwiązuje to ów problem w optymalny sposób.

Stosowanie zawsze zaawansowanych i najnowszych metod uczenia maszynowego, a szczególnie uczenia głębokiego traktowane jest w badanym świecie jako to, co charakteryzuje uczestników początkujących. Oczywiście muszą oni zaświadczać o swojej autentyczności znajomością wymienionych metod. Jednak przy braku znajomości lub ignorowaniu metod tradycyjnych, w tym statystycznych i analitycznych, a także przy niewielkich umiejętnościach czy doświadczeniu w przetwarzaniu i analizie danych, lub także ignorowaniu tych etapów, początkujący uczestnicy mogą uzyskiwać „słabe" wyniki eksperymentów. Słabe w sensie ilościowych metryk dopasowania modelu do danych, słabe w sensie zbliżania się do rozwiązania rzeczywistego problemu, słabe w sensie kosztów mocy obliczeniowej niezbędnej do pracy modelu w systemie produkcyjnym. Możliwa jest sytuacja, w której model bazujący na zaawansowanej sieci neuronowej pracujący na 2 tys. zmiennych niezależnych daje dopasowane do walidacyjnego zbioru danych na poziomie 98%, jednak zespół DS przedstawia klientowi POC z użyciem znacznie prostszego obliczeniowo modelu pracującego na dwustu zmiennych dającego dopasowanie 96%, ponieważ korzyść wynikająca z szybkości działania i kosztów ewentualnego wdrożenia i utrzymania prostszego modelu na produkcji jest uznawana za wyższą, niż korzyść wynikająca z wyższej miary dopasowania modelu bardziej złożonego do danych. Przypomnijmy, że modelowanie, szczególnie z użyciem uczenia głębokiego i ogromnej mocy obliczeniowych jest w badanym świecie uznawane za fajne, za zabawę, za coś przyjemnego – w przeciwieństwie do raczej nudnej i żmudnej pracy z przygotowaniem i analizą danych przed modelowaniem. Uczestnicy mogą odczuwać coś, co naszym zdaniem możemy nazwać pokusą używania uczenia głębokiego, jako właśnie najfajniejszej metody modelowania danych:



Rysunek 42: Pokusa używania metod uczenia głębokiego a autentyczność uczestnika świata DS – mem internetowy

[tekst mema]
Ja: „To nie wygląda na takie skomplikowane. Może nawet mógłbym to rozwiązać za pomocą regresji logistycznej.” Ja do siebie: „Użyj uczenia głębokiego!”
Źródło:
https://www.facebook.com/artificialintelligencememes/photos/a.1653953787951182/1761188270561066/, dostęp 03.06.2020 r.

Niemniej o **autentyczności, a nawet maestrii uczestnika zaświadcza działanie z użyciem metod najbardziej efektywnych**, najbardziej optymalnie przyczyniających się do rozwiązania problemu, a niekoniecznie metod najbardziej skomplikowanych, zaawansowanych, najciekawszych, najfajniejszych czy najbardziej modnych w danej chwili. Uczenie głębokie, a nawet ML w ogóle, jest naszym zdaniem fetyszyzowane zarówno przez wannabesów i początkujących uczestników badanego świata DS, jak i przez klientów świata DS (por. 6.2.1.2).

Niemniej nawet silnie osadzeni uczestnicy świata DS miewają swoje ulubione języki programowania, ulubione pakiety i ulubione metody, z których mogą korzystać niemal zawsze na wstępnych etapach eksperymentów. W naszych badaniach wskazywano np. na metody z pakietu XGBoost. Dopiero na późniejszych etapach mogą korzystać z szerokiej gamy narzędzi i metod, nawet do etapu customizowania czy budowania narzędzia lub metody na potrzeby konkretnego projektu. Jest to zgodne z iteracyjnym podejściem do pracy w DS i uznawane za zgodne z wartościami badanego świata. Pamiętajmy przy tym, iż bazujące na

uczeniu maszynowym rozwiązania produkcyjne nierzadko składają się z wielu zespolonych modeli wytrenowanych z użyciem kilku różnych metod i wielu narzędzi.

Szczególnie krytykowani przez uczestników osadzonych w badanym świecie są początkujący lub wannabesi nie mający podstaw matematycznych w zakresie np. będącej fundamentem uczenia maszynowego algebry liniowej, a znający jedynie uczenie głębokie z kursów MOOC:

> R: (...) to co ja obserwuję w Polsce, to firmy wolą zatrudnić słabeuszy, jakich mogą znaleźć na rynku i tak jakby nie patrząc na to że potrzebne jest **jakieś** doświadczenie [pauza] **niż** zlecić to na zewnątrz. To jest najczęstszy przypadek.
> *B: Mhm, czyli nie chcą tych danych nie chcą żeby dane wychodziły na zewnątrz?*
> R: Yyy, dwa czynniki. Paranoicznie nie chcą, żeby dane wychodziły na zewnątrz, w sumie nie ważne czy te dane rzeczywiście komuś mogą się przydać czy nie, to jest jedna sprawa, a druga sprawa **nie widzą różnicy**, bo wszystkim się wydaje że wystarczy **kręcić** tymi śrubkami i to wszystko i to załatwia sprawę. Wtedy używam takiego przykładu [pauza] eee, kto sądzisz, że jest lepszym diagnostą? Ktoś kto pracował w **szpitalu**? Przez, nie wiem [pauza] **20 lat**? I spotkał się z najróżniejszymi przypadkami czy ktoś kto [pauza] **obejrzał sobie** wszystkie odcinki doktora House'a.
> *B: No tak [z uśmiechem].*
> R: I jak się poda taki przykład, to powoduje że [pauza] klienci się **trochę** zastanowią że nie do końca tak jest że ktoś sobie potrzaskał sieci, eee, po kursie online,
> *B: No taak [z uśmiechem].*
> R: Tylko jest jednak, aaa, doświadczenie ma jakieś znaczenie. Nie mówię o tym, że doświadczenie musi być **ogromne**, ale tak jakby, cały czas pokutuje to, weźmiemy sobie taką skrzyneczkę, będzie to sieć neuronowa i ona to załatwi wszystko wszystko. {wywiad 5}

Widzimy takie praktyki także jako obronę granic autentycznego DS i własnej pozycji – także pozycji rynkowej, gdyż Rozmówca podał przykład rozmowy z klientem – uczestników silnie osadzonych w DS. W kontekście zaświadczania o autentyczności jest to przykład powołania na doświadczenie w „rozwiązywaniu realnych problemów".

Jakie problemy są w społecznym świecie DS uznawane za realne? Przypomnijmy (por. część 5.2.2.), że działanie podstawowe uczestników badanego świata może być wykonywane komercyjnie, akademicko lub hobbystycznie, zaś sam fakt pracy w firmie lub na uczelni nie ma znaczenia w perspektywie uznawania się i bycia uznawanym za uczestnika badanego świata. Określenie data scientist jest jednak raczej nazwą stanowiska pracy w podmiocie komercyjnym niż nazwą „stosowną" dla każdego uczestnika / uczestniczki badanego świata. Biorąc powyższą konstrukcję jako odzwierciedlającą perspektywę uczestników świata DS twierdzimy, że **za realne problemy uznaje się takie, których rozwiązanie ma wartość dającą się wyrazić w pieniądzach**. Jest to nasze, autorskie ujęcie – w dyskursie badanego świata mówi się raczej o rozmaitych korzyściach utylitarnych: szybciej, taniej, w większej skali, łatwiej itp., odnosząc się wprost do kategorii optymalizacji, efektywności i sprawczości.

Sądzimy jednak, że o autentyczności i o maestrii uczestników jednoznacznie zaświadcza pieniężna wartość rozwiązywanych problemów – zysk finansowy płynący z dokonanej zmiany, ze zwiększonej efektywności jest, naszym zdaniem, traktowanych jako mierzalna metryka sukcesu w projekcie DS. Tym samym **autentyczny uczestnik powinien legitymizować się doświadczeniem w realizacji zyskownych projektów komercyjnych**. Podkreślmy, że nie chodzi o wysokość indywidualnych zarobków. Zarobki nie zaświadczają o autentyczności, bowiem ponadprzeciętnie wysoką pensję może uzyskiwać osoba oceniana negatywnie w kontekście sprawczości (por. określenie „bullshit artists" w części 5.4.2.), w kontekście ciekawości lub samorozwoju (np. w branży finansowej por. 5.4.3.-4.) albo wynagradzana za pracę nie będącą działaniem podstawowym (np. prowadzenie szkoleń). Chodzi zatem o efektywność i sprawczość osiąganą dzięki zrealizowanym projektom, a w szczególności dzięki wdrożonym modelom. Prestiż doświadczenia w zrealizowanym wdrożeniu jest bardzo wysoki, bowiem większość projektów DS kończy się niepowodzeniem, a wiele spośród tych udanych na poziomie POC czy MVP nie zostaje ostatecznie wdrożonych produkcyjnie. Pamiętajmy, że celem części projektów DS w ogóle nie jest wdrożenie produkcyjne. W różnych sposobach organizacji pracy małej firmy lub działu DS w dużej firmie może być i tak, że osoby zajmujące się DS nie mają nic wspólnego z realizowanym – bądź nie – wdrożeniem. Traktując ostrożnie tezy o formowaniu się jakichś „struktur" czy „hierarchii" wtajemniczenia w świat społeczny wydaje się nam, że bliskie perspektywie uczestników badanego świata, a przynajmniej jego dominującej pozycji, może być następujące rozumienie doświadczenia w rozwiązywaniu realnych problemów jako elementu zaświadczania o autentyczności:

- ktoś, kto potrafił wykonać „coś działającego" tzn. zrealizować projekt DS od początku do końca, niezależnie czy w działalności komercyjnej, akademickiej czy hobbystycznej, może być uznany za uczestnika świata DS (przy użyciu właściwych narzędzi, uczestnictwie w dyskursie świata i pozytywnej ocenie w odniesieniu do wartości świata);
- ktoś, kto udowodnił, iż potrafi „dowozić" (por. 5.4.1.) tzn. zrealizował kilka projektów DS w środowisku komercyjnym (przy co najmniej powyższych założeniach) może być uznany za prawdziwego data scientista – sądzimy, że tyczy się to osób posiadających co najmniej kilka lat doświadczenia zawodowego;
- za uczestnika osiągającego poziom maestrii może być uznana osoba, która (przy co najmniej powyższych założeniach) legitymizuje się kilkoma produkcyjnymi

wdrożeniami zrealizowanych projektów komercyjnych o dającej się oszacować wartości zysku dla klientów.

„**Robię to od dziecka**" jest strategią zaświadczania o autentyczności, odwołującą się zarówno do doświadczenia, jak i cech osobistych. Polega na deklarowaniu, że osoba interesowała się i zajmowała się matematyką lub informatyką już w wieku kilku / kilkunastu lat życia.

> [imię i nazwisko prelegenta]
> Człowiek, któremu tata zamiast bajek opowiadał zagadki matematyczne {autoprezentacja prelegenta na meetupie, obserwacja 4}
>
> R: (...) I to się okazuje, że żeby to mieć to dosyć dużą część życia trzeba się tymi rzeczami zajmować. Czyli to są ludzie, którzy nie kopali piłki na podwórku, nie grali w gry, tylko **siedzieli** i na przykład ich ciekawiło dlaczego coś działa a dlaczego coś nie działa w matematyce, tak? No i to są **dziwacy**, znaczy jak się przekroi społeczeństwo to mało jest takich osób, tak mi się wydaje. (...) mamy dosyć dużo ludzi którzy programowali już w gimnazjum czyli to nie było tak że się w szkole nauczyli tylko mieli taki wewnętrzny push żeby iść w tą informatykę, tak? {wywiad 2}
>
> R: (...) padło pytanie, jakie miałaś marzenie w dzieciństwie. I w sumie czasami to tak jest, że jak usłyszysz reakcję innych, to dopiero rozumiesz, że to nie jest naturalne [z uśmiechem]. Ja zawsze chciałam być na styku, w takim sensie, mój tata bardzo szybko zaszczepił we mnie miłość do Exceli, ja **uwielbiałam** rozliczać się z wszystkiego w Excelu, więc ja już na bardzo wczesnym etapie podjęłam decyzję, że ja chcę się zajmować technologiami i wspomagać finanse, yy wtedy. Ograniczyłam się do finansów wtedy, teraz to wyekstrapolowałam na cały biznes. Myślę, że jestem jedną z nielicznych osób, które w gimnazjum wiedziały na jaki kierunek studiów idą. Więc ja dość szybko byłam ukierunkowana, ja nie będę kryła, że to jest po prostu moja pasja. Ja męża mam również w tej samej branży, więc wiesz nasze rozmowy wieczorne są głównie o tym [śmiech]. Wiesz, my w piątek sobie siadamy i odpalamy Andrew Ng [z uśmiechem] i my to lubimy. Ja naprawdę się tym strasznie pasjonuje i nie kryję (…) {wywiad 22}

Tego typu deklaracje pojawiają się także w innych środowiskach i innych kontekstach związanych z technologiami informacyjnymi, np.

> Po czym poznać, że dana osoba 'czuje' postęp? Najłatwiej po tym, co robiła za młodu. Wybitni przedsiębiorcy często bawią się technologią i podejmują różne inicjatywy na długo, zanim wpadną na pomysł, że mogliby z tego zrobić sposób na życie. Reid Hoffman dostał pierwszą pracę (w wieku dwunastu lat) po tym, jak pokazał jednemu z twórców gry, która mu się podobała, instrukcję z zaznaczonymi elementami, które jego zdaniem można by poprawić. Nie zrobił tego, by sobie dorobić - po prostu zależało mu na ulepszeniu gry. Marc Benioff miał piętnaście lat, kiedy sprzedał swój pierwszy program (How To Juggle) i założył firmę, żeby tworzyć gry na Atari 800. Larry Page zbudował drukarkę z klocków Lego (igłową, ale zawsze) (Schmidt i in. 2014:212).

Sądzimy, iż popularność tej strategii w badanym świecie można z jednej strony wyjaśniać tym, że DS jest to w Polsce młody świat społeczny (por. 5.1.4.) składający się

przeważnie z młodych ludzi (por. 5.1.2.). Rzadko można spotkać osoby z wieloletnim doświadczaniem w realizacji działania podstawowego. Powołanie się na zainteresowania i zajęcia kontynuowane od dzieciństwa ma zatem wskazywać na długość – w sensie ilości spędzonych godzin – doświadczenia w obcowaniu z technologiami, niekoniecznie tymi uznawanymi za właściwe w DS, a zatem zaświadczać o swobodzie w ich użytkowaniu, o posiadaniu cennego, bo nie możliwego do nabycia inaczej niż przez doświadczenie repertuaru wiedzy milczącej (*tacit knowledge*). Z drugiej strony opisywana strategia jest popularna dlatego, że odwołuje się do takich wartości świata DS, jak samorozwój, samodzielność, ciekawość i racjonalność. Jeżeli ktoś uczył się programowania i startował w konkursach matematycznych mając 11 lat w roku 2002, kontynuował aktywność na tych polach, a w 2017 zaczął pracę jako data scientist, to dla innych uczestników badanego świata będzie to komunikat, że już od dziecka dużo pracował samodzielnie, uczył się i rozwijał. Szczególnie należy pamiętać, jak w Polsce wyglądały technologie informacyjne i dostęp do materiałów edukacyjnych dotyczących programowania w 2002 roku (nawet nie mówiąc o programowaniu dla dzieci); nasz bohater najpewniej korzystał w dużej mierze ze zdobywanych z trudem źródeł anglojęzycznych. Jasnym komunikatem będzie również, że to osoba ciekawa, „prawdziwie zainteresowana" (por. 5.4.7.), przy założeniu, że nikt nie przekupywał ani nie przymuszał jedenastolatka do konkursów matematycznych i programowania. Komunikat dotyczy również tego, że bohater już jako dziecko myślał logicznie, z użyciem kwantyfikacji, racjonalnie. Potencjalnie wolał zatem przeznaczyć wolny czas na matematykę i programowanie, bowiem lepiej czuł się w tej racjonalnej sferze, niż np. uprawiając sport z rówieśnikami czy bawiąc się / uczestnicząc w życiu towarzyskim. Tutaj jednak opisywana strategia zaświadczania o autentyczności staje się niejednoznacznie akceptowana, bowiem łączy się z, wielokrotnie w naszej pracy przywoływaną, kategorią nerda, lub może tylko ze stereotypowym i utrwalonym w popkulturze wizerunkiem nerdów. Opisał to E. Babbie, w kontekście obaw studentów socjologii przed nauką statystki:

> Płeć? Zgaduję, że pewnie był to chłopak – tak w każdym razie twierdziła większość moich studentów. (…) Marvin [fikcyjny szkolny geniusz matematyczny] z reguły jest kiepsko zbudowany (albo za chudy, albo za gruby). Zwykle nosi okulary i w ogóle sprawia wrażenie słabeusza. Przez cały okres chodzenia do szkoły zapraszany jest na (średnio) 1, 2 imprezy, na których nikt z nim nie rozmawia. Ma fatalną cerę. Jest niedostosowany społecznie, nikt nie chciałby być na jego miejscu i w ogóle budzi raczej litość niż zazdrość. W miarę jak rozmawiałem o Marvinie z moimi studentami, stawał się dla mnie coraz bardziej oczywiste, że większość z nich wiąże sprawność w matematyce z jego godną pożałowania charakterystyką. (…) Wszystko to, co powiedziałem o Marvinie, odpowiada powszechnemu stereotypowi – ale tylko stereotypowi: to nie ma nic wspólnego z biologią. Są kobiety świetne w matematyce. Atrakcyjni ludzie potrafią liczyć nie gorzej niż nieatrakcyjni. (…) Tak więc jeśli należysz do

> tych, którzy uważają się za „niezbyt dobrych z matematyki”, to być może chodzi o twój podświadomy strach, że jeśli w kursie statystyki pójdzie ci lepiej niż źle, to, powiedzmy, dostaniesz pryszczy (Babbie 2006:479–80).

Nerd ma dla części uczestników świata DS pozytywne konotacje, może być kategorią pozytywnej autoidentyfikacji jako osoby technicznej, a także zaświadczać o byciu „prawdziwie zainteresowanym”. Dla innych konotacje są neutralne lub negatywne, występuje czasem odcinanie się od bycia nerdem jako kimś, kto przesadnie skupia się na technologiach (narzędziach lub metodach) pracy z danymi, a zbyt mało uwagi poświęca na rozwiązywaniu prawdziwych problemów.

To, co opisujemy jako strategię zaświadczania o autentyczności pomaga zrozumieć przemilczane technologie społecznego świata. Podkreślmy, że przemilczane lub co najmniej mało problematyzowane technologie w świecie DS, to te znane od kilkudziesięciu lat (SQL, terminal i język bash, podstawy HTML/CSS/JavaScript).

> *B: Ja rozumiem, że to jest tak oczywiste, że wiesz, yy ale czy na ekonometrii uczyli cię korzystać z terminala?*
> R: No nie, na ekonometrii na pewno nie.
> *B: To skąd to umiesz?*
> R: [pauza] Nie wiem, ja nie umiem sobie nawet przypomnieć, w którym momencie się nauczyłam tego. Niee, to raczej w ogóle przed przed studiami, na pewno grubo, grubo. Dużo wcześniej.
> *B: Ok, widzę, że jest szok, cieszę się [ze śmiechem]*
> R: Tak, bo zadajesz takieeee [z uśmiechem], wiesz [pauza]
> *B: Jasne że dla ciebie to jest oczywiste.*
> R: Ale to jest słuszne pytanie, boo nie wszyscy przychodzą z backgroundem IT na dzień dobry.
> *B: To nie jest oczywiste dla ludzi, którzy którzy słyszą pierwszy raz w życiu o data science, nie?*
> R: Nie, no to tak, tylko że widzisz, znajomość terminala dla mnie w ogóle w IT, jeśli chcesz coś robić, no to jak najbardziej takich komend linuksowych, y SQL jak najbardziej, znajomość, zrozumienie w ogóle baz danych relacyjnych, co to są relacyjne bazy danych. Rozumienie co to są dane, dane ustrukturyzowane, nieustrukturyzowane, takich podstawowych pojęć, yyy znajomość Pythona i wszystkich tych bibliotek, czy tam R, związanych z data science. Wydaje mi się, że już z takich podstaw no to,
> *B: A yy takie takie webowe rzeczy HTML, JavaScript? To ma jakieś znaczenie?*
> R: Nie do końca. Nie musisz tego znać. Ja się bawiłam teraz w JavaScript, bo sobie rozwijałam właśnie [nazwa projektu realizowanego hobbystycznie], ale JavaScript jest tak nieprzyjemna dla mnie, że ja nie szłam dalej w ten język. No a HTML [z uśmiechem] wiesz, **ja jak miałam 10 lat** jak tworzyłam pierwsze strony, więc bardziej dlatego to znam, ale nie, nie jest to podstawa. {wywiad 22}

Jeżeli bowiem jest tak, że obecnie liczne grono uczestników poznało pewne technologie będąc dziećmi i posługuje się nimi od kilkunastu czy dwudziestu kilku lat, to te technologie są dla nich przezroczyste i w konsekwencji przemilczane, zgodnie z następującą logiką – skoro ja i całe moje otoczenie korzystamy z nich od zawsze, to nie ma czego problematyzować. Tym samym raczej nie problematyzują tych technologii nawet ci uczestnicy, dla których owe

technologie nie są przezroczyste i nie „robią tego od dziecka” – możliwe, że w ramach strategii nie podkopywania własnej autentyczności.

Sądzimy, że „robię to od dziecka” może w badanym świecie zaświadczać o tzw. prawdziwym zainteresowaniu DS także dlatego, że jest strategią odwołującą się do osobistych, może nawet tzw. wrodzonych, jakoby uwarunkowanych biologicznie zdolności, predyspozycji czy zainteresowań obejmujących matematykę i informatykę:

> *B: OK yyy, jak to się stało, że poszłaś na takie studia?*
> R: Ja w zasadzie miałam jedynie dylemat czy pójść na informatykę, czy na matematykę, tak. Na informatykę na politechnikę, na matematykę na uniwersytet iii to były dwa miejsca gdzie składałam papiery. Miałam **zawsze** zamiłowanie do matematyki i wręcz przeciwne uczucia do przedmiotów humanistycznych [z uśmiechem], więc tu wątpliwości nie było. Nie bardzo widziałam się w jakieś innej dziedzinie. (...) Natomiast faktycznie, umysł ścisły odziedziczyłam po rodzicach i taką smykałkę. {wywiad 24}

Psychologowie wskazywali, że w szeroko pojętym zachodnim kręgu kulturowym matematyka jest czy była uznawana za trudną i jednocześnie niekoniecznie ważną w życiu codziennym – w USA sprzedawano lalki Barbie z nagranym zdaniem „matematyka jest trudna” (*math class is hard*) – oraz, że osiągnięcia matematyczne przypisywane są zaś w większym stopniu zdolnościom (*aptitude*) niż wysiłkowi włożonemu w naukę (Ashcraft 2002). Zakładamy, że w Polsce jest podobnie.

> Przez ćwierć wieku polscy uczniowie nie zdawali matury z matematyki (obowiązkowy egzamin maturalny z matematyki zdawany był po raz pierwszy po wieloletniej przerwie w roku 2010). Można było skończyć liceum, nie rozumiejąc podstawowych pojęć matematycznych. Dlatego matematyka nabrała cech wiedzy elitarnej (Skura i Lisicki 2018:52).

Zatem bez wątpienia strategią zaświadczania o autentyczności w badanym świecie może być powołanie się na zainteresowanie, zdolności, wiedzę i umiejętności matematyczne. Jednak w dyskursie świata DS różne pozycje różnie akcentują wagę znajomości matematyki, choć rzecz jasna nie jest obecna pozycja lekceważąca matematykę. Za **próg wejścia do świata DS może być przyjmowany poziom znajomości matematyki tak od poziomu zaawansowanego egzaminu maturalnego jak i od algebry liniowej / statystyki matematycznej**. Nawet w książce przedstawiającej podstawy ML dla osób nie-technicznych, dla osób raczej zajmujących się biznesem niż chcącym wejść do świata DS, stwierdzono, że temat „może zrozumieć każdy, kto podczas studiów miał semestr matematyki lub algebry liniowej” (Foreman 2017:17).

Stereotypowo w badanym świecie mówi się o silniejszym nastawieniu na znajomość matematyki/statystyki w subświecie użytkowników języka R, zaś silniejszym nastawieniu na

umiejętności koderskie w subświecie używających Pythona. Taka wizja jest przesadnie uproszczona, co opisujemy bardziej szczegółowo w części 6.2.3. Również bez wątpienia strategią zaświadczania o autentyczności może być powołanie się na zainteresowanie, zdolności, wiedzę i umiejętności informatyczne, a szczególnie koderskie. Powtórzmy jednak do znudzenia – data scientist nie jest programistą (*software engineer / developer*) i zaawansowane umiejętności programistyczne, np. płynne posługiwanie się niskopoziomowymi językami programowania nie zaświadcza o autentyczności uczestnika. **Próg wejścia do świata DS stanowi raczej znajomość R lub Pythona pozwalająca realizować projekty od początku do końca**. Standardy pisania kodu są różne w różnych subświatach, w różnych firmach i zespołach DS, ale ogólnie rzecz ujmując prawdziwy uczestnik musi samodzielnie i efektywnie radzić sobie ze wszystkimi zadaniami koderskimi na każdym etapie iteracyjnej realizacji projektu. Podobnie data scientist nie jest matematykiem-teoretykiem, a o autentyczności nie zaświadcza znajomość ani abstrakcyjnych dziedzin matematyki, ani nawet zaawansowanych szczegółów matematycznych w algorytmach uczenia maszynowego. To drugie może być podstawą do osiągnięcia poziomu maestrii w posługiwaniu się metodami i narzędziami ML, ale samo w sobie nie będzie w badanym świecie wysoko cenione z uwagi na małą sprawczość.

Podsumowując, **uczestnicy badanego świata zaświadczają o autentyczności poprzez kategorię bycia osobą techniczną**.

> *B: A macie w teamie, jakąś osobę która jest waszym przełożonym? Jak to jest, jest jakaś taka hierarchia?*
> R: Tak, jest jeden główny [krótka pauza] główny manager, który jest w stopniu chyba dyrektor ale nie pamiętam, przepraszam jak źle pamiętam [z uśmiechem] iii
> *B: Nie, no wiesz, jeżeli jeżeli dla ciebie to nie ma jakiegoś takiego większego znaczenia, no to to to, dla mnie to jest ok [z uśmiechem]*
> R: [śmiech] Iii oprócz tego jest jeszcze [krótka pauza] e dwóch managerów. No, tak mniej więcej wygląda wygląda struktura.
> *B: A ci managerowie też są data scientistami?*
> R: **Tak**, wszyscy którzy są w naszym zespole są **technicznymi** osobami, nie ma takich osób, które na przykład zajmują się tylko ymm product managmentem, albo tylko są product ownerami. Nie, wszyscy muszą **usiąść** [ze śmiechem] i wszyscy wszyscy programują, wszyscy zajmują się data science. {wywiad 20}

Może to oznaczać różny miks zainteresowań, zdolności, wiedzy, umiejętności matematycznych i informatycznych. Ten miks musi być jednak taki, by pozwalał na efektywne i samodzielne wykonywanie działania podstawowego – pracy na danych cyfrowych. W środowisku związanym z Politechniką Warszawską trwały dyskusje dotyczące tłumaczenia nazwy data science na język polski – ostatecznie kierunek nazwano „Inżynieria i

Analiza Danych". Argumentacja na rzecz tłumaczenia data science jako „danetyka", już nie jako kierunku studiów, ale jako dziedziny, jest przykładem na postrzeganie autentyczności uczestnictwa w badanym świecie – w kategoriach „osoby technicznej" oraz w utylitarnej wartości realizowanych projektów:

> Wydaje się, że nazwy nauk (w szczególności tych, z którymi mamy styczność najczęściej) powstające poprzez dodanie końcówki -logia są częściej humanistyczne lub przyrodnicze niż techniczne (np. filologia, psychologia, socjologia, politologia, biologia, geologia, paleontologia). Stąd końcówka -logia jest w pewnym stopniu nacechowana tymi dziedzinami. Dlatego **danologia rodzi we mnie odczucie, że jest to nauka humanistyczna o danych, w której prowadzi się rozważania filozoficzne o tym, czym są dane, jaką rolę odgrywają w życiu, itd. Dlatego zupełnie nie pasuje to do technicznej natury tej dziedziny, w której na co dzień się programuje, pracuje z różnymi technologiami, stosuje lub opracowuje algorytmy matematyczne**. (...) Tłumaczenie Data Science jako danologia w moim przekonaniu wypacza również samą naturę tej dziedziny. (...) **Bez kontekstu biznesowego - samo stosowanie algorytmów i wykorzystywanie jakichś technologii - to nie jest data science**! Dlatego sprowadzanie data science do nauki jest całkowicie błędnym kierunkiem. To jest drugi z poważnych powodów, dla których **o niebo lepszym tłumaczeniem jest danetyka - jest ona wolna on jednoznacznego interpretowania jej jako nauka (a po przyswojeniu sobie tego terminu, naturalnie będzie kojarzyć się z biznesem). Dlaczego? Ponieważ końcówka -tyka jest dużo częściej stosowana w przypadku praktycznych nauk, które jednocześnie są gałęziami biznesu**. Spójrzmy na informatykę - czy czytając ten termin myślimy o jakiejś "nauce"? Ja automatycznie myślę o branży IT, chociaż oczywiście jest to też nauka. Robotyka - moje pierwsze skojarzenie to stosowane w przemyśle mechanizmy i inżynierowie konstruujący użytecznie maszyny/roboty. Ale jednocześnie robotyką można nazywać gałąź nauki. Astronautyka - moje pierwsze skojarzenie to praca inżynieryjna związana z lotami w kosmos. Oczywiście to też jest nauka. Dlatego **danetyka wpisuje się w tę prawidłowość - mi nasuwa ona skojarzenie z "inżynierami" danych, którzy działają w biznesie. Gdy usłyszymy od kogoś z biznesu, że pracuje w danetyce, to brzmi to rozsądnie - od razu czuć, że pracuje w obszarze technicznym w dziale/zespole związanym z danymi** [podkr. RŻ] (Ryciak 2019).

Mniej jednoznaczną strategią zaświadczania o autentyczności jest powołanie na utylitarną wartość zrealizowanych projektów. Zasadniczo do tego można sprowadzić to, co wyżej nazywaliśmy rozwiązywaniem realnych problemów – ale pamiętajmy, że i ta kategoria bywa w badanym świecie uznawana za biznesową kliszę i część uczestników woli mówić właśnie o dostarczaniu wartości (bo np. w trakcie realizacji dopiero ustala się problem lub problem zostaje porzucony na rzecz innego, bardzie wartościowego por. 5.2.3). W badanym świecie nie występuje legitymizacja tylko poprzez odniesienie do tego, jaką wartość utylitarną miały zrealizowane projekty. To, że ktoś generował poważne zyski dla firmy jako np. specjalista w zakresie marketingu (tzw. ekspert dziedzinowy) nie zaświadcza przecież, że jest prawdziwym data scientistem. Taka osoba może przejść na stanowisko data scientist, a jej ekspercka znajomość dziedziny może być uznawana za zaletę w tej roli zawodowej (opisywaliśmy to wcześniej jako jedną ze ścieżek wejścia do świata DS), niemniej by być w badanym świecie postrzegana jako data scientist taka osoba musi zaświadczyć o swoich

kompetencjach technicznych w zakresie narzędzi i metod uznawanych przez uczestników świata DS za odpowiednie. Osoby nie-techniczne, jak właśnie eksperci dziedzinowi czy różnego rodzaju role odpowiedzialne za np. zarządzanie, sprzedaż, marketing, HR, księgowość, produkcję, administrację, są często przez uczestników badanego świata przyporządkowywane do szerokiej kategorii osób biznesowych czy po prostu biznesu. Mianem biznesu przeważnie określa się w badanym świecie klientów – zarówno wewnętrznych i zewnętrznych – ale określenie to może być stosowane także wobec przełożonych czy osób wspomagających realizacje projektów od tzw. strony biznesowej (jak np. project manager, product owner, marketingowcy), o ile uważa się je za osoby nie-techniczne.

Powrócimy w tym miejscu do **kategorii nerda w DS** i powiążmy to z areną dotyczącą wykonywania działania podstawowego: czy uczestnicy realizują działanie podstawowe żeby rozwiązać problem (dostarczyć wartość utylitarną), czy zrobić model z użyciem danych (dostarczyć układ technologiczny)? Naszym zdaniem w badanym świecie dominuje pozycja pozytywnego nastawienia do kategorii nerda (jako osoby technicznej i „prawdziwie zainteresowanej"), a także pozycja uznająca rozwiązywanie problemów za cel działania podstawowego, co w wybranych aspektach jest ze sobą sprzeczne. Mające charakteryzować nerdów niskie umiejętności komunikacyjne (lub tylko brak zainteresowania tą sferą pracy, por. 5.3.4.) mogą przecież być przeciwskuteczne w rozwiązywaniu problemów tzw. rzeczywistych, wysokiego poziomu (nie problemów w rodzaju „nie działa mi ta linijka kodu" czy „import danych trwa zbyt długo"), bowiem w sytuacjach innych niż stawianie problemów samemu sobie, będzie to zawsze polegać w pierwszej kolejności na komunikacji z innymi ludźmi i zrozumieniu ich kategorii – na czym polega cel, jaki stan będzie oznaczał osiągnięcie celu, na czym będzie polegał sukces a na czym porażka itp. Z innej strony, mające charakteryzować nerdów myślenie racjonalne, logiczne i abstrakcyjne będzie skuteczne (a może tylko postrzegane jako skuteczne przez pozycję uznającą nerdowatość za pozytywną cechę) w rozwiązywaniu problemów w tym sensie, że sprzyja koncentracji na samym problemie, a nie na ludziach, np. tym, że ktoś w firmie klienta boi się zmian, które jego zdaniem przyniesie skutecznie wdrożony model uczenia maszynowego.

W naszym przekonaniu najważniejsze jest jednak, że nerd jest kategorią kojarzoną w badanym świecie z raczej niewielką znajomością biznesu i niewielkim nim zainteresowaniem. Zatem z osobami biegłymi technicznie i zorientowanymi silnie na techniczne aspekty realizowanych projektów, a pomijającymi aspekty utylitarne. Zdaniem niektórych

uczestników świata DS najbardziej pożądanym połączeniem jest jednocześnie bardzo dobra znajomość narzędzi, metod, biznesu i umiejętności komunikacyjnych:

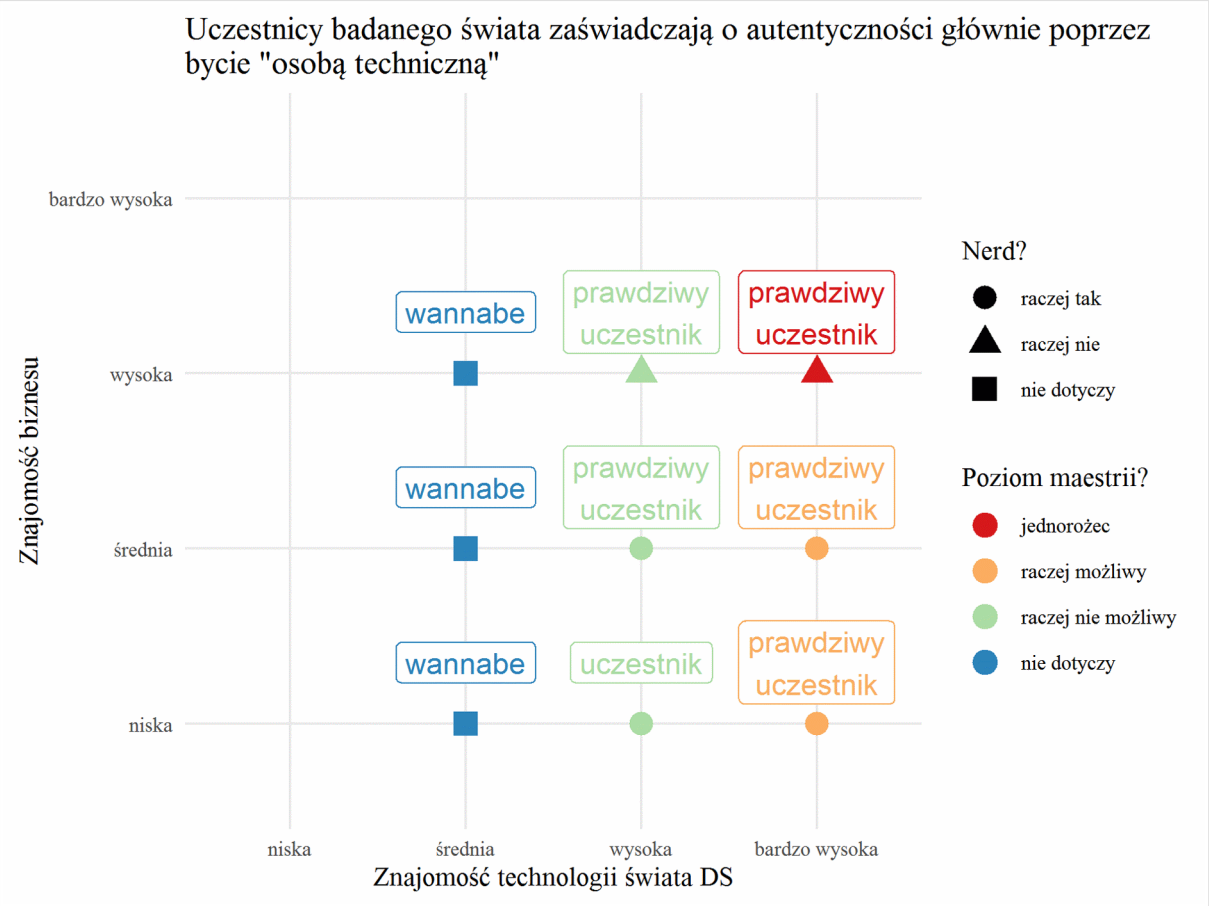
> R: (...) siadam z nimi i też pomagam im w tym, jak oni mają opowiadać biznesowi o tym co robią, żeby to było zrozumiałe dla biznesu. To jest często bardzo duży problem. Zwłaszcza dla nich jest dużym problemem oddzielić to co jest benefitem biznesowym, a co technologicznym, czyli to, że oni wybrali najlepszy algorytm i jest wysokie accuracy[260], no to to nie jest to jak do końca sprzedajesz biznesowi, tylko raczej mówisz o wynikach, jak to będzie. Chociaż coraz więcej biznes, no jednak ten poziom typu co to są metryki sukcesu, rozumie. Ale jakie są wszystkie algorytmy, no to nie wszyscy. Ale musisz umieć to wytłumaczyć, więc ja w tym bardzo im pomagam, żeby jednak oni umieli to wytłumaczyć, no bo nie oszukujmy się, też data scientist powinien umieć biznesowo wytłumaczyć co on zrobił. To często nie jest łatwe, oni muszą się dostosować do tego poziomu, jaki biznes ma jaką wiedzę technologicznie.
> *B: A jak to wynika z twojego doświadczenia, czy dużo jest takich osób w data science, które właśnie widzą ten cel biznesowy i potrafią o nich rozmawiać, czy raczej jest dużo takich osób, które widzą accuracy i które widzą metodę, że teraz zrobimy, nie wiem, GAN-y i będzie tak super?*
> R: Osoby, które rozumieją biznes to są **perełki**. Powiem wprost. Ja nie kryję też, że ja tak też miałam, dlatego miałam tyle ofert, bo jak ktoś ze mną rozmawiał i widział, że ja rozumiem biznes i jestem w stanie opowiedzieć o biznesie, to było takie [pauza] **ok**, mało takich osób, starajmy się wziąć. Tak, nadal jest mało. To też wynika z tego, że większość data scientistów, mimo wszystko to są osoby, które wyrosły z programistów, a prawda jest taka, że i teraz jest coraz większy taki, znaczy taki też powinien być trend, że te osoby z wiedzą biznesową powinny też się kształcić bardziej w data science i to wtedy są najlepsi data scientiści, którzy mają taką wiedzę biznesową. {wywiad 22}

Zatem chociaż osoby legitymizujące się jednocześnie bardzo wysoką znajomością technologii i wysoką znajomością biznesu (por. rys. 43.) mogą być uznawane za wyjątkowo wartościowych data scientistów, tzw. „jednorożce" czyli cenne, rzadkie, bez mała mityczne hybrydy, to osoby biegłe technologicznie (oczywiście w zakresie właściwych narzędzi do wykonywania działania podstawowego w świecie DS), a nie posiadające wysokiej znajomości biznesu także definiuje się w badanym świecie jako prawdziwych uczestników (rys. 43). Z drugiej strony samo bycie nerdem nie zaświadcza o byciu uczestnikiem świata DS, ani o autentyczności uczestnictwa. Mówiąc językiem badanego świata: pewne cechy nerdów korelują z byciem data scientistem, ale nie powodują go (Kuncewicz 2019c).

Na rysunku 43. – mapie pozycyjnej – prezentujemy syntezę powyższych ustaleń w pięciu wymiarach: 1. znajomości technologii świata DS, 2. znajomości biznesu, 3. autentyczności uczestnictwa (od wannabesów, nie uznawanych za uczestników, przez uczestników do prawdziwych uczestników), 4. poziomu maestrii, 5. odniesienia do kategorii

260 *Accuracy* w nadzorowanym uczeniu maszynowym to dopasowanie odpowiedzi modelu do wartości zmiennej zależnej, wyrażone jako stosunek wszystkich odpowiedzi do odpowiedzi poprawnych (czyli zgodnych z wartościami zmiennej zależnej), np. 0,9 lub 90%.

nerda. Na tej mapie redukujemy znajomość narzędzi i metod do jednego wymiaru (1.), pomijamy zaś uczestnictwo w dyskursie świata DS.



Rysunek 43: Zaświadczanie o autentyczność poprzez znajomość technologii świata DS a znajomość biznesu z uwzględnieniem poziomu maestrii uczestnika i kategorii nerda – mapa pozycyjna

Uwaga 1: etykiety barwne znajdują się ponad punktami danych (np. „uczestnik": znajomość technologii świata DS – wysoka i znajomość biznesu – niska)
Uwaga 2: brak punktu danych na mapie oznacza pozycję nieobecną w dyskursie społecznego świata DS (np. znajomość technologii świata DS – niska i znajomość biznesu – bardzo wysoka)
Źródło: opracowanie własne za pomocą GNU R

W proponowanym ujęciu za nerdów raczej nie są uznawane osoby o wysokiej znajomości biznesu. Osoby uznawane za reprezentujące poziom maestrii muszą legitymizować się bardzo wysoką znajomością technologii badanego świata (np. zbudować kilka własnych pakietów czy dowieść praktycznej znajomości kilku różnych języków programowania), zaś poziom znajomości biznesu nie ma w tym aspekcie większego znaczenia. Znajomość biznesu może pomagać w zaświadczaniu o autentyczności u uczestników początkujących, o niższej (aczkolwiek i tak wysokiej) znajomości technologii. Przykładowo za autentycznego uczestnika będzie raczej uznawana osoba będąca magistrem informatyki i posiadająca dwa lata doświadczenia zawodowego, jedno wdrożenie

produkcyjne i potrafiąca rozmawiać z „biznesem” (klientami nie-technicznymi) niż osoba przygotowująca się do obrony doktoratu z matematyki w zakresie rozwinięcia sieci neuronowych typu GAN, a nie mająca doświadczenia komercyjnego. Bez wątpienia także osoba z doświadczeniem komercyjnym może faktycznie nie interesować się tzw. biznesem i cechować się niską jego znajomością. Być może z punktu widzenia pracodawców i właścicieli firm tzw. nerdzi są postrzegani jako bezpieczni, bezkonfliktowi pracownicy, którzy nie będą potrafili przejąć inicjatywy biznesowej, władzy w firmie / dziale / zespole, nie podkupią intratnego klienta do swojej firmy prowadzonej „na boku”, ogółem nie są groźni w sensie biznesowym, bo biznes ich nie interesuje. Jakoby wystarczy tymże nerdom dać dostęp do dobrego sprzętu komputerowego, wysoki budżet szkoleniowy, dawać ciekawe łamigłówki do rozwiązywania, pozwolić przychodzić do pracy w klapkach i nie zmuszać do rozmów z klientami, a obie strony będą zadowolone?

Chcemy przede wszystkim pokazać to, że w badanym świecie znajomość technologii jest uznawana za pierwotną w tym sensie, że bez niej nie można zostać uznanym za uczestnika. Znajomość biznesu czy bardziej abstrakcyjnie ujęte dostarczanie wartości utylitarnej jest uznawane za wtórne wobec technologii w tym sensie, że jest osiągane dzięki zastosowaniu „właściwych technologii” i osiągane wraz z doświadczeniem w stosowaniu tychże technologii. W tym sensie wskazywano nam, że **osiąganie celów o charakterze technologicznych i celów utylitarnych może być dla data scientisów tym samym**:

> [R opowiada ogólnikowo w języku technicznym o ostatnim projekcie – tajemnica służbowa]
> *B: Mhm, ok. Yyy, a możesz powiedzieć do czego to jest potrzebne, tak ostatecznie?*
> R: Y, to będzie część większego produktu, który będzie wypuszczony, y ale ten projekt [z uśmiechem], ten produkt jeszcze nie jest dostępny dla, dla ogólnie, dla powiedzmy dla konsumenta, więc nie wiem jak bardzo mogę wejść w szczegóły znowu.
> *B: Ok, rozumiem. Nie, bo wiesz co, pytam o to z tego powodu, że mmm słyszałem [krótka pauza] jakby dwie rzeczy, od od od data scientistów. Są tacy ludzie którzy mówią, że to co my robimy, no to musi przynieść korzyści biznesowe i to jest najważniejsze, że my nie robimy sztuki dla sztuki, tak.*
> R: Aha
> *B: A inni mówią bardziej, że o my to tam musimy mieć accuracy takie a takie, bardziej mówią o jakichś takich metrykach typowo technicznych, tak. Dlatego, dlatego zapytałem do czego to jest potrzebne, bo jestem ciekaw czy dla ciebie to ma znaczenie do czego to jest potrzebne, czy bardziej dla ciebie ma znaczenie to, no że to jest właśnie przekształcanie języka naturalnego za pomocą jakiejś, jakiejś metody?*
> B: [westchnienie] Myślę że [krótka pauza] nasz, nasi, osoby postawione na tych wysokich szczeblach, osoby decyzyjne starają się znaleźć jakiś balans pomiędzy tym, pomiędzy tym żeby te produktu które my dostarczamy, żeby one były state of the art, żeby wyciągały jakąś, żeby dawały wyniki, które można powiedzieć że to są dobre wyniki, ale też yy starają się tak prowadzić projektami, żeby one przynosiły biznesową korzyść, więc ja bym powiedziała że, dla mnie [krótka pauza] ja się zgadzam z takim podejściem iii dla mnie to powinno być w **jakiś sposób** zbalansowane, bo [krótka pauza] myślę też że to jest też trochę związane, jedno jest

> zależne od drugiego, boo jeżeli ten produkt nie będzie osiągał dobrych wyników, nie będzie dawał jakiejś dobrej dokładności, too on też nie może przynieść później korzyści biznesowej, bo to będzie, nie wiem brzydko mówiąc, bubel trochę [ze śmiechem] [pauza] Ja to tak widzę.

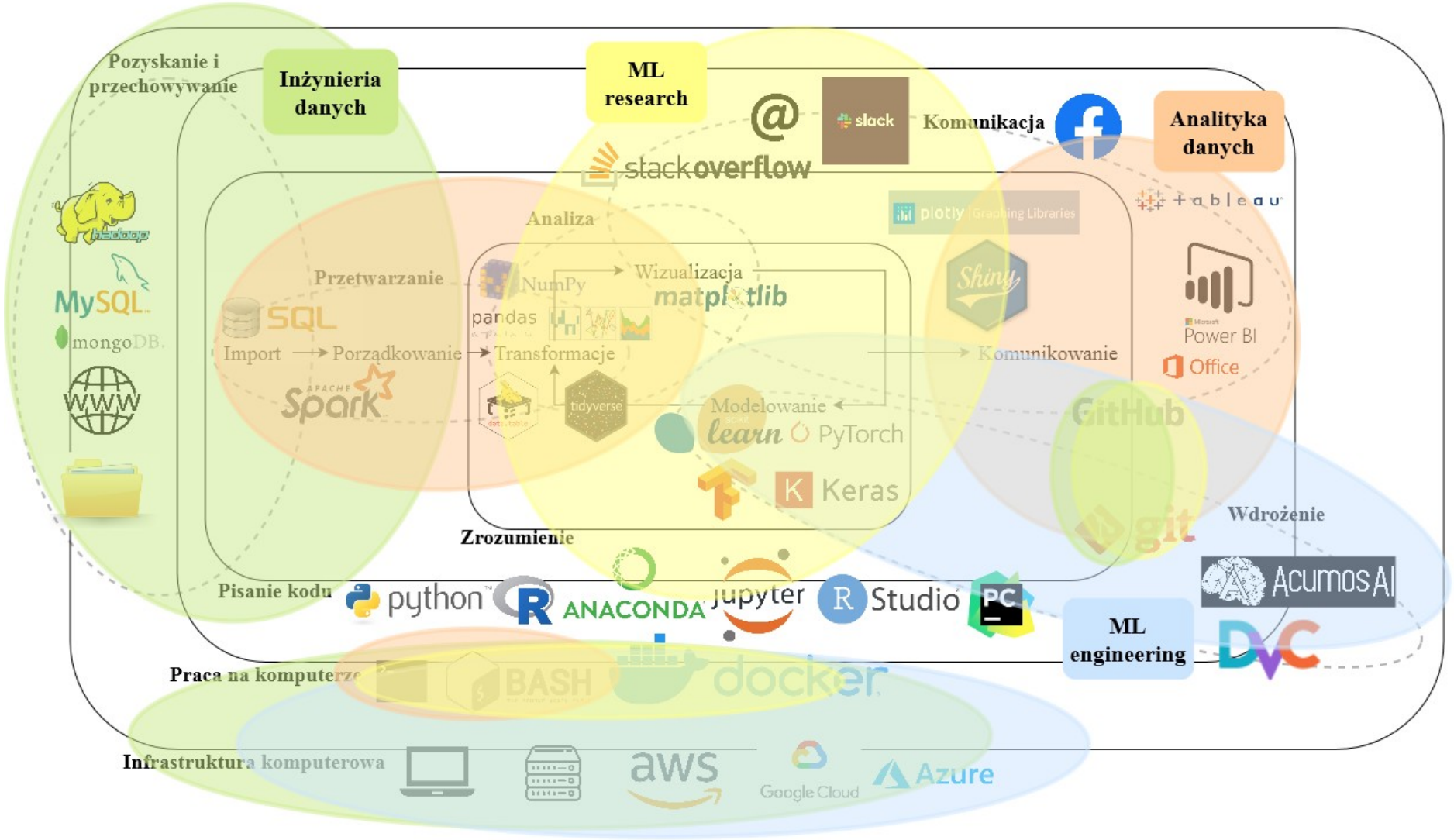
Choć wiele uwagi poświęciliśmy strategiom zaświadczania o byciu prawdziwym data scientistem, to musimy powiedzieć, że dla części uczestników, a nawet uczestników silnie osadzonych „data scientist" nie jest kategorią autoidentyfikacji. Częściowo ze względu na pojawianie się na rynku pracy stanowisk określanych jako fake (dosł. fałszywa) data science, uznawanych za faktyczne zadania analityka danych, częściowo chyba ogólnie jako wyraz procesów segmentacji, a szczególnie profesjonalizacji badanego świata **już od około 2018 r. w Polsce widzimy utratę prestiżu określenia data scientist**. Niektórzy uczestnicy mogą wręcz unikać tego określenia, jako np. nieostrego, coraz bardziej pospolitego czy powszechnie fałszowanego na rzeczy określeń koncentrujących się wokół uczenia maszynowego – ML engineer oraz ML researcher.

6.1.3. Segmentacja – profesjonalizacja

ML engineer, ML researcher to obok tzw. **inżynierii danych** subświaty profesjonalne badanego świata DS. Dwa pierwsze z wymienionych widzimy jednoznacznie jako wyodrębnianie się specjalizacji z DS. Inżynieria danych powstała zaś na przecięciu trzech obszarów – po pierwsze tego, co nazywamy w naszej pracy światem DS, po drugie, big data w rozumieniu technicznym, a po trzecie, dość stabilną częścią tradycyjnej informatyki, czyli bazami danych. Wyróżniamy jeszcze inny segment / subświat **analityki danych**, będący polem starszym i nieco bardziej ustabilizowanym niż DS. Ten segment chcemy opisać jako wchłonięty i przekształcony w części uznawanej za „podzbiór" świata DS. Inna część analityki danych jest przez uczestników świata DS postrzegana jako zupełnie odrębna od DS. Widzimy jasną granicę światów wyznaczoną przez właściwy sposób wykonywania działania podstawowego – pisanie kodu. Ta część analityki danych, która domyślnie wykonuje pracę na danych za pomocą interfejsów graficznych jest poza DS, zaś ta kodująca może być postrzegana jako właśnie wspomniany podzbiór. Subświaty ML engineerów / researcherów i inżynierów danych to subświaty domyślnie kodujące.

Skoro stos technologiczny jest w naszym ujęciu podstawowym obszarem odniesienia w praktykach zaświadczania o autentyczności w świecie społecznym DS, to jest także podstawą w wyznaczaniu granic społecznego świata DS oraz jego subświatów. Na rys. 44.

prezentujemy szkic granic omawianych subświatów w odniesieniu do stosu technologicznego świata DS oraz działania (por. rys. 36 i tab. 2).



Rysunek 44: Subświaty / segmenty społecznego świata DS w Polsce – inżynieria danych, analityka danych, ML research, ML engineering w odniesieniu do stosu technologicznego i działania podstawowego DS

Źródło: opracowanie własne za pomocą draw.io, logotypy pobrane z oficjalnych stron www dostawców, inne rysunki na licencjach zezwalających na użycie
Uwaga: porównaj z rys. 36, 24 oraz tab. 2.

W badanym świecie panuje względna zgodność co do tego, że **analityka danych różni się od DS brakiem stosowania metod uczenia maszynowego**. Tym samym analityka, posługując się podobnymi narzędziami do przetwarzania i analizy danych w bardzo małym stopniu nastawiona jest na budowanie „czegoś działającego", a w większym stopniu na raporty lub narzędzia do automatyzacji raportowania czy eksplorowania danych (np. dla zarządu firmy z użyciem Microsoft Power BI lub Tableau). W porównaniu do takiego „klasycznego" ujęcia analityki danych DS to

> wyjście poza paradygmat klasycznych systemów BI (*business intelligence*) w kierunku użycia tzw. złotej pętli decyzyjnej, gdzie wygenerowana propozycja nie trafia do menedżera, ale jest wykonywana automatycznie (Surma 2017:43).

Działanie podstawowe może być tu wykonywane zarówno w języku Python jak i w R. Analityka jest w badanym świecie uznawana za „podzbiór" DS w tym sensie, że jakoby prawdziwy uczestnik świata DS potrafiłby sobie poradzić z zadaniami analityka – stanowią one wstęp do elementarza uczestnika badanego świata. Analityka postrzegana jest zatem jako łatwiejsza, mniej sprawcza, mniej ciekawa, mniej rozwijająca niż DS. W badaniach terenowych spotkaliśmy się z opiniami, że w najbliższych latach niemal całość pracy polegającej na opisowej analizie danych / raportowaniu może zostać zautomatyzowana. Zdobycie pracy kodującego analityka bywa postrzegane jako wstęp do pracy jako data scientist. Informowano nas także, że osoby zatrudniane na stanowiska data scientist bez doświadczenia zawodowego (w większych organizacjach tzw. junior) wykonują faktycznie zajęcia analityków danych. Analitycy danych raczej otrzymują wynagrodzenia niższe niż data scientiści, zaś dla analityków publikowanych jest więcej ofert pracy (por. rys. 18) – przy czym należy pamiętać, że analityka danych posługująca się głównie arkuszami kalkulacyjnymi i językiem SQL nie jest uznawana za część świata DS. Granica między kodującą analityką a data science jest bardzo rozmyta, bo choć mówi się o braku stosowania metod ML w analityce, to często zdarza się, że tzw. analitycy stosują pewne metody ML, a tzw. data scientiści często stosują analizy opisowe / eksplorację / testowanie hipotez. Problem granic jest tu dyskutowany np. pod kątem tego, czy modele liniowe lub logitowe należy uznać za ML, czy znane od dziesięcioleci „proste" algorytmy ML (m.in. drzewa decyzyjne) należy uznać za charakterystyczne dla DS.

Twierdzimy, że w dyskursie nie-technicznym, np. w HR, w akademii, w szeroko pojętym biznesie, trwa proces zbliżania się zakresów pojęciowych analityki danych i data science, co przez silnie osadzonych, „prawdziwych" uczestników świata DS może być

postrzegane jako utrata prestiżu określenia data scientist. Rysunek 44. jest w tym kontekście raczej przedstawieniem perspektywy takich silnie osadzonych, zaryzykujmy stwierdzenie – konserwatywnie postrzegających DS uczestników. Dla innych uczestników, a także osób spoza świata DS (np. reprezentujących biznes czy HR) analityka danych i DS może być po prostu tożsama. Niemniej wskazywano nam jednoznacznie, że to, co kiedyś było nazywane analityką było / byłoby już w 2018–2019 nazywane DS.

> *B: No dobra dyktafon już mi działa. Mmmm dobra to, to zacznijmy od czy ty jesteś data scientist?*
> R: Aammm to znaczy zależy jak to rozumieć. Ty jak pewnie sam wiesz najlepiej ilu ludzi w branży ogólnie analitycznej tyle definicji czym jest data science i czym jest data scientist. Ee, jakby **formalnie** nie jestem, nie mam tego w tytule swojego stanowiska. Eee, ale myślę, że jeśli chodzi o zadania które wykonuję to w wielu firmach zostałoby to nazwane jako data scientist. Eeee i podejrzewam, że gdyby teraz moja firma szukała kogoś na podobne stanowisko, też nazwałaby to, to, że szuka data scientist. {wywiad 16}

Ponieważ to modelowanie jest jakoby „najfajniejszą", najbardziej „seksowną", ciekawą, rozwojową i sprawczą częścią pracy data scientistów, to obecnie w badanym świecie rośnie prestiż subświatów koncentrujących się wokół uczenia maszynowego – ML engineer oraz ML researcher, zaś w komunikacji marketingowej i w ogóle biznesowej mówi się raczej o sztucznej inteligencji niż o analityce, o big data czy nawet data science.

Rozmówczyni {wywiad 8} wskazała, że aplikując do pracy w kraju Europy Zachodniej zauważyła, że DS było tam postrzegane węziej niż w Polsce, gdzie prawdziwe data science to raczej tzw. full-stack data science. W tamtym kraju DS opisywano raczej zgodnie z polskim rozumieniem kodującego analityka danych, a nie data scientista:

> *B: No dobra, no to yyy pytanie, najpierw takie, czy ty jesteś data scientist?*
> R: Teoretycznie tak mam na umowie o pracę, ale jakby data scientist to jest wydaje mi się bardzo takie rozległe jednak określenie, eee i dużo osób, ostatnio w ogóle aplikuję do różnych prac w [nazwa zachodnioeuropejskiej stolicy], i tam to jest rozdzielone na trzy różne funkcje tak naprawdę. I jest właśnie data scientist, i to jest osoba która bardziej zajmuje się przygotowanie raportów, statystykami dla klientów, i tak dalej, jest machine learning engineer, osoba która zajmuje się implementowaniem modeli różnych, czytaniem paperów i implementacji tego do wszystkich programów i całego pipeline'u. Jeszcze jest machine learning researcher, czyli osoba która tylko tak naprawdę czyta papery i zajmuje się researchem i pisaniem swoich modeli, swoich, swoich, swoich paperów tak naprawdę.
> *B: Mhm, OK.*
> R: W Polsce to jeszcze nie jest tak zaawansowane więc jakby wszyscy są data scientist. Więc jakby musimy się zajmować trochę wszystkim. {wywiad 8}

ML engineer jest zatem rolą profesjonalną odpowiedzialną za produkcyjne wdrażanie modeli i utrzymanie ich działania. Działanie podstawowe będzie tu wykonywane np. w językach Python, Scala, a także językach niskopoziomowych jak C++.

Nazwa wskazuje, że jest to stanowisko inżynierskie, zorientowane bardziej na stabilność rozwiązań niż eksplorację i eksperymenty. Rola ta może być postrzegana w badanym świecie jako bardziej prestiżowa niż DS z powodu większej sprawczości – koncentruje się przecież na wdrożeniach i utrzymaniu, a także jako bardziej wymagająca i sprzyjająca samorozwojowi w obszarze znajomości technologii (także kojarzonych z tradycyjnym programowaniem, np. znajomości niskopoziomowych języków programowania). Ten subświat w znacznie większym stopniu niż DS zajmuje się problemami infrastruktury obliczeniowej (por. rys. 44), jej kosztów czy czasu działania. Może to wskazywać na przesunięcie w badanym świecie właśnie w kierunku wyższego wartościowania wdrożeń modeli niż modelowania (jako eksperymentowania z metodami i ich parametrami). Szczególnie, że to drugie staje się coraz łatwiejsze, nawet do momentu, w którym nie wymaga samodzielnego pisania kodu (np. na platformach typu KNIME, RapidMiner, DataRobot, Dataiku czy platformie firmy Peltarion). Subświat ten zdecydowanie nie jest postrzegany jako „podzbiór" DS, a jako jego wyspecjalizowana część, w zakresie inżynierskim zaś także wykraczającą poza rozumienie DS (nawet szerokie, tzw. full-stack). W prezentowanym nieco niżej, dłuższym fragmencie wypowiedzi Rozmówca {wywiad 19} wskazywał, że ML engineer to rola skoncentrowana na technologicznych aspektach projektów DS (ale nie przechowywania i dostępu do danych – to raczej inżynieria danych), zaś data scientist na metodologicznych. Jak wcześniej wielokrotnie wspominaliśmy czasami praca zespołu / działu lub małej firmy zajmującej się DS jest i taka, że klient wewnętrzny czy zewnętrzny otrzymuje zaakceptowaną wersję modelu np. w pliku .h5 i wdraża (lub nie) siłami własnego zespołu / działu IT. Zatem trudno nam określić, czy opisywany subświat raczej wypączkował ze świata DS, czy powstał na przecięciu DS i jakiegoś segmentu „tradycyjnej" informatyki (Strauss 1984:125, za: Kacperczyk 2016:50).

ML researcher jest rolą profesjonalną odpowiedzialną za śledzenie nowości, tworzenie rozwiązań pośrednich / hybrydowych i rozwijanie autorskich metod uczenia maszynowego. Działanie podstawowe będzie wykonywane raczej w Pythonie, lecz gama języków programowania może być szeroka, choć raczej nie obejmuje R; być może za technologie do wykonywania działania podstawowego należałoby uznać zapis matematyczny i język ludzki (niemal zawsze j. angielski), a na niższym poziomie abstrakcji np. edytor LaTeX[261], czyli technologie do pisania artykułów naukowych w dziedzinie ML. Osoby takie nierzadko mogą być równolegle akademikami. Prestiż tego stanowiska postrzegany jest głównie w kontekście bardzo wysokiego poziomu znajomości najdrobniejszych szczegółów

261 A właściwie język składu TeX (*TeX typesetting language*), por. https://www.tug.org/begin.html, dostęp 02.07.2020 r.

metod uczenia maszynowego, w tym tak teoretycznych szczegółów matematycznych jak i detali informatycznych – poziomu znacznie wyższego, niż potrzebuje data scientist do „efektywnego rozwiązywania prawdziwych problemów". ML researcher może być postrzegany w badanym świecie jako rola bardzo elitarna, bliska nauce akademickiej, nadzwyczajnie ciekawa, ale także nadzwyczajnie sprawcza – tworząca nowe metody, potencjalnie dla milionów użytkowników. Subświat ten, podobnie jak ML engineering, jest postrzegany w świecie DS jako specjalizacja tak daleka, że wykraczająca nieco poza jego granice. Widzimy tu proces gdzieś pomiędzy pączkowaniem a przecinaniem się światów, tym razem świata DS i akademickich badań nad uczeniem maszynowym.

Podział na ML engineering / research jest inny, niż historycznie wcześniejszy i dość mało w Polsce popularny podział w DS, zwany typ A (*analysis*) / typ B (*building*). O ile typ B można by utożsamiać z ML engineeringiem, to typ A raczej z rolą usytuowaną na granicy DS i analityki danych, zaś na pewno nie z ML researchem. Przyjrzyjmy się wypowiedzi jedynego z naszych Rozmówców, który określał siebie jako ML researcher:

> R: (...) robię to w trzech równoległych trakach. Z jednej strony to jest moja działalność stricte naukowa i pisanie artykułów naukowych i udział w konferencjach naukowych itd., druga noga, to jest oczywiście dydaktyka, staranie się być jak najbardziej na bieżąco, jeżeli chodzi o dydaktykę, no a trzecia to jest właśnie moje zaangażowanie takie w projekty komercyjne, gdzie **tam** dopiero się okazuje, czy rzeczy, które tutaj robimy mają jakikolwiek sens czy nie, czy komukolwiek są przydatne. Bo koniec końców, jak ktoś z prywatnej kieszeni jest gotów zapłacić złotówkę za wykonaną prace, to to jest jak gdyby potwierdzenie, że to jest przydatne. Nie ma takiej oczywistej walidacji, jeżeli chodzi o działalność taką stricte naukową czy dydaktyczną.
> *B: A czy pan nazywa siebie data scientist?*
> R: Nie. To może być też wynik jakiegoś takiego mojego osobistego balansu, ale ten termin data scientist mi się kojarzy z [pauza] jakby to dobrze określić? Ja bym w ten sposób nazwał osobę, która jest biegła w algorytmach, która ma podstawy, jeżeli chodzi o statystykę i uczenie maszynowe, natomiast nie posiada kompetencji takich czysto inżynierskich związanych na przykład z pracą wdrożeniową, z tym że najlepszy możliwy algorytm tak długo, dopóki on pozostaje tylko jakimś skrypcikiem, jest kompletnie bezużyteczny. On nie ma żadnego znaczenia, jaki on jest dobry, dopóki się go nie sproduktyzuje, co oznacza tak naprawdę obłożenie go dziesiątkami warstw tego, co jest wymagane dla inżynierii, yy oprogramowania, żeby to mogło został włączone bezpiecznie do produktu, żeby to zostało przetestowane, zintegrowane, bez tego to wszystko jest tak naprawdę bezużyteczne. I ten termin data science to mi bardziej kojarzy się z osobami, które wykonują takie badania eksploracyjne, może trochę wizualizacji, gdzieś tam na pograniczu statystyki, gdzie być może wychodzi z tego jakiś prototyp, albo przynajmniej jest jakieś wskazanie kierunku o!, ta rodzina algorytmów prawdopodobnie nadaje się do rozwiązania naszego problemu, ta rodzina nie ale koniec końców, oprócz pytania o to, czy dana rodzina algorytmów się nadaje [pauza] może tak, żeby tak naprawdę stwierdzić, czy dany algorytm się nadaje, czy nie, sam fakt przebadania go eksperymentalnie i stwierdzenie o!, tutaj dostaje 90% precyzji, a ja tutaj dostaję 88% precyzji to być może nie ma to żadnego znaczenie i jeżeli jeden algorytm zajmuje sekundę, a drugi się wykonuje w 15 milisekund. I wtedy 2% straty dokładności nie ma żadnego znaczenia, bo tak naprawdę produkcyjnie liczy się tylko i wyłącznie mamy ścisłe ograniczenie na czas

odpowiedzi. Więc, to brzmi trochę pejoratywnie, ale ja dobrze rozumiem potrzebę istnienia osób, które realizują takie data science, bo oni rzeczywiście są w stanie bardzo zawęzić pole przeszukiwania, po to, żeby nie próbować wszystkich możliwych sposobów. Natomiast ja bym siebie plasował już trochę bardziej po stronie też takiego twardszego podejścia do problemu i próby też produktyzacji tego zupełnie na poważnie.
B: Też słyszałem nazwę na przykład machine learning engineer, czy yyy machine learning researcher, nie wiem, czy one w ogóle w Polsce się zadomowiły, yy chyba niekoniecznie?
R: Próbujemy je wprowadzić.
B: Nie wiem czy coś takiego byłoby panu bliższe? Ja też nie mam żadnej definicji tego, ja staram się dowiedzieć od ludzie, czy taka definicja jest?
R: Gdybym ja miał siebie jakąś jedną etykietką określić, to bym powiedział machine learning researcher. To by było to, co mi najbardziej pasuje. ML engineer to już jest osoba, która być może nie będzie tak dobrze rozumiała, mówiąc kolokwialnie, **bebechów** algorytmu, natomiast będzie w stanie go w perfekcyjny sposób wprowadzić do produktu. Czyli to już są potrzebne bardzo wysoki kompetencje, czysto programistyczne i inżynierskie, rzeczy, których nie da się uzyskać na przykład pracując [krótka pauza] no, nie istnieje pojęcie ML engineer w środowisku akademickim, bo po prostu tutaj ludzie są za słabi programistycznie. Tu naprawdę trzeba pójść i przez 3 lata dzień w dzień, non stop programować, żeby się tego nauczyć. To jest trudna sztuka. Jak ktoś naprawdę będzie świetny i jak do tego dochodzi cały stos technologii, z którym trzeba być na bieżąco, które się cały czas zmieniają. I jeszcze trzeba na co dzień siedzieć z DevOps-ami, którzy jeszcze swoje dorzucą, że jest taka infrastruktura, takie ograniczenia, takie środowiska do testowania, do ciągłej integracji itd., więc to są kompetencje, których nie ma w ogóle w świecie akademickim, to są ludzie, którzy muszą się nauczyć bardzo dobrze programować, i oni też na tyle dobrze rozumieją algorytm, że są w stanie na przykład robić takie dostosowanie parametrów, algorytmu itd., natomiast być może sami z siebie nie wymyślą nowego algorytmu, albo będą trochę traktowali te algorytmy, jako taką czarną skrzynkę. (…) {wywiad 23}

Inżynieria danych (*data engineering*) **tożsama jest z big data w sensie technicznym**. Data engineer cechuje się porównywalnym prestiżem i zarobkami co data scientist, ale nazwa ta określa węższy wycinek zakresu obowiązków niż data scientist. Działanie podstawowe data engineera koncentruje się wokół przechowywania i dostępu do danych, także przetwarzania ich oraz problemów dotyczących infrastruktury technicznej. W sensie działania podstawowego nie ma on nic wspólnego z analizą ani z modelowaniem danych. Używane są rozmaite języki programowania, rozmaite technologie bazodanowe oraz rodzaje infrastruktury (por. 5.3.2 - 3). Ten raczej segment niż subświat inżynierii danych, a może po prostu świat oddzielny od DS, współpracuje z data science głównie w zakresie udostępniania zbiorów danych, a częściowo także przy wdrożeniach. Współpraca np. data engineera z ML engineerem może być niezbędna do przygotowania tzw. pipeline'u (dosł. rurociągu), którym dane „płyną" w systemie produkcyjnym. Sądzimy, że opisywany segment najlepiej opisać byłoby jako profesjonalny subświat informatyki zajmującej się bazami danych, przecinający się ze światem DS. Inżynieria danych uznawana jest w badanym świecie za obszar względnie dobrze zdefiniowany w porównaniu do DS:

B: Mhm, mhm, ok. No i jeszcze mamy [krótka pauza] ten zawód data engineera*, tak? Który,*

znaczy ja już teraz zrozumiałem mniej więcej że to jest inny zawód, niż data scientist, aaa ale ale, jakbyś jeszcze mógł mi to, to przedstawić. Czym to się różni?
R: Eee czyli, data engineer tak jak ja to rozumiem [śmiech], bo często to się może rozjeżdżać, chociaż to mi się wydaje jest **w miarę** już ustandaryzowane [krótka pauza] to jest ktoś kto się zajmuje zbieraniem danych, yyy czy do hurtowni danych, czy jak to się teraz nazywa data lake'u, co raz częściej jest to w chmurach, czyszczeniem ich oraz po prostu tworzeniem automatycznych procesów, które do tego służą. **Gdzieś** jakaś platforma na przykład internetowa generuje nowe dane, gdzieś już są zbierane i potem po kolei ileś tam kroków są automatyzowane coś tam się dzieje, żeby te dane były w jakimś stopniu **przeczyszczone** i udostępnione ludziom którzy potem się nimi zajmują, data scientistom, ludziom zajmującym się bardziej Bussines Intelligence, itd.
B: Mhm, ok. A w twojej firmie też są takie stanowiska data engineerów?
R: Tak. Takie stanowiska raczej są już w **każdej** większej firmie, która no ma jakiś już, od dłuższego czasu zajmuje się tym, no bo to jest w jakimś stopniu wiedza dość, yyy noo [pauza] znaczy zakres obowiązków jest na tyle duży, że to może spokojnie pochłaniać kilka etatów tak na prawdę. To nie jest też coś co, [cmoknięcie] pojedynczy data scientist, czy powiedzmy może być, część obowiązków pracy jednej, dwóch osób zajmujących się normalnie data science. I w dużej firmie gdzie ilość danych jest już gigantyczna, generowanych i to wymaga, **specjalnej** nawet wiedzy, żeby **w miarę** to **wydajnie** i czy to w miarę szybko, czy **kosztowo**, też względnie tanio przetworzyć. {wywiad 19}

W obszar inżynierii danych bywa jednak włączany zakres produkcyjnego wdrażania modeli uczenia maszynowego, zaś profesję ML engineera niektórzy widzą jako wyłaniającą się właśnie z inżynierii danych. Być może nie jest to pozbawione sensu. Zauważmy jednak, że przywołane niżej źródło nie jest aktem samowiedzy badanego świata DS, a raportem biznesowym, w którym terminy np. „umiejętności matematyczne", „rozwijanie algorytmów" nie przystają do rozumienia roli ani inżyniera danych, ani ML engineera w społecznym świecie DS:

> Na granicy światów. W przestrzeni pomiędzy podejściem „teoretyczno-naukowym" prezentowanym przez data scientist a „praktycznym" data engineerów coraz częściej powstaje miejsce dla inżynierów specjalizujących się w uczeniu maszynowym. Machine learning enginneer to zwykle programista albo ekspert posiadający solidne zaplecze technologiczne – raczej inżynier niż naukowiec. Łączy obydwa światy, więc posiada bardzo zbliżone umiejętności zarówno do data scientist, jak i data engineer, zwłaszcza statystyczne, matematyczne i analityczne. Jego podstawowym zadaniem jest projektowanie, optymalizowanie i utrzymywanie modeli uczenia maszynowego, a także rozwijanie powiązanych z nimi algorytmów (Gontarz 2019:9).

Rzecz jasna w badanym świecie buduje się zespoły realizujące to, co nazywamy w naszej pracy projektami DS, wraz z wdrożeniami. Zespół złożony z data scientista, ML engineera i ML researchera wskazywano jako wymarzony:

> R: To dobrze widać w przypadku teraz ostatnimi laty bardzo popularnych głębokich sieci neuronowych, gdzie mnóstwo ludzi potrafi to poskładać, potrafi zbudować sobie, zdefiniować sobie taki stos, zbudować sobie taką sieć, podefiniować poszczególne warstwy, uruchomić uczelnię itd. zrobić sobie optymalizację hiperparametrów. Natomiast, jakby tak naprawdę ich zapytać, co tam się w środku dzieje, albo spróbować poprosić ich, żeby napisali to równanie,

> które tam się propaguje przez tą sieć, to już nie. To już będzie wymagało też innego rodzaju kompetencji. Więc tak bym to rozróżniał. Ja jestem wielkim fanem odróżniania data science od yy od to od jest machine learning engineer i machine learning researcher. Ostatnio spotkałem się z terminem MLOps, na zasadzie jak jest DevOps, to jest MLOps, to by było chyba to samo, co machine learning engineer tak naprawdę. Bo wydaje mi się, że dobrze jest, żeby organizacje, szczególnie firmy prywatne, rozumiały te różne zbiory kompetencji i jakby budując swoje [pauza] parę razy zdarzyło nam się doradzać czy robić taki głos doradczy dla firm, które chciały zorganizować swój proces, i zacząć stosować ML i tam bardzo silnie ich uczulaliśmy na to, że wynajęcie na pół etatu 4 doktorantów z uczelni do niczego nie doprowadzi. To będzie bezsensowne przepalanie zasobów, i marnowanie czasu. To nie ma żadnych szans na sukces. I staraliśmy się im wytłumaczyć, że właśnie ta specyficzna pozycja, że złożenie dobrego zespołu, który będzie zawierał kogoś od data science, komu można szybko zlecić zrób tą wizualizację, odpal ten algorytm, siamten algorytm, zobacz taką metodę [niezrozumiałe] i zrób to na próbce, zrób to na pliku CSV, zrób to na jakimś zrzucie sprzed miesiąca i niekoniecznie musisz kombinować, jak napisać dobrą końcówkę do jakiegoś mikroserwisu, który akurat prosto z produkcji ściął. No no no data science nie będzie umiał tego kompletnie zrobić, będzie siedział i na to patrzył – co to w ogóle jest? Z drugiej strony jest ktoś, kto jest takim ML researcher, no to będzie ktoś, kto będzie zupełnie na bieżąco i będzie mówił [zmienia głos] dobra, no to może spróbujmy tu, a tutaj taki model opublikowano, a tutaj taki, a tutaj na GitHubie jest taka biblioteka, może tego spróbujmy, no i wreszcie ci ludzie na dole od ML, od inżynierii ML-a, którzy tak naprawdę to wszystko razem potem sobie spinają. To jest taki dream team, jeżeli chodzi o budowę takiego środowiska. {wywiad 23}

Organizacja pracy w zespołach DS (które mogą nazywać się np. zespołami AI czy analitycznymi), może być różna. Słynna Cassie Kozyrkow, Chief Decision Scientist w Google pisze nawet o dziesięciu różnych rolach w perfekcyjnym zespole tego typu (Kozyrkov 2018). Nawet nazwa jej stanowiska – decision scientist – bywa postrzegana jako profesjonalny subświat data science lub jako obszar graniczący z DS, ale nie-techniczny, raczej statystyczny lub naukowy i skoncentrowany na wspomaganiu decyzji. Jeszcze inaczej postrzegany jest nie-techniczny, w Polsce przez nas nie spotkany segment o nazwie analytics translator – rola rozumiejąca ML, lecz nie kodująca, skoncentrowana na biznesowych aspektach projektów i komunikacji z klientami (Henke, Levine, i Mcinerney 2018).

W badanym świecie dyskusja o granicach i zakresach opisywanych subświatów dopiero się zaczyna, mieliśmy nawet okazję brać udział w meetupie poświęconym temu właśnie zagadnieniu {obserwacja 38}. Nie jest to zgodne z celem rozprawy, ale można by opisać spór o kształtujące się granice tych profesjonalnych subświatów jako szeroką arenę w której stronami są np. świat DS, szeroki świat informatyki, świat marketingu, świat HR, świat decydentów. Z punktu widzenia uczestników świata DS nierzadko przecież formalne nazwy stanowisk, sposoby wynagradzania pracowników, organizowania zespołów i ich pracy czy prowadzenia projektów DS nijak mają się do tego, jak oni postrzegają swój świat społeczny i wyróżnione subświaty wraz z ich specyfiką.

> [rozmowa dotyczy tego, że DS obejmuje bardzo szeroki zakres i jest różnie nazywane]
> *B: Hmmm, a jak odróżniasz eee, jak odróżniasz te rzeczy, tak? No bo yyy, rozumiem że data science jest no powiedzmy, yyy graniczy z BI, graniczy z machine learning engineering, czy z tym decisive science. Czy mógłbyś mi powiedzieć jak odróżniasz te rzeczy od data science?*
> R: Znaczy ja bym **nawet** powiedział że [krótka pauza] znaczy bo to są często definicje jak wygląda opis danego stanowiska w firmie [krótka pauza] co jest czym. Aaa ogólnie stanowiska, znaczy stanowiska data scientistowe, mam wrażenie że się rozciągają **najszerzej**. Czyli tak naprawdę w nie wpadaa **wszystko**. Czasami masz na prawdę data scientista, który robi Business Intelligence, w firmie, to co **najczęściej** było określane jako Business Intelligence [krótka pauza] czasami, nawet to sięga zakres pracy data scientista do data engineeringu, jeżeli nie ma dedykowanego zespołu. A już takie kwestie właśnie jak powiedzmy analityka danych, czyli to powiedzmy google analytics, aaa właśnie też, już [krótka pauza] Znaczy powiedzmy to jest takie też mocno, często zamienne, nie? Tak na prawdę jak nie ma polskiego odpowiednika data scientist, no to chyba najlepszym jest, analityk danych. Eee [pauza] no i to jest właśnie data scientist to, w takim stopniu właśnie obejmuje też te kwestie właśnie decisive science, czyli to jest ta najbardziej **podstawowa** kwestia w wielu firmach, że po prostu ludzie którzy mają duży zbiór danych mieli wyciągnąć statystyki a ktoś inny podejmuje decyzje, no i to jest to właśnie machine learning engineering, czyli to jest tworzenie modeli. Co do rozgraniczenia powiedzmy **u mnie** w firmie, eea no to [pauza] data scientiści to są bardziej ci co się **skupiają** na aspektach metodologicznych, statystycznych, **badawczych** w danym projekcie, a **inżynierowie**, czy to machine learning inżynierowie, powiedzmy w takiej nomenklaturze bardziej szerszej, czy jak to się u nas nazywa ze względów, korporacyjnych because of reasons [pauza] reasearch inżynierowie, to są znowu z kolei ludzie skupieni bardziej na aspektach **technicznych**, typu utrzymanie produkcji, oo co tam było, optymalizacji kodu, automatyzacji. Granicą byłoby skupienie, czy się ktoś skupia na metodologii bardziej [pauza] czy właśnie, **granicą** powiedzmy w tym przypadku byłoby czy dla kogoś głównym punktem zaczepienia jest eee [krótka pauza] metodologia, wiedza w jakimś stopniu teoretyczna, tak nazwijmy to roboczo, czy yyy kwestie technologiczne. {wywiad 19}

Twierdzimy, że **w badanym świecie panuje powszechne przekonanie, że DS znajduje się na wczesnym etapie rozwoju w niemal każdym sensie** – technologicznym, specjalizacji ról profesjonalnych, świadomości potencjału biznesowego, adaptacji w rozmaitych branżach, regulacji prawnych itd. Można postrzegać to jako kolejną legitymizację badanego świata; ma ona głównie charakter neutralizacji. Nastawiona jest raczej na zewnątrz niż do uczestników tego świata, zatem do audytorium nie-technicznego. Szczególnie to „wczesne stadium rozwoju" dotyczyć ma Polski, postrzeganej w badanym świecie jako technologicznie i biznesowo peryferyjna, zapóźniona, z niskimi budżetami (przypomnimy powszechną w polskim DS praktykę pracy dla klientów zagranicznych). Wielokrotnie spotkaliśmy się z porównaniem obecnego (w czasie realizacji naszych badań) stadium rozwoju DS w Polsce do tego, jak (w ujęciu stereotypowym) wyglądał i działał polski internet na początku lat 90. Pisano wówczas statyczne strony w języku znaczników HTML, uruchamiane przez modemowe łącza na stacjonarnych komputerach. Emaile były czymś egzotycznym, nie istniała skuteczna wyszukiwarka internetowa ani media społecznościowe, a użytkownicy, czyli tzw. internauci, stanowili niemalże subkulturę złożoną z informatyków,

geeków i naukowców. Dziś strony internetowe może budować każdy np. za pomocą szablonów Wordpressa, wiele osób jest połączonych z siecią 24 godziny na dobę za pomocą smartfona, a konta w mediach społecznościowych mają nie tylko noworodki, ale i domowe zwierzęta. Skoro tak rozwinęły się technologie i wzorce korzystania z internetu, to „my" (audytorium nie-techniecznie) mamy sobie poprzez analogię wyobrazić, jak rozwiną się technologie i wzorce korzystania z produktów bazujących na uczeniu maszynowym.

6.2. Areny społecznego świata data science

Arena jest polem, na którym toczy się istotny dla świata społecznego spór i zdaniem A. Straussa oznacza niezgodę na kierunek działania podejmowany przez świat społeczny lub jego segment (1993:227). Kacperczyk zwróciła uwagę na to, że Clarke utożsamia świat społeczny z arenami i sugeruje że analiza aren jest niezbędną częścią analizy społecznych światów (Kacperczyk 2016:572-73). Zgadzamy się w pełni z takim podejściem.

Poprzez zaangażowanie w arenę społeczne światy i ich segmenty czy subświaty realizują swoje interesy i walczą o własne pozycje, warunkujące przetrwanie i rozwój. Arena jest dyskursywną (a czasami i fizyczną) przestrzenią spotkania i „pragmatycznej troski o daną sprawę" (Kacperczyk 2016:42). Społecznego świata nie sposób zrozumieć w odosobnieniu od innych światów, z którymi spotyka się na arenie – badacz powinien opisać stanowiska, sojusze i „wrogów" spierających się na arenie, a także to, co i jak chcą w owym sporze osiągnąć dla siebie. Badanie aren pomaga dotrzeć do wewnętrznej złożoności społecznych światów. Procesy np. segmentacji, legitymizacji społecznych światów mogą być najlepiej zrozumiane właśnie na arenach. Areny pojawiają się na każdym poziomie życia społecznego: od makro przez mezzo i mikro do pojedynczych aktorów – uczestników społecznych światów (ibidem). My widzimy areny w dużej mierze także jako spory o wartości.

> Cokolwiek robimy jako aktorzy społeczni i uczestnicy życia społecznego – wypowiadamy się zawsze w imieniu jakiegoś społecznego świata (lub jego segmentu). Również w dyskusjach reprezentujemy zawsze jakiś społeczny świat i wypowiadamy się z pozycji jego uczestnika. **W społecznym świecie dyskusja nigdy nie wygasa i powoduje, że podlega on nieustannym wewnętrznym przemianom. Pojawiają się w nim areny – miejsca sporne, generujące polaryzację opinii, segmentację świata i wydzielanie się subświatów** [podkr. RŻ]. Ich uczestnicy zaczynają widzieć inaczej podstawowe działanie lub wykonują je w odmienny sposób, wytwarzając wąską specjalizację danej aktywności, rozwijają nową technologię i szukają uzasadnień i legitymizacji dla własnego wyjątkowego sposobu działania (Kacperczyk

2016:18).

W tej części realizujemy ostatnie z postawionych celów szczegółowych rozprawy. Jest to określenie i opis głównych aren sporów oraz uchwycenie pełnej sytuacji badania (*the full situation of inquiry*). Właściwe jest to domknięcie uchwycenia pełnej sytuacji badania. Odnosimy się często do części 6.1. oraz całej części 5. wprowadzając kilka zupełnie nowych wątków. Raczej dążymy do nagłośnienia i objaśnienia wcześniej już sygnalizowanych kwestii, do syntezy naszych ustaleń i do podsumowania wykonanego opisu społecznego świata DS. Bez wątpienia każda z aren, które opisujemy – a także te, które tylko wspominamy – mogłaby zostać ujęta w oddzielnej rozprawie. Za pomocą map miejscami proponujemy spojrzenie z lotu ptaka (Uri 2015:146), niemniej zaznaczmy, że staramy się oddać głównie perspektywę uczestników świata DS. Badania etnograficzne prowadziliśmy przecież niemal wyłącznie w ich środowisku, zatem pisząc o innych światach zaangażowanych w areny raczej (re)konstruujemy perspektywę świata DS. Cała część 6.2. domyka ujęcie pełnej sytuacji badania, na którą łącznie składają się rozdziały: 2. poświęcony przeglądowi literatury akademickiej, gdyż naukowców traktujemy jako zaangażowanych w arenę; 3. i 4., tutaj nasze badania wraz z ich ramą teoretyczną i metodologią stanowią wyraz właśnie takiego zaangażowania, a te rozdziały pokazują usytuowanie badacza (RŻ) w odniesieniu do świata DS; wreszcie 5. i 6., jako prezentacja wyników naszych badań i analiz.

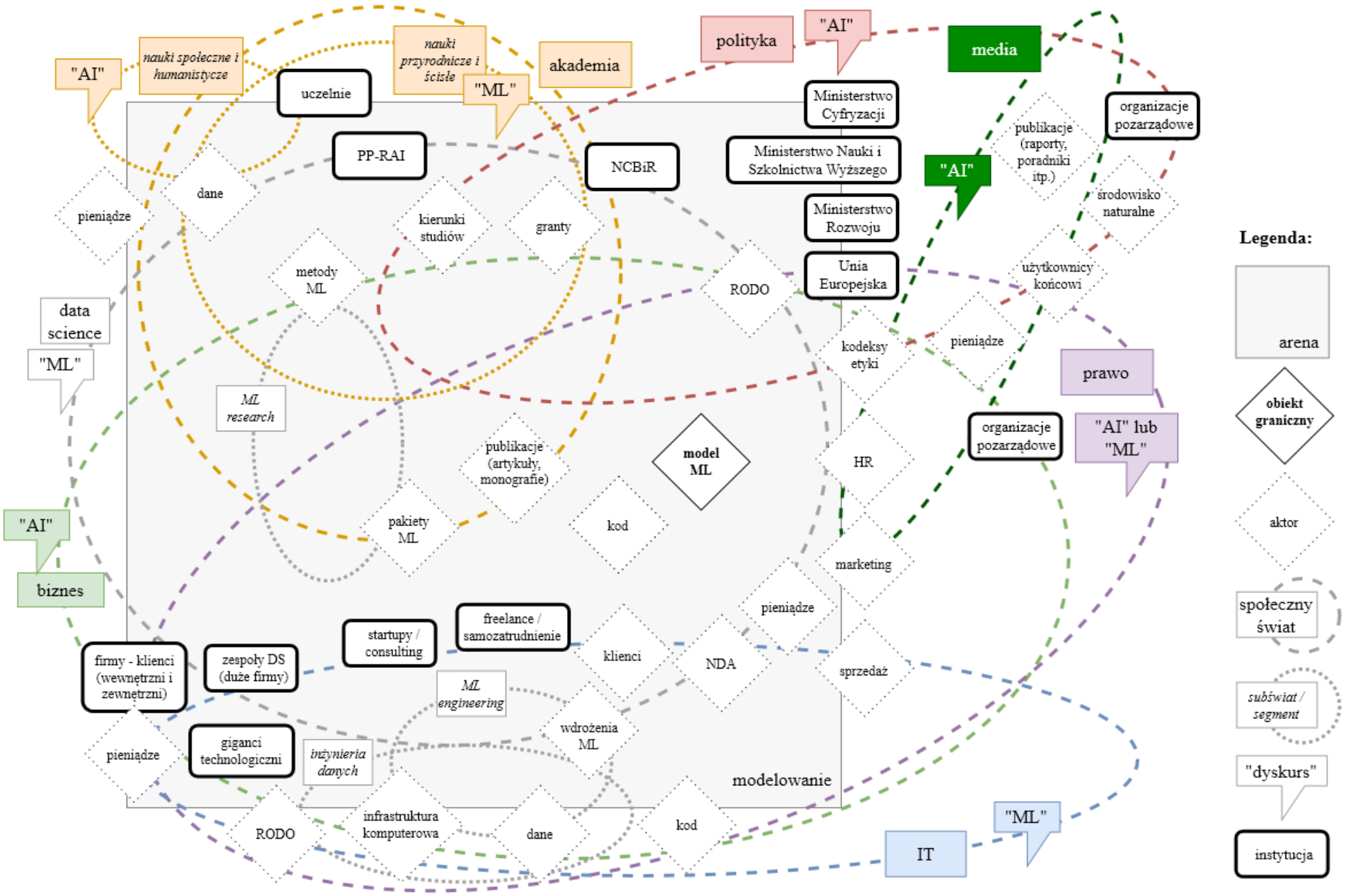
6.2.1. Modelowanie

Modele uczenia maszynowego widzimy jako **obiekty graniczne**. W ujęciu Susan L. Star i Jamesa R. Griesemera (1989) obiekt graniczny jest czymś, co znajduje się na granicy społecznych światów, na ich przecięciach lub w pęknięciach. „Coś" może być tak materialne jak i niematerialne, konkretne jak i abstrakcyjne, np. urządzenie, program komputerowy, lek, akt prawny, pojęcie. Różne światy czy subświaty zaangażowane w arenę mają różne interesy, wyrażające się w różnych sposobach definiowania i interpretowania obiektu granicznego. Strony areny spotykają się „wokół" obiektu granicznego, każda ma wobec tegoż obiektu rozmaite wymagania, będące wyrazem owych, nierzadko sprzecznych interesów. Problemem pojęcia obiektu granicznego może być to, czy wywołuje on podziały między światami / subświatami i krystalizowanie się areny, czy też to podziały wewnętrzne społecznego świata / światów (a zatem potencjalna arena) „szukają" obiektów, które mogą stać się graniczne – tzn. obiektów pozwalających zarysować odmienność interesów i legitymizować swoją pozycję w dyskursie, forsując własną definicję obiektu (Kacperczyk 2011:192, 2016:40–41). W

opisywanej arenie skłaniamy się ku pierwszemu sposobowi ujęcia ontologii obiektu granicznego, tzn. sądzimy, że to komercyjne wdrażanie modeli uczenia maszynowego jest faktem, który zapoczątkował krystalizowanie się owej areny.

Modelowanie jest oczywiście obiektem granicznym także „wewnątrz" badanego świata DS, o czym pisaliśmy już wiele w odniesieniu do działania podstawowego, wartości, procesów, nie odwołując się bezpośrednio do pojęcia obiektu granicznego. W skrócie: modelowanie jest nieodłącznym elementem składającym się na działanie podstawowe. Nie można być uczestnikiem badanego świata, nie wykonując modeli ML. Znajomość rozmaitych metod i narzędzi do modelowania, a także doświadczenie w komercyjnym tworzeniu modeli odróżnia uczestników słabiej od silniej osadzonych w badanym świecie. To modelowanie (a właściwie jego udział w obowiązkach służbowych oraz to, na jakich aspektach modelowania rola zawodowa jest skoncentrowana) jest tym, co odróżnia kodujących analityków danych od data scientistów, a tych drugich od ML engineerów i ML researcherów. To modelowanie jest „wisienką na torcie", pracy data scientistów, czymś najciekawszym, najfajniejszym, najbardziej seksownym. To modelowanie, a właściwie wdrażanie modeli daje sprawczość w świecie zewnętrznym, to ono zwiększa efektywność w wybranym wycinku rzeczywistości. Jednocześnie samo modelowanie nie wskazuje na uczestnictwo w badanym świecie – osoba skoncentrowana tylko na budowaniu modeli ML może być zarówno akademikiem jak i graczem w Kaggle lub wannabesem.

Na rys. 45. prezentujemy modelowanie jako arenę, w której wokół modelu ML jako obiektu granicznego trwa spór społecznych światów – DS, biznesu, akademii, IT, mediów, polityki i prawa:



Rysunek 45: Modelowanie jako arena – mapa społecznego świata / areny w Polsce
Źródło: opracowanie własne z użyciem draw.io

Względnie wspólne są interesy światów technicznych, czyli DS oraz IT (jako szeroko pojętej informatyki nie-akademickiej). Światy te współpracują w miejscu znajdującym się „blisko" granic świata DS, czyli we wdrażaniu modeli ML. W zakresie uzyskania dostępu do danych i ich wstępnego przygotowania czy tzw. pipeline'u (czyli przepływu danych w procesie – rurociągu) oraz w zakresie różnych zadań dotyczących infrastruktury komputerowej zdarza się tak, że świat DS korzysta z usług świata IT. W innych przypadkach wszystkimi tymi czynnościami mogą zajmować się uczestnicy świata DS, a w większych działach / zespołach ludzie wyspecjalizowani w ML engineeringu lub inżynierii danych. Szczególnie ci drudzy raczej postrzegani jednak jako świat IT niż DS. Na rys. 45 ujęliśmy to jako subświat DS jedynie „stykający się" z granicą DS, a faktycznie będący „wewnątrz" świata IT; rzecz jasna pamiętając o schematyczności mapy i płynności oraz porowatości granic społecznych światów. Charakterystycznymi aktorami w obu tych światach technicznych są dane, kod, klienci i pieniądze. Są to zatem wspólne interesy finansowe, tj. światy DS i IT często zarabiają na tych samych projektach DS, realizowanych dla klientów wewnętrznych / zewnętrznych w ramach działań instytucji – różnego rodzaju firm. Sądzimy, że taka działalność komercyjna jest kluczowym czynnikiem przetrwania i rozwoju zarówno dla świata IT jak i DS. Dalej opiszemy drugorzędny, ale niezbędny czynnik przetrwania i rozwoju dla świata DS, czyli współpracę ze światem akademickim. Klientami projektów DS są czasami instytucje publiczne, jednak w Polsce skala takiej działalności jest na tyle mała, że nie ujęliśmy jej jako elementu zaangażowania świata polityki w opisywaną arenę. Kończąc wątek wspólnych interesów światów DS i IT musimy przypomnieć, że startupy, zespoły DS czy pojedynczy freelancerzy mogą realizować projekty i zarabiać niemal bez wsparcia świata IT. Przypomnijmy, że większość projektów DS kończy się na różnych etapach eksperymentów i nigdy nie dochodzi do produkcyjnych wdrożeń. Nie mamy żadnych informacji na temat tego, aby uczestnicy świata DS byli wynagradzani tylko za efekty wdrożeń czy w ogóle za doprowadzenie do wdrożenia – powszechnie występuje wynagradzanie nieuzależnione od etapu zakończenia projektu.

W obu światach technicznych obiekt graniczny definiowany jest oczywiście w sposób techniczny, czyli jako model uczenia maszynowego (por. dyskurs „ML" na rys. 45). Podkreślmy jednak stanowczo, że wybór metody oraz przygotowanie wytrenowanego, zwalidowanego i ocenionego jako „dostatecznie dobry" w odniesieniu do wymagań projektu modelu jest kontrolowane niemal wyłącznie przez świat DS. Ze strony świata IT pojawiają się negocjacje związane z inżynieryjnymi aspektami działania modelu, jak czas działania do

wygenerowania odpowiedzi modelu, jego stabilność, powtarzalność działania, ilość zużywanej mocy obliczeniowej, akceptowane przez model formaty danych itd. Raczej nie ma pomiędzy tymi światami negocjacji o aspektach metodologicznych, czy dotyczących przygotowania modelu. Dla świata IT modele ML są właściwie gotowymi, czarnymi skrzynkami. Powodem tego nie jest w żadnym razie brak rozeznania technicznego, ewentualnie brak rozeznania metodologicznego, ale naszym zdaniem, nie w tym rzecz. Świat IT jest zobowiązany do innych, inżynieryjnych zadań, a sfera przygotowania modelu jest po stronie świata DS i nie leży w interesie ludzi z IT zagłębianie się w nią.

Skoro na arenie modelowania interesy finansowe światów technicznych realizowane są w instytucjach, czyli niemal wyłącznie w firmach komercyjnych, to oczywiście **arena ta jest w dużej mierze kontrolowana przez światy biznesu oraz prawa**. Aktorem biorącym udział w każdym projekcie komercyjnym badanego świata jest formalne zobowiązanie tajemnicy, często w formie tzw. NDA (*non-disclosure agreement*, por. rys. 45). Podpisują je poszczególni uczestnicy projektu. Umowy typu NDA to ukonstytuowane zobowiązanie (*commitment*) uczestników świata DS wobec organizacji / instytucji i osób, które płacą za usługę (por. część 5.3.2). Tajemnica raczej nie obejmuje kodu, szczególnie jeżeli część rozwiązania ujęta została przez data scientistów w pakiet, to najczęściej jest on publikowany na zasadzie otwartości (*open source*). Nie znane są nam przypadki, by tajemnica obejmowała metody modelowania w sensie teoretycznym, ani ujęte w kod / pakiet. Tym samym, upraszczając nieco, światy biznesu i prawa nie mają władzy nad kodem i metodami ML. Raczej mają władzę nad przygotowanymi, wytrenowanymi modelami ML – te mogą być objęte tajemnicą i w przypadku modeli komercyjnych sądzimy, że nie ma praktyki publikowania gotowych modeli w open source; nie byłoby to zgodne z interesem świata biznesu.

Bez wątpienia świat biznesu ma dużą, może nawet największą na opisywanej arenie władzę nad danymi, zaś świat prawa walczy o uzyskanie władzy. Prawo jest w tym wypadku częściowo rzecznikiem polityki, a częściowo i pośrednio rzecznikiem (jedynym z bardzo niewielu rzeczników) użytkowników końcowych będących przedmiotem działania wdrożonych produkcyjnie modeli ML. Pamiętajmy, że ci „użytkownicy" nie zawsze będą ludźmi – wdrożenie może przecież służyć np. klasyfikacji innych obiektów (np. książek, filmów, ogórków, por. 5.2.4.), choć może i w tych wypadkach człowiek tylko pozornie nie jest przedmiotem; na pewno i tak jest ostatecznie adresatem odpowiedzi systemu bazującego na modelu ML. Nie ma zaś wątpliwości, że „użytkownikiem" jako przedmiotem działania

owych systemów są ludzie w przypadkach np. banków, firm windykacyjnych, zajmujących się edukacją, zdrowiem, marketingiem, ubezpieczeniami, telekomunikacją – w tych wypadkach na podstawie danych dotyczących ludzi modelowanie są ludzkie charakterystyki i działania.

Znaczący dla walki o władzę nad danymi aktor zaczął funkcjonować dokładnie od 25. maja 2018. To dzień, od którego w Polsce stosowane jest tzw. RODO (w jęz. angielskim GDPR – *General Data Protection Regulation*), czyli rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27. kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (czyli wcześniejszego, ogólnego rozporządzenie o ochronie danych). RODO na celu ma ujednolicenie i zwiększenie ochrony danych osobowych obywateli Unii Europejskiej. Wprowadza ono dla obywateli np. możliwość wystąpienia do administratora „swoich" danych o usunięcie tychże (tzw. prawo do bycia zapomnianym) (Czapska 2018; Ministerstwo Cyfryzacji RP 2018a). W badanym świecie DS to prawo do bycia zapomnianym jest dyskutowane m.in. w kontekście sytuacji, w której już po przygotowaniu modelu ML osoba występuje o usunięcie „swoich" danych. Czy w takiej sytuacji legalne będzie korzystanie z modelu, w którym w procesie przygotowania był użyty rekord z zapisem danych owej osoby, czy model należałoby uczyć i walidować ponownie, na danych bez rekordu, do którego usunięcia zobowiązuje RODO {obserwacja 46}? Model wyuczony i zwalidowany na określonych zbiorach danych byłby przecież inny, gdyby wykorzystano inne zbiory danych. Poza tym ponowne przygotowanie modelu jest kosztowne, a żądania usunięcia danych mogą się powtarzać. To potencjalnie prowadzi do poważnych konsekwencji, stawiających pod znakiem zapytania opłacalność używania modeli ML w zastosowaniach dla danych osobowych (a właściwie każdych algorytmów, w których parametry modelu są estymowane z danych, nawet regresji liniowej czy logitowej), zatem zdecydowanie takie rozwiązanie prawne nie leży w interesie światów IT, DS i biznesu. Obecnie, w lipcu 2020 r. nie mamy żadnych informacji, by istniała jednoznaczna interpretacja obowiązujących przepisów w opisywanej sprawie.

Spór o sposoby definiowania tego, czym w ogóle są dane osobowe można by opisać także jako arenę – bez wątpienia jest to temat na niejedną rozprawę socjologiczną. Wydaje nam się, iż zaangażowane są w nią te same światy, pewnie część opisywanych przez nas instytucji i aktorów (por. rys. 45.), ale raczej poza badanym światem DS. Z naszych badań wynika, że problemy dotyczące legalności danych, z których korzystają uczestnicy świata DS

są postrzegane jako problemy świata DS tylko przez niektórych uczestników – osoby posiadające firmy lub zarządzające nimi, osoby zarządzające działami / zespołami DS. W akademii raczej takie problemy podnoszą także osoby na stanowiskach profesorskich, poza tym problem ten jest zdecydowanie mniej obecny. W środowisku komercyjnym osoby „szeregowe" zajmujące się DS postrzegają problemy prawne jako nie-techniczne oraz nie-biznesowe, zatem raczej nie leżą one w ich obszarze zainteresowania. Osoby „szeregowe" uważają, że odpowiedzialność za kwestie legalne spoczywa na działach prawnych (w dużych firmach wewnętrzny dział prawny bywa określany np. jako „bezpieka" {obserwacja 14}), osobach zarządzających wyższego szczebla, a także klientach[262]. Także komercyjnie, ale i akademicko oraz hobbystycznie używane są różne źródła danych, np. dane z wolnego dostępu (*open access*), dane z rozmaitych systemów przemysłowych czy internetowych, ale nie będące danymi osobowymi (nawet, gdy dotyczą ludzi np. z Google Analytics); oczywiście są to dane zarówno ustrukturyzowane jak i nieustrukturyzowane. Czasami dane są ściągane z sieci za pomocą metod web scrapingu, i akurat ta praktyka jest w badanym świecie powszechnie uznawana za mogącą prowadzić osoby / firmy / uczelnię do nieprzyjemnych konsekwencji prawnych i stosowana jest albo w przypadkach nie budzących wątpliwości, albo potajemnie. Nie jesteśmy w stanie wypowiedzieć się o charakterystyce ani o skali takich praktyk. Z pewnością w naszych badaniach świata DS generalnie nie mieliśmy dostępu do informacji o nielegalnych praktykach dotyczących danych. Jeden raz w rozmowie nieformalnej wskazano nam, że wcześniej stosowane przez tę osobę praktyki posługiwania się danymi „chyba były nielegalne" lub „chyba obecnie zostałby uznane za nielegalne", więc lepiej o tym nie rozmawiać. Podkreślmy wyrażone wahanie – w walce o władzę nad danymi świat prawa jest zawsze opóźniony w stosunku do możliwości technicznych i praktyki biznesowej, więc czasami nie wiadomo, co jest legalne a co nie, ponieważ liczne kwestie są prawnie nie uregulowane.

Bardziej znaczącym dla świata DS elementem regulacji prawnych związanych z RODO jest tzw. prawo do wyjaśnienia. To „prawo" nie jest precyzyjnie określone, ale kilka pozycji w opisywanym sporze (tzn. modelowaniu jako arenie) stoi na stanowisku, iż wynika ono z RODO. Chodzi o zapisy w sekcji 4 – „Prawo do sprzeciwu oraz zautomatyzowane podejmowanie decyzji w indywidualnych przypadkach", na które składają się artykuły 21 i 22

262 Powtórzymy, że 78% badanych polskich „firm AI" w pytaniu wielokrotnego wyboru wskazało, że dane na potrzeby projektu zapewnia im klient, 77% że dane zapewniają oni sami, 73% wskazuje na korzystanie z danych darmowych w wolnym dostępie, zaś najmniej (31%) deklarowało zakup danych (Borowiecki i Mieczkowski 2019:42–43; por. 5.3.2).

– „Prawo do sprzeciwu" oraz „Zautomatyzowane podejmowanie decyzji w indywidualnych przypadkach, w tym profilowanie". Faktycznie w Polsce tzw. prawo do wyjaśnienia[263] obowiązuje od 4. maja 2019 tylko i wyłącznie w odniesieniu do decyzji kredytowych. Podkreślmy, że komercyjne modele oceny zdolności kredytowej powstawały już w latach 50. i choć Przanowski uważa, że są one „prekursorami" współczesnych modeli w dziedzinie data science (2014:13-15), to w społecznym świecie DS w Polsce raczej nie uznaje się oceny zdolności kredytowej za DS. To zastosowanie modelowania jest nie tylko starsze niż DS, ale przede wszystkim bazuje na innych metodach i narzędziach niż te, uznawane za „właściwe" w badanym świecie. Pewna część osób zajmujących się różnego rodzaju modelowaniem – nie tylko zdolności kredytowej – w instytucjach finansowych posługuje się np. językiem R i może być uznawana za uczestników świata DS, ale zdarza się to rzadko. Sądzimy, że uczestnicy świata DS przeważnie postrzegają modelowanie w instytucjach finansowych jako odmianę „tradycyjnej" analityki danych i statystyki, dodatkowo raczej nudną i konserwatywną, bo przez regulacje prawne ograniczoną do łatwo dających się wyjaśnić metod statystycznych – raczej nie można w tej dziedzinie stosować najmodniejszego w DS uczenia głębokiego ani nawet złożonych modeli klasycznego ML (np. SVM, lasów losowych) czy zestawów wielu modeli (*model ensemble*); prawie nie ma w tej dziedzinie pracy z danymi nieustrukturyzowanymi ani używania dużych infrastruktur obliczeniowych w tzw. „chmurze". Wejście w Polsce opisywanego prawa do wyjaśnienia decyzji kredytowej opisuje jako swój

263 Ustawa z dnia 21 lutego 2019 r. o zmianie niektórych ustaw w związku z zapewnieniem stosowania rozporządzenia Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych) (Dz.U. 2019 poz. 730):
Art. 70a
1. Banki i inne instytucje ustawowo upoważnione do udzielania kredytów na wniosek osoby fizycznej, prawnej lub jednostki organizacyjnej niemającej osobowości prawnej, o ile posiada zdolność prawną, ubiegającej się o kredyt przekazują, w formie pisemnej, wyjaśnienie dotyczące dokonanej przez siebie oceny zdolności kredytowej wnioskującego.
2. Wyjaśnienie, o którym mowa w ust. 1, obejmuje informacje na temat czynników, w tym danych osobowych wnioskującego, które miały wpływ na dokonaną ocenę zdolności kredytowej.
3. W przypadku przedsiębiorców bank i inna instytucja ustawowo upoważniona do udzielania kredytów może pobierać opłatę za sporządzenie wyjaśnienia, o którym mowa w ust. 1. Wysokość opłaty powinna być odpowiednia do wysokości kredytu.
4. Przepisy ust. 1–3 stosuje się odpowiednio do przedsiębiorcy ubiegającego się o pożyczkę pieniężną.
Art. 105a ust. 1c
1a. Banki, inne instytucje ustawowo upoważnione do udzielania kredytów, instytucje pożyczkowe oraz podmioty, o których mowa w art. 59d ustawy z dnia 12 maja 2011 r. o kredycie konsumenckim, a także instytucje utworzone na podstawie art. 105 ust. 4, mogą w celu oceny zdolności kredytowej i analizy ryzyka kredytowego podejmować decyzje, opierając się wyłącznie na zautomatyzowanym przetwarzaniu, w tym profilowaniu, danych osobowych – również stanowiących tajemnicę bankową – pod warunkiem zapewnienia osobie, której dotyczy decyzja podejmowana w sposób zautomatyzowany, prawa do otrzymania stosownych wyjaśnień co do podstaw podjętej decyzji, do uzyskania interwencji ludzkiej w celu podjęcia ponownej decyzji oraz do wyrażenia własnego stanowiska.

sukces Fundacja Panoptykon, której działalność jest przykładem zaangażowania organizacji pozarządowych w arenę (por. rys. 45):

> Zmiany w prawie bankowym, wprowadzające prawo do wyjaśnienia decyzji kredytowej dla każdego, to jeden z sukcesów, które osiągnęliśmy dzięki wytrwałości naszego zespołu: przez kilka miesięcy składaliśmy opinie, uczestniczyliśmy w komisjach sejmowych i spotkaniach roboczych, przekonując przedstawicieli rządu i banków, że taka zmiana jest potrzebna (Fundacja Panoptykon 2019).

Tzw. prawo do wyjaśnienia – w sensie nieprecyzyjnego przepisu RODO – jest co do zasady próbą prawnego uregulowania obiektu granicznego opisywanej areny, czyli modelu ML. W sensie technicznym całą władzę nad przygotowaniem modelu ML ma świat DS. To od konkretnych osób piszących kod zależą kluczowe szczegóły, to w gronie osób piszących kod zapadają decyzje techniczne. Kształt modelu jest negocjowany z klientami w obszarze biznesowym, z IT w obszarze inżynieryjnym, prawnikami w kwestiach legalnych, rzadziej z akademią w zakresie metod, ale nikt poza zespołem DS czy wręcz konkretnym człowiekiem piszącym kod nie ma władzy by zadecydować o wyborze metody ML, o użyciu konkretnego przekształcenia na danych, o zastosowaniu konkretnej wartości hiperparametru. Chyba największy wpływ na kształt modelu (mówimy o środowisku komercyjnym) mają klienci, których uwagi (w rodzaju np. model działa za wolno; jest za dużo fałszywych alarmów; system jest za drogi, bo zużywa x zasobów infrastruktury obliczeniowej; segmentacja użytkowników według modelu jest zupełnie inna niż nasza segmentacja z badań marketingowych) są w zespołach DS przy ewentualnej współpracy z IT tłumaczone na szczegóły techniczne i służą do optymalizowania przygotowywanego rozwiązania w kolejnych iteracjach. Zatem z punktu „szeregowych" data scientistów sytuacja z tzw. prawem do wyjaśnienia wygląda podobnie, jak z innymi przepisami wynikającymi z RODO czy dowolnych aktów prawnych[264] – nie zajmują się tym, ponieważ ich praca ma charakter techniczny, w razie potrzeby dostosowują się do zaleceń przełożonych i prawników własnych lub klienta. Dla DS jako działalności akademickiej lub hobbystycznej to prawo nie ma zastosowania.

Aktywnie w walce o „wyjaśnialne" modelowanie ML uczestniczą raczej osoby bardzo silnie osadzone w badanym świecie, jak Przemysław Biecek. Istnieje akademicki obszar badań i rozwoju uczenia maszynowego, który jest odpowiedzią właśnie na powstające i

264 Dowolnych aktów prawnych, ponieważ osoby zajmujące się DS w Polsce mogą pracować dla klientów na całym świecie (por. część 5.1.3). Wielokrotnie wskazywano nam w badaniach, że dana osoba / zespół DS nie pracuje w ogóle dla klienta polskiego. Około 1/3 badanych przez Fundację digitalpoland podmiotów (firm AI w Polsce) uzyskuje dochód głównie – w co najmniej 70% – ze zleceń zagranicznych (Borowiecki i Mieczkowski 2019:6, 20).

przyszłe regulacje prawne, odpowiedzią na kontrowersje etyczne stosowania modeli ML oraz odpowiedzią na zapotrzebowanie klientów biznesowych na zrozumiałość modeli (do czego jeszcze wrócimy). Nazywa się ten obszar XAI (*eXplainable Artificial Intelligence*, dosł. wyjaśnialna sztuczna inteligencja). P. Biecek m.in. publikował w tej dziedzinie artykuły naukowe[265] oraz pakiety (np. DALEX) i współtworzy wyposażone w chatbota narzędzie wyjaśniana modeli ML (Biecek 2018d; Kuźba i Biecek 2020; Staniak i Biecek 2018), wypowiadał się o XAI na wydarzeniach świata DS (Biecek 2018f) i {obserwacja 46}, prowadził w tym zakresie zajęcia na Politechnice Warszawskiej i Uniwersytecie Warszawskim we współpracy z firmą McKinsey Digital (Biecek 2020). Tego rodzaju działalność w dziedzinie XAI widzimy jako wspólną walkę światów DS i akademickich badań nad ML (w subświecie nauk ścisłych) o wspólne interesy i pozycje, warunkujące ich przetrwanie i rozwój – szczególnie świata DS. Rozwiązania techniczno-metodologiczne, np. w formie pakietów takich jak DALEX, mają przecież pozwolić na wyjaśnianie nawet bardzo złożonych modeli ML, zatem zapobiec „wykastrowaniu" świata DS z najbardziej efektywnych, najbardziej sprawczych, być może najbardziej dochodowych metod zaawansowanego ML, uczenia głębokiego i zestawów modeli (*model ensemble*). Bez względu na to, skąd pochodzi żądanie wyjaśnienia – czy jest ono wymuszone przez legislację, czy postulowane w mediach lub w świecie polityki (być może w formie rzecznictwa w imieniu marginalnie traktowanych użytkowników końcowych), czy jest ono wymagane przez coraz bardziej wyedukowanych klientów nie mających zaufania do systemów bazujących na ML (częściowo także z obawy przed konsekwencjami prawnymi czy wizerunkowymi), a więc i chęci inwestowania w systemy będące tzw. czarnymi skrzynkami. Metody XAI są zorientowane na zachowanie komercyjnych zastosowań niezbędnych, konstytutywnych dla świata DS technologii – czyli modeli ML – w obliczu żądań o wyjaśnialność. Bez wątpienia wyjaśnialność w sensie prac w dziedzinie XAI leży także w interesie akademii, bowiem jako obszar nowy i modny w Europie może być obiecujący w perspektywie indywidualnych karier akademickich, jak i budowania renomy instytucji poprzez publikacje naukowe, granty, udział w konferencjach, przedmioty i kierunki studiów, może udział w debacie publicznej i wpływ na światy oraz możliwości współpracy ze światami polityki, prawa i szczególnie biznesu. W ogóle przetrwanie i rozwój każdego rodzaju badań nad ML czy właściwie nad dowolnymi metodami i narzędziami związanymi z DS leży w interesie akademii. Punktami przecięcia ze światem DS są przecież nie tylko metody i pakiety (same w sobie niekomercyjne), ale i granty

265 „Praca została dofinansowana w ramach projektu RENOIR w ramach umowy grantowej Marie Skłodowskiej-Curie nr 691152 oraz grantu NCN Opus 2016/21 / B / ST6 / 02176" (Biecek 2018d:4).

(np. NCN lub NCBiR), doktoraty wdrożeniowe, publikacje naukowe. Oczywiście są osoby, określające się np. jako ML researcher, które łączą karierę akademicką i komercyjną, i to nie tylko w XAI. Bez wątpienia wyjaśnialność, w sensie praktycznej możliwości wyjaśnienia działania (wdrożonego produkcyjnie modelu) konkretnemu operatorowi systemu bazującego na modelu ML lub konkretnemu użytkownikowi – przedmiotowi działania takiego systemu, leży w interesie biznesu. O ile oczywiście jest dostatecznie łatwa, efektywna i optymalna kosztowo w porównaniu z ryzykiem i kosztem konsekwencji prawnych, wizerunkowych lub etycznych.

Nauki społeczne i humanistyczne także włączają się w spór o rolę RODO w regulacji systemów bazujących na danych i modelach. Wskazywano m.in. że

> w RODO brakuje precyzyjnego języka i explicite wyrażonych praw (*rights*) oraz zabezpieczeń (*safeguards*), zatem pojawia się ryzyko, że to prawo będzie "bezzębne" (*toothless*) [tłum. własne RŻ] (Wachter, Mittelstadt, i Floridi 2017).

W artykule zawarto wiele słów krytyki wobec omawianego rozporządzenia, wskazując przede wszystkim na duże pole do jego sprzecznych interpretacji, a nawet samo to, czy prawo do wyjaśnienia w ogóle rozwiązuje problem transparencji i odpowiedzialności / rozliczalności (*accountability*) (ibidem). W ogóle twierdzimy, że humanistyka i nauki społeczne uczestniczą w omawianej arenie modelowania tylko jako głos krytyczny. Dzieje się tak zarówno w sensie krytyki metodologiczno-epistemologicznej (por. część 2.2.1. tej rozprawy), gdzie czasem po tej samej stronie krytyki stają przedstawiciele nauk ścisłych i przyrodniczych, i przede wszystkim krytyki konsekwencji społecznych (2.2.3.), w tej sferze widzimy oczywiście znaczący wkład nauk prawnych. Raczej nie ma podstaw, by mówić o udziale w arenie tych przedstawicieli humanistyki i nauk społecznych, którzy sami używają modeli ML w swojej pracy naukowej – i tak rzadko są to rzeczywiście modele ML, raczej rozmaite sposoby zbierania, przetwarzania i analizy danych, a jeżeli już, to według nas model ML nie stanowi obiektu granicznego dla tej części subświata (czemu daliśmy wyraz na rys. 45). Marginalny udział w opisywanej arenie mogą brać ci badacze społeczni, zajmujący się DS jako badanym wycinkiem rzeczywistości, jak my. Niemniej rozwój, choć może niekoniecznie przetrwanie świata DS leży w interesie nauk społecznych i humanistyki w ogóle. Świat DS dostarcza temu subświatowi akademii nowinek i modnych tematów: nowych narzędzi i strategii metodologicznych oraz pola badań i refleksji krytycznej w zakresie metodologii i epistemologii, konsekwencji społecznych, ekonomii, prawa, etyki, filozofii.

Wracając do wątku własności danych, już pod koniec lat 90. w humanistyce wskazywano, że cyfryzacja doprowadzi do coraz to nowych wyzwań prawnych; zasadnicze pytanie miało dotyczyć tego, do kogo będzie należała wiedza (Lyotard 1997:33). Zgadzamy się z Hararim, który w charakterystyczny dla siebie makrohistoryczny i jednocześnie popularnonaukowy sposób stwierdza, że władza gigantów technologicznych staje się w groźna kontekście własności danych, i to w kilku aspektach:

> Jeśli dane znajdą się w rękach zbyt niewielu ludzi – ludzkość podzieli się na różne gatunki. Wyścig po dane już trwa, a prowadzą w nim tacy giganci, jak Google, Facebook, Baidu i Tencent. Na razie wydaje się, że wielu z nich przyjęło model biznesowy „handlarzy uwagi". (...) Jednak ci giganci z branży IT prawdopodobnie mierzą znacznie wyżej niż którykolwiek z wcześniejszych handlarzy uwagi. W rzeczywistości nie chodzi im wcale o sprzedawanie reklam. Jest raczej tak, że przykuwając naszą uwagę, potrafią gromadzić ogromne ilości danych na nasz temat, które są warte wielokrotnie więcej niż dochody z reklam. Nie jesteśmy ich klientami – jesteśmy ich produktem. W perspektywie średnioterminowej ten skarb w postaci danych otwiera drogę radykalnie odmiennemu modelowi biznesowemu, którego pierwszą ofiarą będzie sam przemysł reklamowy. Ten nowy model opiera się na przeniesieniu władzy z ludzi na algorytmy, w tym także władzy dokonywania wyborów i kupowania różnych rzeczy. Z chwilą gdy algorytmy zaczną wybierać i kupować za nas, tradycyjna branża reklamowa padnie. (…) Na dłuższą metę, gromadząc dostateczną ilość danych i dysponując wystarczającą mocą obliczeniową, informatyczni giganci będą mogli zhakować najskrytsze tajemnice życia, a następnie wykorzystać tę wiedzę nie tylko do dokonywania za nas wyborów albo do oddziaływania na nas, lecz również do przeprojektowywania życia organicznego i do tworzenia form życia nieorganicznego. Sprzedawanie reklam może być konieczne do podtrzymania tych gigantów na krótką metę, często jednak dokonywane przez nich oceny aplikacji, produktów i spółek opierają się na pozyskiwanych przez nie danych, a nie na generowanych przez nie przychodach pieniężnych. (...) Nie mam absolutnej pewności, że informatyczni giganci otwarcie myślą o tym w takich kategoriach, jednak ich działania wskazują na to, że wyżej cenią gromadzenie danych niż same dolary i centy (Harari 2018:111-12).

Pisaliśmy (część 6.1.2.1), że Harari przyjmuje za prawdę to, co opisujemy jako jedną ze strategii legitymizacji badanego świata DS. Nie oznacza to jednak, że jego tezy są bez sensu – pomijając futurystyczne wizje i skupiając się na etnograficznym „tu i teraz" – bez wątpienia słusznie wskazuje on na rzeczywisty problem olbrzymiej władzy gigantycznych firm technologicznych nad danymi. Nie będziemy powtarzać argumentacji i historii takich, jak słynna „afera" Cambridge Analytica (por. część 2.3.3). Ta władza i wpływ na opisywaną arenę modelowania jest ogromny, także w innych aspektach.

Giganci technologiczni wspierają najpopularniejsze[266] na świecie pakiety open source do modelowania ML. TensorFlow (TF), nakładka Keras (i mniej popularny TensorFlow.js) są wspierane przez Google, zaś PyTorch (i mniej popularne Caffe) przez Facebook. Giganci

266 W Polsce na używanie TF wskazało 69% badanych firm, Keras 49%, zaś PyTorch/Torch – 38% (Borowiecki i Mieczkowski 2019:39–40; por. 5.3.1.)

technologiczni oferują także popularne[267] usługi udostępniania mocy obliczeniowej w tzw. chmurze – Microsoft Azure, Amazon Web Services (AWS) i Google Cloud Platform (GCP). Przykłady zaangażowania gigantów technologicznych w arenę modelowania można by mnożyć: od kursów MOOC, oferowanych nie tylko przez wymienione firmy, ale także np. IBM, przez inne narzędzia np. popularny do zastosowań dydaktycznych notatnik programistyczny Google Colab. Nie zapominajmy o tym, że nawet jeżeli MOOC czy dowolny kurs / szkolenie / studia oferuje jakaś inna firma czy uczelnia, to uczy pracy na narzędziach technologicznych gigantów – inaczej nie jest to popularny kurs. Do Google należy także wielokrotnie wspominana platforma Kaggle, za pomocą której ludzie uczestniczą w konkursach modelowania ML. W prasie technologicznej wskazywano, że kilka lat temu dołączanie gigantów technologicznych do społeczności open source było nowością. Firmy te mogłyby przecież tworzyć zamknięte (*proprietary*), komercyjnie oprogramowanie do ML. Uznały jednak, że ten model nie sprawdzi się w przypadku DS. Firmy takie jak Google i Facebook wspierają rozwiązania open source zapewne po to, by zachować możliwość szybkiego wprowadzania najnowszych rozwiązań metodologicznych, płynących do świata DS z akademii (ewentualnie przez subświat ML research). Inaczej środowisko DS szybko porzuciłoby „przestarzałe" oprogramowanie zamknięte, wypuszczone na rynek przez technologicznego giganta. Inną zaletą wspierania oprogramowania open source jest to, że taki rodzaj oprogramowania przyciąga talenty z obszaru DS do gigantów technologicznych. Talenty, dla których ważny jest jednocześnie samorozwój, wyzwania, sprawczość i otwarty dostęp do wypracowanych rozwiązań (por. część 5.4. o wartościach świata DS). Takie „talenty" przyciągają z kolei innych data scientistów do pracy dla technologicznych gigantów, już nawet niekoniecznie w rozwijaniu narzędzi open source, ale w ogóle w zespołach / działach DS realizujących projekty dla tych gigantów. Autor pakietów Caffe i Caffe2, Yangqing Jia, pracował nad TF będąc zatrudnionym w Google, a później przeszedł do Facebooka (Arakelyan 2017). Przypomnijmy, że wśród najsłynniejszych w świecie DS naukowców w Google pracowali m.in. Andrew Ng i Petre Norvig, zaś w Facebooku np. Yann LeCun (por. część 2.1. rozprawy). Giganci technologiczni są w badanym świecie DS uznawani za wybitnie atrakcyjnych pracodawców, choć niektórzy uczestnicy wskazują, że nie podobają im się pewne ich działania. Podczas wywiadów, mówiąc np. o problemach etycznych związanych z modelami ML, wielokrotnie wspominano nam historie opisane w części 2.2.3. tej rozprawy, jak niesławny model używany do selekcji CV do pracy w Amazon.

267 Według raportu Fundacji digitalpoland z obliczeń w „chmurach" Azure, AWS, GCP korzystało 76% badanych firm (pytanie umożliwiało wielokrotny wybór, por. 5.3.3.)

Rok 2018, szczególnie w związku z tzw. skandalem Cambridge Analytica był przełomowy, ponieważ od tamtego czasu kwestie naruszeń prywatności i etyczność wykorzystywania danych weszły do głównego nurtu dyskursu publicznego. Cała sprawa koncentruje się jednak na problemie dostępu do danych, własności danych i celu ich użycia, a nie na modelach. Problemy etyczne dotyczące używania modeli ML są zdecydowanie poza głównym nurtem dyskursu publicznego, niemniej są znane i co najmniej od 2018 są na agendzie w opisywanej arenie modelowania. Zarzuty etyczne płyną ze strony światów: mediów (czasami w roli rzecznika użytkowników końcowych), polityki, rzadziej prawa lub nauk społecznych i humanistycznych. Mogą one płynąć także od biznesu-klienta. Odpowiedzi widzimy dwojakie. Tą techniczną, ze strony świata DS i akademii (nauk ścisłych), jest opisywane wyżej XAI. Odpowiedzią nie-techniczną, charakterystyczną przede wszystkim dla świata biznesu są tzw. kodeksy etyczne, czasami rady/komisje ds. etyki. Widzimy kodeksy etyczne jako aktora usytuowanego na peryferiach opisywanej areny modelowania (por. rys. 45), aktora mającego styczność ze światem DS, ale ulokowanego na przecięciu światów biznesu, mediów i polityki (pewnie można by opisać kodeksy etyczne jako obiekt graniczny dla „areny wewnętrznej" tych trzech światów). Kodeksy etyczne mają też styczność ze światem prawa, ale nie są aktami prawnymi, więc prawnicy przyjmują role pośrednie zarówno w tworzeniu (głównie we współpracy z biznesem) jak i krytyce (np. we współpracy z mediami, ewentualnie naukami społecznymi i humanistyką lub wręcz jako naukowcy) takich publikacji. Przytaczaliśmy już krytykę ze strony nauk społecznych i humanistyki (2.2.3.5. tej rozprawy), gdzie jako rozwiązania wyłącznie deklaratywne opisywano kodeksy etyczne, kodeksy zasad itp., oraz certyfikaty lub zezwolenia. Powtórzmy jedynie, że zgadzamy się z tezą, iż takie publikacje stanowią przede wszystkim działanie wizerunkowe firm, które w myśl liberalnej poprawności politycznej muszą pokazać światom mediów, polityki, prawa czy wreszcie klientom i użytkownikom, że dbają o prywatność, brak dyskryminacji itp. Nastawione jest to na budowanie albo odzyskiwanie zaufania do świata biznesu przez inne światy, od których zależy przetrwanie i rozwój rozmaitych firm. Trafna wydaje nam się również teza (por. Wagner 2018:84), że firmy, a szczególnie technologiczni giganci, chętnie tworzą np. kodeksy czyli ogólnie miękkie rozwiązania deklaratywne, aby uniknąć twardych regulacji prawnych. Bez wątpienia brak regulacji prawnych może leżeć w interesie wielu firm zarabiających dzięki modelom ML i traktowany być przez nie jako stan pożądany, zaś regulacje prawne postrzegane jako przeszkoda w działaniu biznesu.

W polskim świecie DS różnego rodzaju miękkie rozwiązania deklaratywne postrzegane są przez uczestników jako coś „spoza" świata DS; coś, co ma znikomy lub żaden wpływ na opisywaną arenę modelowania i na sam badany świat społeczny. W sierpniu 2020 r. nie mamy informacji, aby w polskim świecie DS czy jakimkolwiek jego segmencie istniał dokument w rodzaju np. kodeksu etycznego postępowania data scientistów. Istnieją dokumenty zwane kodeksami postępowania (*code of conduct*), tworzone przez organizacje dość powszechnie uznawane za należące do świata DS (nie tylko polskiego), ale nie dotyczą one modelowania, są tworzone głównie jako kodeksy dla konferencji (PyData 2018; Why R? Foundation 2018). Podobną treść mają kodeksy postępowania organizacji i inicjatyw nastawionych na włączanie do DS mniejszości genderowych (*gender minorities*) (por. PyLadies 2019; R-Ladies 2019; WiMLDS 2018). Istnieje kodeks postępowania profesji data science (*Data Science Code of Professional Conduct*), który mógłby się wydawać względnie oddolnym, „wewnętrznym" dla świata DS dokumentem, ale w badanym przez nas świecie społecznym nie słyszeliśmy o nim nigdy – został przygotowany około 2014 przez mało wpływowe i nieobecne w Polsce stowarzyszenie (Data Science Association 2020). Propozycję profesjonalnej przysięgi Hipokratesa dla data scientistów przytoczyła też O'Neil (2017a:277). Jej treść[268] po załamaniu rynków w 2008 r. stworzyli Emanuel Derman i Paul Wimott, specjaliści od inżynierii finansowej. Choć książka O'Neil jest znana w polskim świecie DS, to z wspomnianą przysięgą nie spotkaliśmy się w trakcie badań nigdy – sądzimy, że z perspektywy uczestników świata DS nie ma ona w badanym świecie ani w opisywanej arenie modelowania żadnego znaczenia. Podobnie z perspektywy uczestników w badanym świecie raczej nie mają znaczenia tego rodzaju dokumenty[269] powstające lokalnie, poza światem DS. W Polsce poradniki/raporty poświęcone etyce tzw. sztucznej inteligencji przygotowały np. organizacje pozarządowe: Klub Jagielloński z Fundacją Centrum Cyfrowe (Tarkowski,

268 „Będę pamiętał, że to nie ja stworzyłem świat oraz że nie musi on dopasowywać się do moich wyliczeń. Będę śmiało używał modeli do określenia wartości, nie będę jednak ulegał nadmiernej fascynacji matematyką. Nigdy nie będę dążył do elegancji modeli kosztem zgodności ze stanem rzeczywistym bez wytłumaczenia, dlaczego tak postąpiłem. Nie będę stwarzał wobec klientów korzystających z moich modeli fałszywego poczucia pewności co do ich trafności. Zamiast tego będę wyraźnie informował o ich założeniach oraz brakach. Rozumiem, że moja praca może mieć ogromny wpływ na społeczeństwo oraz ekonomię, a którego w znacznej części mogę sobie nie zdawać sprawy" (O'Neil 2017a:277).

269 Przypomnijmy, że dla części uczestników mają znaczenie publikacje przygotowane przez podmioty spoza świata DS, np. zawierające wyniki badań związanych z branżą DS w Polsce – wielokrotnie cytowane przez nas w części 5.1. (Borowiecki i Mieczkowski 2019; Gontarz 2019).

Paszcza, i Mileszyk 2020) [270] i Fundacja Panoptykon (Szymielewicz i Obem 2020) [271]; w drugim dokumencie zacytowano wypowiedź P. Biecka i podano link do strony MI^2 DataLabu, zajmującego się m.in. XAI, ale wskazuje to jedynie na rozeznanie autorek w temacie, a nie na znaczenie dokumentu dla świata DS. Z nieco większym zainteresowaniem traktowane mogą być takie dokumenty, jak tzw. Biała Księga sztucznej inteligencji (Komisja Europejska 2020), szczególnie przez właścicieli firm zajmujących się DS i osoby na stanowiskach kierowniczych w DS, ponieważ tego rodzaju dokumenty wskazują na potencjalny kierunek rozwoju regulacji prawnych oraz przyszłe kierunki publicznego finansowania projektów badawczych / naukowych, we współpracy firm i akademii w tzw. AI. Dokumentów dotyczących potencjalnego lub realnego finansowania badań nad AI istnieje już wiele, jak wymieniono np. w raporcie z pierwszego Zjazdu Polskiego Porozumienia na Rzecz Rozwoju Sztucznej Inteligencji (PP-RAI, por. rys. 45):

> Panelistka zwróciła jednocześnie uwagę, na znaczenie europejskiego partnerstwa publiczno-prywatnego (Big Data Value Association - BDVA), które odgrywa kluczową rolę przy definiowaniu tematów związanych z Big Data w ramach programu H2020. Następnie p. Szołucha zaprezentowała szacunkowe budżety przeznaczone przez KE dla poszczególnych obszarów tematycznych. Zgodnie z zapowiedziami KE inwestycje w projekty B+R w latach 2018-2020 mają wynieść ok. 1.5 mld Euro. Według panelistki te inwestycje KE mają przygotować grunt do przyszłych inwestycji, planowanych na nową perspektywę budżetową, na lata 2021-2027. Zgodnie z zapowiedziami KE, obok Programu Horyzont Europa (następcy Programu Horyzont 2020), utworzony zostanie nowy program, Program Cyfrowa Europa (ang. Digital Europe), w ramach którego mają być realizowane wdrożenia i testy w środowiskach przemysłowych. Planowany budżet programu to ponad 9 miliardów Euro. Jak podkreślała podczas swojego wystąpienia Pani Szołucha, jednym z pięciu strategicznych obszarów tego programu ma być sztuczna inteligencja, z budżetem 2.5 miliarda Euro na działania związane z upowszechnianiem wdrożeń sztucznej inteligencji w całej gospodarce i społeczeństwie europejskim (Czarnowski i in. 2018:11–12).

Sam dokument taki jak tzw, Biała Księga jest w naszej opinii w komercyjnym DS traktowany jako nie mający znaczenia, szczególnie dla codziennej pracy „szeregowych" data scientistów. Może on być uznawany za potencjalnie wpływowy raczej przez przedstawicieli świata biznesu:

> Biznes aprobuje to, że Komisja [Europejska] chce regulować wyłącznie zastosowania SI „wysokiego ryzyka", czyli np. w medycynie czy transporcie. „Biała księga" pomija jednak sprawę zasadniczą: odpowiedzialności za szkodę oraz krzywdę wyrządzoną przez rozwiązania oparte na sztucznej inteligencji – zauważa Aleksandra Musielak, ekspertka z Departamentu Prawa Gospodarczego Konfederacji Lewiatan w rozmowie z naszym portalem

270 Tekst pod tytułem zawierającym powszechne na opisywanej arenie odniesienie do popkultury – brytyjskiego, dystopijnego serialu sci-fi *„AlgoPolska": 11 postulatów, których realizacja pozwoli uniknąć mrocznych wizji rodem z „Black Mirror".* Działanie sfinansowane ze środków Programu Rozwoju Organizacji Obywatelskich na lata 2018-2030.

271 Tekst pt. *Sztuczna inteligencja non-fiction*. Materiał powstał dzięki wsparciu Samsung Electronics Polska.

[sztucznainteligencja.org.pl] (Zagórna 2020).

Swoje wytyczne, co do etyki AI, przygotowała także – poza tym epizodem nieobecna na arenie (rys. 45.) – strona Kościoła rzymskoatolickiego (Pontificia Accademia per la Vita 2020b). Dokument „Wezwanie Rzymu o etykę sztucznej inteligencji" (*Rome call for AI ethics*) sygnowali w dniu 28. lutego 2020 przewodniczący Papieskiej Akademii Życia (*Pontificia Accademia per la Vita*) Vincenzo Paglia, prezes firmy Microsoft Brad Smith, wiceprezes IBM John Kelly III, dyrektor generalny ONZ ds. Wyżywienia i Rolnictwa Dongyu Qu i włoska minister ds. innowacji technologicznych Paola Pisano. W ceremonii uczestniczył przewodniczący Parlamentu Europejskiego David Sassoli, odczytano list Papieża Franciszka, wystąpił wielokrotnie przez nas cytowany filozof L. Floridi, zaś sygnatariusze wezwania wyrazili chęć współpracy w celu promowania „etyki algorytmów" (*algor-ethics*), czyli etycznego wykorzystania AI zgodnie z sześcioma[272] zasadami (Mróz 2020; Pontificia Accademia per la Vita 2020a). Kościół rzymskokatolicki oraz w ogóle kościoły, związki wyznaniowe i instytucje religijne są – lub może były do niedawna – tym, co Clarke określiła jako „światy nieobecne, jakich można byłoby oczekiwać, że się pojawią [na opisywanej arenie modelowania]" (Kacperczyk 2016:573). Może ma trochę racji Harari, twierdząc iż współcześnie religie nie radzą sobie w odpowiadaniu na problemy techniczne ani polityczne, ale potrafią budować interpretacje kształtujące tożsamość:

> W XXI wieku zatem religie nie zajmują się wprawdzie sprowadzaniem deszczu, leczeniem chorób ani budowaniem bomb – ale zaczynają ustalać, kim jesteśmy „my", a kim „oni", kogo powinniśmy leczyć, a kogo bombardować (Harari 2018:120).

Dodaje przy tym, że takie problemy jak np. rozwój AI wymagają rozwiązań na poziomie globalnym, w czym żadna z religii nie może być pomocna, gdyż każda uważa tylko swoją perspektywę za słuszną. Tym samym zdolność religii do kształtowania tożsamości postrzega on jako część problemu, a nie coś, co przyczynia się do jego rozwiązania (Harari 2018:115-122). Niemniej i tak poza wspomnianym dokumentem nie widzimy dotychczas

272 (Mróz 2020):

1. Transparentność – rezultaty działania algorytmów SI [Sztucznej Inteligencji] powinny być w pełni możliwe do wyjaśnienia;
2. Włączenie – powinny być uwzględnione potrzeby wszystkich ludzi, by wszyscy mogli na równi korzystać z możliwości rozwoju;
3. Odpowiedzialność – ci, którzy projektują i budują SI, powinni postępować odpowiedzialnie i transparentnie;
4. Bezstronność – algorytmy SI nie powinny działać według jakichkolwiek uprzedzeń, gwarantując przez to sprawiedliwość i ochronę ludzkiej godności;
5. Niezawodność – systemy SI powinny umożliwiać niezawodne działanie;
6. Bezpieczeństwo i prywatność – systemy SI muszą działać w sposób bezpieczny i z poszanowaniem prywatności użytkowników.

prób budowania takich kształtujących tożsamość interpretacji ze strony religii – w sensie instytucji, bo media związane z konkretnymi wyznaniami, także w Polsce, publikowały już teksty na ten temat (Jóźwiak 2019; Merritt 2018). Wydaje nam się to zaskakujące, ponieważ sądzimy, iż zajęcie stanowiska o charakterze tożsamościowym, moralnym, wobec zastosowań tak rozbudzającej dyskurs publiczny technologii tzw. sztucznej inteligencji leży w interesie każdej religii instytucjonalnej. Zdecydowane stanowisko mogłoby przecież być elementem wpływu na politykę i prawo, jak instytucjonalne religie czynią w rozmaitych sprawach. Wymieniony dokument Papieskiej Akademii Życia jest próbą wejścia Kościoła rzymskokatolickiego na omawianą arenę.

Swoje stanowisko w sprawach etyki AI publikowały niemalże na całym świecie przeróżne podmioty państwowe, prywatne, organizacje pozarządowe, międzynarodowe konsorcja, uczelnie i inne. Dokumenty mają różny charakter – od zbioru zasad, jakie przygotowano np. z udziałem Elona Muska i Stephena Hawkinsa w Asilomar (Future of Life Institute 2017), przez dokumenty bardziej nastawione na prezentowanie wartości i wizji po tzw. strategie rozwoju AI w różnych krajach – przygotowywane są nawet zestawienia i krytyczne analizy dziesiątków takich dokumentów (Alekseichenko i in. 2018; Fjeld i in. 2020; Greene i in. 2019; Rothenberger, Fabian, i Arunov 2019; Taylor i Dencik 2020). Takie deklaratywne publikacje można by traktować jako obiekty graniczne dla mini-aren usytuowanych na obrzeżach opisywanej areny modelowania. Twierdzimy jednak, że dla świata DS mają one nikłe znaczenie. Zdecydowanie największy wpływ – choć i tak bardzo mały – mają dokumenty dotyczące uczestników świata DS właśnie nie jako uczestników świata, a jako pracowników określonych organizacji. Wówczas są oni zobowiązani formalnie, by co najmniej otwarcie nie występować przeciwko stanowisku pracodawcy. Niemniej i tak wielokrotnie spotykaliśmy się z interpretowaniem takich firmowych kodeksów czy wytycznych jako wyłącznie działań wizerunkowych, nierzadko robionych „na wyrost", czasami nieadekwatnych do pracy data scientistów:

> *B: A też słyszałem różne firmy, no nie wiem, Google nawet ma jakieś takie swoje kodeksy etyczne związane no z tak zwaną sztuczną inteligencją, już wiem że to jest śmieszne określenie, yy, ale czy wy macie coś takiego? Jakiś kodeks etyczny, że te modele to muszą być jakieś tam? Czy? No nie wiem*
> R: Śmiesznie że się pytasz o to [śmiech] bo mieliśmy taką sytuację że byliśmy w trakcie dopiero przygotowywania pierwszego case'a, pierwszego POCa. Dostaliśmy od szefa naszego szefa **wielgachny** dokument właśnie na temat etyki przy budowaniu takich rozwiązań z zakresu machine learning i sztucznej inteligencji. Także taki wielgachny dokument do nas przyszedł, i musieliśmy się z nim zapoznać i musimy się do niego stosować, mimo że [śmiech] na razie te rozwiązania trochę raczkują w sumie. To już taki gigantyczny kodeks już mamy, powstał.
> *B: A to kto napisał taki gigantyczny kodeks?*

R: To, to przyszło z [nazwa firmy, w której istnieje zespół DS], więc dokładnie nie wiem kto to napisał, nie znam autora. Być może to powstało na potrzeby tych innych grup które tam istnieją, a z którymi my nie mamy specjalnego kontaktu mimo że one coś tam więcej robią. [pauza] No ale wydaje mi się że tam na papierze istnieje dużo więcej niż w rzeczywistości [śmiech] no to też stąd taki papier powstał.
B: Aha, no dobra, czyli powstał u was, po prostu u was w firmie, tak? Gdzieś tam, gdzieś tam wcześniej? Mmm, i z tego co rozumiem to yyy to wyprzedza rzeczywistość, tak? To
R: Z mojej perspektywy, tak. Natomiast mówię być może istnieją tam wewnątrz firmy w [siedziba firmy poza Polską] zespoły, które robią to na jakimś bardziej zaawansowanym etapie, niż my w tej chwili. Nie wiem dokładnie. Nie dopytywałam się o to. {wywiad 18}

W naszych badaniach często wskazywano, że w przypadku osób pracujących w dużych firmach z jednej strony istnieje w organizacji ogólny kodeks postępowania / etyki, z drugiej pracują prawnicy, troszczący się o zgodność działalności z przepisami w odpowiednim kraju. Data scientiści budując i walidując modele zwracają uwagę na tzw. biasy (od *bias*, dosł. obciążenie, skrzywienie) traktując to jako problem techniczny wynikający ze specyfiki użytego zbioru danych, metod ML i konkretnego sposobu dostrojenia modelu. Powtarzają się opinie o kodeksach jako elemencie strategii wizerunkowej, szczególnie w przypadku gigantów technologicznych pracujących nad odbudową zaufania i dopasowaniu się do głównego nurtu politycznej poprawności po wpadkach ujawnionych w mediach:

B: Ok. [pauza] A słyszałem o takich rzeczach, yyy [krótka pauza] no przykładowo w Google'u tak, jest coś takiego jak kodeks etyki machine learningu, macie coś takiego? (...)
R: Takiego czegoś jak czysto kodeks machine learningu nie ma, są takie powiedzmy to zgaduje że Google zrobiło to po to żeby **wypchnąć** to jako że się tymi aspektami przejmują, jako że to jest jedna z tych firm, eee co ma powiedzmy, no mamy pewne, czasami problemy z naruszeniami eee ochrony danych osobowych. Więc oczywiście o to chodzi z tym, ale też o takie rzeczy właśnie jak, yee posiadanie zgody **użytkowników** na przetwarzanie ich danych i tak dalej, i właśnie danych z, z zachowaniem anonimowości i tak dalej.
B: Ale to zgaduję że macie od tego dział prawny?
R: Ee, tak. Znaczy jedno że dział prawny, plus drugie że są ogólnie takie ogólne nie tylko dostosowane do machine learningu kwestie właśnie związane z, eee właśnie z ochroną danych osobowych, po prostu ogólnie do **całości** działania, firmy.
B: Mhm, ogólne przepisy związane ze zgodami na wykorzystywanie danych, tam zgodności z RODO, tak, tego typu rzeczy to są?
R: No dokładnie. Na wszystko musi być zgoda użytkownika, na, zależy czego tam dane dotyczą, jeżeli jest wymagana.
B: A też, też słyszałem o taki rzeczach, mm, że firmy robią kodeksy, na przykład że nie może być jakichś tam obciążeń w modelach, typu no nie wiem kobiety, mężczyźni, czy to w waszej działalności ma jakieś znacznie?
R: Znaczy [krótka pauza] hmm w naszym przypadku najczęściej nawet nie wiemy czy ktoś jest kobietą czy mężczyzną. Więc, znaczy takie, w sensie powiedzmy biasy **mogą** wychodzić, w modelach, nawet jeśli nie ma na wejściu informacji że to jest kobieta albo mężczyzną. W sensie biasy że [pauza] jeżeli podstawimy losowe dane tylko z tą informacją, typu typowe dane, nie, jakoś wylosowane, to że gdzieś, coś powiedzmy trochę inaczej, jakieś różnice powiedzmy dla kobiet mogą wchodzić, ale na przykład jak nie mamy informacji, no to, albo nie możemy tego robić, to trzeba też odbiasować model, to nie jesteśmy w stanie. Na przykład zgaduję coś takiego, że z taką polityką wyszedł Amazon, pewnie [krótka pauza] po tym jak mu wyszło, że ten model od rekrutacji. {wywiad 19}

Twierdzimy, że w Polskim świecie DS nie tworzy się jak dotąd kodeksów DS czy modelowania, choć oczywiście za problem uznaje się możliwość dyskryminowania czy jakkolwiek pojętego krzywdzenia użytkowników końcowych, ponieważ kodeks to nie jest „coś działającego". **Praca nad kodeksem byłaby w świecie DS raczej uznana za bezwartościową, ponieważ takich deklaratywnych rezultatów nie uważa się za coś, co rozwiązuje rzeczywisty problem**. Jako rzeczywisty problem uznawane są bezapelacyjnie tzw. **biasy, traktowane jako problem techniczny**, a więc rozwiązywalny przede wszystkim **za pomocą kodu**. Nie-techniczne, deklaratywne, rozwiązania problemu „biasów" w modelach mają więc dla uczestników świata DS podobny status, jak przygotowywane przez biznes materiały marketingowe, sprzedażowe, wychwalające oferowane rozwiązania z użyciem pojęcia sztucznej inteligencji lub ewentualnie zachwalające pracę przy rozwijaniu tejże – to wszystko tylko opakowanie dla rozwiązań bazujących na ML, to tylko slajdy w PowerPointcie (por. 6.2.1.1).

W technicznej książce poświęconej XAI, a dokładnie analizom wyjaśniającym modele ML (*Explanatory Model Analysis*), P. Biecek i T. Burzykowski na wstępie zaproponowali trzy wymogi, które powinien spełniać każdy model predykcyjny (Biecek i Burzykowski 2020). Te jawnie wzorowane na trzech prawach robotyki Isaaca Asimova z wydanej w 1942 książki sci-fi „Zabawa w berka" wymogi to (ibidem):

- Walidacja predykcji. Dla każdej predykcji modelu należy być w stanie zweryfikować, jak silne są dowody wspierające tę predykcję.
- Uzasadnienie predykcji. Dla każdej predykcji modelu należy być w stanie zrozumieć, które zmienne wpływają na predykcję i w jakim stopniu.
- Spekulacja predykcji. Dla każdej predykcji modelu należy być w stanie zrozumieć, jak ta zmieniłaby się, gdyby zmieniły się wartości zmiennych ujętych w modelu.

Zdaniem autorów (ibidem; Figure 1.1: Shift in the relative importance and effort [symbolically represented by the shaded boxes] put in different phases of data-driven modelling) obecnie w ML standardem jest rozpoczynanie pracy od prostej, eksploracyjnej analizy danych, następnie przygotowania wielu rodzajów modeli, na koniec zaś wykonywana jest walidacja modeli za pomocą procedur technicznych, jak podział zbioru danych na treningowe i testowe oraz walidacja krzyżowa (*cross validation*), której wyniki służą iteracyjnemu poprawianiu modeli. Być może w nieodległej przyszłości zarówno eksploracja (autoEDA) jak i modelowanie (autoML) zostaną w wielu zadaniach technicznych

zautomatyzowane, zaś walidacja będzie odbywać się z użyciem metod XAI oraz zasad etyki (*ethics*) i sprawiedliwości (*fairness*). Autorzy wskazują na wiele przykładów negatywnych konsekwencji społecznych wdrażania modeli będących tzw. czarnymi skrzynkami, które opisywaliśmy w części 2.2.3. tej rozprawy. Podkreślmy jednak, że w pracach takich, jak ta (Biecek i Burzykowski 2020) wymogi wobec modeli predykcyjnych nie są traktowane jako rozwiązanie deklaratywne, a jako wyjaśnienie założeń stojących za proponowanymi rozwiązaniami technicznymi. Są to wymagania wobec modeli, więc nie mogą być traktowane jako kodeks postępowania dla osób zajmujących się DS i nie mówią nic ani o domenie, w której budowany jest model, ani o potencjalnych konsekwencjach jego wdrożenia.

Być może należy spodziewać się powstania w Polsce „środowiskowego" kodeksu postępowania w DS, który stanie się kolejnym obiektem granicznym mini-areny w opisywanej arenie. Będzie to mini-arena raczej „wewnątrz" badanego świata DS. Wielokrotnie deklarowano nam bowiem indywidualną wrażliwość na kwestie etycznych konsekwencji rezultatów pracy data scientistów, na prywatne lektury (jak choćby *Broń matematycznej zagłady* O'Neil czy popularnonaukowe prace Hararigo), na niechęć do pracy w obszarach, które Rozmówczynie i Rozmówcy postrzegali jako bardziej niż inne wątpliwe etycznie. Niemniej w naszych badaniach nie znajdujemy sygnałów o kształtowaniu się etyki DS o charakterze zbiorowym w badanym świecie. Etyka, w sensie oceny rezultatów działalności pracy, jest dość jednoznacznie postrzegana jako indywidualna sprawa każdego z uczestników i w żadnym razie nie podlegają oni w tym zakresie zbiorowym zobowiązaniom wobec społecznego świata. Zupełnie inaczej niż w przypadku wartości tego świata czy zaświadczania o autentyczności. Wydaje nam się, że niektórzy uczestnicy mogą uważać za coś naturalnego stan, w którym „jeszcze" nie powstał w Polsce taki kodeks, w myśl popularnej legitymizacji badanego świata, czyli powołaniu się na tzw. wczesne stadium rozwoju DS. Wskazywano nam, że na razie kodeksy powstają w firmach na zasadzie reakcyjnej. Sądzimy, że wolno traktować to jako wyraz nastawienia biznesu, tak jak świata DS, głównie na efektywność – po co robić kodeks, o który nikt nie prosił? Po co naprawiać coś, co nie jest zepsute?

> *B: Chciałem jeszcze cię zapytać o to od czego właściwie zaczęliśmy, czyli od tej etyki związanej właśnie z algorytmami z automatyzacjami. Jak to jest z tą etyką, ale w twojej pracy? Czy masz, nie wiem, jakiś kodeks etyczny albo nie wiem co?*
> R: Wiesz coo, nie mamy.
> *B: Czy masz jakieś swoje, swoje zasady?*
> R: Oczywiście, w pracy myślę, że każdy z nas ma jakiś swój kodeks etyczny.
> *B: No dobra. Aaa ale spisanego jakiegoś kodeksu, firma ma jakiś kodeks?*

R: Myślę, że każda firma ma jakiś kodeks. Czasami on powstaje, zwłaszcza przy takich startupach, to to często powstaje z potrzeby. Podam ci realny przykład, który występuje w wielu firmach IT. Firmy IT bardzo często na poziomie startupu mają kompletną dowolność dresscode`u i śmieją się, jak to korporacje mają dresscode. I wszystko jest fajnie, do momentu, kiedy nie zaczynasz się skalować i twoja firma nagle nie zajmuje się dużymi klientami, a twoi programiści, którzy są widoczni, przychodzą ci do pracy normalnie w majtkach i to jest realny case, przyszedł chłopak w slipkach, no w majtkach. No to w tym momencie wprowadzasz dresscode **ubierajcie się**. Nadal to nie jest jakiś ostry dresscode, to nie jest garsonka i szpilki, ale trochę kultury. Ja uważam, że w skali zwłaszcza, procesy **są** potrzebne. Jeśli chodzi o etykę, to też często są takie rzeczy typowo, raczej musisz być miły dla innych, respektować ich. Zwłaszcza teraz jak pracujemy w tak międzynarodowym środowisku, no to jednak dyskryminacja, akurat nie spotkałam się mówiąc szczerze, tyle pracuję w międzynarodowym środowisku i nie spotkałam się z dyskryminacją związaną z narodowością nigdy. Więc jest taki podstawowy kodeks etyczny, ale ja uważam, że cały czas brakuje i też o tym debatowałam z moim znajomym, że trochę brakuje tego kodeksu etycznego, jeśli chodzi o tworzenie rozwiązań data science. No bo to nie jest tak, że najpierw siadasz i uwzględniasz to od razu, raczej jest tak, że to wiesz, gdzieś tam może się o tym wspomni i na to powinni patrzeć i dostawcy, ale też przede wszystkim myślę też te osoby biznesowe, dla których się buduje te rozwiązania, co będzie etyczne, co będzie nieetyczne. No oczywiście w firmach jest to troszeczkę inaczej niż na przykład tak jak w Stanach są rozwiązania do osądzania ludzi, bazujące na sztucznej inteligencji, które wiemy już, że miały baiasy na przykład co do ludzi czarnoskórych, związane z historycznymi danymi. Więc no ta etyka jest szalenie ważna, ale jak mam ci szczerze powiedzieć, to dla mnie nadal to jest taki temat, o którym ja bardziej dyskutuję niż wiem i wydaje mi się, że my nadal jesteśmy na tym etapie takiej dyskusji, jak to zrobić dobrze, no i dobrze, że jesteśmy na takim etapie, zacznijmy to wdrażać, żebyśmy w pewnym momencie się nie obudzili, czy te wszystkie rzeczy są etyczne, czy jednak powinniśmy się cofnąć, a jak wiemy, cofa się strasznie ciężko.
B: Ale to uważasz, że to jest rzeczywiście ważne, w takich typowo biznesowych zastosowaniach, gdzieś tam w produkcji czy coś takiego? No bo booo ja rozumiem, że w sądownictwie, tak, no to jest superważne, ale tak yy głęboko w biznesie?
R: Jak weźmiesz sobie największe korporacje na świecie [pauza] i ile one mają danych o tobie, one wiedzą **idealnie**. Przynajmniej one wiedzą z kim ty mieszkasz, co kupujesz, co robiła dzisiaj twoja partnerka, żona dziewczyna. Czego wy razem szukaliście, gdzie jedziecie, no potrafi rekomendować, no w tym też musi być pewna etyka, tak? Więc no tak, jak najbardziej. Wiesz, te duże korporacje mają **ogromny** wpływ na nasze życie, więc jak najbardziej.
{wywiad 22}

Zwracamy uwagę na fakt, iż postrzeganie firmowych kodeksów przez uczestników świata DS jako działań wizerunkowych nie musi oznaczać, że kodeksy takie są negatywnie ocenianie. Podkreślmy jednak, że Rozmówczyni zaznacza swoją wrażliwość etyczną – sama „czuję gdzieś te zasady, (...) żeby ten produkt który tworzymy, eee nie odzwierciedlał nie wiem, czyichś przekonań" {wywiad 20}. Zatem z jednej strony uczestnicy mogą nie uznawać kodeksów za sprawcze, z drugiej jednak za względnie potrzebne, ale też niekoniecznie, bowiem wystarcza im indywidualne poczucie etyki:

B: Mhm, ok, rozumiem. Mmm a chciałbym jeszcze spytać, bo słyszałem o tym że, zaczynają firmy tworzyć takie, kodeksy etyczne, kodeksy postępowania dla [pauza] no nie wiem jak to powiedzieć, czy dla data scientistów, czy to są raczej takie kodeksy, że AI ma być tam jakieś, że ma nie dyskryminować, czy słyszałaś o czymś takim i czy wy macie coś takiego?
R: E, słyszałam że takie rzeczy właśnie występują, że są podejmowane jakieś kroki w dużych

> korporacjach typu Amazon, typu Facebook, żeby AI, nie miało jakiegoś biasu, ale nie wydaje mi się żeby coś takiego istniało, w naszej firmie. Mamy jakiś ogólny, kodeks etyczny, ale nie kojarzę żebyśmy mieli coś specyficznie odnośnie AI.
> *B: Ok. Czyli to jest taki kodeks w ogóle dla całej korporacji?*
> R: Tak. To jest, to jest, to jest jakiś kodeks postępowania y jak, yy, jak, yyy nie wiem, trudno mi powiedzieć, straciłam wątek. Ale tak, jest jest, nie jest kodeks bezpośrednio skierowany pod AI, które będziemy robić.
> *B: Rozumiem, a czy taki kodeks w ogóle do czegoś by się tobie przydał?*
> R: Mmm [krótka pauza] myślę że nie, powiedzmy ja **czuję** gdzieś te zasady, wiem że, że są jakieś delikatne sprawy, na które trzeba zwrócić uwagę [krótka pauza] że, żeby nie wiem, żeby ten produkt który tworzymy, eee nie odzwierciedlał nie wiem, czyichś przekonań, czy coś takiego. Myślę że może przydałby się innym osobom które [krótka pauza] nie czują czegoś takiego, nie czują takiej potrzeby. Eee mnie osobiście nie, ale uważam że to jest jakaś dobra praktyka, którą firmy powinny zacząć stosować.
> *B: Mhm, mhm, ok.*
> R: Boo później kiedy się okazuje że ktoś czegoś takiego nie stosował i to wypływa, eee do jakiejś większej publiczność, to to jest wielki wstyd dla firmy i i i jakaś czarna plama, na ich, ich opinii.
> *B: Mhm, mhm. Tak, tak wizerunkowo to bardzo źle działa potem, tak?*
> R: Tak. Moim zdaniem tak. Się jakoś, yyy niezręcznie się później, później o tej firmie mówić. Tak jak właśnie był przypadek, w Amazonie, żee AI dyskryminowało kobiety przy wyborze CV i jakoś, to bardzo zmieniło ogółem opinie o tej firmie, znaczy nie miałam nigdy dobrej opinii o Amazonie [z uśmiechem] ale pogorszyła się. {wywiad 20}

Wyżej pisaliśmy o dokumentach mających stanowić strategie rozwoju AI w konkretnych krajach czy obszarach gospodarczych. Tworzenie takich dokumentów jest dość popularnym w świecie polityki sposobem uczestniczenia w opisywanej arenie modelowania. Polskie prace w tym zakresie (Borowik i in. 2018; Koloch i in. 2017; Kroplewski i in. 2019; Marczuk i in. 2019; Ministerstwo Cyfryzacji RP 2018b) są zdecydowanie entuzjastyczne. W przygotowywaniu publikacji brały udział osoby reprezentujące wspominane już organizacje pozarządowe (jak Fundacja digitalpoland i Klub Jagielloński) czy uczelnie wyższe.

Jak wspomnieliśmy w części 5.1. rozprawy, Ministerstwo Cyfryzacji RP twierdzi, iż w UE brakuje 400 tys. pracowników na „rynku danych" i luka ta będzie się powiększać (Borowik i in. 2018). Wskazano, że „niezbędne jest, abyśmy w Polsce rozwijali dziedziny związane ze zbieraniem, analizowaniem oraz przetwarzaniem danych" zaznaczając, że

> (..) spośród dostępnych dla nas możliwości rozwojowych najkorzystniejsze jest inwestowanie w gospodarkę opartą o dane. Polska gospodarka może odnieść ponadprzeciętne korzyści dzięki zwiększonej produktywności i nowym modelom biznesowym (Borowik i in. 2018).

Program gospodarki opartej o dane nazwano Przemysł+ (ibidem). W innym dokumencie Ministerstwo Cyfryzacji posługuje się głównie terminem AI, niemniej całość strategii bazuje na twierdzeniu, że **gospodarka oparta na danych stanowi kluczowy element rozwoju Polski i będzie „główną siłą napędową wzrostu gospodarczego i**

tworzenia miejsc pracy, nie tylko w obszarze zaawansowanych technologii, lecz całej gospodarki” (Ministerstwo Cyfryzacji RP 2018b:7). Odniesiono się do szeregu międzynarodowych strategii rozwoju AI[273] – są to dokumenty Unii Europejskiej, Organizacji Narodów Zjednoczonych i grup G-7 oraz G-20 (Ministerstwo Cyfryzacji RP 2018:104-109). Współtwórczyni omawianego dokumentu wskazuje w popularnonaukowym cyklu „Uniwersytet Łódzki komentuje”, że stosowanie w biznesie rozwiązań bazujących na danych jest jednoznacznym źródłem przewagi konkurencyjnej:

> Sukces osiągną bowiem te spośród nich [firm i marek], które będą w stanie reagować, być może w czasie rzeczywistym, umiejętnie łącząc dane, treści, kanały i technologie (Kaczorowska-Spychalska 2019).

Ministerstwo Cyfryzacji stwierdziło także, że AI zarówno zlikwiduje, jak i utworzy miejsca pracy. Aby Polska znalazła się po tej stronie równania, po której powstaną nowe miejsca pracy, Ministerstwo Cyfryzacji postuluje istnienie ponad 700 firm budujących AI w Polsce do roku 2025. Stwierdzono, że „AI przetransferuje bogactwo do tych krajów, które będą potrafiły ją budować i kontrolować” (Ministerstwo Cyfryzacji RP 2018:22–23). Powodów, dla których Polska musi budować własne AI jest, jak napisano, wiele, ale za najważniejsze uznano dwa:

> [1] Dane to zasób. Na razie potrafimy go zdobyć ale jeszcze słabo potrafimy go analizować. Jeżeli nie będziemy mieli własnych rozwiązań AI, staniemy się krajem „surowcowym”. Będziemy dostarczać dane, ale będziemy zależni od tych, którzy potrafią je zmienić w wysokoprzetworzony i dochodowy produkt. „Scenariusz Wenezueli” [2] Spójrzmy prawdzie w oczy: sztuczna inteligencja będzie eliminowała kolejne zawody. Które, jak szybko i w jakim stopniu, to oddzielna dyskusja. To co ważne, to na każde zlikwidowane 1000 miejsc pracy, sztuczna inteligencja wygeneruje około 1280 nowych (Gartner - „AI and the future of work” Grudzień 2017). Jeżeli nie będziemy budować naszego AI, to praca zniknie w Polsce, ale pojawi się w innym kraju. Tak jak rewolucja IT zabrała pracę w wielu krajach, ale wykreowała ją w Indiach czy… w Polsce. (Ministerstwo Cyfryzacji RP 2018b:22)

Owe 720 firm ma odzwierciedlać potencjał intelektualny i biznesowy Polski w porównaniu do innych krajów budujących AI, przyjęto bowiem, że w 2025 nasz kraj ma znajdować się między 20 a 25 percentylem najlepszych w tej dziedzinie krajów na świecie. Podkreślono, że chodzi o 720 firm „budujących AI, a nie tylko wdrażających”, wyjaśnienia tego nie znajdujemy w dokumencie. Oceniono liczbę owych budujących AI firm na 30 podmiotów w roku 2019[274] i mają one generować łączne przychody w kwocie około 0,35 mld

273 Przegląd takich strategii na świecie przygotowała Fundacja digitalpoland (Alekseichenko i in. 2018)

274 Przypomnijmy, że digitalpoland ocenia ich ilość wielokrotnie wyżej, pisząc o 264 ankietowanych firmach AI (Borowiecki i Mieczkowski 2019).

PLN. W roku 2025 przychód tej branży ma wynosić 8,3 mld PLN, zatem 24 razy więcej (Ministerstwo Cyfryzacji RP 2018b:22–23). Zachowano zatem zbliżony przychód na firmę[275].

> W pierwszej kolejności Pan Robert Kroplewski [Pełnomocnik Ministra Cyfryzacji do spraw społeczeństwa informacyjnego[276]] rozszerzył informacje na temat nowych inicjatyw międzynarodowych oraz aktualnego stanu prac nad tzw. europejską strategią sztucznej inteligencji. Podsumował swoje doświadczenia z udziału w różnych międzynarodowych grupach eksperckich – wskazał także na skuteczność działań ze strony ministerstw, przykładowo powołanie do grona wysokiego szczebla ekspertów AI przy OECD (tzw. AIGO). W jego opinii prace nad wspólnymi dokumentami strategicznymi UE są w różnych stopniach zaawansowania, z perspektywą ukończenia ich w przyszłym roku, a wiele zagadnień jest jeszcze w fazie dyskusji ekspertów. Dlatego polskie środowisko ma ciągle szansę na wypracowanie własnych propozycji oraz przedstawienie ich także w takich grupach eksperckich. Podkreślił, że Rząd RP interesuje się coraz bardziej budową strategii rozwoju Sztucznej Inteligencji w Polsce. W tej chwili, tematyką sztucznej inteligencji zajmują się głównie trzy ministerstwa, tj.: [1] Ministerstwo Przedsiębiorczości i Technologii, w kontekście innowacji gospodarki, [2] Ministerstwo Nauki i Szkolnictwa Wyższego, w zakresie głównie związanym z finansowaniem prac badawczo-rozwojowych (zwłaszcza NCBiR) oraz kształceniem kadr, [3] Ministerstwo Cyfryzacji, w zakresie możliwych zastosowań przez administrację publiczną oraz w kontaktach z Komisją Europejską (Czarnowski i in. 2018:7).

W badanym świecie DS na krajowe strategie rozwoju AI i podobne deklaracje zwracają uwagę raczej tylko osoby silnie osadzone w badanym świecie, nierzadko (a czasem jednocześnie) pełniące funkcje kierowników zespołów DS lub właścicieli firm oraz posiadające stanowiska profesorskie na uczelniach państwowych. Nie jesteśmy w stanie wypowiedzieć się o trendach, co do oceny działania Ministerstw RP w omawianym zakresie – raz tylko mieliśmy okazję porozmawiać z osobą wyrażającą krytyczne opinie. Zdaniem Rozmówcy panuje chaos i wzajemne zwalczanie się frakcji w partii rządzącej, przy jednoczesnej orientacji na samorozwiązanie się problemów gospodarczych, najpewniej dzięki słynnej, niewidzialnej ręce rynku. Rozmówca upatruje roli państwa głównie w finansowaniu kształcenia ludzi w zakresie ML na uczelniach wyższych:

> *B: A jak to wygląda z pana perspektywy w, w yyy Polsce? Ja będę mówił branża data science, bo tak się przyzwyczaiłem. Nie wiem, czy to jest dobre określenie, czy nie za bardzo, ale czy w ogóle mamy taką branżę w Polsce? Jak to się kształtuje pana zdaniem?*
> R: Moim zdaniem nie jest źle, o czym może świadczyć fakt, że próba znalezienia kogokolwiek kompetentnego na rynku graniczy z niemożliwością, rynek jest całko całkowicie wysuszony. Zdecydowanie budzi się świadomość w firmach, ona jeszcze cały czas chyba jest trochę sterowana takim hypem, ale generalnie jak tylko na jakiejś konferencji się pojawi hasło AI, to ludzie natychmiast biegną. Na pewno co przeszkadza, jakieś jakieś takie ruchy ze strony rządowej i to jest bardziej szamotanina, żeby budować jakąś strategię cyfryzacji.
> *B: Właśnie, bo Ministerstwo coś tam opublikowało, prawda?*
> R: Nie wiem, czy pan będzie chciał to wykorzystać, czy nie, ale prawda jest taka, że niestety są

275 Dla 30 firm zysk 0,35 mld PLN to około 11,67 mln PLN na firmę. Dla 720 firm przy zysku 8,3 mld PLN to odpowiednio 11,53 mln PLN.

276 https://www.gov.pl/web/cyfryzacja/robert-kroplewski1, dostęp 06.08.2020 r.

> dwa Ministerstwa, właściwie trzy, które próbują grać w tą grę, to jest Ministerstwo Cyfryzacji, Ministerstwo Gospodarki i Ministerstwo Nauki i Szkolnictwa Wyższego, i one są obsadzone przez szczerze nienawidzące się frakcje PiS-owskie. Oni oni wewnątrz nienawidzą się bardziej, niż wszystkich pozostałych dookoła, więc oni nawzajem sobie torpedują wszystkie pomysły, jakie mają. Jak Ministerstwo Cyfryzacji z czymś wyjdzie, to natychmiast Gowina ich ubije, jak Gowin wymyśli coś, to natychmiast Ministerstwo Cyfryzacji, a Ministerstwo Gospodarki jeszcze się ucieszy. Miała być jakaś jedna wspólna szkoła doktorska w całej Polsce z AI. Pojawiły się dwa razy takie hasła, zapadła cisza. Miały być kierunki zamawiane z AI, znowu cisza. Pomijając ogólnie burdel, jaki ta ekipa organizuje, czy to w ramach tej nieszczęsnej reformy czy w ogóle w swoich działaniach, to nie widać tutaj **żadnej** spójności, żadnej myśli przewodniej. To jest takie na zasadzie, wrzućmy trochę pieniędzy może coś z tego powstanie. A niestety nie da się tego rozwiązać tylko siłami przedsiębiorców, bo niestety oni potrzebują tego początkowego motorka, w postaci ludzi, którzy będą to robili. A ich jest najzwyczajniej w świecie za mało. I tak kształcimy za mało informatyków i informatyczek, tylko część z nich jest kształcona po linii jakiegoś takiego ML-a, nawet bardzo niewielka cześć, Więc nie ma tego zasobu ludzkiego, który byłby w stanie być tym katalizatorem zmiany. Ja miałem doświadczenia z czterema dużymi firmami, które próbowały i próbują nadal rozpędzać ten swój wózek ML-owy, i muszę powiedzieć, że nie wygląda to jakoś dobrze. Bo to bardziej jak działanie po grudzie. Wszyscy rozumieją, że to jest konieczność, że to jest przyszłość, że my musimy i tak dalej i tak dalej, aaaa ale jak przychodzi co do czego, no to jest trudne. To jest takie poruszanie się po omacku, to jest **duża**, droga inwestycja. {wywiad 23}

Kończąc tę część rozważań o arenie modelowania (rys. 45.) przyjrzyjmy się dalszej części nie przerywanej pytaniami badacza wypowiedzi Rozmówcy, w której ponownie podkreśla on rolę szkolnictwa wyższego oraz wskazuje na makroregionalną specyfikę badań i zastosowań modeli ML. Europę charakteryzują działania w obszarze XAI oraz badań futurystycznych, USA to zastosowania komercyjne wzorowane na sukcesach technologicznych gigantów, zaś Chiny to zastosowania w robotyce i automatyce, nastawione na zwycięstwo w wyścigu o fabryki bez robotników ludzkich:

> R: (...) Pokazywanie nam przykłady ze świata, że tu się nam udało, bo mamy takie firmy, które są czysto AI, na przykład Uber, zresztą to nie jest firma taksówkowa, to jest firma oparta na ML-u. w Polsce już takich firm też zaczyna być trochę, nie wiem, jakiś tam Głodny.pl na przykład. Tak na dobrą sprawę to działa tylko i wyłącznie, czy powinno funkcjonować tylko na zasadzie ML, bo ta firma nie produkuje żadnej innej wartości tak naprawdę, a ma szansę być ogromną rzeczą. Patrzę sobie na Chińczyków, w zeszłym roku trzy największe inwestycje AI były z Chin, pierwsza trójka. Alibaba nie jest jedyną rzeczą, którą oni tam robią. Największa amerykańska inwestycja była na 4 miejscu. Co jak gdyby jest zrozumiałe, bo Chińczycy doskonale wiedzą, że jeżeli ten ML czy ten AI rzeczywiście wystartuje, to wystartuje nie w postaci aplikacji internetowych, tylko to będą fabryki, które będą same produkowały. A w momencie, kiedy fabryki będą same produkować, to nie będzie żadnego powodu, żeby te fabryki, żebyśmy my dzisiaj, w sensie Zachód, produkowali to w Chinach. Będzie można za darmo przenieść z powrotem całą produkcję do Europy i do Stanów Zjednoczonych. Chińczycy nie są głupi, oni to doskonale rozumieją, że dla nich to jest być, albo nie być. Więc albo będą mieli pierwsi, za darmo, całkowicie zautomatyzowane fabryki, albo wypadną z gry, więc inwestują jak szaleni, ale inwestują nie tylko w start-upy, programy rządowe itd., ale potężnie inwestują w ludzi. Na przykład jest teraz program wsparcia dla chińskich doktorantów, którzy wracają z amerykańskich uczelni, to dotyczy najlepszych uczelni amerykańskich, i ktoś, kto zrobi doktorat w obszarze, gdzieś w okolicach AI, ludzie robią doktoraty młodo, to Chińczycy nie dość, że oferują bardzo dobrze płatną posadę z powrotem w Chinach, to jeszcze poza

> samymi pieniędzmi oferują od razu stanowisko takie jawnie kierownicze. Co jak gdyby zmienia perspektywę patrzenia młodego człowieka, bo ten młody człowiek mówi sobie tak, skończyłem na przykład na Uniwersytecie Berkeley, zrobiłem doktorat i teraz chcą mnie na przykład w Facebooku. I pójdę, i będę junior ML researcher w Facebooku, i będzie fajnie, bo będę miał świeże owoce i kawę i będzie jak w Google'u, będę przez 10 lat siedział i będę miał tego samego szefa, i będę robił cały czas te same rzeczy na Facebooka, i będę zarabiał na tyle, że będzie mnie stać na dwupokojowe mieszkanie w San Francisco, a z drugiej strony, jak wrócę, to będę żył jak król. Po prostu nie dość, że mam inne perspektywy na życie, startuję z innego punktu, to nawet jak ta pensja w sposób bezwzględny będzie mniejsza, no to przecież 100 tys. dolarów rocznie w Chinach, to jest tyle, co milion dolarów w Stanach. I ci ludzie tłumnie wracają, po prostu ssanie z powrotem do Chin jest gigantyczne. My jak na razie jedyne, co robimy to eksportujemy po 300 tys. zł w głowie każdej osoby, która wyjeżdża z Polski i po ukończeniu darmowych studiów tutaj, wyjeżdża i pracuje za granicą. [pauza] Więc, mmmm wracając do początku pytania, wydaje mi się, że jest zdecydowany ruch nawet na takiej zasadzie, że **nawet** jakieś czynniki rządowe **kojarzą** te hasła, wiedzą, co to jest AI i ML, jest jakaś potrzeba, to też jest ewidentnie pokazywane w w różnego rodzaju projektach europejskich, eeee. Ciekawe jest też to, jak bardzo różnią się domeny zastosowań. Jak bardzo widać, jak się różnicują, dywersyfikują się badania prowadzone w Stanach, w Europie i w Chinach, gdzie w Chinach to na prawdę to jest ML plus robotyka, jako jako absolutnie nr 1., Europa z kolei **suuuper mocno** stawia nacisk na po pierwsze taki general AI, czyli takie bardziej futurystyczne badania, ale też bardzo silnie jest na temat human AI, albo explainable AI, czyli ten aspekt społeczny, ee decyzji algorytmicznych. Natomiast Amerykanie, to jest generalnie [niezrozumiałe, śmiech] kolejny Facebook, kolejny Twitter, kolejny Instagram. Polskie uczelnie zaczynają eeee, czuć wiatr w żagle, większość uczelni otwiera kierunki. Czasami to są specjalizacje, czasami to są dodatkowe studia podyplomowe, ee my na przykład teraz powzięliśmy dosyć duży wysiłek i w październiku otwieramy całkowicie nowy kierunek studiów. Równoległy do informatyki, który się będzie nazywał AI. Oczywiście nazwa jest taka bardziej marketingowa, natomiast będziemy uczyli takich bardzo solidnych podstaw ze statystyki, matematyki, programowania i uczenia maszynowego plus trochę tej sztucznej inteligencji, na zasadzie takiej, żeby jednak wyjść poza, mieć jakieś elementy wnioskowania logicznego, jakieś elementy wyjaśnialności, plus zastosowanie, ale też chcemy zaprosić ludzi z automatyki, robotyki, elektroniki, żeby też trochę pokazywali, jak ML może wejść do świata Internetu Things, i w jaki sposób można tworzyć sobie takie autonomiczne, jak wyposażać takie autonomiczne zbiory agentów w jakieś tam formy inteligencji. (…) {wywiad 23}

Na koniec opisu „z lotu ptaka" areny modelowania (rys. 45) zaznaczmy, że naszym zdaniem prawie nieobecny w dyskursie jest użytkownik końcowy i środowisko naturalne. Obaj wymienieni aktorzy są na opisywanej arenie traktowani przede wszystkim przedmiotowo, każdy raczej jako zasób niż podmiot mający interesy i wpływy na arenie. Choć w dyskursie świata DS i wśród jego biznesowych rzeczników mówi się o danych jako zasobach, to dane nie biorą się znikąd. Często są zapisem działań ludzkich, działań osób płacących swoimi danymi w tym specyficznym barterze w zamian za bezpłatne usługi, np. Facebookowi czy firmie Google. Słabo obecnymi w dyskursie aktorami, sprowadzonymi do roli zasobów są także nisko opłacani „klikacze" np. zatrudniający się przez Mechanical Turk. Ich niewykwalifikowana praca polega przede wszystkim na wzbogacaniu danych, tzw. etykietowaniu danych, które mają zasilać modele nadzorowanego uczenia maszynowego. W badanym świecie DS powstają jednak rozwiązania techniczne i metodologiczne pomagające

zmniejszyć ilość ręcznego etykietowania danych (Ratner i in. 2017), niemniej nie można tego uznać za wzmacnianie pozycji „klikaczy" na opisywanej arenie. Wskazywano w tym kontekście także na postkolonialny charakter obszaru tzw. sztucznej inteligencji – w naszym ujęciu właściwie tożsamej z areną modelowania – ponieważ niewidoczna praca (*ghost work*) klikaczy wykonywana jest często przez mieszkańców byłych kolonii anglojęzycznych, w warunkach ograniczonego prawa pracy, a w niektórych przypadkach przez osoby odbywające karę pozbawienia wolności lub osoby ubogie. Z jednej strony ma to wspierać rozwój ekonomiczny oraz elastyczność zatrudnienia w krajach postkolonialnych, z drugiej utrwala model dystrybucji wiedzy i pracy – pracy niskopłatnej, nie wykwalifikowanej i przy ograniczonych prawach pracownika (Mohamed i in. 2020:10). Wielu słabo obecnych w dyskursie aktorów składa się na pełen łańcuch dostaw systemów bazujących na modelach ML – od wykorzystujących duże ilości zasobów środowiska naturalnego „chmur" obliczeniowych do nisko płatnej pracy górników w skrajnych warunkach, wydobywających metale na komponenty infrastruktury komputerowej (Portmess i Tower 2015; Joler i Crawford 2018).

Czasami użytkownik końcowy jest równoznaczny z klientem (użytkownik jako ktoś, kto korzysta z systemu bazującego na uczeniu maszynowym, np. handlowiec, który przed spotkaniem uruchamia aplikację, gdzie widzi historię kontaktu z kontrahentem, oraz maszynowo wygenerowaną klasyfikację tego kontrahenta oraz także maszynowo wygenerowane trzy potencjalne produkty do zaproponowania mu). W takim przypadku faktycznie jest klientem, czyli płaci firma zatrudniająca handlowców. Niemniej z naszych badań wynika, iż w badanym świecie DS za dobrą praktykę uznano by konsultacje z handlowcami już na etapie tworzenia systemu i zbieranie od nich informacji zwrotnej także po wdrożeniu systemu. Zatem użytkownicy są to jak najbardziej obecni i mają sprawczość. Z drugiej strony czasami za użytkownika należy uznać osobę na pozycji kontrahenta z powyższego przykładu, tzn. kogoś, kto nie tyle korzysta, a „z niego" korzysta i „na nim" działa system oparty na modelach ML – z historycznych danych takich kontrahentów modele dają odpowiedzi dla konkretnego podmiotu. Taki model użytkownika występuje powszechnie i w dużej skali w przypadku technologicznych gigantów, jak np. Facebook, Google, LinkedIn (należący do Microsoft), Spotify, Netflix, gdzie modele odpowiadają za tzw. personalizowane treści rekomendowane konkretnym użytkownikom końcowym. To personalizowanie to owoc liberalnego kultu wolnej, racjonalnej jednostki – zatem trzeba komunikować produkty, usługi i generalnie marketing zwraca się do jednostki, bo wyborca wie lepiej, bo klient ma zawsze rację, bo ludzie sami o sobie decydują (por. Harari 2018:281). Takie założenia o jednostce nie

oznaczają jednak, aby głos tego rodzaju użytkowników końcowych był znaczący w opisywanej arenie.

Niemniej być może w konsekwencji obecnych w głównym nurcie dyskursu publicznego głosów podnoszących kwestie np. prywatności użytkowników końcowych, w dużej mierze od 2018 r. – po słynnej „aferze" Cambridge Analytica oraz w związku z wprowadzeniem GDPR w Unii Europejskiej – użytkownicy końcowi poprzez rzeczników takich jak organizacje pozarządowe, media, polityka, a w konsekwencji prawo będą zdobywać nieco większą niż obecnie sprawczość na dalece nierównej w tym względnie arenie modelowania. Sądzimy, że ideologia danetyzacji przestaje być przyjmowana bezkrytycznie przez wszystkie zainteresowane strony (por. Dijck van 2014:200–201). Wołanie użytkowników o prywatność dostrzegły już firmy i włączyły zapewnienia dotyczące dbałości o ich prywatność do swoich komunikatów marketingowych. Od 2019 czyni to nie tylko Apple, od dawna znane z podobnych zapewnień co do oferowanych przez firmę urządzeń; czynią nie tylko firmy, dla których prywatność jest podstawą oferowanych usług (np. DuckDuckGo, Protonmail, Mozilla, Brave); czynią to także Facebook i Google (Lerman i O'Brien 2019; Robertson 2019). Problematyka pojawiła się już w mainstreamie popkultury – platforma Netflix emituje serial „Społeczny dylemat (*The Social Dilemma*)" w reżyserii Jeffa Orlowskiego, gdzie pojawiają się wielokrotnie przez nas cytowane spostrzeżenia w rodzaju „jeśli nie płacisz za produkt, to ty jesteś produktem" (Obem 2020).

Oczywiście użytkownicy końcowi mogą domagać się czegoś dalece innego, niż swojej prywatności, a prywatność nie jest jedynym obszarem, w którym mamy do czynienia z konsekwencjami społecznymi wdrożeń modeli ML (a prywatność związana jest raczej z pozyskiwaniem i użyciem danych, a nie z modelowaniem) czy w ogóle działalności świata DS. Pamiętajmy także, że użytkownicy nie jest to jednorodna, nie-techniczna grupa. Użytkownikami końcowymi systemów bazujących na modelach ML są nie jakieś szare masy, o których – mamy nadzieję, że nie – wypowiadamy się tu z wyższością i troską, ale to tak samo socjologowie, politycy, data scientiści, prawnicy i przedstawiciele rozmaitych światów społecznych, organizacji i innych społecznych całości.

W naszym ujęciu, być może niesłusznie, nieobecna jest także popkultura. Chyba wolno byłoby traktować **popkulturę jako tło dla areny modelowania**. Właśnie z popkultury, z science fiction, jak np. współczesnego serialu brytyjskiego „Czarne lustro" (*Black Mirror*), czy nieco starszych filmów np. „Terminator", „Raport mniejszości" wyobrażenia o tzw. sztucznej inteligencji nierzadko czerpie nie tylko nie-techniczny użytkownik końcowy, ale w

różnym zakresie i stopniu czynią to uczestnicy każdego z wyróżnionych na arenie modelowania światów, łącznie z DS i IT.

Oczywiście **pozycją nieobecną w dyskursie** opisywanej areny modelowania jest to, że **co do zasady nie należy lub nie warto w ogóle używać modeli ML**. Taka pozycja nie reprezentowałaby interesów żadnego ze światów / subświatów zaangażowanych w arenę.

W kolejnych dwóch częściach (6.2.1.1. i 6.2.1.2.) przyglądamy się bardziej szczegółowy przecięciu światów DS i biznesu, szczególnie w kontekście działań podejmowanych pomiędzy data scientistami i ich rzecznikami, przedstawicielami firm-usługodawców (tzw. klientami wewnętrznymi) i firm-klientów, oraz osobami chcącymi wejść do świata.

6.2.1.1. Sztuczna inteligencja to slajd w PowerPointcie

Z punktu widzenia uczestników świata DS, do dyskursu o modelowaniu, właściwego dla ich społecznego świata, należy pojęcie uczenia maszynowego (ML), oczywiście obok innych terminów technicznych dotyczących danych, kodu i różnego rodzaju narzędzi oraz metod. Podobnie postrzegają oni dyskurs światów IT oraz nauk przyrodniczych i ścisłych w akademii. Tam także używane jest pojęcie ML. Sądzimy, że uczestnicy świata DS zauważają posługiwanie się zarówno pojęciem ML jak i sztucznej inteligencji (AI) w świecie prawa. Inne światy nie-techniczne, czyli nauki społeczne i humanistyczne w akademii, biznes, media i polityka raczej postrzegane są jako posługujące się pojęciem AI (por. rys. 45). **Widzimy pojęcie AI jako tzw. opakowanie** (Kacperczyk 2016:45) **dla systemów bazujących na modelach ML oraz w ogóle dla działalności świata DS**.

> Szczególną rolę pełni tu przykuwające uwagę „opakowanie" społecznego świata (*package* – Fujimura, 1997) rozumiane jako zestaw narzędzi, metod, wartości i ideologii wyrażonych w unikalnym języku, zawierający specyficzną terminologię i wyznaczający charakter praktyk organizacyjnych. Tego rodzaju „opakowanie" umożliwia utrzymanie społecznego świata poprzez dostarczanie klarownych środków obrony świata i sposobów definiowania podstawowego działania. Wyjątkowo atrakcyjne „opakowanie" może przyciągać inne światy do wspólnego zaangażowania w określone działanie, o ile przemawia do ich identyfikacji i odwołuje się do ich wartości (Kacperczyk 2016:45).

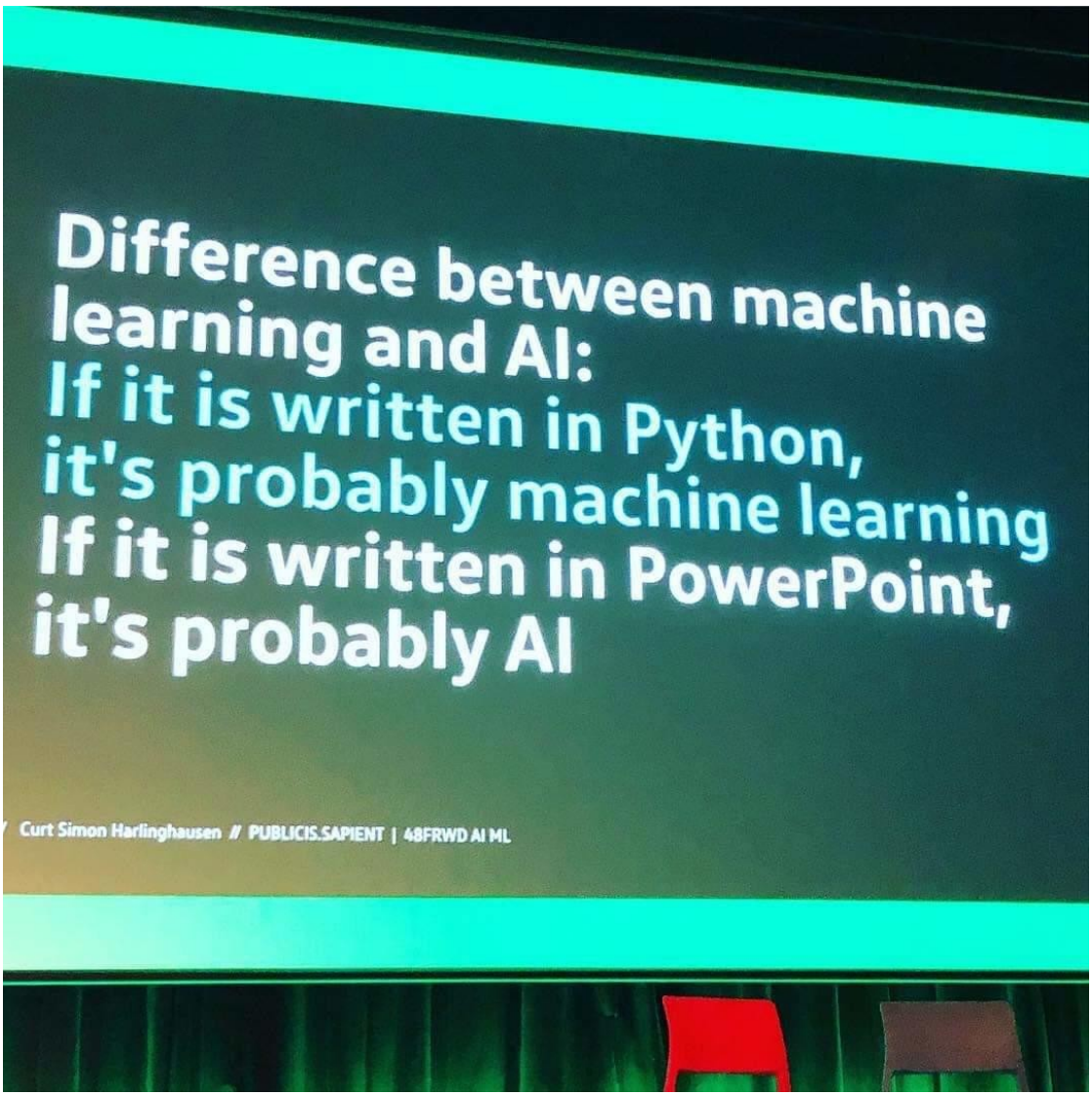
Uczestnicy świata DS raczej nie posługują się pojęciem AI w swoim gronie. Naszym zdaniem, gdy mówią o AI, wskazuje to na żart lub ironię. Pojęcie to stosują w komunikacji z innymi światami nie-technicznymi zaangażowanymi w arenę. W kontekście działań komercyjnych, pojęcie AI jest stosowane głównie przez biznesowych rzeczników świata DS. Owi rzecznicy to osoby na stanowiskach sprzedażowych i marketingowych, które formułują

komunikaty do klientów wewnętrznych lub zewnętrznych projektów DS, czasami w tej roli występują także osoby będące uczestnikami świata DS, np. kierujące zespołami / działami DS. Autor podręcznika do nauki podstaw ML dla osób nie-technicznych w następujący sposób uzasadnił cel książki:

> Jako konsultant z łatwością zdobywałem klientów za pomocą kartki papieru i prezentacji PowerPoint, ponieważ nie odróżniali oni sztucznej inteligencji (AI) od analityki biznesowej (BI) lub analityki biznesowej (BI) od licencjackiego stopnia naukowego BS [*Bachelor of Science*]. Napisałem tę książkę, ponieważ chciałem zwiększyć rzeszę osób rozumiejących i potrafiących implementować techniki analizy danych (Foreman 2017:16).

Rzecznikami mogą być także osoby zajmujące się zasobami ludzkimi (HR), formułujące komunikaty do ludzi zainteresowanych pracą w DS, czy bardziej konkretnie budowaniem modeli ML – początkujących uczestników świata DS i wannabesów. Zatem AI jako opakowanie dla działalności świata DS (a właściwie najbardziej atrakcyjnego, ciekawego, sprawczego, seksownego itp. jej wycinka, czyli modelowania ML) jest opakowaniem przygotowanym zarówno dla klientów – szeroko rozumianych jako osoby, które płacą za projekty i pracę data scientistów – jak i dla osób zainteresowanych pracą w DS.

Jaskrawym przykładem tego, że uczestnicy świata DS postrzegają AI jako nie-techniczne opakowanie dla swojej, technicznej działalności polegającej przede wszystkim na pisaniu kodu do przetwarzania, analizy i modelowania danych cyfrowych jest żartobliwa sentencja, która zdobyła środowiskową popularność na początku 2019 roku:



Rysunek 46: AI jako opakowanie dla działalności świata DS – żartobliwa sentencja

[Różnica pomiędzy uczeniem maszynowym a AI: Jeżeli to jest napisane w Pythonie, to prawdopodobnie uczenie maszynowe. Jeżeli jest napisanie w PowerPointcie, prawdopodobnie to AI] Źródło: https://www.linkedin.com/feed/update/urn:li:activity:6501030890754314240/, dostęp 03.09.2020 r.

Tę „sentencję" wygłosił Curt Simon Harlinghausen, który nie jest uczestnikiem świata DS, a przedsiębiorcą w branży internetowej, podczas wystąpienia na konferencji nie-technicznej (48forward 2019). Jak sądzimy, miało to być komunikatem nastawionym na edukowanie potencjalnych klientów zamawiających usługi bazujące na modelach ML. Niemniej „sentencja" przyjęła się w badanym świecie DS. W naszych rozmowach w pierwszej połowie 2019 r. powtarzała się wielokrotnie, zarówno z inicjatywy badacza jak i Rozmówców:

> *B: Mhm, ok. A ty czym konkretnie się zajmujesz? W twojej pracy?*

R: W mojej pracy jestem w zespole, który szumnie nazywa się AI Core Team[277] [śmiech]
B: AI Core Team? Tak?
R: Tak, tak dokładnie [ze śmiechem].
B: Ok, a dlaczego, a dlaczego szumnie przepraszam?
R: Nie to kolega dosłownie w zeszłym tygodniu coś takiego powiedział że yyy generalnie jak piszesz kod to jest machine learning ale na prezentacji to się nazywa AI. Nie no śmieję się z tego AI że to wszędzie teraz się pojawia takie supermodne hasło w każdym zespole, w każdej firmie, na każdej prezentacji, a de facto taka prawdziwa sztuczna inteligencja to jeszcze nie została stworzona. To jakby. {wywiad 18}

B: A słyszałem yyy coś takiego, w formie żartu, żeee jeżeli to jest machine learning, no to jest pewnie napisane w Pythonie,
R: [śmiech]
B: wiesz o czym mówię?
R: Tak, a AI to jest w PowerPointcie [śmiech]
B: No właśnie. I możesz mi wytłumaczyć ten żart?
R: Yyy moim zdaniem to jest dlatego właśnie, że AI lubią tego pojęcia używać, osoby od marketingu, bo ono brzmi egzotycznie, ono, ludzie mają jakieś wyobrażenia o sztucznej inteligencji wykreowane przez filmy i media, więc dlatego PowerPoint to się kojarzy z takim dynamicznym managerem który próbuje ci coś sprzedać, a machine learning brzmi już trochę nudniej i to są właśnie te osoby, które siedzą i klepią kod w Pythonie. Ja tak to widzę.
B: Mhm, mhm, mhm, ok. Czyli to raczej marketingowcy mówią o tych, robotach które tutaj wszystko będą same i tak dalej, tak, a ci od kodu to tak nie sądzą, chyba?
R: Ja mam, takie wrażenie patrząc po nie wiem, jakimś środowisku w którym się obracam. [pauza] Chciałabym może żeby, chciałabym może żeby AI było tak mądre jak, jak ludzie próbują sprzedać że jest. {wywiad 20}

R: (...) Tu może gwoli wyjaśnienia, bo bardzo często mylone są te dwa pojęcia sztuczna inteligencja i uczenie maszynowe, to ponieważ to jest mimo wszystko Politechnika, tutaj mamy dosyć przyziemne na to spojrzenie. Jest taki złośliwy żart, który krąży po korytarzach, że jeśli sztuczna inteligencja, to prawdopodobnie to jest w Power Pointcie, jeżeli uczenie maszynowe to w Pythonie, i to trochę oddaje rzeczywistość. Ta sztuczna inteligencja to jest często też tak na etapie, bo byśmy chcieli, bo jakieś wizje, natomiast w praktyce, to co na co dzień działa, co da się zakodować w [niezrozumiałe], to jest uczenie maszynowe. (…) {wywiad 23}

Zatem AI jest w badanym świecie traktowane jako termin nie-techniczny, stosowany głównie przez biznesowych rzeczników świata DS w celu sprzedawania usług czy systemów, w których używane jest uczenie maszynowe. Jak dalej pokażemy, nawet nie musi to być ML. Terminem „AI" opakowywane są technologie nie związane z ML, bazujące na zaprogramowanych regułach, ale np. rozpoznające język naturalny, jak chatboty. W sensie technicznym nie jest to błąd, bowiem ML jest uznawany za subdyscyplinę AI (Goodfellow i in. 2016:9; Russel i Norvig 2009:2; Raschka 2018:26).

Nie ma w badanym świecie jednoznacznej zgody co do tego, czy używanie rozmytego, obrośniętego popkulturowymi skojarzeniami terminu AI w komunikacji ze światami nie-technicznymi leży w interesie świata DS. Praktyka ta jest dość powszechna. W raporcie

277 Nazwa zmieniona z zachowaniem słownictwa zbliżonego do oryginały, w celu zachowania poufności wywiadu.

londyńskiej firmy MMC Ventures wskazano, że wśród ponad 2 800 europejskich startupów twierdzących, że zajmują się AI, w około 40% nic na to nie wskazuje – „nie znaleziono żadnego potwierdzenia, by AI stanowiło ważną część oferowanych produktów" (Schulze 2019). Opisana praktyka ma być stosowana w celu podniesienia atrakcyjności startupów w oczach inwestorów (ibidem).

Pod koniec naszych badań (maj 2019) spotykaliśmy się z opiniami, iż bez AI jako opakowania dla DS nie da się funkcjonować na rynku. Zdaniem Rozmówczyni {wywiad 25} osoby zajmujące się DS muszą deklarować, iż zajmują się AI, bowiem termin DS nie wzbudza zainteresowania klientów. Za „fasadą", jaką jest opisywane opakowanie, kryją się zarówno firmy, których działania Rozmówczyni oceniała pozytywnie, jak i takie, gdzie poza opakowaniem nie widzi ona wiele (na co wskazują także wyniki finansowe tych „fasadowych" firm jakoby zajmujących się AI):

> *B: Rozumiem. Chciałem jeszcze poruszyć temat tej branży w Polsce. Bo ja w ogóle posługuję się tym terminem branża data science, yyyy. Nie wiem, czy to jest szczęśliwy termin, czy nieszczęśliwy. Ale proszę mi powiedzieć, czy z pani punktu widzenia w ogóle jest w Polsce taka branża?*
> R: Myślę, że tutaj dużo bardzie trafnym terminem rzeczywiście byłoby to AI, dlatego że **wszyscy** próbują się pod nie podpinać. Jeśli chodzi o typową branżę w sensie w tej chwili to jest termin modny i ktokolwiek nie zajmuje się rzeczywiście data science, to w zasadzie deklaruje, że zajmuje się AI. Nie dlatego że uważa to za bardziej trafny termin, tylko dlatego że to jest w tej chwili nośne i jak mówi, że zajmuje się data science, to nikt go nie słucha. Więc wszyscy musimy deklarować, że zajmujemy się AI. Czy jest taka branża w Polsce? Są jej powijaki. Jest kilka firm, które są na rynku rzeczywiście od lat, które próbują coś zrobić. Opinie o nich są bardzo różne, to bardzo różnie bywa. Są różne fenomenalne start-upy, które wystartowały z taką pompą, że naprawdę wrażenie było, że za chwilę przebiją granicę i polecą na cały świat. Większość z nich dość szybko padła. Są inne firmy, które nie robią wokół siebie dużo szumu, a robią dobrą robotę, ciężko zaprzeczyć, że takie są. Są ogromne korporacje, które mają w Polsce centra analityczne, chociażby [nazwa firmy A], [nazwa firmy B] próbuje w tej chwili zbudować kolejny zespół w Warszawie. (...) Są firmy takie jak [nazwa firmy C], które próbują zrobić jakiś wielki produkt na skalę światową. No niestety [cmoknięcie] w momencie kiedy próbujemy z nimi współpracować, eee a współpracujemy z nimi częściej niż by się mogło wydawać, okazuje się, że od środka wygląda to zupełnie inaczej. Oczywiście są to prywatne sprawy firmy, w związku z czym nie będę o takich rzeczach mówić w szczegółach. Natomiast ładna fasada często jest tylko i wyłącznie ładną fasadą. Wystarczy popatrzyć na wyniki finansowe tych firm. {wywiad 25}

Trzeba jednak rozróżnić odbiorców takich, bazujących na terminie AI, komunikatów na dwie grupy – po pierwsze, osób zainteresowanych pracą jako data scientiści; po drugie, klientów. W pierwszym przypadku sądzimy, że uczestnicy świata DS raczej negatywnie oceniają używanie terminu AI przez swoich rzeczników, w tym wypadku pracowników działów HR / rekruterów. Z naszych badań wynika, że ocena jest negatywna z jednej strony w tym sensie, że data scientiści nie chcą w swoich zespołach osób przyciągniętych do pracy

obietnicami o AI, jak również wśród uczestników jest dość powszechne przekonanie o tym, iż lepiej unikać firm rekrutujących nachalnie pod hasłem „AI" z małą ilością szczegółów technicznych w ofertach pracy. Pamiętajmy jednak, że termin AI jest stosowany w nazwach grup meetupowych zajmujących się DS, a przyciąganie na spotkania osób zainteresowanych pracą w DS jest jednym z celów firm wspierających organizacje meetupów, bowiem np. buduje świadomość tego, jacy pracodawcy w tej branży są dostępni na lokalnym rynku. W przypadku komunikacji z klientami, stosowanie terminu AI jest oceniane ambiwalentnie w badanym świecie DS. W naszej opinii trudno wskazać nawet coś na kształt różnych pozycji w dyskursie na ten temat. Jest raczej tak, że większość uczestników zajmujących się komercyjnym DS ocenia te praktyki jednocześnie jako leżące i nie leżące w interesie swojego świata społecznego, a nawet węziej, w interesie swoim i swoich firm/zespołów/działów. Przyjrzyjmy się wypowiedzi osoby, która w obecnej pracy nie pisze kodu, zajmując się prowadzeniem projektów i kontaktem z klientami (zatem pełni rolę biznesowego rzecznika świata DS), niemniej uważa się za tzw. osobę techniczną i ma doświadczenie w wykonywaniu działania podstawowego świata DS:

> *B:No dobra, a czy śmieszy Cię żart sztuczna inteligencja jest napisana w Power Pointcie, a ML w Pythonie?*
> R: Nie, bo ja robię sztuczną inteligencję w Power Pointcie. Nie no żartuję [z uśmiechem]. Zawsze jak robię slajdy i prezentację, to mówię no, zajmuję się AI [z uśmiechem]. [pauza] Trochę tak jeeeeest, bo my lubimy trochę rozdmuchać
> *B: To nie jest żart, tylko to jest prawda?*
> R: No, my lubimy rozdmuchać według mnie. Czyli lubimy lubimy rozdmuchać, że coś jest większ bo **są** pewne realne zagrożenia. Jak na przykład kwestia odpowiedzialności. To jest takie realne zagrożenie. Na pewno **część** prac zostanie zautomatyzowana, ale czy większość przez sztuczną inteligencję? Wiele rzeczy zostanie zautomatyzowanych przez **proste** skrypty, RPA, naprawdę nie nie bardzo zaawansowaną analitykę.
> *B: RPA? Co to takiego?*
> R: Robotic Process Automation, czyli tak naprawdę [pauza] takie mirrorowanie, co robią ludzie. RPA yyy po prostu działa tak, jak osoba, która widzi GUI, tak? Widzi Graphical User Interface i ona jest w stanie po prostu zrobić dokładnie te same ruchy.
> *B: Ale to to działa na zasadzie yyyy if-else, tak?*
> R: Tak.
> *B: A nie jakiegoś tam ML-a?*
> R: Dokładnie. I wbrew pozorom to **to** jest najwięcej w tej chwili automatyzacji. Ten RPA, te przeróżne skryptowe też automatyzacje. I one często zamykają nawet **całe działy**, bo oczywiście to jest efektywniejsze, szybsze i oczywiście przede wszystkim nie popełnia błędów. Bo to jest też to, algorytm się nie męczy [z uśmiechem]. Jak by to nie zabrzmiałooo, my jesteśmy bardziej podatni na błędy. Nadal to, na jakim etapie jest sztuczna inteligencja, to ona umie się zająć **bardzo** wąskim wycinkiem. Ona nie nie za zaadresuje holistycznego problemu. Więc jeśli masz holistyczny problem, który się składa z czterech mniejszych fragmentów, to ty dostaniesz wszędzie insighty. Tutaj, tutaj, tu tu tu, o co chodzi, ale to ty sam będziesz musiał sobie to połączyć. My nie jesteśmy jeszcze na tym etapie i nie wiem, kiedy będziemy na tym etapie, że ona będzie umiała naprawdę złapać szerszy kontekst. To na razie nie jest tak. Na razie my jej podajemy przykłady no na przykład w deep learningu i ona się uczy na tych przykładach.

B: Jeszcze tutaj o coś Cię dopytam yyy, a potem wrócimy do Ciebie. Chciałem zapytać, czy ten medialny szum wokół AI nie jest tak, yyyyy że się opłaca tobie i w ogóle ludziom z nazwijmy to już szeroko ludziom z Twojej branży?

R: Czy mi się opłaca, no oczywiście. Tak jak w ogóle całe IT mi się opłacało [ze śmiechem]! No tak, no. Nie da się ukryć, że **zawsze**, jeśli idzie za tym szum medialny, to się opłaca, ale yyy trzeba spojrzeć i na blaski, i na cienie. Blaskiem jest to, że rzeczywiście teraz wszyscy chcą AI, ale cieniem jest to, że idziesz do firmy i oni oczekują, że generalnie nie nie no wiadomo co stworzysz. Więc trzeba też czasami studzić ten zapał i mówić, fajnie, guys, ale to nie jest ten etap, to nie będzie tak. Druga sprawa **bardzo ciężko** jest zrozumieć firmom też, że data science, machine learning to jest eksperymentowanie. A eksperymentowanie wiąże się z porażkami. I ciężko im czasami zrozumieć, że to jest tak, że nie zawsze będzie działać idealnie algorytm. Zabawne jest to, że my wiemy, że jako ludzie popełniamy błędy, ale jak mówimy, że algorytm ma accuracy na poziomie 90%, to jest takie [zmienionym głosem] jak to, 10 casów robi źle? I wy mówicie, że to jest gotowe rozwiązanie? Oczywiście, teraz trochę ekstrapoluję, ale rzeczywiście są takie rzeczy, że rzeczywiście, jak czasami idziemy do firm, to troszeczkę trzeba studzić, a nawet mam wrażenie, że w tej chwili już w tych dużych firmach to trochę narasta taka frustracja. Co innego jest **obiecywane**, co innego **dostarczone**. I to nie chodzi mi o frustrację na dostawców, tylko bardziej na ten przegięty szum medialny i jakieś takie ciśnięcie z góry słuchaj, wszyscy ej-aja robią, to my też musimy. Bo to jest naprawdę. Ja wiem, że lubi się żartować, że o, w korporacjach to ktoś przychodzi i mówi taka technologia, to teraz znajdź mi projekt do tego. Tak jest, no. Naprawdę tak jest czasami.

B: To jest bardzo ciekawe. Czyli chcą jakąś technologię, żeby mieć technologię? Nie od problemu, tylko od technologii?

R: Wiesz co, bardziej dlatego, bo inni używają, działa, zobaczmy. Oczywiście cel sam w sobie jest OK no, fajnie, zobaczmy jak działają te nowe technologie. Ale czasami to jest szukanie wręcz na siłę albo właśnie użycie technologii, która spełnia swój cel, jak blockchain dla jakości danych, ale czy to jest optymalne finansowo [z uśmiechem]?

B: No właśnie, mówiłaś o tym blockchain, ja sobie tutaj zapisałem, bo mam taki pomysł, że chcę napisać, żeeeee niektóre rzeczy, przyjmijmy, że niektóre technologie, robią się takim fetyszem w biznesie. Że sama ta technologia jest super i sama ta technologia jest pożądana, a nie to, co ona zrobi. Słyszałem takie historie, że że że klient musi usłyszeć, że zrobimy machine learning, bo inaczej jest niezadowolony.

R: To jest prawda to jest prawda, ale wiesz co? Tak jest z każdą technologią. Kaaaaażda technologia ma taki cykl. Jak wchodziło big data, to wszyyyyscy big data, to nic, że nie miałeś tych big data albo po co ci one były, albo utrzymanie tych danych było bez sensu, bo nie wiedziałeś, co z nimi zrobić. Ale musiałeś. No tak jest z każdą technologią. Ale wiesz, to możemy patrzeć i na poziomie firmy, i na poziomie nas, ludzi. Co, każdy chce nowego iPhone'a w sensie ekstrapoluję każdy. Ale jest jakiś taki trend. Tak, no rzeczywiście w firmach. Ale to też wynika z tego, że żadna firma nie chce być do tyłu, zwłaszcza z tych dużych. Oni nie chcą być do tyłu, oni wiedzą, że jakieś technologie mają pewne możliwości i eksplorują je czasami zbyt wcześnie, zwłaszcza data science jeszcze tak dwa lata temu, to było głównie POCe, które nie szły dalej i to z różnych względów. Naprawdęęę, przeróżnych względów. Od takich no braku zaufania ludzi po takie, że wiesz, nieumiejętność wytłumaczenia tego biznesowi przez osoby dowożące, bo to też jest ważne, tak. Jednak ten aspekt taki edukacyjny jest bardzo istotny. Chociaż na dzień dzisiejszy ja mam wrażenie, że przez ostatnie półtora roku – ale ja jestem mocno zbiasowana, więc jakby trzeba brać na to poprawkę, bo ja się obracam po prostu cały czas w tym kręgu osób ja mam wrażenie, że to się naprawdę [krótka pauza] AI, sztuczna inteligencja, przez ostatnie półtora roku, to już każdy w firmie musi widzieć o co chodzi chociaż, generalnie rozumieć koncept.

B: Ale rozumiem, że przez cały czas musisz edukować klientów i tak trochę sprowadzać ich na ziemię, tak?

R: No to zależy też jakich. Niektórzy już są mocno wyedukowani. Dlatego mówię to już widać na dzień dzisiejszy, że oni są mocno wyedukowani, tak. Ale raczej, ja bym powiedziała, to jest sprowadzanie na ziemię, a czasami po prostu od razu trzeba grać jednak w te otwarte karty.

> Słuchajcie, taka jest możliwość, to można dostarczyć i przede wszystkim jednak data science to eksperymentowanie, ewoluowanie, cały czas ewoluowanie, żeby **rzeczywiście** to spełniało to założenie biznesowe. Wydaje mi się, że ja to widzę też, że coraz bardziej komfortowo też biznes się zaczyna z tym czuć, z tymi rozwiązaniami. {wywiad 22}

Wygórowane, bazujące na niejasnych wizjach i niewiedzy technicznej oczekiwania klientów być może przyciągają ich do rozpoczęcia współpracy z firmami oferującymi usługi związane z modelowaniem danych, co raczej leży w interesie badanego świata DS, bo traktowane jest jako popyt. Niemniej już na wczesnych etapach wiele projektów kończy się, a klienci mogą być rozczarowani brakiem rozwiązania problemu biznesowego, brakiem wdrożenia, nierzadko mając trudności z akceptacją eksperymentalnego charakteru projektów DS i tego, że nawet najbardziej dopracowane systemy bazujące na ML nie mają 100% skuteczności. Jednocześnie problem biznesowy może być rozwiązany bez użycia ML, a za pomocą rozwiązań z zakresu inżynierii oprogramowania (za pomocą zapisywania reguł, tzw. if-else), jednak wówczas nie jest to w badanym świecie uznawane za projekt DS – nawet kiedy produkt ma elementy przetwarzania języka naturalnego, jak np. chatbot czy asystent głosowy (choć wówczas często będzie nazywany „AI"). Rozmówczyni {wywiad 22} uważa, że klienci stają się coraz bardziej wyedukowani – zaryzykujmy stwierdzenie, że zaczynają postrzegać DS coraz mniej jako magię, a coraz bardziej jak majsterkowanie.

Pisaliśmy już o przekonaniu uczestników badanego świata co do tego, że DS będzie się wciąż dynamicznie rozwijać, mieć coraz większy wpływ na świat, choć pewnie opadnie nieco entuzjazm wobec najmodniejszych obecnie terminów, jak np. AI (por. 5.4.3). Niemniej dominujące w badanym świecie jest przekonanie, że „prawdziwa", równa czy wyższa ludzkiej sztuczna inteligencja jest jeszcze niezmiernie daleko, o ile w ogóle kiedykolwiek będzie możliwa. Rozmówczyni {wywiad 18} wskazała, że jak najbardziej DS rozwija się dynamicznie, należy być na bieżąco by utrzymać się w branży, jednak po pierwsze, projekty DS są wykonalne raczej w dużych, bogatych i zaawansowanych technologicznie firmach o wysokiej „kulturze danych", głównie w największych miastach; po drugie, składające się na opakowanie dla ML (i innych technologii) obietnice związane z „AI" pozostaną w sferze deklaracji – w PowerPointcie – co najmniej do momentu powstania komputerów kwantowych:

> *B: To właściwie wszystko, znaczy to właściwie wszystkie pytania jakie mam, jakie mam przygotowane. Yyy, czy coś chciałabyś dodać? Do naszego wywiadu?*
> R: Yyy no nie wiem tam mówiłeś że może byśmy pospekulowali na temat przyszłości [śmiech] w ogóle tego zawodu.
> *B: No! Bardzo chętnie. Bardzo chętnie.*

R: Bo tak jak ja obserwuje teraz trendy przynajmniej w firmie [nazwa kraju z Europy Zachodniej], dużej, która ma na to budżet to [pauza] no stety albo niestety, małe firmy się tym nie zajmują, bo nie mają na to pieniędzy. Więc wiedziałam że gdybym chciała zmieniać pracę na przykład no to opcja będzie tylko w Warszawie, Krakowie, Wrocławiu, ewentualnie Poznaniu. Nie ma co raczej marzyć o mniejszych miastach no i tylko że to się w dużych firmach na razie rozwija i pewnie tak, tak to zostanie. [pauza] No i że raczej to będzie wymuszone, że będzie to istniało tylko w firmach które, w których jest wysoka jakość, że tak powiem wysoka kultura danych.
B: Aha, co to znaczy?
R: [niezrozumiałe] że spróbowaliśmy, czy tam robiliśmy projekt dla firmy w której rzeczywiście dane były przetrzymywane tylko i wyłącznie w Excelach [śmiech] innych na każdy miesiąc. I okazało się że jest tam mnóstwo pułapek i zrobienie jakiejś sensownej analizy na takich danych było niemożliwe. Po prostu, bo jakby, nie wiem, za mało było tych danych, one się strasznie zmieniały, przede wszystkim było tam, widzieliśmy dużo jakiś dziwnych zmian które nie były w tych danych odzwierciedlone. Takie że na przykład nie wiem mieliśmy skok szeregu o 30-40% pytamy się pracownika, a bo tutaj na przykład jakiś tam nie wiem, dodatkowa promocja się zaczęła, no i w danych nie mieliśmy nigdzie informacji o tej promocji więc potem dodanie tego było niemożliwe. Eee, no więc te tematy będą się tylko kręciły pewnie w dużych firmach które mają te dane jakoś sensownie ustrukturyzowane. Eee, a co do samego rozwoju [śmiech] AI, to myślę że przez najbliższe lata, jeszcze się nie musimy niczym martwić, raczej AI nie działa
B: Czyli pozostanie w PowerPointcie tak?
R: Tak, pozostanie w PowerPointcie i raczej nasza pozycja tu nie jest zagrożona. Możemy zacząć się martwić jak powstaną komputery kwantowe, to trochę się sytuacja może zmienić [śmiech]. Yy, no ale właśnie fajne jest to że ten, [pauza] bardzo dynamicznie się to rozwija, a że się zmieniają w bardzo szybkim tempie i narzędzia i możliwości [pauza]. Choćby teraz jakiś czas temu właśnie też na potrzeby projektu robiłam analizę bibliotek machine learningowych. No to też widać że biblioteki które powstały parę lat temu już są u schyłku, a niektóre się bardzo intensywnie rozwijają. I to wszędzie, w całej tej branży są takie tendencje, że narzędzia niektóre będą się szybko dynamicznie rozwijały, a takie które powstały i nie są rozwijane to będą zanikać, to nie jest tak jak z Excelem, że będzie wieczny [śmiech]. Yyy, no ale myślę że cały czas branża, działka, ta działka będzie się prężnie rozwijała i mam nadzieję że uda mi się w niej utrzymać. {wywiad 18}

Powyższy fragment jasno ilustruje oczywisty wniosek, że **na arenie modelowania** (rys. 45.) **największą władzę nad przetrwaniem i rozwojem zaangażowanych w nią światów i aktorów mają aktorzy o najwyższym kapitale ekonomicznym i technologicznym.**

6.2.1.2. Magia i majsterkowanie

Klienci, wannabesi – osoby chcące wejść do świata DS, a nawet początkujący uczestnicy mogą postrzegać uczenie maszynowe jako coś magicznego. Model ML staje się fetyszem, który sam w sobie ma rozwiązywać realne problemy biznesowe (głównie z perspektywy klientów) lub stanowić o byciu prawdziwym data scientistem (raczej z perspektywy wannabesów / początkujących). Dla uczestników silniej osadzonych w badanym świecie modelowanie z zastosowaniem ML (czy w ogóle projekty DS) to praca majsterkowicza, który sam wyposaża swój warsztat i chałupniczo przerabia narzędzia, używa ich czasem niezgodnie z pierwotnym przeznaczeniem, pracuje metodą prób i błędów.

Wykonanie modelu ML, który pomaga w rozwiązaniu realnego problemu, wymaga sporo majsterkowania – czasami żmudnego i męczącego, czasami pomysłowego i nieszablonowego, zawsze zaś trochę eksperymentalnego, gdzie rezultaty są niepewne.

Modele ML trenuje się na konkretnych zbiorach danych, mających specyficzne obciążenia, tzw. biasy. Przypadkowe współwystępowanie cech w zbiorze danych treningowych prowadzi do błędów w działaniu gotowego modelu, niemniej takie błędy nie jest łatwo zauważyć osobom przygotowującym modele. Słynny w badanym świecie DS eksperyment (Ribeiro, Singh, i Guestrin 2016:1142–43) dotyczył właśnie tego zagadnienia. Celowo przygotowano model nadzorowanego uczenia maszynowego tak, aby dawał obciążoną odpowiedź, a następnie oceniano działania tego modelu przez studentów (zaznajomionych z ML). Model miał klasyfikować zdjęcia wilków i psów husky. Zbiór danych treningowych przygotowano tak, by wilki zawsze prezentować na śniegu, zaś husky zawsze bez tegoż. Walidacja wykazała, że zdjęcia zawierające śnieg lub inne, jasne tło przy dolnej krawędzi klasyfikowane były przez gotowy model jako „wilk", zaś nie zawierające takich cech jako „husky" – bez względu na kolor zwierzęcia czy pozę w kadrze. Następnie zapytano 27 studentów studiów II stopnia, mających ukończony co najmniej jeden przedmiot z zakresu ML:

- czy ufają w prawidłowe działanie tego modelu w realnym świecie?
- dlaczego?
- w jaki sposób model rozróżnia zdjęcia wilków i psów husky?

Badanym pokazano 10 odpowiedzi modelu, gdzie 8 było poprawnych, jedna pokazywała husky na śniegu zaklasyfikowanego jako „wilk", jedna zaś wilka bez śniegu zaklasyfikowanego jako „husky". Wypowiedzi 10 z 27 studentów oceniono jako wskazujące na zaufanie do modelu, zaś 12 z 27 badanych miało sądzić, iż śnieg na zdjęciu jest jedną z cech, których „wyuczył się" model. Następnie studentom pokazano graficzne wyjaśnienie odpowiedzi modelu (rys. 47):

Rysunek 47: Obciążony model ML "wilk czy husky?" – błędna odpowiedź modelu i wyjaśnienie

(a) husky zaklasyfikowany jako „wilk”
(b) wyjaśnienie
Źródło: Figure 11: Raw data and explanation of a bad model’s prediction in the „Husky vs Wolf” task (Ribeiro, Singh, i Guestrin 2016:1143)

Po wyjaśnieniu graficznym pytania powtórzono. Na zaufanie do działania modelu wskazywały tylko 3 badane osoby, zaś śnieg jako cechę wyuczoną przez model wskazało aż 25 z 27 badanych. Zdaniem autorów, metody wyjaśniania klasyfikacji konkretnych przypadków przez modele ML (rys. 47.) są przydatne dla uzyskania wglądu w modele i dla oceny, czy ufać ich działaniu i dlaczego (Ribeiro, Singh, i Guestrin 2016:1143).

Eksperyment ten pokazuje także, iż w sensie technicznym nie ma wątpliwości, że trudne jest uzyskanie modelu ML dającego wysoki odsetek poprawnych odpowiedzi i mającego wysoką zdolność tzw. generalizacji, czyli nie „zbiasowanego” ani przeuczonego, „nauczonego na pamięć” {wywiad 8} współwystępujących cech w konkretnym zbiorze danych. W sensie realizacji celu całego projektu DS, będzie to bez porównania trudniejsze. Realizacji projektu, którego małą częścią jest praca nad przygotowaniem modelu – a przy uwzględnieniu szeregu czynników organizacyjnych, biznesowych, kulturowych, formalnych czy infrastruktury technicznej – produkcyjne wdrożenie i utrzymanie wypracowanego modelu ML tak, by ostatecznie przynosił klientowi wartość dodaną jest bardzo, bardzo trudne. Niemniej wielokrotnie spotkaliśmy wśród uczestników badanego świata opinie, iż **zarówno klienci jak i wannabesi oraz początkujący uczestnicy świata DS postrzegają modele ML jako coś, co w magiczny sposób rozwiązuje realne problemy**. Magiczny, czyli nie dający się racjonalnie wyjaśnić. Magiczny, czyli sam w sobie cudownie działający, na zasadzie –

weź problem, dodaj ML, problem rozwiązany. To alternatywny system ciągów przyczynowo-skutkowych, w którym imponujący efekt osiągany jest tajemnie, nie wiadomo jak.

Widzimy to jako fetyszyzację modeli ML, a wśród klientów nawet fetyszyzację samego ich opakowania, czyli AI. Przyjmujemy rozumienie fetyszy zaproponowane przez (Thomas i in. 2018), bazujących na koncepcjach Wiliama Pietza, Freuda, Karola Marxa i Davida Graebera (ibidem:4):

- Fetysz jest obiektem materialnym przesiąkniętym możliwościami (*capabilities*), które nie są z natury rzeczy właściwościami lub funkcjami samego przedmiotu.
- Te nadmiarowe możliwości są generowane w punkcie styku pomiędzy osobami o różnych pozycjach i tym samym rozszerzają one zakres ich wkładu (*outcomes*) społecznego, kulturowego i ekonomicznego.
- Ten wkład społeczny, kulturowy i ekonomiczny jest błędnie rozpoznany jako obietnica fetyszyzowanego obiektu lub zastąpiony tym obiektem.
- To zastąpienie lub błędne rozpoznanie jest samo w sobie skuteczne: fetyszyzowany obiekt umożliwia coś, co bez niego by nie zaszło.

Twierdzimy, że fetyszyzowanie modeli ML, a szczególnie konkretnych metod, jak najbardziej modnego uczenia głębokiego (*deep learning*) jest w badanym świecie traktowane jako coś charakterystycznego dla wannabesów lub początkujących uczestników. W przypadku uczestników silniej osadzonych w świecie DS dążenie do ciągłego stosowania uczenia głębokiego we wszystkich projektach przy pomijaniu innych metod ML oraz metod statystycznych i analitycznych jest oceniane zdecydowanie negatywnie – jako brak profesjonalizmu. Ślepa koncentracja wannabesów i początkujących uczestników świata DS na deep learningu czy ML jest zatem przeciwskuteczną strategią zaświadczania o autentyczności uczestnictwa w badanym świecie. Używanie uczenia głębokiego czy w ogóle ML może być dla wielu z takich początkujących osób fetyszem, ale samo przez się nie sprawia, że człowiek zostaje uznany za uczestnika świata DS przez innych uczestników. Postrzeganie wannabesów i początkujących w badanym świecie wydaje nam się takie, że te osoby uznają deep learning (lub cały ML) za magiczny artefakt, czyniący zwykłego człowieka magikiem – data scientistem. Przy tym problemy czyszczenia i przygotowania danych, czasami analizy danych i związane z tym narzędzia, jak np. SQL uważają nie tylko za nudne, ale i nieważne w pracy data scientistów. Jak mówiliśmy wielokrotnie, wśród silnie osadzonych uczestników świata DS rzeczywiście te problemy bywają uznawane za nudne (szczególnie przygotowanie

danych), ale ogromnie ważne dla realizacji projektu DS – jest to żmudna, brudna ale niezbędna robota, w przeciwieństwie do fajnego, seksownego i modnego modelowania. Poza tym samo modelowanie, choć fajne, nie ma w sobie niczego magicznego, a wymaga wiele majsterkowania. Jest jeszcze „gorzej" dla uznania ich autentyczności w badanym świecie, gdy wannabesi lub początkujący uważają, że ML w magiczny sposób rozwiązuje rzeczywiste problemy. Jest to poważny występek przeciwko wartościom badanego świata. Twierdzimy też, że magiczne postrzeganie jakiejkolwiek metody czy narzędzia używanego w świecie DS jest sprzeczne z niezbędnym elementem świadczącym o byciu autentycznym uczestnikiem – z byciem tzw. osobą techniczną.

Elementem bycia autentycznym uczestnikiem świata DS jest używanie metod ML, ale w autentycznym uczestnictwie chodzi o opanowanie szerokiego zakresu narzędzi i metod oraz rozwiązywanie rzeczywistych problemów (por. 6.1.2.2). To jest postrzegane jako cenne, bowiem ma przyczyniać się do zwiększania sprawczości i efektywności uczestnika świata DS, a także ma świadczyć o dążeniu do samorozwoju, i o ciekawości względem nowych narzędzi oraz zwiększać swobodę działania (por. 5.4). Widzimy wyraźnie omawiany wcześniej (część 5.3.1.) na przykładzie wyboru języka programowania postulat dobierania narzędzi do problemu. Najbardziej cenieni w badanym świecie DS są przecież uczestnicy, którzy potrafią dobrać najbardziej optymalne do problemu narzędzia i metody, a znacznie mniej tacy, którzy do kolejnych problemów stosują wciąż i wciąż te same. To podejście nie tylko redukuje sprawczość i efektywność, ale także świadczy o braku samorozwoju, oraz braku ciekawości względem nowych narzędzi. Nisko cenieni są też tacy, którzy wybierają narzędzia kierując się wyłącznie modą lub marketingiem. Choć może to wskazywać na ciekawość i dążenie do samorozwoju, to nie wskazuje na efektywność ani sprawczość. Jest to podejście traktowane w świecie DS jako wyraz braku zainteresowania rozwiązaniem rzeczywistego problemu. W badanym świecie powtarzane jest powiedzenie „kiedy masz tylko młotek, wszystko zaczynasz traktować jak gwoździe" (*I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail*) (Maslow 1966:15). Można by uznać je za motto silnie osadzonych uczestników świata DS – przyjrzyjmy się wypowiedziom takich osób:

> *B: Pierwsze moje pytanie. Czy pani jest data scientist?*
> R: Tak.
> *B: Mhm, i co to właściwie oznacza?*
> R: Oznacza to analizowanie danych z pomocą wszelkich dostępnych na świecie, zgodnie z różnymi zasadami wiedzy, metod, zarówno statystycznych, jak i bardzo podstawowych, w zależności od tego, jakie są konkretne dane i jakiego rodzaju analizy potrzebują. Nie ma tutaj

> żadnych ustalonych standardów, w tym sensie, że korzystamy z tego, co naprawdę jest potrzebne. Jeśli ee wystarczy nam zwykła analiza, podstawowa związana z obrazowaniem danych, to to jest to, czym się posługujemy eee jeżeli potrzebne są zaawansowane metody machine learningu, to to jest to, po co sięgamy. W zależności od konkretnej sytuacji, od konkretnych potrzeb, od konkretnych danych, tak? Najlepiej jest wykorzystywać jak najprostsze metody, dostosowane do konkretnej sytuacji. {wywiad 25}
>
> *B: OK, a tych projektach data science czy zawsze używacie uczenia maszynowego, ee we wszystkich projektach? Czy*
> R: Nie. Nie używamy uczenia maszynowego we wszystkich projektach bo wiemy co to jest uczenie maszynowe. (...) Stosujemy uczenie maszynowe kiedy coś nam to daje. Prowadzimy projekty z jasnym nastawieniem biznesowym, czyli możemy używać nawet bardzo **bardzo** prostych narzędzi. Nie musimy zawsze robić czegoś skomplikowanego, data science to nie jest tylko używanie rzeczy, których nie rozumieją przeciętni ludzie. {wywiad 7}

Odwołania do magii nie są skonstruowaną przez nas kategorią analityczną – pojawiała się ona *in vivo* w badaniach:

> R: Ale dokładnie to trudno mi powiedzieć, eee, właściwie w tej chwili doszło do tego, że jak się nie napisze że jest się data scientistem i jak się nie napisze że się używa machine learningu i sztucznej inteligencji toooo ludzie nie wierzą że się potrafi usprawnić ich biznes przy wykorzystaniu danych. Natomiast kiedy to się zaczęło?
> *B: Aha?*
> R: Jest mi bardzo trudno powiedzieć.
> *B: Rozumiem.*
> R: Myślę, że nie więcej niż 10 lat temu.
> *B: Ale czy to znaczy, że teraz klienci oczekują takiej nazwy, data science?*
> R: Ooo, to jest mi trudno powiedzieć. Wydaje mi się, że tak, booo to jest takie słowo klucz, albo taki termin klucz że aha, data science jest to taki gość, że mmmm, ma jakieś tam takie skomplikowane narzędzia i tymi narzędziami coś tam sobie kręci, wychodzą z tego modele i te modele automagicznie usprawniają biznes. Najlepiej jednak, żeby to była sztuczna inteligencja, i część klientów tego oczekuje. Żeby była sztuczna inteligencja, żeby było żeby była informacja o machine learning, tak, my usprawnimy twój biznes z wykorzystaniem AI.
> *B: Mhm*
> R: Natomiast ci klienci, którzy wiedzą, że chodzi o podejmowanie decyzji w oparciu o dane i to nie do końca ma znaczenie, jak wykorzystamy te dane, czy tam jest ta sztuczna inteligencja czy tam jest regresja liniowa, a może zrobimy zwykłą tabelkę, ci klienci, którzy to wiedzą, no tooo [pauza] nie mają klapek na oczach i nie sądzą że jak im zrobimy AI, to załatwi ich wszystkie problemy.{wywiad 5}
>
> *B: No dobra, no to na czym [pauza] Wyobraź sobie, że pierwszy raz słyszę takie słowo data scientist i chciałbym wiedzieć na czym to polega? Co to znaczy?*
> R: [śmiech] Ojej, to jest osoba która, bierze dane od klienta, które leżą prawdopodobnie u niego gdzieś na serwerze i nic z nimi nie robi. Eee i robi z nimi magię i przedstawia ją w postaci raportu, co ten klient ma zrobić z tymi danymi. Co ma wdrożyć do swojego systemu, jak może je wykorzystać, jak może zoptymalizować to jak jego biznes działał. Tak w skrócie.
> *B: Mhm, ok. Ale jak to robi z nimi magię?*
> R: No tak, o to właśnie chodzi, nie? Trzeba zobaczyć te dane tak naprawdę, porobić parę statystyk co w nich jest, zobaczyć i przygotować je odpowiednio. I to tak naprawdę jest 80% albo 60% całej roboty, zabawa z tymi danymi, jakby czyszczenie ich i tak dalej [pauza] Znaczy zależy też co chcemy osiągnąć nie? Musimy zobaczyć na dane i później odpowiednio do tego dobrać [pauza] jakiś model jak chcemy je przetworzyć, znaczy co będzie działało. {wywiad 8}

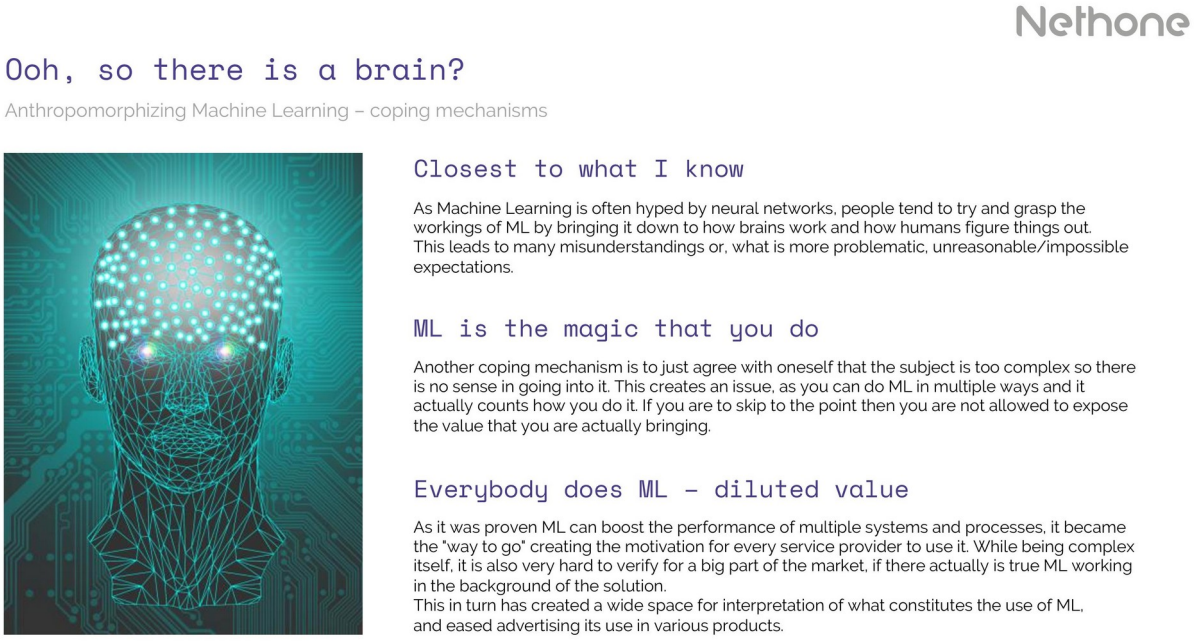
Deklarowanie, że technologia działa magicznie, ma budzić skojarzenie imponującej i bezproblemowej funkcjonalności. Funkcjonalności, w której efekt końcowy jest zdumiewający, a środki, za pomocą których efekt został osiągnięty, są nieistotne, a nawet tajemne (*inscrutable*). Już na początku lat 70. pisarz Arthur C. Clarke, stwierdził, że „każda wystarczająco zaawansowana technologia jest nie do odróżnienia od magii". Działanie „magiczne" to zatem jasna konotacja pozytywna. Oznacza, że technologia jest zaawansowana, wygodna, pozwalająca korzystającym z niej ludziom nie przejmować się szczegółami technicznymi, a wyłącznie cieszyć się z efektów jej używania. Cechą definiującą magiczne technologie jest także bezkosztowość. Działanie magii jest pomniejszone o niekorzystne elementy, np. walkę i wysiłek. Odwołanie do magii polega zatem nie tylko na zapewnieniu alternatywnego systemu związków przyczynowo-skutkowych, ale także na odwróceniu uwagi od metod i zasobów, w sensie technicznym niezbędnych do uzyskania określonego efektu (Elish i Boyd 2018:67:68).

Wskazywano nam, że wiedząc o fetyszyzowaniu magicznych, zaawansowanych technologii po stronie klientów, data scientiści mogą celowo używać bardziej skomplikowanych narzędzi czy metod, niż wymaga tego projekt. Zdarza się także praktyka opakowywania relatywnie prostych rozwiązań w deklarowane „zaawansowanie", także w komunikacji zespołu DS z klientem wewnętrznym:

> R: Operacyjna praca polega na tym, że dostajemy tak zwany use case czyli po prostu jakieś problemy, zapytania od, z grupy, z grupy matki że tak powiem. Taki telecom [śmiech] mogę to zdradzić, to nie jest tajemnica. Naszym jakby pracodawcą, budżetodawcą jest [nazwa firmy] od nich dostajemy, dostajemy problemy, które mamy za pomocą narzędzi najlepiej jakiś takich bardziej zaawansowanych rozwiązywać. Bardziej zaawansowanych bo to potem oczywiście ładnie na prezentacjach wygląda [śmiech] ale ogólnie rzeczywiście posługujemy się takimi machine learningowymi rzeczami by to i rozwiązać, a potem jeszcze jakoś ładnie sprzedać. {wywiad 18}

Tak jak opisaliśmy wyżej, w kontekście opakowywania ML czy w ogóle DS w pojęcie AI, tak samo zaczarowywanie tej dziedziny poprzez odwołania do magii nie jest jednoznacznie uznawane za sprzyjające interesom badanego świata. Być może część rzeczników świata DS uważa, że marketing i sprzedaż projektów ML opakowywanych w AI i zaczarowywanych leży w interesie biznesowym. Wydaje nam się jednak, że sytuacja zmieniła się w czasie i to, co prowadziło do pozyskiwania kolejnych klientów w roku 2017 było nieskuteczne w roku 2019. Klienci stają się coraz bardziej wyedukowani i coraz bardziej rozczarowani, bowiem wiele niespełnionych obietnic co do magicznych, inteligentnych systemów funkcjonuje w świecie biznesu. Już od początku 2019 r. w świecie DS wzmacnia

się pozycja, w której świadomi, racjonalni klienci są postrzegani jako sprzyjający interesom badanego świata. Przyjrzyjmy się fragmentowi prezentacji wygłoszonej na jednym z największych wydarzeń polskiego świata DS (rys. 48), gdzie prelegent – uczestnik świata DS – przedstawiał taki właśnie punkt widzenia. Wyjaśniał, że klienci są przesyceni komunikatami marketingowymi, zaczarowani opakowaniem modeli ML oraz poziomem ich skomplikowania technicznego. Dlatego klienci, by pojąć sytuację, w której projekt DS jest realizowany w ich firmie, posługują się kilkoma strategiami (por. rys. 48), które data scientiści powinni rozumieć. Bazując na tym rozumieniu, zdaniem prelegenta, należy dbać o przejrzystość całego projektu DS jak i poszczególnych modeli ML, komunikować się z klientami bez użycia języka technicznego, rozmawiać raczej o finansowych, a nie technicznych metrykach sukcesu. Należy także używać metod XAI, bowiem pomagają budować zaufanie klienta do budowanych modeli. Taka strategia ma przyczyniać się do powodzenia projektu, to jest zakończenia go wdrożeniem przynoszącym zyski, a więc spełnienia oczekiwań klienta biznesowego {obserwacja 46}.



Rysunek 48: „Ooch, to tutaj jest mózg?" – antropomorfizacja jako strategia radzenia sobie z rozumieniem ML przez klientów

Źródło: Wystąpienie Aleksandra Kijka z firmy Nethone na konferencji DSS 2019 {obserwacja 46}, slajd w wersji rozszerzonej udostępniony Remigiuszowi Żulickiemu przez Aleksandra Kijka

[tekst]
Najbliższe temu, co znam ja:
Ponieważ szum medialny wokół ML często dotyczy sztucznych sieci neuronowych, ludzie mają tendencję do pojmowania działania ML poprzez sprowadzenie go do tego, jak działa mózg i jak ludzie rozgryzają sprawy. Prowadzi to do wielu nieporozumień lub, co jest bardziej problematyczne, do nierozsądnych/nierealnych oczekiwań.

ML to magia, którą robisz Ty:
Innym mechanizmem radzenia sobie [z rozumieniem ML przez klientów] to przyznanie sobie, że temat jest zbyt skomplikowany, by był sens się w niego zagłębiać. Prowadzi do problemu, w którym, choć potrafisz używać ML na wiele sposobów i ma znaczenie, jaki to sposób, to jeżeli masz przejść do rzeczy nie możesz wyeksponować wartości, jaką faktycznie wnosisz

Wszyscy robią ML – rozcieńczona wartość:
Ponieważ udowodniono, że ML może zwiększyć wydajność wielu systemów i procesów, stał się on „właśnie tym sposobem", motywując wszystkich usługodawców do używania ML. ML jest sam w sobie skomplikowany, a w dużej części rynku bardzo trudno jest zweryfikować, czy pod spodem rozwiązania faktycznie pracuje prawdziwy ML. To z kolei stworzyło szeroką przestrzeń dla interpretacji tego, co stanowi o wykorzystaniu ML i ułatwiło reklamę jego zastosowania w różnych produktach.

Majsterkowanie nie jest określeniem używanym in vivo w społecznym świecie DS. To nasza kategoria analityczna wzorowana na klasycznym ujęciu *bricolage'u* (majsterkowania) i *bricoleur* (majsterkowicza) Claude'a Lévi-Straussa. Konsultowaliśmy zastosowanie tej kategorii do opisu pracy świata DS w wielu rozmowach nieformalnych, zyskując akceptację uczestników badanego świata. Szczególnie pomocny w konsultacjach był komiks z serii „xkcd" autorstwa Randalla Munroe, pokazany przez badacza (RŻ) na jednym z meetupów {obserwacja 33}, a znany nie tylko w DS, ale i innych światach technicznych (por. Zaród

2018). Postaci komiksu rozmawiają o w pełni zautomatyzowanym procesie zbierania i analizy danych (*pipeline*; dosł. rurociąg – w polskim świecie DS nie tłumaczone), i komentują, że ten proces to domek z kart, zbudowany z przypadkowych skryptów, który zawali się przy jakimkolwiek dziwnym wsadzie (*input*). Nasi rozmówcy reagowali na komiks lub jego opis albo rozbawieniem, albo konstatacją „niestety to prawda", czasami łącząc majsterkowanie z wczesnym stadium rozwoju świata DS (por. 6.1).

W socjologii wielokrotnie stosowano kategorie majsterkowania / majsterkowicza w odniesieniu do działalności w obszarze nowoczesnych technologii informacyjnych, nowoczesnej nauki i inżynierii. Zgadzamy się z tym, że także w DS występuje działalność polegająca na:

- używaniu gotowych materiałów i dostosowywania ich do swoich potrzeb, także niezgodnie z pierwotnym przeznaczeniem (Szpunar 2012:79–80);
- pragmatycznej manipulacji narzędziami i aparaturą, w celu uzyskania działających i reprodukowalnych układów, przyjmującej postać wypróbowywania różnych konfiguracji materiałów i technik, czemu nie musi towarzyszyć refleksja teoretyczna (Afeltowicz i Pietrowicz 2008:45–46).

Podobnie jak Zaród, potwierdzamy (na gruncie DS) tezy Gabrieli Coleman i Johana Söderberga co do tego, że tworzenie oraz modyfikacja narzędzi są niezbędne w stawaniu się i byciu uczestnikiem badanego świata (co zaobserwowano w tak różnych zbiorowościach profesjonalnych, jak inżynierowie, hakerzy, badacze i drwale), oraz że dla uczestników badanego przez nas świata – tak jak dla hakerów – ważna jest nie tylko wolność modyfikowania narzędzia, ale i uznanie autorstwa (Zaród 2018:340). Praca świata DS jest dla nas majsterkowaniem jeszcze szerzej pojętym: majsterkowaniem zarówno w sferze narzędzi, w sferze metod, jak i sferze praktycznego celu projektów DS, a nawet w sferze prawnej i etycznej. Majsterkowanie w DS nie jest jednak czymś mitycznym, przeciwstawnym nauce i inżynierii, jak chciał Lévi-Strauss. Jest specyficzną jakością, odróżniającą DS od działania w nauce akademickiej, jak i w inżynierii (np. oprogramowania, *software development*). We współczesnych ujęciach kategorię majsterkowania oczywiście odnosi się do nauki i inżynierii (por. Afeltowicz i Pietrowicz 2008), jednak z naszych badań wynika, iż uczestnicy świata DS odróżniają swój świat od tych obszarów także poprzez odwołanie – nie wprost – do bycia światem „bardziej majsterkującym". Znajduje to wyraz w postrzeganiu tych obszarów, jako

różniących się od DS w sferze takich wartości, jak ciekawość, samorozwój, swoboda, samodzielność (odmiennie dla obszaru nauki i obszaru IT, por. rys. 38).

6.2.2. Data scientistka

Autorzy wielokrotnie cytowanego przez nas raportu AI Now Institute (część 2.2.3.) podkreślają, że specyficzna kultura firm technologicznych i zajmującego się tzw. AI środowiska naukowego nie sprzyja kompleksowemu podejściu do problemów etyki systemów AI. Twierdzą, że systemy AI utrwalają (*perpetuate*) uprzedzenia i dyskryminacje panujące także w samych firmach. Autorzy mówią o braku różnorodności, szczególnie płciowej, i powtarzających się praktykach nękania (*harassment*), wykluczania i nierówności płacowych, które są nie tylko szkodliwe dla pracowników tych firm, ale wpływają na oparte na AI produkty, które te firmy tworzą (Crawford i in. 2018:42).

Opisywaliśmy już znany, także w badanym świecie DS, przykład defaworyzującego kobiety systemu, używanego w procesie rekrutacji do firmy Amazon (Crawford i in. 2018:38-39). Wskazuje się też na ilościową dysproporcję płci w branży, oczywiście nie tylko w Polsce. Na trzech najważniejszych konferencjach poświęconych uczeniu maszynowemu w 2017 r. kobiety stanowiły 12% uczestników, a sytuacja nie zmieniła się zasadniczo od lat 90. XX w., kiedy w badaniu osób publikujących w poświęconym AI czasopiśmie naukowym „IEEE Expert" kobiety stanowiły 13%. W zespołach zajmujących się tzw. AI w największych firmach sytuacja jest podobna. Kobiety stanowią odpowiednio 15% i 10% pracowników w Facebooku i Google'u. Wśród absolwentów studiów magisterskich w dziedzinie informatyki (*computer science*) w USA było 18% kobiet (Crawford i in. 2018:38). To akurat nie pokazuje całości obrazu, ponieważ uczestnicy świata DS rekrutują się spośród absolwentów wielu kierunków (por. 5.1.2. i 6.1.1). Podobnie wskazują na to badania rynku pracy w USA. Tylko 20% respondentów ukończyło studia informatyczne (*computer science*), około 25% data scientistów to matematycy lub statystycy, 20% jest absolwentami nauk przyrodniczych, a aż 35% innych dyscyplin (Burtch 2018:20). Frakcja kobiet ogółem jest jednak zbliżona do podawanych przez AI Now, i wynosi 15%, przy czym jest najwyższa dla grupy z najkrótszym doświadczeniem zawodowym i spada wraz z liczbą lat doświadczenia (ibidem:28-29). Analiza ankiet związanych z polskim światem społecznym data science wskazuje na udział kobiet od 5% do 30% w zależności od źródła danych (por. rys. 7.) i podobną zależność udziału kobiet pod względem wieku (rys. 11). Niemniej tezy autorów raportu o odzwierciedlaniu kultury firm technologicznych w ich produktach mogą być słuszne. O tym,

że technologia uosabia wartości, badacze z dziedziny STS wypowiadali się już od dawna (Friedman i Nissenbaum 1996; Nissenbaum 2001; Winner 1980), zaś w swoim niepowtarzalnym języku o technologii jako utrwalonym społeczeństwie Bruno Latour pisał w 1991 r. (Latour 2013). W innej pracy Crawford także zajmowała się tym problemem (Crawford 2016). Jak dalej pokażemy, podobne przekonanie występuje w badanym świecie DS. Działania nastawione na włączanie kobiet i innych mniejszości genderowych do świata DS uzasadniane są m.in. w kategoriach ulepszania budowanych produktów / realizowanych projektów. Za takim uzasadnieniem stoi przekonanie, że skład zespołów DS ma wpływ na charakter budowanych przez zespoły rozwiązań, i że te rozwiązania będą bardziej różnorodne (*diverse*) i w konsekwencji „lepsze" wówczas, gdy w zespołach będzie więcej niż obecnie osób reprezentujących mniejszości (z naszych badań wynika, że w Polsce chodzi głównie o kobiety, nie widzimy innych mniejszości w dyskursie).

W sensie jakościowym to, że kobiety stanowią 10% – 20% osób uczestniczących w badanym świecie oznacza, że **kiedy w zespole DS pracuje kobieta, to niemal zawsze jest to jedna kobieta wśród kilku mężczyzn**. Z naszych badań wynika, że w Polsce zespoły złożone z większej liczby kobiet, w większości z kobiet lub z samych kobiet powoływane są raczej przez uczestniczki organizacji i inicjatyw nastawionych na włączanie mniejszości genderowych do DS (jak R-Ladies, PyLadies, WiMLDS), np. w celu uczestniczenia w hackatonie lub jednorazowo na warsztatach / szkoleniach. W komercyjnej sferze DS takie zespoły zdarzają się bardzo rzadko. Z zespołem DS, na który składały się trzy kobiety pracujące jako data scientistki i trzech mężczyzn na stanowiskach biznesowych, spotkaliśmy się tylko raz podczas konferencji {obserwacja 26}. Ten zespół z firmy McKinsey, jednej ze sponsorujących konferencję, w trakcie swojej prelekcji werbalnie, bezpośrednio podkreślał wyjątkowość składu swojego zespołu, zalety pracy w takim różnorodnym zespole DS. W wywiadach swobodnych spotykaliśmy się raczej z wypowiedziami:

> *B: Mhm, okey, a jak ci się ogólnie pracuje w tej branży jako kobiecie?*
> R: Bardzo dobrze. [pauza] Dobrze pod tym względem że znaczy myślę że rzeczywiście mi się zawsze dobrze z mężczyznami pracowało. Nie miałam nigdy jakichś takich problemów, i nie czułam się nigdy specjalnie dyskryminowana. I wręcz widzę że raczej w zespołach, ludzie się cieszą że, że tu mają jakąkolwiek kobietę, że to jest tak że jak jest 5 czy 8 mężczyzn w zespole i rzeczywiście zawsze byłam jedyną kobietą. Niezależnie czy to w tej pierwszej firmie, czy tej konsultingowej małej, czy teraz w tym trzecim zespole, zawsze byłam jedyną. I [śmiech] to było no dość mocno że tak powiem niekonwencjonalne, ale jak robiłam etat i przechodziłam rozmowę rekrutacyjną do teraz tej aktualnej pracy, no to rekruterzy, czyli z działu HR, gość, powiedział mi że [pauza] no mam spore szanse i też ze względu na płeć mam spore szanse [śmiech] że jednak chciałby żeby chociaż jedna kobieta w zespole była. Oczywiście **nigdy** nie chciałam żeby moje powiedzmy mój udział w zespole był oceniany ze względu na płeć, tylko na

> umiejętności, ale mimo wszystko zawsze się dobrze w takim środowisku czułam. {wywiad 18}
>
> *B: OK. Ale to teraz w swoim zespole jest pani jedyną kobietą?*
> R: Tak i nic nie wskazuje na to żeby było nas więcej. {wywiad 25}

Istnieją, także w Polsce, organizacje i inicjatywy nastawione na włączanie do DS mniejszości genderowych (*gender minorities*): kobiet tzw. cisseksualnych, czyli identyfikujących się z żeńską płcią przypisaną (*ciswomen*), kobiet i mężczyzn transseksualnych (*trans women*, *trans men*), osób niebinarnych (*non-binary, genderqueer*), osób agenderowych (*agender*) (DataCamp 2018; PyLadies 2019; R-Ladies 2019; WiMLDS 2018). Zauważamy zniuansowany język określający tożsamość płciową/genderową – połączyliśmy określenia obecne głównie w materiałach R-Ladies i DataCamp. Aktywistka i data scientistka Shaikh Reshama wskazuje, że bardziej różnorodne i aktywne na rzecz różnorodności jest środowisko osób posługujących się językiem R niż językiem Python (Reshama 2018). Niemniej w Polsce największe konferencje użytkowników Pythona w DS / ML i użytkowników R (PyData 2018; Why R? Foundation 2018) posiadają niemal identyczne kodeksy postępowania (*code of conduct*), gdzie szczegółowo wymieniono akceptowane zasady:

> PyData ma na celu zapewnienie udziału w konferencji wolnej od prześladowań każdemu bez względu na płeć, orientację seksualną, tożsamość płciową i ekspresję, niepełnosprawność, wygląd fizyczny, rozmiar ciała, rasę czy religię. Nie tolerujemy nękania uczestników konferencji w żadnej formie. Cała komunikacja powinna być odpowiednia dla profesjonalnej publiczności, w tym dla osób o różnym pochodzeniu. Seksualny język i zdjęcia nie są odpowiednie dla żadnego miejsca konferencji, w tym wystąpień. Bądź miły dla innych. Nie obrażaj ani nie poniżaj innych uczestników. Zachowuj się profesjonalnie. Pamiętaj, że nękanie i seksistowskie, rasistowskie lub wykluczające dowcipy nie są odpowiednie dla PyData. Uczestnicy naruszający te zasady mogą zostać poproszeni o opuszczenie konferencji bez zwrotu kosztów według własnego uznania organizatorów konferencji. Dziękuję, że pomogliście uczynić z tego przyjemne i przyjazne wydarzenie dla wszystkich (PyData 2018).

Jednocześnie najważniejsza, odbywająca się corocznie od 1987 konferencja w dziedzinach AI i uczenia maszynowego „Neural Information Processing Systems", znana jest jako NIPS. Słowo „nips" jest m.in. slangowym określeniem sutków (*nipples*), jak i obraźliwym określeniem osób pochodzenia japońskiego (od Nippon – Japonia w jęz. japońskim) (Hu 2018). Konferencja jest popularna: około 2,5 tys. biletów na edycję 2018 wyprzedano w niecałe 12 minut (Koch 2018). W 2018 jeszcze przed konferencją z inicjatywy John Hopkins University wystosowano petycję wzywającą do zmiany nazwy (Shead 2018; Simonite 2018). Wskazywano na seksistowskie incydenty w poprzednim roku: nieformalna impreza przed rozpoczęciem konferencji nazwana była akronimem TITS (co slangowo

oznacza cycki) (Hu 2018). Elon Musk wzbudził aplauz mówiąc ze sceny: „no NIPS without TITS" (nie ma sutów bez cycków) (Koch 2018), zaś nazywający się „zróżnicowanym zespołem" pracownicy firmy Dessa nosili i sprzedawali koszulki z napisem „My NIPS are NP-hard" (moje suty są NP-twarde[278]) (Koch 2018; Simonite 2018). Rada konferencji w odpowiedzi na petycję najpierw przeprowadziła ankietę wśród uczestników z ostatnich pięciu lat, a w jej wyniku postanowiła nie zmieniać nazwy (Koch 2018; Simonite 2018). Ostatecznie, po kolejnej petycji z inicjatywy Animy Anandkumar – dyrektorki badań firmy NVIDIA[279] nazwę jednak zmieniono na NeurIPS (Koch 2018; The Neural Information Processing Systems Foundation Board of Trustees 2018). Konferencja wprowadziła też kodeks postępowania (*code of conduct*) pierwszy raz w 2018 (Merity 2017; NeurIPS 2018).

Działania wspominanych grup, przynajmniej przez najbardziej nagłośnioną pozycję w dyskursie opisywanej areny dotyczącej data scientistek, są oceniane pozytywnie. Rozmówcy argumentowali, iż kobiety mają trudniejszy niż mężczyźni dostęp do świata DS z powodu rozmaitych stereotypów i wieloletniej socjalizacji raczej do ról nie-technicznych. Grupy nastawione na włączanie mniejszości genderowych mają dawać kobietom (głównie wannabesom i początkującym) poczucie przynależności, bezpieczeństwa, dostarczać przykładów innych kobiet silnie osadzonych w badanym świecie (*role models*):

> *B: Rozumiem, no dobra. Wiesz co no to kończąc już powoli, chciałem cię spytać o coś takiego [pauza] czy data science i analiza danych to jest męski zawód?*
> R: [śmiech] Nie, nie wydaje mi się żeby w jakikolwiek sposób, znaczy no, zależy co rozumiemy przez to czy to jest męski zawód, nie? No bo to że jest, że jest dysproporcja płciowa w zawodach technicznych i na politechnikach i ogólnie w dziedzinach ścisłych, no to jest fakt, i tego, temu się nie da zaprzeczyć, ale nie ma moim zdaniem, nie ma nic w data science co by powodowało że to miałoby być bardziej męskie, tak, broń Boże. Znaczy no ja mimo, znaczy mimo wszystko no, bo ja jestem zasadniczo mocno przeciwny jakiemukolwiek tego typu, kategoryzacji czy właśnie, eem no nie wiem, to się po angielsku nazywa gatekeeping, ale nie wiem jak to jest po polsku, żeby. Jestem bardzo przeciwny żeby komuś zabraniać, albo utrudniać pracę w jakimś zawodzie, bo jest innej płci, czy innej nie wiem innej rasy, czy pochodzi z innej klasy społecznej, tak więc to. Nie wydaje mi się że, no nie, że to no nie, data science nie jest typowo męskie, ale no jest tam więcej mężczyzn tak jak w każdej dziedzinie ścisłej niestety nadal.
> *B: Nie, no bo pytam o to z tego względu, bo widziałem niedawno że powstał taki projekt Women in Machine Learning and Data Science i to jest coś takiego jak R-Foundation, że no właśnie dają kasę na jakieś meetupy, czy czy, czy tego typu rzeczy. No i też jest to R-Ladies*

278 Lub „NP-trudne", bo żart ma jednocześnie konotacje anatomiczne i matematyczne. Te drugie dotyczą tzw. NP-trudnych problemów obliczeniowych (*NP-hardness; non-deterministic polynomial-time hardness*). Tłumaczymy jako „twarde", by podkreślić, że w przywołanych źródłach zaznaczano seksistowski wymiar żartu.

279 Firma będąca producentem m.in. popularnych procesorów graficznych, tzw. GPU, które standardowo służą do wyświetlania wysokiej jakości grafiki (np. trójwymiarowej w grach komputerowych). W uczeniu maszynowym, szczególne deep learningu, obliczenia wykonuje się często na GPU zamiast na CPU, czyli podstawowym procesorze komputera, z uwagi na znacznie krótszy czas trenowania takiego samego modelu na GPU niż na CPU.

R: No tak R-Ladies jest bardziej duże no.
B: No i właśnie się, właśnie, właśnie się zastanawiam, no czy kobiety są dyskryminowane i w odpowiedzi na to są tego typu inicjatywy, czy czy no nie wiem, jakoś zbytnio to upraszczam?
R: Mhm, mhm, znaczy no, no tak. Moim zdaniem wszelkie tego typu, mmm inicjatywy no są odpowiedzią na taką [pauza] no **systemową i kulturową dyskryminację kobiet**, tak, w tego typu dziedzinach, no czy ogólnie no, ogólnie w życiu społecznym i wszędzie. Więc no jak najbardziej, no to się nie bierze znikąd tak, no ale tego typu inicjatywy są odpowiedzią na to że, że no nasze społeczeństwo, no nie tylko tutaj mówię o naszym polskim społeczeństwie, ale o większości społeczeństw na świecie działa tak że no nie wiem, dziewczynom i kobietom jest jakoś trudniej zacząć się interesować w ogóle na przykład dziedzinami ścisłymi, eee więc no to, no to sięga bardzo głęboko nie, w sensie no moim zdaniem, tak? No bo to nie tylko jest, to nie jest dyskryminacja na poziomie zawodowym, no to jest dyskryminacja taka kulturowa i systemowa która gdzieś tam się zaczyna w przedszkolu, tak? Więc no to, te inicjatywy moim zdaniem bardzo dobrze że powstają i, no iii tylko trzeba życzyć powodzenia, no bo robią dobre rzeczy, tak? No bo to faktycznie, tak to wygląda że [pauza] znaczy no też jest, też jest, kurde no to, znaczy to można gadać i gadać, no ale. To też tak wygląda że nie wiem, na przykład jest jest meetup eRowy, ja też trochę mam sobie do zarzucenia w kontekście naszego meetupu [nazwa miejscowości], no bo, pewnie można było więcej robić w kierunku jakiegoś tam zachęcania bardziej dziewczyn czy kobiet do tego, do uczestniczenia. No bo w momencie jak przychodzisz, jesteś kobietą i przychodzisz na meetup, znaczy nie wyobrażam sobie, tylko hipotetyzuję, i nie jestem sobie tego w stanie wyobrazić w 100%, ale przychodzisz na meetup i jest tam nie wiem 30 osób, z czego 28 to goście, no i się kończy meetup, jest zaproszenie na piwko, no i ci goście sobie siadają i sobie tam w swoim gronie gadają o rzeczach. No to, no to na pewno się też można czuć jakoś wykluczonym nie? Więc też te inicjatywy pomagają no chociażby w tym żeby przyjść na spotkanie gdzie jest większość kobiet i po prostu mogą się wtedy, no mogą się one wtedy czuć bezpieczniej i jakoś tak bardziej komfortowo iii, no właśnie kiedy wiedzą że się nie spotkają, no bo [krótka pauza] kurwa no, naprawdę o tym by można gadać i gadać. No ale no, w sensie wyobrażam sobie że, że na przykład nawet na naszym meetupie może być że jak przyjdzie jakaś dziewczyna która na przykład jeszcze się nie zna w ogóle na tym temacie, iii no to na przykład może się bać zadać jakieś pytanie prelegentowi, bo spotka się eee wiesz ze spojrzeniem ze politowaniem ze strony jakichś tam 5 gości którzy się tym zajmują od 3 lat nie, no a jak przyjdzie na spotkanie gdzie jest 30 kobiet to będzie to jakoś bardziej, bardziej dla niej komfortowe. No i też te, też te inicjatywy, no R-Ladies na pewno działają właśnie na takim polu podstawowym że robią warsztaty dla początkujących też w dużej mierze, nie? Żeby zachęcić w ogóle do zaczęcia jakiegoś zainteresowania tym tematem, więc to też jest mega spoko, no bo takie meetup no to też się charakteryzują tym że często, często są już dla ludzi którzy siedzą w jakimś temacie, nie? Więc to jest mega spoko też. {wywiad 9}

Spotkaliśmy także w tym kontekście nawiązanie do kategorii nerda, postrzeganego m.in. jako osoba mało „towarzyska", zainteresowana raczej abstrakcjami lub technologiami niż ludźmi, a będącą punktem odniesienia dla świata DS (por. rys. 43). Kobiety o tego typu charakterystyce są, zdaniem Rozmówczyni, jeszcze bardziej niż podobni mężczyźni poddawane presji otoczenia, muszą wykazać się jeszcze większą niż oni wytrwałością w rozwijaniu swoich jakoby „nienormatywnych" w ogóle, a szczególnie już dla kobiet nietypowych zajęć i zainteresowań:

B: A jak pani sądzi, czy coś dają te inicjatywy grup właśnie przeznaczonych dla kobiet, typu R-Ladies czy tam PyLadies? Są też w Polsce
R: Tak, myślę, że tak, ponieważ to ośmiela kobiety do tego, żeby się z kimś identyfikować. Zawsze silniej czujemy się w grupie. Oczywiście, no to jest trudne dla dziewczyn z małych

> miasteczek, one, bo mają do tego kiepski dostęp. Natomiast jeśli ktoś jest z dużego miasta, gdzie taka grupa funkcjonuje mniej lub bardziej sprawnie, no to zawsze można się poczuć członkiem jakiejś społeczności. Nie czuje się człowiek taki, taki **wyautowany**. Wie pan, ja mam wrażenie, że naprawdę większość kobiet, które dzisiaj pracują zawodowo w tych technicznych zawodach tak w miarę wysoko, znakomita większość z nas, gdyby nas poddać diagnostyce psychiatrycznej, okazałoby się, że mamy spektrum autyzmu. Takie jest moje podejrzenie, to nie jest poparte niczym. Natomiast takie jest moje podejrzenie, bo kto normalny oparłby się tym wszystkim nagabywaniom, namowom, przekonywaniom? [pauza] Musiało być coś, co spowodowało, że to wszystko nam poszło gdzieś mimo uszu. No bo nie ma siły. Ja jestem Aspergerem i spodziewam się po prostu, że większość z nas ma tego rodzaju jakieś tam, ee, tam, jesteśmy nieneurotypowe. W przypadku mężczyzn matematyków czy informatyków takich, z którymi mam do czynienia to też to obserwuję, bo ich jest po prostu więcej. To są jednostki, które sprawiają wrażenie zupełnie normalnych. Większość z nas, tak jak patrzę, ma jakieś swoje skrzywienia. One są nieszkodliwe, one nam pozwalają jakoś tam funkcjonować jako dorosłym, ale większość z nas jako dzieci, jak się tak porozmawia, czuła się kompletnie wyautowana, z takiego czy innego powodu nie była w tej takiej maistreamowej grupie, tylko gdzieś z boku. Takżeee tego się spodziewam, że większość kobiet, które się dzisiaj przebiły, gdzieś tam ma coś za uszami trochę inaczej. Nie jest to szkodliwe, pozwala nam to normalnie funkcjonować, ale pozwoliło nam też dojść tu, gdzie jesteśmy pomimo tego, że byłyśmy kierowane, no, może ty jednak byś została fryzjerką, kosmetyczką, może pójdziesz do szkoły uczyć matematyki. No nie, nie pójdę, to mnie nie interesuje.
> *B: Rozumiem. Ale to nie dotyczy tylko tej no relatywnie nowej branży data science, tylko w ogóle tych techniczno-matematycznych dziedzin, tych dziedzin STEM tak zwanych, tak?*
> R: Tak mi się wydaje. {wywiad 25}

Na czym polega arena dotycząca data scientistek – kobiet w świecie społecznym DS? **Spór toczy się wokół tego, jak w badanym świecie traktować kobiety**. Obiektem granicznym tej areny jest płeć. Najbardziej widoczna i pozornie jedyna pozycja (a z pewnością najbardziej nagłośniona, poprawna politycznie, a jak sądzimy także z uwagi na poprawność bez obaw wyrażana badaczowi) w dyskursie badanego świata to, jak wspominaliśmy, dążenie do budowania bardziej różnorodnych zespołów DS, które mają budować lepsze produkty i realizować lepsze projekty, bo wypracowują rozwiązania jakoby właśnie bardziej różnorodne, nie „zbiasowane" w kierunku perspektywy młodych, białych mężczyzn – grupy ilościowo bezwzględnie dominującej. W naszej opinii jest to pozycja reprezentowana np. przez członków grup nastawionych na włączanie mniejszości genderowych do DS (lub tylko przez rzeczników tych grup), jak i przez biznesowych rzeczników świata DS (bez wątpienia można wyróżnić osoby zajmujące się HR, raczej także ludzi odpowiedzialnych za wizerunek firm i za marketing). Pozycja ta ma więc uzasadnienie ekonomiczne, utylitarne, nakierowane na efektywność, zatem jest zgodna z wartościami świata DS. Odwołuje się także do wartości humanistycznych, takich jak etyka, w kontekście odpowiedzialności etycznej za sprawczość, jaką mają ludzie realizujący projekty DS i budujący elementy systemów bazujących na ML. Różnorodne zespoły mają przyczyniać się do tego, że budowane rozwiązania, przede wszystkim modele ML, będą mniej

dyskryminować czy krzywdzić osoby z rozmaitych mniejszości. Niemniej osią argumentacji jest uzasadnienie utylitarne. Przyjrzyjmy się wypowiedzi Rozmówczyni, będącej organizatorką jednego z meetupów nastawionych na włączanie mniejszości genderowych do DS:

> *B: Dlaczego to jest w ogóle ważna sprawa, prawa kobiet w data science?*
> R: Prawa kobiet w data science? Znaczy generalnie
> *B: Bo to jest ta sprawa tak?*
> R: Wiesz cooo, ja [krótka pauza] tak jak mówiłam na początku że jestem daleka od etykietek, jak mówiliśmy o data scientistach
> *B: Ok, tak.*
> R: To myślę że tutaj patrzę podobnie, czyli generalnie eee prawa kobiet, tak, ale spojrzałabym na o szerzej, generalnie prawa ludzi, komfort pracy, rozwój technologii w kierunku etycznym? Chodzi o to że, mmm w zasadzie tutaj są dwie kwestie, tak jakby spojrzeć nawet na to z biznesowego punktu widzenia, to firmy które wdrażają tak zwane diversity mają wysoki ten współczynnik różnorodności pod pod różnymi aspektami, to są jednocześnie te firmy które mają najwyższą eee, skuteczność yyy gospodarczą. [niezrozumiałe] robię kalki czasem z angielskiego, no bo te raporty są po po angielsku, tak
> *B: Jasne.*
> R: Ale chodzi o to że, że że to są te firmy, które na przykład w poziomie innowacyjności eee przodują między innymi dlatego że [westchnienie] algorytmy, które tworzą eee osoby tak, zajmujące się machine learningiem od odzwierciedlają tak na prawdę te biasy[280], czyli te różne uproszczenia, czy skrzywienia eee percepcyjne, czy przetwarzania informacji, które mają ludzie. I co raz częściej się okazuje że jeżeli na przykład, nie wiem czy czytałeś o tych algorytmach chociażby dotyczących [niezrozumiałe] rozpoznających twarze, nie identyfikowały osób czarnoskórych jakooo yyy, jakoo ludzi. To w kontekście samochodów autonomicznych jest nie do przecenienia, więc tu nawet nie chodzi tylko o kwestię tego co nazwiemy dyskryminacją w takim a nie innym sensie i takie dyskusje teoretyczne, tylko to że mmm, większość ludzi w data science albo buduje produkty, które mają realny wpływ na otaczający ich w jakiś sposób świat, albo pracuje naukowo nad problemami które też nas dotyczą, bądź będą dotyczyły za, za jakiś czas. No i pytanie czy chcemy tworzyć społeczność ludzi odpowiedzialnych, eee którzy no nie boją się tych różnych dylematów tak, etycznych, społecznych i jakoś dyskutują o kwestiach trudnych, czy nie. I wydaje, ja staram się na to szerzej właśnie patrzeć, w taki sposób tak, bo to nie chodzi tylko o o o prawa kobiet, tylko o o, o genera, o generalnie [ze śmiechem] oo prawa ludzi, tak. Ja bym chciała żeby nie było z tym takiego problemu jak problem praw kobiet, żeby to było zupełnie naturalne, żee funkcjonują pracownice [ze śmiechem] o wysokich kompetencjach, czy pracownice uczące się wdrażające po prostu na takich samych zasadach nie, no niezależnie od, właśnie nie wiem, płci, pochodzenia, orientacji seksualnej, poziomu niepełnosprawności i cokolwiek sobie yyy wybierzemy. Natomiast no wiem, że nie żyjemy w utopii tak, czy w realiach utopijnych i też w takim świecie nie, nie, nie będę żyć, więc stąd mmm próba yyy działania takiego, która chociażby trochę nas do tego przybliża, żeby osoby które miały, które mają trudniej z jakiegoś powodu, mmm mogły mieć wyrównane yyy wyrównane te możliwości, na tyle na ile to jest realne. {wywiad 21}

W dalszych rozważaniach tę pozycję nazywamy pozycją różnorodności.

W polskim świecie DS faktycznie nie spotkaliśmy się z problematyzowaniem innych mniejszości niż kobiety. Istnieją natomiast prace dotyczące defaworyzowania w DS z uwagi

280 Od słowa *bias*, dosł. obciążenie, skrzywienie.

na rasę i pochodzenie, co nazywane jest formą kolonializmu. Chodzi o defaworyzowanie nie tylko w ramach szeroko opisywanego w naukach społecznych (por. 2.2.3.) działania wobec użytkowników końcowych, czy tzw. klikaczy, ale także wobec osób zajmujących się DS (Crawford, West, i Whittaker 2019; Mohamed i in. 2020). Stanowisko wyróżnionych autorów także odpowiada pozycji „różnorodności".

Przeciwstawna pozycja zakłada doskonałą merytokrację, w której nie liczą się takie cechy jak płeć, a tylko i wyłącznie kompetencje techniczne i metodologiczne konkretnych osób. Jest to także zgodne z wartościami badanego świata DS. Pozycja „różnorodności" jest z tej pozycji krytykowana za sprzeciwianie się merytokracji, za aktywne wspieranie np. kobiet, wspieranie wypływające z samego faktu bycia kobietą. Z tej pozycji to, że w DS i w ogóle w technologiach jest mało kobiet postrzega się jako różnorako uwarunkowany fakt, którego nie można zmienić decyzjami np. w dziale HR. Nie występują żadne sygnały o postrzeganiu którejkolwiek płci jako lepszej / gorszej w DS – płeć ma być cechą nieistotną:

> *B: Ok. A jeszcze mnie to ciekawi, jak to jest z płciami, u ciebie w pracy?*
> R: [pauza] Mmmm, no z płciami jest, znaczy nie ma 50 na 50, yaaa **ale** jest powiedzmy coś koło 20, 30% kobiet, co myyyyślę że względem odsetku **kobiet**, które ogólnie są w branży, ee myślę jest nawet nadreprezentacją. Mmm, no teraz na przykład przy tej rekrutacji, ee zgłosiła się chyba w sumie, na 40 cefałek jedna kobieta tylko, która jak dostała wstępne zadanie, mmm przestała się odzywać.
> *B: Przestała się odzywać?*
> R: Tak.
> *B: Ok.*
> R: Więc no nawet jakbyśmy chcieli podwyższać, yy no to nie zbyt, znaczy nawet jakbyśmy na upartego chcieli wziąć kobietę, na podstawie tej rekrutacji to nawet by się nie dało.
> B*: Mhm, czyli co, jest jest bardzo ciężko znaleźć, kobiety, tak?*
> R: **Tak**. [pauza] Mmm, myślę że w branży jest [pauza] tylko pytanie właśnie ile może być w branży, bo na przykład w Poznaniu działa, mmm ten meetup [nazwa organizacji wspierającej mniejszości genderowe w DS]. I **tam** uczestników, w sensie mężczyzn, jest zwykle 70, koło 80%.
> *B: Mężczyzn 80%?*
> R: W trakcie meetupów, w sensie wydarzeń.
> *B: Jasne.*
> R: I powiedzmy tam no, można by się spodziewać nadreprezentacji kobiet, nie? w sensie względem tego co jest w branży. Eea [długa pauza] czekaj jeszcze tutaj w jedno miejsce spojrzę
> *B: Ok.*
> R: W sensie, bo założyłem **swój** meetup ostatnio. I nawet nie zrobiłem żadnego wydarzenia jeszcze, a już mi się do niego zapisało 57 osób.
> B: O, fajnie. A jak się nazywa?
> R: [nazwa meetupu].
> *B: Ok.*
> R: Mogę ci nawet linka podesłać jak chcesz.
> *B: Będę śledzić.*
> R: [robi coś na komputerze]
> *B: O fajnie, dzięki. [otwiera link]*
> R: I tu na przykład sobie możesz przejrzeć jakie osoby się zapisały, nie? Prawie sami

> mężczyźni. Z kobiet jedna, dwie, czy tam, ale to są dwie organizatorki [nazwa organizacji wspierającej mniejszości genderowe w DS]. Eee, trzy [długa pauza] cztery [pauza] pięć, sześć, siedem, osiem. No. Osiem na 58 osób, to kobiety.
> *B: Tak. Rozumiem. Hmm, ok, ale czy, czy chciałbyś żeby było więcej kobiet?*
> R: [pauza] Znaczy, pffffff moim zdaniem to jest **każdego** wybór ścieżki życiowej, co chce robić.
> *B: Mhm, ale czy, czy dla ciebie w pracy, ma to jakieś znaczenie, czy czy czy zatrudnisz kobietę czy faceta?*
> R: Dla mnie nie. I mam nadzieję że nigdy nie będę **musiał**, na przykład decydować się że zatrudnię eee gorszą kobietę, niż inny kandydat mężczyzna, tylko dlatego że jest kobietą.
> *B: Mhm, ale to. Ok. Czy to znaczy że jest jakaś presja na kobiety, że nie wiem, twój szef chce kobiety?*
> R: Nie ma na szczęście, ale są, słyszałem że inne firmy mają mocne preferencje żeby zatrudniać, typu żeby wyrównywać. Znaczy nie tylko w data science, tylko ogólnie w technologii i tak dalej, nie?
> *B: Mhm, ale to ma być jakieś wizerunkowe, czy [pauza] nie wiem o co to chodzi?*
> R: [pauza] Znaczy dla **części** firm mi się wydaje że wizerunkowe. Na przykład w **nauce** to już jeden profesor mówił z 10 lat temu, że w grantach musi się tłumaczyć, dlaczego nie ma po równej liczbie mężczyzn i kobiet. Po prostu, skąd ma napłodzić tych kobiet.
> *B: Skoro ich nie ma, tak?*
> R: Znaczy w sensie na informatyce na przykład, u mnie na kierunku było [pauza] 5 kobiet na 130, 140 osób. {wywiad 19}

Osoby stojące na pozycji „doskonałej merytokracji" krytykują takie praktyki, jak kierowanie dofinansowania do konferencji / szkoleń DS tylko dla kobiet, podczas gdy mężczyźni płacą pełne stawki (DataCamp 2018). Powszechnie krytykowane są praktyki nieformalnego lub formalnego nacisku na to, by w składzie zespołu DS znajdowały się kobiety (co czasami nazywane jest kwotą, od *gender quota* – np. zespół HR zaproponował, żebyśmy dążyli do kwoty 30% kobiet w dziale DS). Te i podobne praktyki uznawane są z punktu widzenia opisywanej pozycji za niesprawiedliwe, nierówne. Uznawane są za obniżanie poprzeczki dla wybranej grupy, a w konsekwencji za działanie nie leżące w interesie świata DS, bo obniżające jego elitarność, jakość pracy, poziom biegłości technologicznej poprzez dopuszczanie do uczestnictwa osób, które na „normalnych" warunkach nie weszłyby do świata. Oczywiście z pozycji „różnorodności" owe „normalne" warunki są takie, że kobietom jest znacznie trudniej wejść do świata, obniżanie poprzeczki jest wyrównywaniem szans, rzekoma doskonała merytokracja jest tylko stwarzaniem pozorów egalitarności w imię zachowania elitarnego statusu grupy dominującej, a większa różnorodność leży jednoznacznie w interesie świata DS z uwagi na opisywaną już zależność między różnorodnym składem zespołów DS a lepszymi produktami i projektami.

> *B: Yyy, ok. Z tego co wiem, to data science jest męską branżą, w takim sensie ilościowym. Mówi się o kilkunastu procentach kobiet w populacji. A jak tobie się pracuje jako kobiecie? Czy to w ogóle ma jakieś znaczenie, płeć?*
> R: Wiesz coo, muszę ci dać trochę backgroundu, bo to też ja mam może niepopularne zdanie.
> *B: Interesuje mnie oczywiście twoje zdanie.*

R: Dlatego ci powiem, ale dam ci background, żebyś wiedział. Ja mam trójkę braci. Ja się wychowywałam w męskim towarzystwie, ja jak siadam z kumplem to on mówi nie rozumiem kobiet. A ja mówię, no, to jest nas dwoje [z uśmiechem]. Głównie mam też, no raczej obracam się w bardzo męskim towarzystwie i rzeczywiście więcej mam tych nawyków od płci przeciwnej niż od mojej płci. I teraz tak, **jest** mniej kobiet, co nie jest dziwne, bo po prostu historycznie, już abstrahując od branży, historycznie rola kobiety jednak społeczna rola była trochę inne, teraz trochę ewoluuje ta rola, ale ja bardzo nie lubię czegoś takiego, że teraz się mówi o boże, **nadal** mamy **tylko** 20% kobiet bo powinniśmy spojrzeć na czas, na przestrzeń czasową, że kilka lat temu to było 2-3%, tak? To, że my mamy teraz 20%, to jest sukces. Jeśli my zaczniemy teraz tak naciskać, jak zaczynają naciskać niektóre kobiety czy niektóre nurty, to my to tylko zepsujemy i ja to już mocno widzę. To, że firmy narzucają kwotę, że mają mieć 50% kobiet, to niszczy, to strasznie niszczy, bo po pierwsze budzi frustrację w mężczyznach. Ja uważam, że jedyną wykładnią powinny być kompetencje, nie płeć. I w momencie, kiedy mężczyźni widzą, że nie awansują, bo kwota jest taka a nie inna, mimo tego, że mają kompetencje. To jest frustrujące i ja im się nie dziwię. Druga sprawa, są też kobiety, które to wykorzystują. Ale z drugiej strony spójrzmy, no jest więcej mężczyzn. Czy spotkały mnie nieprzyjemności duże z tego powodu, że jestem kobietą, na mojej przestrzeni kariery? Tak. Ze strony mężczyzn i ze strony kobiet, także z obu tych stron. Ja zarówno pracowałam dla firm, które bardzo promowały kobiety, jak i w których było bardzo mało kobiet i był większy nacisk na mężczyzn i uważam, że i tu, i tu są dobre i złe strony. Ja uważam, że my powinniśmy ewolucyjnie po prostu do tego podejść i tak rozsądnie, naprawdę bazując na danych. Czyli tak, jest mniej kobiet, ale trend jest taki, że jest ich coraz więcej i tak z ciekawostek właśnie ta firma, w której właśnie teraz pracuję, nie ma kwoty. Nie ma kwoty i jest 27% kobiet, co pokazuje, że po kompetencjach tylko i wyłącznie, zostało wybranych 27% kobiet i to pokazuje, że rzeczywiście jakby ten trend idzie w górę i ja to widzę po wszystkich statystykach, że to idzie. Ja nie lubię nacisku. A czy kobietom jest trudniej czy łatwiej? To ciężko powiedzieć. To bardzo zależy też, nie? Ja mam też silne zdanie i raczej umiem je wypowiedzieć, więc mi jest łatwiej. Czy to jest kwestia płci czy nie? Kwestia bardzo dyskusyjna. Mi jest z tym dobrze. To, że jest więcej kobiet, fajnie, ja też promuję, współzałożyłam też w Polsce [nazwa organizacji wspierającej mniejszości genderowe w DS]. Aktualnie nie jestem tam czynnie uczestnicząca, po prostu ze względu na braki czasowe, ale dziewczyny fantastycznie działają, w ogóle [nazwisko liderki organizacji] wymiata. I fajnie, że tak się wspieramy i że jest więcej kobiet, ale po prostu nie róbmy tego **na siłę**. To mi przeszkadza mówiąc szczerze, bardzo nie lubię tego, że to jest robione na siłę, bo ja tylko odczuwam tego negatywne skutki, wynikające z frustracji właśnie kobiet, mężczyzn. Zwłaszcza, że powiem ci szczerze, że coraz bardziej widzę, że jednak mężczyźni mówią kurczę, chcielibyśmy więcej kobiet, bo też widzimy w leadershipie, jest nawet taka książka teraz, Harvard Business Review wydał, a à propos właśnie leadershipu mężczyzn i kobiet. Kurczę, nie nie pamiętam nazwy teraz i sobie nie przypomnę. To też czytałam i to też była taka bazująca na statystykach przeróżnych, no i ja powiem jedno, warto mieć i kobiety, i mężczyzn, tak samo jak w ogóle warto mieć różne osoby, bo po prostu mamy różne podejścia, a jak wszyscy będą mieć to samo podejście, to tak naprawdę nie dojdziemy do najlepszego pomysłu, bo jednak ten proces wymyślania najlepszych pomysłów polega trochę na ich killowaniu[281]. Ja uwielbiamm [pauza] ja mam takiego szefa, do którego przychodzę z pomysłem i mówię [imię], rzucam ci pomysł do killowania. I on mnie bombarduje tam po prostu. Bombarduje i dzięki temu jesteśmy w stanie wyłuskać najlepszy pomysł. I tak to powinno wyglądać i rzeczywiście różnica spojrzeń kobiet, mężczyzn, jest. (...) Więc mówię, fajnie, że jest ten trend z kobietami. Myślę, że jak rozsądnie do tego podejdziemy to będzie ich coraz więcej. Druga sprawa, ja uważam też, że nie powinno się dawać presji. Jeśli któraś kobieta chce cieszyć się, zostać w domu i się w tym spełniać super, niech to robi. To też jest rola społeczna, więc jak najbardziej powinna czuć ten komfort, że może to zrobić. {wywiad 22}

281 Od *kill*, dosł. zabić

Obie pozycje – „różnorodność" i „doskonała merytokracja" – są zgodne z wartościami świata DS, jednak widzimy je jako wzajemnie sprzeczne stanowiska. Sprzeczne jest, by jednocześnie aktywnie promować i wspierać różnorodności oraz nie zauważać np. płci i stawiać tylko na kompetencje. Pozycja „różnorodności" koncentruje się na kategoriach pozakompetencyjnych (jak płeć), zaś pozycja „doskonałej merytokracji" postuluje wyciszenie, wymazanie takich kategorii ze świata społecznego. Pozornie te pozycje są zgodne, bowiem zgodnie z wartościami badanego świata jedna pozycja nie krytykuje drugiej w pełnej rozciągłości. Osoby z pozycji „różnorodności" także zgadzają się ze stanowiskiem, że merytokracja w świecie DS jest dobra, i że nie należy np. zatrudniać słabiej wykwalifikowanej kobiety kosztem lepiej wykwalifikowanego mężczyzny i mogą argumentować, że obecne wspieranie mniejszości służy temu, by w przyszłości można było pomijać kwestię płci – por. wcześniej fragment {wywiad 21} „ja bym chciała żeby nie było z tym takiego problemu jak problem praw kobiet, żeby to było zupełnie naturalne, żee funkcjonują pracownice o wysokich kompetencjach". Osoby z pozycji „doskonałej merytokracji" za to zgadzają się z tezą, iż bardziej różnorodne zespoły mogą realizować lepsze, bardziej różnorodne projekty i przyczyniać się do budowy lepszych, mniej „zbiasowanych" produktów. Zatem przez różnorodny skład zespołów (mówi się czasem nie o płci, ale np. o różnorodnym wykształceniu, doświadczeniu zawodowym, używanym stosie technologicznym itd.) jak najbardziej osiągana może być wyższa efektywność i nawet nie tylko wyższa, ale i bardziej etycznie wrażliwa sprawczość – która i tak przekłada się przecież na korzyści utylitarne jak zadowolony klient, brak kłopotów prawnych, zyski finansowe, ale także przyczynia się do zmieniania świata na lepsze (por. część 5.4.2. rozprawy).

Te dwie pozycje w dyskursie na opisywanej arenie pozostają bez związku z płcią uczestniczących w badanym świecie DS osób, które zajmują owe pozycje.

Opisany spór odbywa się na tle starych, jawnie w badanym świecie potępianych stereotypów o kobietach, jako jakoby naturalnie gorzej predestynowanych do obszarów ścisłych, technicznych. Jeżeli w świecie DS istnieją osoby przyjmujące ten stereotyp jako swój pogląd, to są one ukryte, nie zabierają głosu na arenie, nie można traktować ich jako pozycji w dyskursie. Bez wątpienia jednak uczestniczki badanego świata wciąż spotykają się z osobami obu płci reprodukującymi opisany stereotyp. Rozmówczynie opisywały sytuacje bycia stereotypowo traktowanymi jako osoby nie-techniczne ze strony osób ze świata biznesu (klientów wewnątrz / zewnętrzne jak i koleżanek i kolegów z innych działów) oraz akademii (jako studentki przez profesorów, jako współpracowniczki przez koleżanki i kolegów po

fachu). Co najmniej na całej arenie modelowania (por. 6.2.1.) **występuje zjawisko podważania kompetencji technicznych kobiet wywodzone z samego faktu bycia kobietą**. Przyjrzyjmy się dłuższej wypowiedzi Rozmówczyni, osoby silnie osadzonej w badanym świecie i posiadającej kilkanaście lat doświadczenia zawodowego w akademii i biznesie.

> *B: Chciałem jeszcze spytać o taką rzecz. Bo czytam, że generalnie w branży data science jest około 15% kobiet, tak? Niektórzy piszą o 20%. A jeszcze, jak się dowiedziałem, jest pani absolwentką informatyki, gdzie też się mówi o o o bardzo małej liczbie kobiet. Chciałem zapytać, jak się pani pracuje, jak się pani czuje w takim w cudzysłowie męskim świecie?*
> R: Ciężko. **Zawsze** było ciężko. Tak, jak pan mówi. Na studiach zaczęło nas studia pięć, skończyłyśmy chyba trzy.
> *B: A pięć na ile na roku osób?*
> R: Przyjęto około setki, eee o ile dobrze pamiętam. W każdym razie skończyło nas studia w ogóle 25 osób. Więc to były takie studia, które rzeczywiście mocno przesiały towarzystwo. No i prawda jest taka, że dziewczyny nigdy się nie ciągnęły w ogonie. Niestety, nasz świat jest taki, jaki jest i spotkałyśmy się z baardzo nieprzyjemnymi sytuacjami. Był na przykład prowadzący, który w momencie, gdy nie działały porty USB, takie stare komputery, e z przodu miały porty USB, mieliśmy coś dawać, jakieś programy, które napisaliśmy w domu. Wetknął USB do portu, okazało się, że port nie działa i facet sobie pozwolił na uwagę [zmienionym głosem] no, to tak jak z babą, o jednej dziury daje dostęp, do drugiej już nie. No i człowiek się zastanawia, czy ma już iść na skargę, czy nie, prawda? No bo jeżeli pójdę na skargę, to i tak u tego idioty muszę zaliczyć. A jeżeli pójdę na skargę, to będę mieć kłopoty. Więc jest ciężko. Generalnie ja jeszcze jestem osobą niezbyt wysoką, co też nie dodaje mi mocy. Wszystko [pauza] wszystko trzeba tuszować pewnością siebie. Trzeba mówić z dużym przekonaniem, bo w przeciwnym wypadku nikt nam nie wierzy. Na dodatek nasze osądy czy wszystko to, co robimy jest wielokrotnie sprawdzane. Nie to, żebym miała coś przeciwko, ponieważ to tylko i wyłącznie potwierdza, umacnia moją pozycję. Natomiast nawet mój współpracownik z firmy wielokrotnie, i to nie jest trzykrotnie, pięciokrotnie myślę, że co najmniej kilkanaście razy, przyznał się do kilku, ee **sprawdzał**, czy to, co ja mówię jest prawdą. I to jest chłopak, który jest świeżo po studiach. Naprawdę, nie to, żebym mu to zarzucała to nie w tym rzecz, bo skoro za każdym razem się potwierdzało, to jest OK. Ale to jest robione **notorycznie**, regularnie. Tak samo właśnie w firmie moje jakieś tam prace były potwierdzane przez firmy zewnętrzne. I firmy zewnętrzne próbowały wprowadzać inne metody, że na pewno wyjdzie lepiej, że to, że tamto, że coś jeszcze, **nigdy** nie okazało się, żeby ktokolwiek był w stanie osiągnąć lepsze wyniki. Nie umiem powiedzieć, czy te wyniki osiągane przez inne firmy były gorsze, czy były porównywalne. Wiem, że nie były lepsze. Ale proszę sobie wyobrazić, jak się człowiek czuje, kiedy się dowiaduje, bo wiesz, zatrudniliśmy tutaj z zewnątrz firmę, żeby nam to sprawdziła innymi metodami. Ktoś zdecydował, że wywali **kupę kasy** na to, żeby coś potwierdzić, żeby potwierdzić, że moja robota nie jest do bani. Ja z jednej strony rozumiem to ze względów marketingowych, czy generalnie ze względów związanych z wydawaniem grubych pieniędzy. Że mając tego rodzaju potwierdzenie wiadomo, że kierunek jest słuszny, ale z drugiej strony to zawsze działa na głowę, zawsze się w jakiś sposób obrywa, bo to jest tak jak powiedzenie wiesz co, ufamy ci, ale tak nie do końca. Ta branża tak wygląda. Na dodatek w tej chwili pojawia się mnóstwo młodych ludzi. Nie to, żebym miała cokolwiek przeciwko. Natomiast ludzi, którzy właśnie twierdzą, że wezmą coś gotowego z pudełka, z półki zdejmą i to będzie świetne. Okazuje się, że zapomnieli o podstawach matematyki, czyli że każda metoda ma swoje założenia, czyli że jeżeli dane nie są dopasowane do metody, w sensie czysto teoretycznym, że na przykład, nie wiem, jakiś tam model statystyczny wymaga tego, żeby dane miały rozkład normalny. To są eee tego rodzaju rzeczy, takie. Ludzie tego nie sprawdzają. I okazuje się, że nie działa. A najczęściej kryterium wyboru jest cena. Więc jeżeli przychodzi na rynek świeżak tuż po studiach, to siłą rzeczy będzie miał jakąś swoją cenę, bo po pierwsze jest świeży, więc nie może szarżować. A po drugie, wydaje mu się, że zrobi to tak, tak, tak i będzie szybciutko i

będzie z głowy. Nie sprawdził na przykład właśnie założeń i potem są odgórne głosy, że nie, te metody wasze to są w ogóle do bani, tu nic nie działa. No **działa**, tylko po pierwsze nie zawsze, nie ma cudów. Są takie dane, na których się nic nie da zrobić, bo na przykład jest ich za mało, bo na przykład były zebrane w taki sposób, że brakuje informacji, bo to, bo tamto, bo coś jeszcze, a w procesie zbierania danych generalnie ludzie nie chcą, żebyśmy brali udział, bo uważają, że jest to wyrzucanie pieniędzy. To jest kwestia na przykład dwóch, trzech godzin konsultacji, dogadania tego, że potrzeba by było tego, tego, tego i tamtego. I jak proces ustawić, żeby to było po stronie firmy do zrobienia. A potem się nie da. No i tak jak pan mówi, na dodatek, jak wchodzi kobieta, no to w ogóle jest spora wątpliwość, czy ona może mieć jakiekolwiek pojęcie o tym, co mówi.
B: Ale ze strony klienta się z czymś takim pani spotyka, tak?
R: Zazwyczaj już nie, dlatego że jeśli już klient się decyduje na to, żeby współpracować z kobietą, no to jest już w jakiś sposób uodporniony. Natomiast jest mnóstwo sytuacji, w których klient się **w ogóle** nie decyduje na rozmowę ze mną. I wtedy albo robimy w ten sposób, że idziemy na spotkania we dwójkę, gdzie umawiamy się, że po prostu ja się odzywam głównie, że staramy się przekierować rozmowę na mnie, żeby jednak klient się przekonał. Albo klient w ogóle odmawia. Zdarza się i tak.
B: No i, jak rozumiem, z drugiej strony w środowisku analityków na przykład pani praca jest, jak rozumiem, bardziej sprawdzana niż praca kolegi?
R: Tak, zawsze jest weryfikowana co najmniej dwukrotnie. I co ciekawe również przez kobiety. Współpracujemy też z firmą, która też w ramach [nazwa projektu], która pisze cześć oprogramowania tą taką bezpośrednią dla użytkownika. No i wiadomo, integrujemy różne fragmenty oprogramowania naszego z ich oprogramowaniem no i potrzebowali wykonać pewne testy, żeby mieć szacowanie czasu. No i jak dziewczyna zaczęła puszczać testy, aa akurat tutaj też współpracuję z dziewczyną, o ile się nie mylę to też jedyną kobietą u nich w zespole to też dzwoniła do mnie słuchaj, ale czy to możliwe, bo ja na tej próbce danych od ciebie to mi działa w takim i takim czasie. Czy to jest możliwe?. Ja mówię wydaje mi się, że jest możliwe, ale przygotuję ci drugą próbkę danych, odpowiednio większą, żebyś wiedziała też, jak to się skaluje w czasie, żebyś sobie mogła przeliczyć. No i okazało się, że rzeczywiście to tak działa po prostu. Ale puszczała testy. Bo wiesz, bo ja już to sprawdzałam kilka razy. No właśnie **wiem**. Także niestety, jesteśmy przyzwyczajone do tego, że patrzy się nam na ręce dużo bardziej niż komukolwiek innemu. {wywiad 25}

W powyższej wypowiedzi, poza kolejną w naszych badaniach narracją dotyczącą dyskryminacji ze względu na płeć, na polskiej, publicznej uczelni wyższej, uzyskaliśmy jedyny w trakcie badania tak wyraźny obraz praktyki podważania kompetencji kobiet. **Kompetencja kobiet traktowana jest jako wyjątek, tak samo przez mężczyzn jak i inne kobiety**, i to nie tylko w polskim świecie DS, ani nawet nie tylko na arenie modelowania, ale w różnych obszarach technicznych – tak w działaniach hobbystycznych, jak i komercyjnych, także w nauce i edukacji (Zaród 2018:352-353). Kobietom trudniej niż mężczyznom jest uzyskać status autentycznej uczestniczki świata DS właśnie z uwagi na domyślne traktowanie kobiet jako osób nie-technicznych. Nawet przez osoby nie zajmujące się DS ani żadną sferą techniczną, kobieta na stanowisku technicznym może być postrzegana jako ktoś nie na swoim miejscu, wyłącznie z uwagi na płeć:

R: (...) Jedyne co **kiedyś** niebezpośrednio ja doświadczyłam że było spotkanie z klientem, rozmawialiśmy na bardzo poważne tematy, właściwie rozmowę prowadziła ta [imię], moja

szefowa i wtedy usłyszałam jakiś taki komentarz do niej, że tam może by się pani uśmiechnęła czy coś takiego co wydało nam się **bardzo** nie w porządku i że mężczyźnie by mężczyzna czegoś takiego nie powiedział, podczas no profesjonalnego spotkania gdzieś na szczycie omawiania umowy. Więc żyłam w takim trochę Eldorado [śmiech] niczego jakby złego nie doświadczyłam. Ale przy zmianie pracy, doświadczyłam czegoś co na początku było dla mnie **bardzo** niemiłe typu przychodziłam, chodziłam do różnych działów witałam się, cześć jestem [imię Rozmówczyni] jestem z działu analityki i tak i usłyszałam dużo komentarzy w stylu [głośno] co? Dziewczyna w dziale analityki? Nie pomyliłaś się? No co ty! takie że to miejsce było zwyczajowo przez mężczyzn zajmowane i też miałam taką sytuację, ee w czasie prezentacji, że **notorycznie** właśnie jeden, jeden dyrektor działu notorycznie mi przerywał chociaż teoretycznie ja już przeszłam na rolę specjalisty i miałam coś ciekawego do powiedzenia, i notorycznie bardzo niegrzecznie mi przerywał nie, nie wiem czy to było związane że byłam **nowa**, i coś tam gadałam, jak mi przerywał, czy to było związane z tym że jestem kobietą, nieważne. Takie odczułam, yy że może jak bym była mężczyzną to [śmiech]
B: to byłoby inaczej?
R: To byłoby łatwiej. No ale to było tylko kilka takich złośliwych komentarzy, takie co dziewczyna w dziale analityki? powiedzianych głośno. O czym miałam jeszcze powiedzieć? Tak, jeszcze zdarzyło mi się być na meetupie z koleżanką, bo to był w ogóle meetup [pauza] nawet nie pamiętam chyba języka [pauza] Ruby, więc zupełnie z czymś, z czym się zupełnie nie znałam, ale jedna prezentacja była akurat o technologii która mnie interesowała i przyszłam z koleżanką i byłyśmy chyba **jedynymi** kobietami na sali. I było trochę niezręcznie, ktoś tam nas zagadywał i zaczął rozmowę od tego czy przyszłyśmy [długa pauza] czy przyszłyśmy tutaj z polecenia innego meetupu tylko dla kobiet. Jakoś tak była, yyy była niezręczna sytuacja. Nic złego się nie stało i nie, nie. A i ostatnia rzecz o której chciałam powiedzieć, to już bezpośrednio doświadczyłam czegoś, a nie wierzę w złą wolę tylko raczej takie że ktoś, że w innej firmie jest cały dział analityków i to są tacy mężczyźni z piwnicy trochę, i byłam właśnie na spotkaniu z takim **jednym**, gdzie no wiem że z nim też trochę współpracowałam, konwersowaliśmy, mieliśmy coś tam wspólnie robić i na tym spotkaniu on zaczął tak zwany mansplaining[282] czyli zaczął tłumaczyć bardzo proste rzeczy, pojęcia właśnie z analityki webowej **mi**, kiedy ja o tym bardzo dobrze wiedziałam, więc nie wiem po co on to zupełnie robił. Więc doszło do jakiegoś takiego spięcia, że takie trochę podkurwienie trochę, że to bardzo dobrze wiem, możemy **ruszyć dalej**. Więc to mi się zdarzyło, ale nie wierzę w złą wolę tylko wierzę w takie nie wiem może jakieś, nie wiem, może zaćmienie mózgu. A tak
B: Aha, ale to tak jakby się spodziewał że ty nie wiesz tego, tak?
R: Na sali było jeszcze dużo innych ludzi w tym na przykład jeszcze account'ci i jakby i oni rzeczywiście tego nie wiedzieli, ale nie **musieli** wiedzieć [śmiech], bo jakby coś innego mieliśmy ustalać na tym spotkaniu i jakby nikt nie potrzebował żeby on był [pauza]
B: robił wykład o
R: robił wykład teraz czym jest warstwa danych na stronie. No ale, nie odbierałam tego jakoś bardzo personalnie. Chyba **najgorsze** czego doświadczyłam to takie właśnie o dziewczyna w analityce! to było naprawdę nieprzyjemne. Że to nawet, ci ludzie nie mieli nic złego na myśli, ale taki utarty stereotyp że na tym miejscu powinien być facet, takie, mmm. {wywiad 17}

Rozwijamy dalej związek bycia poddawanym kwestionowaniu kompetencji z byciem osobą młodą i niedoświadczoną w DS, na razie przyjrzyjmy się wypowiedzi dyskutującej tę zależność z byciem kobietą:

B: A czy chciałabyś, yy opowiedzieć o jakiejś sytuacji nieprzyjemnej, która ciebie spotkała?
R: Nie chciałabym chyba opowiedzieć, nie chcę ich sobie tak przypominać, ale raczej podtekst

282 *Mansplain* to neologizm, połączenie słów *man* (mężczyzna) i *explain* (wyjaśnienie), oznacza działania mężczyzn mówiących do kobiet w sposób protekcjonalny, zakładający a priori, iż ten mężczyzna ma w danej materii lepsze rozeznanie niż ta kobieta (por. Bridges 2017).

> seksualny. Niestety spotykałam się z podtekstami seksualnymi, ale tak na co dzień z czym się spotykam, ze względu zarówno na mój wiek, jak i to, że jestem kobietą, no ja nie mam tak, że wchodzę do sali i wszyscy myślą [zmienionym głosem] ta to ma wiedzę, to jej posłuchajmy.
> *B: Ale ze względu na wiek?*
> R: Wiesz co, bardzo często jednak różnica wieku między moimi odbiorcami a mną jest znaczna.
> *B: Czy mówimy o tym, że jesteś młoda i młodo wyglądasz?*
> R: Tak. Tak tak tak tak tak tak, dokładnie o to chodzi [pauza] No to jednak tak, trzeba na dzień dobry często udowodnić, jak rozmawiam z innymi kobietami, też z takimi, które bardzo wysoko zaszły, no to też jednak widzą to, że często jest tak, że na dzień dobry trzeba wejść i trochę udowodnić. Ale ja uważam, to znaczy to nie jest złe, to wynika właśnie jakby z tej ewolucji, ze względów czasowych i ja muszę przyznać, że tak jak na początku, jak byłam jeszcze młodsza i byłam taka bardziej [głośno] jak to nie chcecie mnie słuchać!? I próbowałam to siłą zrobić, to przynosiło odwrotne efekty. Teraz, kiedy przychodzę i ktoś podważy moją merytorykę na dzień dobry, ja po prostu staram się dalej robić swoje i powiem ci, że **zawsze** zawsze spotykam się z super odbiorem. To, ile razy, nawet nie to że obraził mnie ktoś jakoś, ale wielokrotnie słyszałam takie przepraszam cię, myślałem, że nie masz takiej wiedzy. Albo przepraszam cię, wziąłem po prostu, nawet dosłownie miałam takie sytuacje jak z mężem rozmawialiśmy często z osobami, że zwracano się do mojego męża, który jest mniej techniczny niż ja i ja zaczynałam odpowiadać na pytanie i mieliśmy taką sytuację, że właśnie jeden chłopak powiedział, mężczyzna, jedna osoba powiedziała przepraszam cię, spojrzałem na ciebie, mówię kobieta, na pewno nie wiesz. Przepraszam cię, bo sam się złapałem na tym, że to zrobiłem. Więc my też musimy wziąć pod uwagę, że to nie jest tak, że ci wszyscy ludzie to robią celowo. To czasami jest po prostu taki kontekst, który przez ileś lat trwał.
> *B: Ale to rozumiem, że tak zachowują się też osoby z różnych biznesów, z którymi ty się spotykasz, że to nie jest nawet kwestia IT?*
> R: Niee nie. Przeróżnie. Bo czy to jest tylko kwestia IT? Nie oszukujmy się, osób na wysokich poziomach jest w ogóle, dużo więcej mężczyzn. Tak jest no, po prostu. Więc tak, ale nie jest to tak dotkliwe i tak straszne jak to jest promowane. Częściej to jest kwestia tego, że po prostu jest na początku taka konsternacja, ale co też mi słusznie kiedyś kumpel powiedział, to nie zawsze jest kwestia tego, że jestem kobietą. To nie zawsze jest kwestią tego, że jestem młodsza. Po prostu kontekst, wejście, dzień, cokolwiek, tak? Oni też czasami muszą się bronić. {wywiad 22}

Zaród (2018:348-349) wskazuje, że w hakerspejsach kobiety nie mogą w pełni uczestniczyć w tzw. laboratorium, czyli miejscu charakteryzującym się: minimalizowaniem kosztów porażki osób eksperymentujących, karnawałem poznawczym i wspólnotą indywidualnej koncentracji. Te „podwójne standardy" oceny kompetencji dla mężczyzn i kobiet polegają, zdaniem Zaróda, na tym, że dla mężczyzn łatwiej, a dla kobiet trudniej oddzielić błąd od osoby. W związku z tym „mężczyźni mają szerszy margines błędu, mogą stawiać ryzykowniejsze hipotezy bez groźby kompromitacji" (ibidem). Na gruncie badanego przez nas świata kwestionowanie kompetencji także można uznać za praktykę co najmniej utrudniającą pełne, autentyczne uczestnictwo w DS. Jak wielokrotnie pisaliśmy, projekty DS polegają w dużej mierze na eksperymentowaniu, a samo działanie podstawowe to w dużej mierze poprawianie własnych błędów, testowanie pomysłów, szukanie wyjścia ze ślepych uliczek, chałupnicze operacje na własnych narzędziach pracy. Właściwie, jeżeli tylko praca uczestnika jest samodzielna i efektywna, to błądzenie i pomyłki co do zasady (być może

raczej „zasada" ta dotyczy mężczyzn) pozostają w zgodzie z wartościami badanego świata, a nawet mogą być pozytywnie oceniane w perspektywie ciekawości i samorozwoju.

Wśród osób z obszarów technicznych, zaangażowanych w szeroką arenę modelowania (6.2.1.) pojawia się wyrażana w dyskursie identyfikacja praktyki kwestionowania kompetencji osób reprezentujących mniejszości, tak w DS jak i w ogóle w IT. Nie mamy informacji, by działo się to w Polsce, lecz przykładowo podczas obywającej się w Portland konferencji programistycznej The Monktoberfest w 2019 r. na ten temat wypowiedział się słynny w badanym świecie H. Wickham. W swojej prelekcji, skierowanej do osób rozwijających oprogramowanie open source mówił, iż ogromnie dużo zawdzięcza on uczeniu się na własnych, publicznych pomyłkach. Podkreślił jednak, że z jego doświadczenia wynika następująca zależność – kiedy jesteś białym mężczyzną i popełniasz błąd publicznie, ludzie mówią „ale on jest odważny", i działa to na twoją korzyść; kiedy zaś robisz to samo jako kobieta lub osoba kolorowa, ludzie mówią „ale głupi" i uznają za osobę niekompetentną (Wickham 2019a). Na tej samej konferencji Abby Fuller, z wykształcenia politolożka, pracująca na stanowisku inżynierskim Principal Technologist w Amazon AWS, wskazywała na ślepą koncentrację branży IT na bycie osobą techniczną, podczas gdy dla niej w IT najważniejsze jest budowanie czegoś działającego, odnoszenie się w swojej pracy do danych ilościowych, ciekawość, chęć uczenia się i popełniania błędów (kategorie są tożsame z wartościami świata DS w naszym ujęciu), a nie jakaś niedookreślona techniczność sama w sobie. Wskazywała, że ona sama była wyznawczynią techniczności – uważała, że jeżeli mówi na konferencjach cokolwiek nietechnicznego to zaszkodzi swojej karierze, bo przecież ciągle musi udowadniać swoje techniczne kwalifikacje jako osoba z mniejszości. Zdaniem Fuller dla większości ludzi w branży IT domyślny stan jest taki, że nie wierzą w kompetencje kobiety i ona musi je udowadniać. Jej zdaniem nie mówią prawdy ludzie twierdzący, że zwracają uwagę tylko na umiejętności koderskie (Fuller 2019), co postrzegamy jako krytykę pozycji „doskonałej merytokracji".

Z pozycji „doskonałej merytokracji", jaką raczej reprezentuje wyżej cytowana Rozmówczyni {wywiad 22}, wskazywano nam, że kwestionowanie kompetencji dotyczy nie tyle kobiet, co osób młodych i niedoświadczonych, a także jest (być może niecelową) strategią klientów w radzeniu sobie z negocjacjami biznesowymi w trudnej merytorycznie materii. Więc, szczególnie jeżeli chodzi o osoby młode i niedoświadczone to podważanie, a raczej testowanie kompetencji ma jakoby racjonalne uzasadnienie, a po prostu w sensie ilościowym bardzo mało jest w DS kobiet starszych i o dużym doświadczeniu (potwierdzają

to nasze analizy por. 11). Zatem relatywnie znacznie częściej można spotkać kobiety początkujące i młode, więc zależność między byciem kobietą, a kwestionowaniem kompetencji tej osoby w dużym stopniu wyjaśniać mają zmienne wieku i liczby lat doświadczenia. Jakoby może wydawać się, że kwestionowane są kompetencje kobiet jako takich, a chodzi o testowanie osób młodych i niedoświadczonych, przy czym najczęściej kiedy spotyka się w DS kobietę, to jest ona właśnie młoda i niedoświadczona. Jak wskazywaliśmy wcześniej {wywiad 25} kwestionowanie kompetencji spotyka także kobiety silnie osadzone w DS, z wieloletnim doświadczeniem, zatem powoływanie się głównie na wiek i doświadczenie w wyjaśnianiu opisywanej praktyki odczytujemy jako charakterystyczne dla pozycji „doskonałej merytokracji". Niemniej bez wątpienia testowanie kompetencji jest praktyką stosowaną też wobec młodych i początkujących uczestników – mężczyzn. Przyjrzyjmy się opisowi prób sił, jakimi był poddawany Rozmówca jako początkujący data scientist, z wykształcenia biolog, który porzucił doktorat:

> *B: (...) ktoś kto ma podobną trochę sytuację do ciebie, to zn, tylko nie jest biologiem, tylko jest ekonomistą w zespole matematyków chyba i informatyków, mówił żeee jakby yy, no jest czasami trochę inaczej traktowany, że o ty to jesteś po uniwerku, nie? Coś w tym rodzaju. Nie wiem czy z czymś takim się spotkałeś wobec siebie, czy nie wiem może wobec kogoś innego?*
> R: Nie, no zupełnie nie. Chociaż **miałem** na początku pewnie taki filtr gdzieś percepcyjny założony na głowę, bo miałem jakieś takie zadanie i tam zaproponowałem jakieś rozwiązanie, mówię no może byśmy do tych danych tutaj dopasowali jakąś krzywą i popatrzyli jak, już nie pamiętam, bo to było kurczę z rok temu, ponad rok temu, jakaś pierwsza z analiz jakie sobie tam powymyślałem, no i jeszcze byłem wtedy takim [pauza] takim szczylem, to musiałem się meldować za każdym razem do kierowniczki jak coś tam planuje zrobić, wdrożyć i wysłać. No i ona powiedziała, że no spoko, spoko aż to skonsultuje z kimś tam nie? ja mówię no dobra, no i potem się pytam tej koleżanki co ona konsultowała, mówię ej to było aż takie, że ona musiała konsultować, czy to w ogóle ma sens, czy coś tam? Po czym ona stwierdziła, że chyba nie, że chyba [krótka pauza] ona ją pytała po to, bo sama już nie pamięta, czy w ogóle, czy to co mówisz jest prawdą [ze śmiechem] mówi no ale jest, to to, to był dobry pomysł, nie? Generalnie sam fakt że ja posiedziałem mocno w tej matmie chwilę przed tym yyy wejściem w rynek tam, czyli jakby troszeczkę się doszkoliłem, to nie było tak że jak ja widzę, jak nie wiem oglądaliśmy te kursy, wspólnie, gdzie gościu tam wyprowadzał jakieś równania różniczkowe i inne całkowania, no to ja nie pytałem, o czym on mówi tak? tylko jakby byłem świadomy tego co tam się dzieje, iii, to nie był problem. Gdzieś miałem, było parę takich rozmów, gdzie jakby aktywnie wziąłem udział i iii i nie plotłem bzdur, więc gdzieś myślę że sobie przetarłem tam tą drogę, żeby być traktowanym poważniej troszeczkę nie, że pomimo tego że jestem biologiem to, też często, eee no mam coś do powiedzenia co, co jest wartościowe i iii to nie jest tak że w ogóle nie rozumiem matematyki, bo jestem człowiekiem z pola, z pola buraków, albo rzepaku i w kaloszach [z uśmiechem]. {wywiad 15}

Kwestionowanie lub testowanie kompetencji może spotykać także osoby bez wykształcenia technicznego lub matematycznego, a szczególnie tych, którzy nie studiowali na politechnikach. Takim osobom, podobnie jak kobietom, trudno jest uzyskać status autentycznych uczestników świata DS właśnie z powodu domyślnego braku wiary w ich

biegłość techniczną ze strony co najmniej części uczestników, legitymizujących się dyplomem politechniki:

> *B: [włącza dyktafon – dogrywamy drugą, nieplanowaną część wywiadu po chwili rozmowy towarzyskiej] Czyli absolwenci, coś jest na rzeczy z absolwentami uniwersytetu?*
> R: Tak, ewidentnie. Większość moich yyy współpracowników jest po polibudzie, a jak nie, no to chociaż po jakimś doktoracie w Nowym Jorku, z fizyki [pauza] więc generalnie są, są takie śmieszki iii są takie żarciki, to jest wydaje mi się bardzo takie uderzające, takie niemiłe pod tym względem, zwłaszcza że to [pauza] ale to nie tylko data science, pójdziesz na jakikolwiek meetup programistyczny i to samo. Wyjdzie gdzieś że jesteś po uniwerku to do razu jest takie, wiesz [wykonuje teatralny gest strzepywania czegoś z ramienia]
> *B: Ale że w ogóle po uniwerku, to jest w ogóle jakaś lipa, czyy?*
> R: Tak. Tak. Jak jesteś po polibudzie, nie zależnie jaki kierunek, spoko, jesteś po uniwerku niezależnie jaki kierunek, chujówka.
> *B: Aha. Ok.*
> R: Ale to jest takie trochę z przymrużeniem oka, nie? Nikt ci jakby nie [pauza] No dobra nie wiem, zależy w którym towarzystwie, w moim, czasami jest bardzo z przymrużeniem oka, czasami troszeczkę mniej nawet, więc tak dziwnie trochę, nie?
> *B: No dobra. Ale no jakiś przykład, czyli jak to wygląda?*
> R: No nie no właśnie, rozmawiamy sobie o technologiach, ja mówię o data science że robię w tym i w tym i w tym, a on mówi aaa no pamiętam ze studiów nie? a ja mówię a no to fajne fajne miałeś studia nie? zazdroszczę, ja nie miałam tego na swoich studiach, a on mówi ha a to co ty nie po polibudzie, to ja mówię nie po uniwerku, po informatyce z ekonometrią, no i wiesz, taki taki gest, że tam, że się nie umywam, nie? Chwilę później mówi nie no żarcik nie? ale w sumie wiesz o chodzi, to to jest takie bardzo między wierszami, ale jednak zostaje gdzieś. Co mam powiedzieć, no miałam taką samą matematykę jak on, taką samą algebrę liniową, rachunek prawdopodobieństwa wszystko to samo, bo wiem co mieli ludzie, moi znajomi na polibudzie, nie? Więc jak dla mnie jest to kompletnie nieuprawnione, a sami uczą się z dupy różnych rzeczy, które wiem że na informatyce na uniwerku są super poprowadzone, nie są tak przestarzałe jak na polibudzie gdzie jakiegoś [niezrozumiałe] się uczą, podczas gdy na uniwerku jest to już jakby dawno wykreślone z programu, więc jak dla mnie jest to strasznie niesprawiedliwe i ten stereotyp takiego niedojdy jest nadawany niesłusznie, znaczy trochę słusznie, trochę nie. Ale wiadomo są wyjątki i tu i tu, więc ja się czuję jako wyjątek który został niesłusznie, jakby wrzucony do worka, ja nie wiem komu
> *B: Ale to dlaczego słusznie?*
> R: Dlatego słusznie, że no więcej ścisłych przedmiotów tam i ludzie generalnie, tam bardziej idą tacy ściśli, więc no jakby, to się rozumie że tam, generalnie jakby więcej jest osób które mogą być programistami, generalnie, nie? No ale to nadal uogólnienie. [pauza] Ja rozumiem to, no sama znam ludzi z uniwerka i wiem że **generalnie to nie są osoby, którzy idą później w programowanie** że raczej to są osoby które generalnie gorzej statystkę rozumieją i lubią na przykład, i ogólnie jakiekolwiek rzeczy z matmy, więc rozumiem to, z tego powodu, ale to nie zmienia faktu, że [krótka pauza] jak już ktoś widzi że ja robię to co robię, no to już w sumie mógłby nie wątpić w to że mam potrzebne do tego wykształcenie i wiedzę i powiedzmy doświadczenie, a nadal to gdzieś tam wychodzi, nie? (...) tutaj nawet nikt nie mówi że po polibudzie ktoś jest programistą, tylko że bardziej po polibudzie to jest ktoś kto jest ścisłym umysłem po prostu, więc to bardziej jakby pasuje do charakteru, do data science i do statystyki, do programowania i do wszystkiego w sumie.
> *B: Mhm, Ok, czyli jeżeli bym, aee jeżelibym szukał pracy w data science, to nie mogę pisać jestem humanistą, uwielbiam czytać wiersze, ale poza tym jestem zajebistym statystykiem i wymiatam w Pythonie?*
> R: Nie no mógłbyś. I mógłbyś dostać prace zzz, jakby bez problemu, bardziej chodzi o te takie żarciki w kuluarach, nie? Zwłaszcza że akurat dużo jest chyba ludzi po filozofii w data science, więc to nie jest takie, wielu pracodawców zwróci uwagę na twoje umiejętności, a nie na to po

czym jesteś na prawdę, są ludzie w ogóle bez studiów u mnie w pracy, przynajmniej jedna osoba
B: W ogóle bez studiów?
R: Tak. No to [imię], my się w ogóle lubimy. Tak jak mówiłam, w ogóle się nie liczy jaki masz papierek, no ale po prostu później to gdzieś wychodzi na meetupach, nie wiem głupich rozmowach w kuchni i tak dalej nie? jak ktoś jest powiedzmy burakiem to akurat to jest pierwsza rzecz z której sobie zażartuje, można tak powiedzieć. [pauza] Więc w tym kontekście dziewczyna po polibudzie ma prawo wyśmiać chłopaka po uniwerku, a nie że chłopak wyśmieje dziewczynę bo jest dziewczyną, nie?
B: Rozumiem. Mhm [pauza] A czy, jest jeszcze jakiśkolwiek tego rodzaju podział? Bądź podobny?
R: Właśnie nic mi więcej nie przychodzi do głowy. Tak to wszyscy są z różnych środowisk zupełnie. Mamy jedną osobę czarnoskóra, to tutaj też na tym gruncie żadnych nie ma rzeczy.
B: Ok. No dobrze. Dobra.
R: Rozumiem że nie spotkałeś się z tym zagadnieniem? Pewnie wszyscy z którymi rozmawiałeś są po polibudzie i nikt na to nie zwrócił uwagi? {wywiad 10}

B: No dobra, a z politechniką i [nazwa uczelni ekonomicznej A – Rozmówczyni jest jej absolwentką]?
R: Yyy z politechniką i [nazwa uczelni ekonomicznej A] to z nas się zawsze śmiali że my matematyki nie umiemy i programować też nie umiemy w sumie [śmiech] ja to się z nich zawsze śmiałam że oni to w ogóle rozmawiać nie potrafią i z nimi się nie da porozmawiać tak normalnie. {wywiad 18}

Dogłębne opisanie sporu dotyczącego traktowania kobiet w DS wymagałoby oddzielnej pracy doktorskiej, lub kilku, zatem dążąc przede wszystkim do ujęcia pełnej sytuacji naszego badania polskiego świata DS poprzestajemy na powyższym szkicu. Ta arena była dla nas szczególnie trudna do zbadania (por. 4.3.1).

Istnieje wiele publikacji (Beede i in. 2011; Breda i Napp 2019; Crawford i in. 2019; Drury, Siy, i Cheryan 2011; Williams i Ceci 2015; Zawistowska 2013), w tym także publicystycznych, jak bestsellerowa „Brotopia" (Chang 2018) poświęconych sytuacji kobiet w branży IT, w nauce i edukacji technicznej, oraz generalnie w obszarze STEM. Przez osoby zajmujące się DS, w porównaniu do branży IT, badany przez nas świat DS postrzegany jest raczej jako wyjątkowo różnorodny, otwarty i sprzyjający kobietom.

6.2.3. Python, R i Excel

Nawet patrząc na rysunek 28. można zauważyć, że **zastosowania Pythona i R różnią się od siebie**. W badanym świecie uznaje się, że **te różnice wynikają z historii rozwoju każdego z tych języków**, a w konsekwencji cech technicznych tych języków, dostępności technologii ułatwiających korzystanie z nich (przede wszystkim pakietów), a także różnic w środowiskach, które preferują posługiwanie się jednym czy drugim językiem programowania. Jak wspomniano (por. 5.3.1.) język R zbudowali statystycy dla statystyków, i zastosowań innych niż operacje na danych jest dla R niewiele. Język R jest też silnie kojarzony z

akademią – zarówno z matematykami/statystykami rozwijającymi metody analityczne, jak i naukowcami stosującymi analitykę do swojego obszaru. Python kojarzony jest raczej z zastosowaniami komercyjnymi, co, jak wskażemy dalej, jest łączone z łatwiejszą niż dla R integracją z systemami IT.

> *B: To chciałbym spytać cię jakie w ogóle są twoje narzędzia pracy?*
> R: [pauza] Głowa, komputer, to chyba tyle [ze śmiechem] Jeśli chodzi o konkretne narzędzia pytasz, czy jakieś software'owe?
> *B: Mhm, mhm.*
> R: No to Python właśnie, [krótka pauza] jest jeszcze eR na przykład, to pewnie sam stosowałeś gdzieś tam w, na uczelni.
> *B: Tak.*
> R: No to też pośrednio go wykorzystujemy, ale mniej dużo bo jest, bardziej chyba, mniej przejrzysty i trochę bardziej jednak skierowany do środowiska chyba mi się wydaje naukowego, a Python jest taki bardziej komercyjny, taki trochę modniejszy, więcej do niego jest narzędzi, więc dlatego Python. No i są tam różne moduły wbudowane no to w sumie wszystko wokół Pythona się kreci.
> *B:Od razu pytanie, dlaczego modniejszy?*
> R: Nie wiem dlaczego, ale są różne rankingi iii, oceniają na podstawie tego, jak dużo wątków jest na forach, i tego jaki jest na przykład wzrost popularności, w sensie wzrost liczby wątków z roku na rok.
> *B: Czyli jest bardziej popularny w data science, tak? Więcej ludzi, więcej ludzi go używa?*
> R: Tak, tak, tak. A jak więcej używa no to i więcej narzędzi się do niego pojawia, no to jak więcej się pojawia to jeszcze więcej ludzi do niego przechodzi i tak koło się zamyka. {wywiad 10}

W powyższej wypowiedzi wyraźnie widać to, co można nazwać bezpośrednim efektem sieciowym – „jak więcej używa no to i więcej narzędzi się do niego pojawia, no to jak więcej się pojawia to jeszcze więcej ludzi do niego przechodzi i tak koło się zamyka" {wywiad 10}. Zauważmy w innym źródle porównywanie R z pakietami statystycznymi, jak SAS czy SPSS, w ilości publikowanych artykułów naukowych (choć od 2017 r. rośnie użycie Pythona) (por. Fig.2a-2f Muenchen 2019).

Python jest językiem programowania ogólnego zastosowania, którego część rozwinęła się w kierunku pracy z danymi. Uczestnicy badanego świata nazywają Pythona np. prawdziwym lub pełnoprawnym językiem programowania, co bywa elementem legitymizowania działań (czy sposobu wykonywania działań). Zauważmy także, że Python jest historycznie krócej niż R wykorzystywany do operacji na danych w sensie zbliżonym do DS. Choć oba języki powstały w latach 90., to R od samego początku służyć miał tym celom, w Pythonie takie „odgałęzienie" powstało później (por. dalej fragment {wywiad 13}). Historię rozwoju języków Rozmówcy przedstawiali następująco:

> R: Znaczy, generalnie jest tak że eR yy powstało jako narzędzie typowe do analizy danych. Wiele lat temu tak? Było **z góry** myślane i pisane do pod analizę danych. Natomiast Python jest

> językiem szerokim językiem programowania, który generalnie zajmuje się programowaniem aplikacji. No i tam oczywiście są głó głównym takim kością niezgody jest fakt, że eR jest [pauza] **jednowątkowy**, nie wiem na ile jesteś technologiczną osobą na ile nie, no w każdym razie w Pythonie można szybciej i więcej danych zanalizować. No ale czy trzeba? Ja uważam że niekoniecznie. Faktycznie, Python będzie szybszy (...). {wywiad 6}

Wskazywano, że R lepiej sprawdza się w sytuacjach, gdzie rezultatem pracy ma być wgląd w dane, zaś Python przy tzw. wdrożeniach produkcyjnych modeli i wtedy, gdy mają być zastosowane metody ML, a szczególnie metody uczenia głębokiego:

> *B: (...) Dlaczego niektórzy robią w eRze a inni w Pythonie? Nie można robić w jednym?*
> R: Ooo! Ooo [śmiech] Wiesz co, no eR jest świetny do takich szybkich analiz, jest, podobno jest dużo łatwiejszy do nauczenia się dla osób które wcześniej nie miały styczności z programowaniem w ogóle.
> *B: Mhm.*
> R: Ale nie wiem, czy to mi się tylko tak wydaje, że właśnie trudniej jest zrobić w eRze eee duży taki projekt software'owy, który ma chodzić gdzieś na serwerze. Który ma chodzić cały czas na przykład, nie? Wydaje mi się że to jest trudniejsze, może się mylę, ale wydaje mi się że tak. Więc to też trzeba po prostu dostosowywać do tego co chcesz, co chcesz osiągnąć nie? co chcesz robić. Jeżeli to mają być po prostu **analizy** danych, to myślę że eR super. Czy ze zdjęciami by sobie poradził, nie wiem. (...) u nas w ogóle w [nazwa firmy] nie używamy eRa, w ogóle. Wydaje mi się że eR jest w takich bardziej analitycznych sprawach, właśnie jakiś marketing na przykład, alboo banki, no coś takiego. Bo wtedy faktycznie jest dużo danych, ale no po prostu robisz tylko raport dla klienta. I eR jest do tego świetny, bo jest szybki. Ma duże, dobre biblioteki do radzenia sobie z danymi taki, ale żeby właśnie zrobić coś co jest dedykowane na serwer to nie wiem czy takie najlepsze, więc my nie piszemy w eRze w ogóle.
> *B: Mhm, Ok. No dobra.*
> R: Znaczy nawet nie wiem czy takie deep learningowe rzeczy jak sieci neuronowe są w eRze. Wiesz może? {wywiad 8}

Zauważmy wykrzyknik w reakcji na pytanie „nie można robić w jednym?" – dla uczestnika badanego świata oczywistym jest, że nie można „robić w jednym". Widać także, że język R uznawany jest za narzędzie łatwiejsze dla osób, które nie miały nigdy styczności z programowaniem. Jak zaraz wskażemy, Python uznawany jest za narzędzie do pracy z danymi łatwiejsze dla osób znających już inny język programowania. Osoby z wykształceniem lub doświadczeniem informatycznym mogą także znać Pythona w związku z innymi zastosowaniami. Jest to jedna z obiegowych opinii, podkreślmy jednak, że początkujących zainteresowanych uczeniem maszynowym bez względu na wykształcenie zachęca się obecnie do rozpoczęcia nauki od Pythona. W toku badań spotkaliśmy osoby, dla których Python dla operacji na danych był pierwszą stycznością z programowaniem w ogóle. Rozmówczyni podkreśla, że należy **dobierać narzędzia do problemu** „trzeba po prostu dostosowywać do tego co chcesz osiągnąć" {wywiad 8}.

Zauważmy też zdania: „eR jest do tego świetny, bo jest szybki. Ma duże, dobre biblioteki (...) nawet nie wiem czy takie deep learningowe rzeczy jak sieci neuronowe są w eRze" {wywiad 8}. Wyraźnie widać, że **przydatność języka do określonych zastosowań oceniana jest przez pryzmat dostępnych pakietów**. Przy tym osoba, która korzysta już w Pythonie z narzędzi spełniających jej potrzeby, nie podejmuje wysiłku szukania innych rozwiązań – takie działanie mogłoby zostać uznane za bezcelowe, bo nieefektywne.

W następnym fragmencie widać zarówno różnice w historii powstania języków Python i R, różnych ich zastosowań, co jest połączone z różnymi środowiskami użytkowników a także z powstawaniem pakietów do innych zastosowań (a mówimy cały czas o open-source, zatem o technologiach rozwijanych, przynajmniej częściowo, oddolnie przez użytkowników).

> R: (...) Znaczy no w Pythonie pisałem parę skryptów w życiu, eee ale to tak tam z konieczności jak na przykład, nie wiem jakiś, potrzebowałem skądś tam pobrać dane, i to coś, (...) to narzędzie miało API, które, z którym można się było połączyć przez Pythona po prostu wygodnie, a przez eRa raczej nie. W sensie że w eRze nie było nic gotowego do tego, do obsługi tego, więc coś tam w Pythonie po prostu napisałem wtedy.
> *B: Yhm, Ok. [pauza] No właśnie, zazwyczaj ludzie mówi o eRze, albo o Pythonie, iii szczerze mówiąc, może mi to wytłumaczysz, dlaczego nie mogą ludzie robić wszyscy w eRze, albo na przykład wszyscy w Pythonie? [z uśmiechem]*
> R: [śmiech]
> *B: Po co są dwa języki do, jak dla mnie do tego samego?*
> R: A no, no to jest dobre pytanie. Znaczy no na pewno **nie są** do tego samego (...) eR był teoretycznie z zamysłem stworzony przez statystyków dla statystyków, więc on po prostu był, no jest językiem który, miał wbudowane pewne metody, wbudowane pewne funkcjonaliści, z których korzystali statystycy. Aaa, a no Python teoretycznie był stworzony jako język po prostu do programowania, aplikacji, czegokolwiek, eee a że wyewoluował tak w stronę (…). To dlatego myślę też, myślę że to jakoś tak trochę z przypadku bo, po prostu nie wiem, ileś tam, piętnaście lat temu ktoś stworzył bibliotekę do Pythona, która też właśnie miała jakieś funkcjonalności związane ze statystyką nie, albo właśnie z drzewami decyzyjnymi. No i jak ktoś by stworzył taką fajną bibliotekę w tym samym momencie do Javy na przykład, albo do nie wiem C#, no to może ten język by teraz był używany przez data scientistów bardziej. Myślę że to trochę przypadek, że się tworzy jakieś środowisko wokół jednego języka, tworzy się jedna, dwie, trzy biblioteki akurat w tym języku a nie innym, no i staje się to popularne i dlatego to tak wygląda, tam 10 czy 15 lat później. No i to, no właśnie Python i eR były stworzone no, przez innych ludzi, w różnych momentach czasu, z innymi założeniami, więc teraz też siłą rzeczy, niektórym bardziej odpowiada eR, a niektórym bardziej Python tak, bo inaczej się ich używa (...).
> *B: Mhm, OK. A czy, czy twoim zdaniem jakoś, jakoś się od siebie różnią ci użytkownicy tak, tak zasadniczo?*
> R: Nooo, znaczy to, znaczy na pewno trochę tak, bo to tak siłą rzeczy wychodzi że ludzie, na przykład matematycy i statystycy, czy też, aaa no socjologowie, czy yyy biolodzy, bioinformatycy to są, korzystają bardziej, częściej z eRa. No i to, znaczy to wynika, no właśnie z jakiejś tam zaszłości historycznej że eR był stworzony przez statystyków do zdań statycznych, a że socjologowie, bioinformatycy, matematycy, potrzebują metod statystycznych [pauza]
> *B: Tak jakby wtórnych?*
> R: Tak. No to, to jakby zaczynają korzystać z eRa, to to jest pewnie wynik tego że kiedyś tam eR się zaczął pojawiać na uczelniach, na kierunkach no nie tylko właśnie na matematyce ale na innych, a no, a Python się nie pojawiał, no bo on nie był pierwotnie do statystyki stworzony, ale

> jednocześnie Pythona się mnóstwo ludzi uczy na studiach informatycznych.
> *B: OK.*
> R: Więc siłą rzeczy te środowiska wyglądają na pewno inaczej. Nie wiem na ile, to by trzeba nie wiem, zrobić badania jakieś, że że wśród użytkowników Pythona jest więcej programistów, a wśród użytkowników eRa jest więcej no tych zawodów które wymieniłem, więc no one się siłą rzeczy, o tyle od siebie różnią, nie? W sensie, myślę też, że dlatego wśród użytkowników Pythona jest więcej ludzi którzy znają też inne języki programowania, bo to są na przykład ludzie po informatyce. A wśród użytkowników eRa są ludzie którzy znają tylko eRa, i no nie potrzebują się innych uczyć, tak? Tak myślę, ale no też są jakieś tam, różne takie, moje przypuszczenia na bazie obserwacji. {wywiad 9}

Fragment „w eRze nie było nic gotowego do tego, do obsługi tego, więc coś tam w Pythonie po prostu napisałem wtedy" {wywiad 9} ilustruje podejście postulujące **dobór narzędzia do problemu**. Znowu chodzi o dostępność pakietów, za pomocą których można do własnych celów, możliwie łatwo i szybko skorzystać z rozwiązania wypracowanego przez kogoś innego. Jak opisaliśmy (por. 6.2.1.2) występuje także zjawisko fetyszyzowania narzędzi lub metod – stają się one (z różnych powodów, do różnych celów i dla różnych osób) ważniejsze niż rezultat ich zastosowania. Wracając do różnic pomiędzy Pythonem a R, pokażmy inny element fragmentu prezentowanego już w części 5.2.1.:

> R: Yyy, w pewnych sprawach Python jest lepszy, w pewnych sprawach R jest lepszy. Wygląda to tak, że ludzie którzy zajmowali się ogólnie data science, niekoniecznie uczeniem maszynowym i to są ludzie z wykształceniem bardziej matematycznym, to używali R. Ludzie, którzy zajmują się uczeniem maszynowym, w szczególności deep learningniem używają Pythona, bo biblioteki są lepsze. Patrząc na to, eee, dlaczego Python zyskał teraz taką popularność, bo coraz więcej programistów bierze się za uczenie maszynowe. A oni już umieli Pythona, Python jest językiem **ogólnego** zastosowania, wobec czego znali go na potrzeby innych projektów, albo po prostu z nim pracują na co dzień, wobec tego przesiadka dla nich jest łatwiejsza. I to jest powód, dla którego Pythona używa się coraz częściej, natomiast kolejna sprawa to produkcyjne wdrożenie modeli w oparciu o Pythona jest łatwiejsze, aaaa, natomiast sama taka analiza powiedzmy ogólna data science [pauza] no to w Pythonie jest dość niezgrabna, pisze się kodu względnie dużo. (...) Więc to jest tak, że na potrzeby takiego ogólnego data science nie ma dużej różnicy między R i Pythonem, poza tym że według mnie Python jest mniej wygodny, yyy, na potrzeby różnych specjalizacji to się na przykład okazuje że do deep learning w Pythonie jest dużo lepszy niż w R, i on jest z R generalnie przenoszony do Pythona albo pod spodem jest Python, natomiast jak się spojrzy na farmakokinetykę, na jakieś zastosowania powiedzmy bardziej naukowe, nie ogólne data science, no to Python jest cieniutki, bo bibliotek ma dużo dużo mniej. Wobec tego jeśli będziemy szukali takich bardziej specyficznych, niestandardowych modeli statystycznych to o niebo łatwiej znajdziemy je w R niż w Pythonie. Natomiast w standardowym uczeniu maszynowym typu modele klasyfikacyjne, regresyjne, xgboost[283] i pokrewne, deep learning i tak dalej, no to w Pythonie jest to często lepiej zrobić. {wywiad 5}

Konkurencyjność omawianych języków wyraźnie pokazuje powyższy fragment „na potrzeby takiego ogólnego data science nie ma dużej różnicy między eR i Pythonem, poza

283 XGBoost to popularny pakiet do uczenia maszynowego z wykorzystaniem *gradient boosting*, można używać pakietu w różnych językach programowania (por. Chen i Guestrin 2016).

tym że według mnie Python jest mniej wygodny", natomiast **komplementarność lub rozłączność** „deep learning w Pythonie jest dużo lepszy niż w eR (…) natomiast jak się spojrzy (...) na jakieś zastosowania (...) naukowe, nie ogólne data science, no to Python jest cieniutki, bo bibliotek ma dużo dużo mniej" {wywiad 5}.

Zatem czy w badanym świecie panuje zgoda, że „do prototypów eR, do wdrożeń Python" {wywiad 13}? Język R ma być łatwiejszy w pisaniu kodu, szybszy do celów uzyskania pierwszego, obiecującego wyniku – czasami nazywanego POC (*proof of concept*) – zaś w kolejnym etapie projektu tzn. wdrażaniu wyników na produkcję preferowany ma być Python, z uwagi na łatwiejszą integrację z systemami informatycznymi, wyższą niż R szybkość i wydajność działania:

> R: Z mojej perspektywy plusem eRa jest to, że on jest bardzo fajny do nauki i do analizy danych, no i takich **podstaw** programowania dla nieprogramistów. Zresztą eR ma kilka takich pakietów gdzie po prostu piszesz jakby słowami te analizy, jest dplyr gdzie obrabiasz tabelki i mówisz przefiltruj, wybierz kolumny tak, zagreguj, no w sensie używasz słów i operujesz na ramkach, tak? Jak używasz ggplota, no to też używasz słów i narysuj mi [niezrozumiałe], narysuj mi wykres, dodaj punkty, w sensie używasz słów, tak? W sensie jest taki, nie musisz być tam obeznany w języku programowania, tylko [krótka pauza] no i są te pipe'y, to już w ogóle, jest bosko. Więc on jest fajny, po pierwsze nie dla matematyków i myślę że niematematycy jeśli nie wchodzą jakoś głęboko w programowanie to nie potrzebują jakby Pythona, no chyba że, staną przed takim projektem gdzie masz terabajty danych i muszą to robić szybko, więc wtedy to trzeba się zastanowić nad jakimiś optymalizacjami, ale tak na ogół to są tabele danych, no to eR jest, w zupełności wystarczy bo jest prosty, jest tam no prostszy, w Pythonie to na, prostszy względem Pythona do nauki, jeśli ktoś nigdy nie miał styczności z programowaniem. Poza tym no eR jest właśnie bardzo fajny do takich szybkich analiz, nie? Dostajesz ramkę danych, szybko piszesz, od razu widzisz plus minus co tam się dzieje, tak? W momencie kiedy to trzeba zacząć jakby, wrzucać na produkcję, zarządzać wersjami, no w sensie żeby to szybko działało, optymalnie z czymś tam łączyć, no to wtedy się pojawia już taki problem że jednak Python. Bo Python jakby całą tą otoczkę programistyczną, ma znacznie lepszą. Python jest po prostu językiem programowania i on, jakby, może tak, eR został stworzony do analizy danych, a Python jest językiem programowania który ma jakby odgałęzienie, w analizie danych, tak? Więc jeśli ktoś się nie zna na programowaniu, a nie potrzebuje tak naprawdę programowania tylko chce sobie poanalizować dane, to raczej będzie bardziej skłonny do eRa, ale w momencie kiedy ktoś potrzebuje właśnie budować jakieś systemy, które potrafią analizować dane, tak, to to jest Python, bo on ma, wszystko co potrzebne żeby budować serwisy, jakieś tam, no masę różnych rzeczy. To jest taki pełnoprawny język programowanie, więc no nie ma wojny, raczej są plusy i minusy. **Do prototypów eR, do wdrożeń Python**, tak, to wszystko. {wywiad 13}

W badanym świecie rzecz jasna nie ma w tej kwestii konsensusu. Jak pokazywaliśmy na początku rozdziału, za lepsze niż Python do wdrożeń mogą być uznawane inne języki, jak Scala czy kompilowane, np. Java lub C++. Przy tym zarówno Python i R są wykorzystywane i do prototypów, i do wdrożeń (Borowiecki i Mieczkowski 2019:37-38).

Zauważmy w powyższym fragmencie, jak Rozmówca podkreśla zalety pakietów R „dplyr" i „ggplot2", będących częścią ekosystemu pakietów „tidyverse" H. Wickhama: „po prostu piszesz jakby słowami (…) nie musisz być obeznany w języku programowania" {wywiad 13}. W ten sam sposób mówi się o zaletach Pythona w ogóle, także poza światem DS – ma być to z założenia język zbliżony do języka ludzkiego (a właściwie angielskiego, pojawia się określenie *plain English*) (Dodge 2018; Garner 2018; Orsini 2014; Reitz 2018). Python ma być bardziej czytelny (*readable*) niż tzw. pseudokod, czyli sposób zapisu np. funkcji, będący czymś bardziej ścisłym niż opis prozą, ale nie będący zapisem w języku oprogramowania (Downey 2017). W Pythonie występuje relatywnie niewiele znaków specjalnych (w porównaniu do kompilowanych języków programowania), jak np. różnego rodzaju nawiasy. Kod ma być czytelny dla ludzi (*readable*), ponieważ uznano, że jest częściej czytany niż pisany. Czytelność kodu, łatwość w jego pisaniu, domyślna obsługa szczegółów typowych dla operacji na danych ma być zaletą, bowiem oszczędza to cenne zasoby kognitywne data scientista, który **zamiast skupiać się na samym kodzie może skoncentrować się na rozwiązywanym problemie** (Wickham 2018f).

Postulowane czytelność i prostota zostały w przypadku Pythona ujęte w spisane standardy, zasady, znane jako „Zen of Python", PEP20 i PEP8. Użytkownicy Pythona, nie tylko w DS, spierają się czy jakieś rozwiązanie w kodzie spełnia te standardy, co nazywa się kodem pytonicznym (*Pythonic*) (Reitz 2018:53–54). Dla języka R także podjęto próby standaryzacji, określenia zasad stylu pisania kodu (Baath 2012; Google 2012), ale w badanym świecie dominuje opinia, że w R niemalże każdy pakiet ma swoją osobną konwencję, oraz że użytkownicy nie wypracowali konsensusu (Muenchen 2014; Wickham 2014a). Wycinkami języka R z bardziej standaryzowanym sposobem pisania kodu są ekosystemy pakietów, jak tidyverse[284]. Naszym zdaniem pokazuje to, że w DS społeczność użytkowników R jest mniej niż społeczność Pythona zainteresowana tzw. dobrymi praktykami programistycznymi, znanymi ze świata IT. Ma być to także argumentem za tym, że język Python jest „lepszy" niż R do wdrożeń – kod w Pythonie jest bardziej standaryzowany, zatem np. ułatwia to komunikację zespołu DS z zespołem programistów podczas wmontowywania modeli uczenia maszynowego w infrastrukturę informatyczną (Borowiecki i Mieczkowski 2019:37–38, 41).

284 Przykładowo w całym ekosystemie działa operator zwany „pipe", zapisywany jako %<% (dosłownie rurka, por. wcześniejszy fragment „no i są te pipe'y, to już w ogóle, jest bosko" {wywiad 13}). Pipe nie działa w wielu pakietach poza ekosystemem, w tym w domyślnych pakietach GNU R, choć niektórzy twórcy spoza ekosystemu zapewniają działanie operatora w swoich pakietach. Konwencja pisania kodu w tidyverse obejmuje też szereg innych wytycznych, jak np. używanie spacji po obu stronach znaku „=" (por. https://style.tidyverse.org/, dostęp 02.02.2020 r.); istnieje pakiet służący formatowaniu kodu zgodnie z tą konwencją (Müller i Walthert 2020).

Relatywnie łatwiejsza niż dla R komunikacja z, ogólnie rzecz biorąc, światem IT to także przykład powodu, dla którego Python postrzega się w badanym świecie jako prawdziwy czy pełnoprawny język programowania. Może to wskazywać na postrzeganie Pythona jako języka dającego uczestnikowi świata DS większą sprawczość, a w konsekwencji użytkownikom Pythona umożliwiać łatwiejsze niż użytkownikom R osiąganie legitymizacji jako autentycznego uczestnika (choć już nie poziomu maestrii, który charakteryzuje raczej znajomość kilku języków programowania, por. 6.1.2.2). Zdaniem jednej z Rozmówczyń popularność Pythona w DS wynika: po pierwsze, z jego czytelności, po drugie, z kopiowania w pakietach Pythona rozwiązań znanych z MATLABa, po trzecie, z pojawienia się tych pakietów w czasie wzrostu zainteresowania tzw. sztuczną inteligencją:

> R: No więc [MATLAB] to jest taki kombajn matematyczny, w którym strasznie dużo można rzeczy zrobić i od jakiegoś odwracania wielkich macierzy, po jakiś nie wiem Lagrange'a i tak dalej. I to jest pogram który jest drogi w swojej licencji, i z tego co kojarzę, jest w ogóle taki ciekawy podcast na ten temat, jak chcesz mogę ci później podesłać, w każdym razie z tego to wiem, że Python zaczął wprowadzać te narzędzia które ma MATLAB, i stał się wersją MATLABa dla osób którzy nie chcą kupować licencji MATLABa, i tym, na tym ugrał właściwie teraz cały rynek, bo brakowało tego, była taka nisza, jednocześnie AI się stało popularne i nagle no na czym to robić, tak? Python po prostu dobrze wszedł z tym, a przy okazji jest łatwiejszy niż eR wydaje mi się po prostu zrozumiały. Każdy kto programował wcześniej w czymś innym to łatwiej mu jest przejść na Pythona, wydaje mi się niż na eR, które jest takie [pauza] hardcorowo statystyczne, a Python jednak nie jest statystyczny w swojej strukturze, tylko bardziej przypomina Javę, czy cokolwiek, więc to pewnie też z tego wynika popularność Pythona. {wywiad 10}

Uznajemy **użytkowników Pythona i R za dwa subświaty świata społecznego data science zaznaczając, że są to subświaty rozmyte i przenikające się**. Część uczestników – mniej więcej między 10% a 25% (por. 5.3.1.) – używa przecież obu języków. Mogą oni preferować użycie jednego lub drugiego, jak i kierować się zasadą doboru „najlepszego" narzędzia do danego problemu. Widzimy jednak specjalizację tych rozmytych subświatów, tak w sensie używanych metod, zastosowań DS, jak i używania specyficznych narzędzi ułatwiających korzystanie z języka preferowanego w każdym z subświatów. Nie jest to wniosek nowy:

> Podział według rodzaju rozwiązywanych problemów odzwierciedla do pewnego stopnia różnice w używanym sprzęcie i technologiach, używanych do tego celu. Zatem różne subświaty tradycyjnie używają różnych maszyn, różnych języków [programowania], i często różnych podejść, rozwijając swoje procedury i aktywności (Kling i Gerson 1978:27).

Istnieje tutaj arena dotycząca tego, jakie elementy języków są konkurencyjne, komplementarne i rozłączne. Inne niż Python i R języki programowania mogą mieć silnie specjalistyczny charakter w sensie zastosowania (jak Scala, Julia, Lua, Lisp, Prolog), mogą

być charakterystyczne dla wybranych obszarów nauki (MATLAB, Octave) lub biznesu (SAS), mogą być językami świata IT użytecznymi w działaniach wspomagających (we wdrożeniach produkcyjnych – np. Java, C++; w tworzeniu interfejsów dla użytkowników wdrożeń – Ruby, JavaScript, nawet HTML i CSS).

Spór R vs. Python na początkowym etapie naszych badań wydawał nam się niezwykle istotny dla badanego świata, obiektem granicznym jest przecież narzędzie do wykonywania działania podstawowego. Spodziewaliśmy się zatem głębokiego podziału na dwa subświaty użytkowników dwóch języków. Szczególnie w formie online wyjątkowo łatwo było nam dotrzeć do rozmaitych materiałów porównujących wady i zalety obu języków, spór wydawał się żywy i intensywny. Podczas obserwacji często spotykaliśmy się z żartami na temat omawianego sporu. W toku wywiadów swobodnych nasza wstępna wizja areny szybko okazała się mylna, a robocze opisy – powierzchowne. Twierdzimy, że:

- spory o konkurencyjność, komplementarność i rozłączność cech języków Python i R oraz tego, który w jakim zakresie jest lepszy, uczestnicy badanego świata interpretują głównie w kategoriach pro-rozwojowych, przyjacielskich sprzeczek, które przyczyniają się do budowania społeczności DS, ewentualnie wskazując na to, iż są one próbami sił nastawionymi na legitymizacje biegłości technicznej spierających się osób;
- te spory widziane online wydają się znacznie bardziej intensywne, niż są one w opinii osób osadzonych w badanym świecie – jest to łatwe, a nawet rozrywkowe czy wręcz zabawne pole do spierania się między uczestnikami i subświatami świata DS;
- spory od 2018 r. straciły na znaczeniu, gdyż język Python zaczął zdecydowanie „wygrywać" w sensie ilościowym – bazy użytkowników jak i dominacji w uznawanym za najbardziej atrakcyjny elemencie działania podstawowego czyli modelowaniu metodami ML i deep learning (głównie pakiety scikit-learn, PyTorch, TensorFlow), zaś R stał się bardziej niszowym językiem cenionym jako dobre narzędzie przede wszystkim do statystyki, analizy i wizualizacji oraz raportowania danych (np. pakiety tidyverse, Markdown, Shiny), niemniej w R także buduje się modele, w tym deep learningowe oraz R ma bazę użytkowników nie identyfikujących się z DS, np. w bioinformatyce lub akademii (Olszewski 2017, 2018).

Odnośnie pierwszego wniosku, przyjrzyjmy się wypowiedzi silnie osadzonego w badanym świecie Rozmówcy, który wskazuje właśnie na pro-rozwojowy, właściwie budujący

społeczność DS, charakter sporu o wybór jednego z dwóch języków do wykonania działania podstawowego. Kolejny raz zauważmy koncentrację silnie osadzonego uczestnika na doborze właściwego narzędzia do problemu (por. 6.1.2.2.), na tzw. agnostycyzm technologiczny, według naszego określenia:

> R: (…) jeżeli ktoś przychodzi z własną technologią, ja zawsze to mówię na rozmowie kwalifikacyjnej, jeśli to jesteś w stanie robić w swojej komórce i na kartce papieru, dla mnie to nie jest problem póki to co robisz jest [pauza] wysokiej jakości i jest projekt zmierza w tą stronę w którą chcielibyśmy żeby zmierzał, nie ma sprawy. To jest wy to jest wiesz, to są takie święte wojny, nie? Każdy to w różnych dziedzinach ma. (…) Nie widzę sensu w ogóle dzielenia tego, natomiast oczywiście wszyscy w tym zawodzie, trzeba być, są ambitni, każdy uważa się za specjalistę w czymś bardzo często jest, więc mamy te swoje małe wewnętrzne zabawy w kłócenie się o wyższości Pythona nad R lub przeciwnie, co moim zdaniem jest fajne bo też powoduje że obie strony się uczą tak? (...) Więc summa summarum no czymś musimy się zajmować oprócz tej analizy danych. O czymś musimy się przy piwie tam dyskutować. {wywiad 6}

Sam udział konkretnej osoby w sporze o wady / zalety Pythona i R w DS jest (a raczej był) nakierowany na legitymizowanie tej osoby jako uczestnika świata DS. Jeżeli osoba bierze udział w sporze, to zaświadcza o, przynajmniej początkującym, poziomie rozeznania w najważniejszych narzędziach do wykonywania działania podstawowego i jakimkolwiek doświadczeniu w wykonywaniu działania podstawowego, a także o byciu osobą rozeznaną w technologii w ogóle. Wydaje się nam, że wraz z rozwojem nieco różnych pakietów i nieco różnych specjalizacji Pythona i R w DS ten spór stał się sporem zwyczajowym, charakterystycznym raczej dla wannabesów i początkujących uczestników niż dla „prawdziwych" uczestników badanego świata. Ze względu na jasną opozycję: R vs. Python, i wielokrotne powtarzanie tych samych argumentów w sieci, uczestnictwo w sporze stało się łatwo dostępnym elementem enkulturacji do badanego świata. Odnosząc się do drugiego wniosku, powtórzmy, że jest to spór znacznie bardziej wyraźny online niż z perspektywy uczestników świata DS, spór podkolorowany przez mechanizmy działania mediów społecznościowych, a w naszej opinii także rzeczników świata DS. Bez wątpienia jeszcze kilka lat temu treści w rodzaju „czy zacząć przygodę z DS od R czy Pythona" były ogromnie popularne w komunikacji rozmaitych szkoleń / kursów / MOOCów skierowanych do wannabesów.

> *B: Czy jest jakaś wojna pomiędzy eRem a Pythonem, czy pomiędzy użytkownikami, czy to jest w ogóle jakieś złe podejście? Czy ja nie powinienem tak o tym myśleć?*
> R: Nieee. Sądzę, że [pauza] yyy kiedyś, czasy kiedy się ludzie jeszcze lubili o coś kłócić to już chyba minęły, ja pamiętam czasy jeszcze jak się kłócono o to która wersja Linuxa jest najlepsza, yyy jakieś jeszcze dawnymi czasy to oczywiście najwięksi internetowi fighterzy kłócili się które sporty walki są najlepsze, czy by wygrał karateka z judoką [z uśmiechem] i tak dalej. Kiedyś,

> teraz się ludzie już właściwie, każdy taki, taka dyskusja yym i właściwie to są już takie no poprzez nostalgię, tak właściwie, tak raczej jestem, nie sądzą żeby to było na poważnie, to raczej teraz to już tak po prostu yyy [pauza] no, wszyscy to traktują to z przymrużeniem oka, przynajmniej poważni, mówię o poważnych ludziach, bo wiesz zawsze są internetowi napinacze co tam właściwie nie używają ani jednego ani drugiego, tylko generalnie yyy eee lubią sobie tak pona byyy, się opowiedzieć za którymś obozem, ale ludzie którzy po prostu normalnie tym pracują z tego, żyją z tego, pracują na co dzień, no to po prostu zdają sobie sprawę że to jest, że zazwyczaj jest to mixem wielu różnych przyczyn [krótka pauza] ten wybór danej technologii yyy i to nie jest takie proste w ogóle nie można, yyy żeby powiedzieć że ta jest lepsza, czy ta jest gorsza, często po prostu jest to po prostu nawet przypadek, iii [krótka pauza] ok, jak trzeba to będą pracować w eRze, jak trzeba to będą pracować w Pythonie. Yyy jedynym problemem z Pythonem, z eRem jest to że jak szukaliśmy kogoś na praktyki to bardzo ciężko było znaleźć kogoś ze znajomością eRa. Yyy i to był duży problem, yy łatwo było znaleźć kogoś kto zna, względnie łatwo było znaleźć kogoś kto zna Pythona, a nawet jak ten, noo to jest problem. {wywiad 14}

Osoby opowiadające się – w sposób nie żartobliwy ani ironiczny – zdecydowanie „po stronie" Pythona lub R, ostro i przede wszystkim niemerytorycznie krytykujące język, którego nie używają, są w badanym świecie oceniane negatywnie. Takie działanie świadczy raczej o byciu wannabesem, ewentualne początkującym, który serio traktuje internetowe spory, a nie rozumie i nie ma doświadczenia w wykonywaniu działania podstawowego, nie rozumie projektów DS, właściwie nie działa zgodnie z wartościami badanego świata. Oczywiście nawet najbardziej doświadczeni uczestnicy świata DS, uznawani za osoby o najwyższym poziomie maestrii, mają swoje jawnie deklarowane preferencje wobec używanego języka programowania do wykonywania działania podstawowego, jednak nie spotkaliśmy się w toku badań z ostrą krytyką „tego drugiego" języka ze strony takich osób. Taka krytyka być może podważałaby status maestra w oczach innych uczestników.

Odnotujmy, że w ramach Pythona (do działań zbliżonych do DS) i R istnieją także wewnętrzne mini-areny narzędziowe, mogą one wykraczać poza świat społeczny DS. Wśród użytkowników R występują np. spory użytkowników ekosystemu tidyverse z osobami korzystającymi z R bez jego użycia (Matloff 2020), a także spory użytkowników tidyverse (lub węziej – pakietu „dplyr") z użytkownikami pakietu „data.table", dotyczące głównie zastosowań w przetwarzaniu danych (BrodieG 2018; Wickham i in. 2019). W ramach Pythona użytkownicy spierają się m.in. o wady i zalety pakietów do ML i uczenia głębokiego: TF i PyTorcha (por. Sopyła 2019) lub Keras + TF a PyTorch (Agarwal 2019; Misal 2018).

Mniej więcej **od 2018 r. Python uznawany jest, przynajmniej w komercyjnej sferze działalności badanego świata, za domyślny język do wykonywania działania podstawowego**. Język R należałoby poznać czy chociażby potrafić się posłużyć nim jako drugim, w konkretnych zastosowaniach np. analitycznych czy statystycznych, posługując się

już swobodnie Pythonem i stając się silnie osadzonym uczestnikiem świata DS. Istnieją rozwiązania pozwalające na używanie obu języków w jednym środowisku programistycznym (Allaire, Ushey, i Tang 2018; Jolly 2018). Pod koniec naszych badań, w pierwszej połowie roku 2019 spotykaliśmy się z opiniami użytkowników R, że uczą się Pythona czy nawet „przesiadają" na Pythona, bo jest to korzystne dla rozwoju ich kariery zawodowej. Python jest zatem postrzegany w badanym świecie raczej jako język bardziej sprawczy niż R. Jest to spowodowane znacznie większą bazą pakietów do ML i uczenia głębokiego, obecnie wysoko komercyjnie cenionych technologii, a także szeregiem cech technicznych samego języka, silnie powiązanych z rodowodem tegoż języka – „pełnoprawnego" języka programowania do ogólnego zastosowania. Język Python „wygrał" z R na opisywanej arenie, ponieważ bardziej odpowiada komercyjnemu kierunkowi rozwoju społecznego świata DS, miał duży napływ początkujących data scientistów rekrutujących się z obszaru IT. Te osoby najczęściej znają Pythona z innych zastosowań. Obszar IT jest ilościowo większy niż obszar matematyki/statystyki akademickiej i posiada większy kapitał finansowy, jak i więcej władzy (Matloff 2017). Ponadto pakiety używane w najbardziej atrakcyjnym i komercyjnie cenionym elemencie wykonywania działania podstawowego wspierane są przez gigantów technologicznych (co jest zbieżne z wyższym kapitałem IT, co opisaliśmy przy arenie modelowania, część 6.2.1.):

> *B: Hm, okey, a kto wygra? eRowcy czy Pythonowcy?*
> R: No powiem ci że po raz pierwszy w tej pracy mam takie duże zderzenie bo ja, bo połowa zespołu używa eRa połowa zespołu używa Pythona i naprawdę widzę gigantyczne różnice. Wydawałoby się że nie ma takich różnic absolutnych, bo do jednego się tego używa. [pauza] A nie, różnice są bardzo duże. Eee [pauza] kto wygra to nie wiem bo rzeczywiście też jest tak że część bibliotek wspiera eRa przykładowo tam, a część wspiera tylko Pythona na przykład yyy PyTorch czy Tensorflow mają tylko interfejs Pythonowy a Keras ma wszystkie, więc w momencie kiedy ja używam eRa i nie znam Pythona, no to już PyTorcha yyy, czy Tensorflowa nie mogę używać. Co jest trochę smutne [śmiech].
> *B: OK, ale to na czym, to na czym polegają te różnice? Powiedziałaś że są wielkie różnice, yy, pomiędzy ludźmi którzy pracują w tych językach, czy pomiędzy czym?*
> R: Nie, pomiędzy możliwościami. No przykładowo pierwszy projekt polegał na tym że musieliśmy robić wizualizacje na mapach, no i, może nie dokładnie dokończyłam research, ale kolega w Pythonie, był w stanie dużo szybciej i dużo ładniejsze takie wizualizacje na mapach dynamiczne zrobić niż ja. Mi to zajęło dużo czasu i nie, nie wyglądało to aż tak dobrze, no ale z kolei w eRze mogłam trochę pooszukiwać w momencie jak mieliśmy bardzo **dużą** ilość danych i w momencie gdy ta sama biblioteka Pythonowa już sobie nie radziła z ilością powiedzmy punktów na mapie to w eRze dało radę jeszcze to obsłużyć.
> *B: Aha, okey, okey.*
> R: A jeśli chodzi o same modele no to w obu językach no to wiadomo podstawowe metody są te same więc pewnie bardziej zaawansowane **też** są te same, [pauza] no ale w sumie różnicę masz na przykład, właśnie jeśli chodzi o, radzenie sobie z dużymi danym. Na przykład w Pythonie była możliwość jeśli nie dało się policzyć jakieś statystyki na jakimś zbiorze bo był zbyt duży, była możliwość z automatu wylosowania jakichś tam próbek i poczynienie statystyki na tych

> próbkach. W eRze musiałabym to losowanie i próbkowanie robić **ręcznie** [parsknięcie]. Także to no jak ktoś się zaczyna uczyć to tych różnic nie ma, tak? No bo te pierwsze podstawowe algorytmy mają dokładnie te same możliwości, natomiast jeśli już dojdzie do szczegółu to powiedzmy że tu czegoś brakuje, tam czegoś nie ma [śmiech], chyba jednak najlepiej na takim wysokim poziomie znać **oba**.
> *B: OK, rozumiem. Yyy, ale czy jest jakaś taka wojna pomiędzy środowiskami? Bo mi się tak czasami wydaje, niektórzy mi mówią że tak, a niektórzy mi mówią że nie. Ja już nie wiem co mam o tym myśleć. [śmiech] Może mi coś wytłumaczysz?*
> R: Yy, nieeeee no myślę że wojny jako takiej nie ma, bardziej to są takie **docinki**, tak jak ja w poprzedniej pracy były docinki między ludźmi po [nazwa uczelni ekonomicznej A] a po politechnice to tak samo były docinki między ludźmi którzy znali Pythona a ludźmi którzy znali eRa. (…)
> *B: Mmm ale jeszcze, jeszcze kończąc wątek tego kto wygra no może tak głupio zapytałem, ale yy czy mm, jaką wróżysz przyszłość językowi eR i językowi Python w data science?*
> R: Znaczy tak jak teraz na to patrzę to obawiam się że to jednak trochę duże korporacjeeee **wymuszą** przyszłość czegoś. No właśnie tak jak teraz jest nadal na machine learningu na deep learning to właśnie rozwijają się różne frameworki typu Tensorflow, PyTorch czy Keras a to wszystko było pierwotnie rozwijane przez Google albo przez Facebook. Mmmm, no więc to co te duże korporacje sobie że tak powiem wymyślą, narzucą, będą finansować, to to się będzie rozwijało, ale no mimo wszystko kiedy jedne i drugie jest open source'owe ale mimo wszystko ludzie z dobrej woli nie piszą tych pakietów i tak zawsze stoi ktoś jeśli chodzi o takie poważniejsze pakiety no to zawsze stoi za tym ktoś kto jednak ma budżet. Tak więc po czyjej stronie będzie ten **budżet** ten wygra [ze śmiechem]. {wywiad 18}

Być może wolno opisać spór pomiędzy językiem Python i R także w perspektywie makro, jako spór o wpływy w świecie DS dwóch obszarów, na których przecięciu powstało DS. Chodzi o informatykę/IT oraz matematykę/statystykę (por. rys. 2.), a z naszej rekonstrukcji sporu na opisywanej arenie wynikałoby, że zwycięża wpływ informatyki/IT, dlatego, że ten bardziej sprawczy, mniej teoretyczny obszar niż matematyka/statystyka uzyskał silniejsze wsparcie trzeciego gracza – obszaru szeroko pojętego biznesu. W amerykańskim środowisku DS tego rodzaju wnioski pojawiły się już kilka lat temu (Matloff 2017). Ta walka jest w naszym przekonaniu kontynuacją walki, opisanej już prawie 20 lat temu jako konkurencja dwóch różnych kultur modelowania (Breiman 2001). Być może pozostajemy pod wpływem praktyk legitymizacyjnych badanego świata polegających na twierdzeniach, iż wciąż mamy do czynienia z wczesnym stadium rozwoju DS (por. część 6.1.3.) – zarówno technologii, branży jak i rynku na takie usługi, także regulacji prawnych, a wszystko to szczególnie w Polsce – ale sądzimy, iż **kierunek rozwoju badanego świata DS może być taki, że w sferze komercyjnej zostanie on niemal całkowicie wchłonięty przez IT**, stanowisko data scientista zniknie na rzecz inżynierów danych, ML engineerów i ML reasercherów, **a systemy bazujące na modelach ML staną się w szerokim dyskursie publicznym tak samo mało magiczne jak strony internetowe czy aplikacje mobilne**, a nawet jak samochody czy lodówki. Obszar analizy danych, ich wizualizacji i raportowania być może stanie się na powrót przede wszystkim sferą nie-techniczną, bardziej biznesową,

być może dalej bazując na takich narzędziach jak arkusze kalkulacyjne, ale prawdopodobnie na elementach języków SQL, R, a przede wszystkim na przyjaznych dla użytkowników narzędziach z GUI jak np. MS Power BI, Google Data Studio lub Tableau.

Jako podsumowanie przytaczamy dłuższą wypowiedź Rozmówcy, definiującego się jako ML researcher, także pracującego w akademickiej informatyce. Rozmówca rekonstruuje całe tworzenie się obszaru DS jako walkę o wpływy pomiędzy informatyką (w której trzeba odróżnić co najmniej obszary baz danych, ML i hardware) a matematyką/statystyką, częściowo innymi dziedzinami obliczeniowymi, a także jako walkę na narzędzia i metody do operacji na danych:

> *B: Ja mam taką yyy powiedzmy hipotezę, że zmieniła się trochę terminologia przez te 10 lat może trochę dłużej, dotycząca takiej zaawansowanej analizy danych, no do potrzeb praktycznych, że chyba początkowo była moda na taki termin data mining, potem termin big data, potem data science i od niedawna sztuczna inteligencja, o której pan wspomniał na początku, bardzo za to dziękuję, bo miałem później o to pytać. Jak pan profesor sądzi, czy jakaś ewolucja nastała, czy to są rzeczywiście nowe rzeczy, czy to jest taki rebranding, co roku jakiś nowy fajny termin?*
> B: Tak, dokładnie. To jest w bardzo dużej mierze, to jest face lifting, to jest niestety tak, że ten kawałek tortu jest ograniczony, a każdy próbuje się na niego wepchnąć. I historycznie rzecz biorąc to było tak, że u podłoża wszystkiego leży statystyka i statystycy patrzą na nas, na informatyków z prawdziwą pogardą, mówią: [zmienionym głosem] przecież my to wszystko od 100 lat wiemy, a wy wyważacie, nie to, że te drzwi są lekko uchylone, to są drzwi od **stodoły**, które są otwarte na oścież, a wy tam biegniecie i krzyczycie, że coś odkryliście. A my im próbujemy udowodnić, że my jednak robimy rzeczy nowe, że nigdy żaden statystyk nie wziął 100 tys. transakcji sklepowych i nie policzył fizycznie, zakupy jakich produktów z czym się korelują, i w jaki sposób. I faktycznie data mining czyli to, co jest tłumaczone jako eksploracja danych, to był świat ludzi od baz danych, którzy dysponowali ogromnymi wolumenami danych i próbowali zobaczyć, jakie metody statystyczne dadzą się zastosować biorąc pod uwagę ówczesne ograniczenia technologiczne, obliczeniowe, no itd., itd. Ale bardzo wyraźnie to byli ludzie pochodzący ze świata baz danych, hurtowni baz danych. Zawsze była część ludzi, która się zajmowała terminem uczenia maszynowego, chociaż to było traktowane trochę po macoszemu, bo chętniej używali terminu decision support system, czyli systemy eksperckie, czy systemy wspomagania decyzji. I oni nagle stwierdzili, że ten termin jest mało sexy, tym bardziej, że część ich tortu zjadali ludzie, którzy się zajmowali ekonomią, bo to też trochę taka teoria decyzji, podejmowania decyzji, troszeczkę w tym jest ekonomii, trochę psychologii, różne rzeczy. No więc oni też się zaczęli rozglądać, bo tutaj ci nam wchodzą na nasze podwórko z jakimś data mining, a ci z kolei zabierają, no więc oni wpadli na pomysł, może odgrzebać ten termin machine learning i my teraz będziemy ML-owcy. Na to oburzyli się ludzie, którzy się zajmowali bazami danych, ale ponieważ popularność terminu data mining trochę zaczęła się wypłaszczać, trzeba było znaleźć coś nowego, no to big data. No to big data, teraz to są już w ogóle tryliony i to się nie mieści w pamięci i większość waszych algorytmów o kant tyłka roztłuc, bo jak algorytm wymaga obliczenia globalnego i byśmy potrzebowali terabajta RAM-u, no to po prostu się nie da. Więc znowu teraz będziemy pracowali nad algorytmami, które się skaluje. Jak się to zaczęło skalować, to nagle wrócili ludzie od sprzętu i mówili: halo, halo, ale my mamy tutaj procesor GPU i my możemy wam wszystko zrównoleglić, i teraz to wszystko będzie równolegle. I się zaczęło obliczanie równoległe i teraz jest big data, ale ona jest on GPU. Więc tu z kolei obudzili się statystycy i matematycy z ręką w nocniku i mówią, ale co z nami? więc **oni wymyślili data science**. Tyle tylko, że oni niestety wszyscy umieli programować tylko w MATLAB-ie, informatycy im powiedzieli, że ten MATLAB to jest kompletny bull shit, nikt

> w tym nie będzie programował. Oni najpierw się obrazili, a potem zrobili sobie język eR. Przebrandowili tego MATLABA, pododawali trochę, zrobili z tego więcej statystyki, i powiedzieli, to już się doskonale programuje, to jest prawdziwy język programowania. Czyli data science się narodziło tak naprawdę w tym środowisku R. Jako statystyka plus troszkę programowania plus wizualizacja. Rzeczywiście po raz pierwszy wizualizacja poszła tak mocno do przodu, pojawiła się ta cała gramatyka dla wizualizacji, potem cały ten ggplot itd. No to znowu wrócili informatycy, którzy zobaczyli statystyków i matematyków, którzy piszą w jakimś języku, który jest nieczytelny, w którym nie ma żadnych standardów, żadnego programowania obiektowego, to wszystko w ogóle kompletnie jest [pauza] jak informatyk patrzy na eRa, to mówi to jest jakiś koszmar zupełny, my weźmiemy normalny język programowania, w którym są pewne standardy programowania, biblioteki są, itd. i my z tego zrobimy swojego eRa. i zaczęli pracować nad Pythonem i zaczęli go rozwiązywać no i teraz zaczęła się ta bitwa, co jest ważniejsze. Na to się nałożyło to, że paru ludzi w Toronto odgrzebało sprzed 20 lat sieci neuronowe i nagle się okazało, że te sieci neuronowe biją wszystkich kompletnie na głowę w przetwarzaniu języka naturalnego do obrazu i dźwięku. No to trzeba było ukuć jakiś nowy termin no to deep learning. Cały ten deep learning na dobrą sprawę to jest połączenie dwóch rzeczy, to jest trochę algebry liniowej i trochę optymalizacji kombinatorycznej i te dwie rzeczy, jak się do siebie doda, to wychodzi deep learning. Ale faktem jest, że to co zmienia całkowicie obraz to jest skala. Co innego to jest zastosowanie tego do niewielkiego zbioru danych, pobawienie się, to jest 1000 przykładów i coś tam z tego wychodzi. A czym innym jest przepracowanie 10 mln zdjęć i nagle pojawienie się modelu, który dostaje zdjęcie i mówi proszę bardzo to jest kot, a to jest płot, a to jest zielone słońce. Ale myślę, że tą ewolucję terminologii nomenklatury trzeba rozpatrywać z jednej strony, jako trochę wyszarpywanie sobie rynku, bo z punktu widzenia ludzie, którzy się tym zajmują, po drugiej stronie, tam gdzie są firmy i gdzie są pieniądze, nie ma tak naprawdę, ani wiedzy, ani potrzeby rozumienia tych subtelnych różnic, więc trzeba przepchnąć swoje. Cokolwiek tam zadziała, jakikolwiek hejt będzie aktualny, to na tej fali po postu popłyniemy. Ludzie, którzy w tym siedzą głęboko, pewnie dostrzegają jakieś subtelne różnice, z mojego punktu widzenie chyba nie ma to aż tak dużego znaczenia, bo wszystko tak koniec końców, jak się na to spojrzy z całkowicie praktycznego punktu, to ma znaczenie dla ludzi, którzy siedzą w akademii, piszą artykuły i próbują siebie wpakować w jakiś taki wąski fragment i zbudować sobie jakieś swoje audytorium i to jest zbiór czasopism i konferencji dla big data, to jest konferencja dla ludzie data science. Natomiast jak się popatrzy na rzeczywisty problem, gdzie przychodzi jakaś firma i mówi taki, siaki, owaki problem do rozwiązania to tam i tak wszystko wraca do jednego wielkiego worka, bo się okazuje, że jest big data, bo się okazuje, że są ogromne wolumeny danych, jest potrzebny taki niskopoziomowy feature engineering i tam trzeba po prostu usiąść i zakasać rękawy i programować. I jest potrzebny ktoś, kto ładnie to zwizualizuje, i jest potrzebny ktoś, kto odpali eksperymenty i jest potrzebny ktoś, kto zna się na przetwarzaniu równoległym, no bo inaczej to się nie skończy itd. Potem to i tak wszystko wraca do jednego dużego kubka. Myślę, że nie ma co specjalnie się tym ekscytować, zawsze będzie kolejne słowo, teraz jest reinforcement learning[285] jako pomysł, że teraz **wszystko** będziemy w ten sposób uczyli, niezależnie od tego, czy to ma sens, czy nie ma sensu. Za chwilę znowu pewnie wybuchną algorytmy ewolucyjne, bo ludzie już kombinują nad tym. Myślę, że to jest bardziej szum informacyjny, niż jakąś prawdziwa wartość jest w tej ewolucji terminów. {wywiad 23}

Różnego rodzaju technologie, które omówiliśmy w części 5.3.1., a które mają ułatwiać korzystanie z języków programowania, oferują szereg udogodnień w zakresie wykonywania działania podstawowego (pisania kodu do operacji na danych) jak i technicznych działań wspomagających (instalacji i zarządzania środowiskiem, publikacji wyników, kontroli

285 Oznacza uczenie maszynowe ze wzmocnieniem (*reinforcement learning*), stosowane np., w AlphaGo z 2015 r., w uproszczeniu polega na tym, że model gra w grę przeciwko sobie (por. Silver i in. 2016, 2017, 2018).

wersji). O ile jednak **udogodnienia do działań wspomagających mogą przyjmować formę interfejsu graficznego do wyklikania operacji, to udogodnienia do działania podstawowego takiej formy nie przyjmują**. Stanowi to dla nas wyraźny wskaźnik tego, że pisanie kodu jest immanentną częścią działania podstawowego w badanym świecie, zatem nasze ujęcie działania podstawowego jako „pisania kodu do przetwarzania, analizy i modelowania danych" (por. 5.2.) oddaje perspektywę uczestników badanego świata.

Stąd, że działanie podstawowe jest wykonywane za pomocą kodu, a nie w interfejsach graficznych, biorą się w świecie DS (nie tylko w Polsce) żarty dotyczące używania niezwykle popularnego Excela do pracy z danymi, przykłady używania tej aplikacji w sposób niezgodny z przeznaczeniem – np. jak bazy danych, jako interfejsu użytkownika, powtarzające się listy problemów z tego typu narzędziem (Gutierrez 2014:110; Hamideh 2015; Smart 2013; Wickham 2018f). Oczywiście chodzi nie tylko o Excela, ale stał on się w badanym świecie synonimem arkusza kalkulacyjnego w ogóle, więc będziemy używać tego określenia – tak samo zatem do działania podstawowego nie służą np. programy Google Sheets, LibreOffice Calc ani Apple Numbers. Jeden z Rozmówców {wywiad 9} pokazał w trakcie wywiadu mem, w którym Python i R walczą, ale zgadzają się co do tego, że „nienawidzą" Excela (rys. 49):

Rysunek 49: Co do tego jesteśmy zgodni: Python i R walczą, ale obaj nienawidzą Excela – mem internetowy

Źródło: https://www.facebook.com/Rmemes0/photos/a.1230204967031792/2078838445501769/, dostęp 09.26.2018 r.

W wywiadach szukaliśmy interpretacji podobnych żartów, uznając je za ważny element świata społecznego (por. Jemielniak 2019:132). Uczestnicy badanego świata podają wiele argumentów technicznych dotyczących tego, że nie da się wykonywać działania podstawowego w arkuszach kalkulacyjnych. Zauważmy, że **używanie np. Excela jako podstawowego narzędzia pracy odgranicza uczestników świata DS od innych osób pracujących z danymi**.

> *B: Dlaczego [pauza] a może nie dlaczego czy data scientiści śmieją się z Excela? Albo z ludzi, którzy korzystają z Excela? Czy mnie się tylko wydaje że tak jest?*
> R: Ja myślę że to chodzi o coś innego
> *B: Jest słowo Excel i wszyscy mają wiesz [B uśmiecha się szeroko]*
> R: [śmiech]
> *B: jest banan po prostu, nie? Dlaczego tak jest? [ze śmiechem]*
> R: Eee, [śmiech] Ja myślę, że głównie chodzi o to, o możliwości techniczne Excela, bo ja wciąż mam wrażenie że niektórym się wydaje że Excelem można zrobić wszystko, a niestety, OK dobra powiem **moje** prywatne takie ten, nie to że się nie śmieję, prowadzę czasami szkolenia z tego żyję z tego to uważam że jest to pierwszy krok, mówię ludziom też że jakby kiedyś potrzebowali to są **lepsze** narzędzia doo do pracy z tym, ale wiele rzeczy można po prostu szybko zrobić w Excelu i nikt mnie nie przekona że zrobi je szybciej w jakimś tam innym narzędziu, nie? Jakiegoś szybkiego pivota czy coś takiego no bez sensu zatrudniać brać nie wiem yyyy jakiś młot ogromny kowalski jak wystarczy młoteczek, nie? Także dobór narzędzia w zależności od tego jakie są potrzeby. (...) Natomiast nie podzielam tego mówię, hejtu na Excela. Może to też wynika z tego żeby się jakoś oddzielić, że OK to są ci co robią Excela, a my robimy tylko w tym, nie? I jeżeli nawet mamy do wykopania małą dziurkę w ziemi to używamy tam kombajnu, nie? (...) Ja jestem jestem, staram się pamiętać skąd wyszedłem yyy od czego wyszedłem i też myślę że dla mnie jest to ważne **szczególnie** że prowadzę te szkolenia Excela i prowadzę je moja główna grupa docelowa to są właśnie początkujący ludzie, eee i cały czas trzymam ten obraz w głowie, ee, kim byłem ja jak nie znałem Excela bo w ten sposób potrafię przez osiem godzin ludziom pokazać jakie możliwości są jakby **analizy** tych danych, oczywiście w ograniczonym zakresie, bez użycia nawet jednej funkcji [śmiech] oni nie muszą nawet pisać żadnej żadnych funkcji znaczy pisać formuł używać żadnych funkcji żeby yy **pomóc** sobie po prostu. {wywiad 3}

Excel jest popularnym, ograniczonym i właściwie jednym z najprostszych możliwych narzędzi do jakiejkolwiek pracy z danymi. Data scientist nie używa Excela do wykonywania działania podstawowego, tak jak kucharz w restauracji klasy Premium nie przyrządza posiłków z użyciem plastikowego noża, może jednak bez uszczerbku na profesjonalnym wizerunku użyć go na biwaku z rodziną. Fanatyczne unikanie lub krytykowanie arkuszy kalkulacyjnych będzie w badanym świecie raczej postrzegane negatywnie, a także jako cechujące wannabesów lub początkujących, chcących w niewłaściwy sposób legitymizować autentyczność swojego uczestnictwa w świecie DS. Definiowanie np. stanowiska pracy w ofercie rekrutacyjnej jako „data scientist" przy wymaganiu przede wszystkim zaawansowanej znajomości Excela jest postrzegane jako tzw. fake data science, czyli podszywanie się pod

świat DS. O tym wszystkim mówi Rozmówca pointując – „[Excel] to nie jest poważne narzędzie” {wywiad 12}:

> *B: Ok. Chciałem cię jeszcze zapytać o [pauza] o Excela, w zasadzie, bo powiedziałeś na początku że niektórzy nazywają się data scientistami, a pracują tylko w Excelu. No rozumiem że dla ciebie to jest ściema, że tacy ludzie nie są data scientistami?*
> R: Nie, to nie jest ściema, eee chodzi o to, że jeżeli udałoby się takie rozgraniczenie stworzyć, jeżeli moje drzewo, mój model w głowie, zobaczyłby że użytkownik umie tylko Excela to powiedziałbym bardziej że jest analitykiem danych, niż data scientistem. Mmm wiadomo no można zrobić jakieś tam wygładzanie wykładnicze w Excelu, czy tam jakieś inne metody, które, no ale, no to są bardzo podstawowe metody, jednak mimo wszystko.
> *B: Mhm, nie ja też pytam, pytam też o to dlatego, bo zauważyłem że [pauza] czy tam na jakimś meetupie, czy na jakiejś konferencji data sciensowej jak ktoś coś powie o Excelu to ludzie się śmieją, zastanawiam się co jest takiego śmiesznego w tym biednym arkuszu kalkulacyjnym?*
> R: Nie, nie ma nic śmiesznego, bo to nie jest tak że to środowisko, to nie jest tak że ja nie korzystam z Excela, bo żeby policzyć 2 plus 2
> *B: Jestem pewien że na ekonometrii korzystałeś dużo z Excela, ale może teraz nie*
> R: No korzystałem dużo z Excela [pauza] ale teraz nie, bo to nie chodzi o to tak że się ludzie śmieją, po prostu [pauza] ja **korzystam** z Excela, ale do takich rzeczy nie wiem jak gdzieś mam, muszę wyjechać, policzyć koszty eee [pauza] no bo tak naprawdę jeśli, można by popatrzeć do czego mógłbym używać Excela. I ja tak naprawę teraz w tym momencie nie jestem w stanie żadnego sensownego wykorzystania Excela znaleźć, to jest [pauza] typowo taki **menadżerski**, narzędzie, bo jeżeli chcę coś sobie zwizualizować, to nie używam, na szybko to użyję Power BI, Tableau, eee albo jakiegoś każdego innego narzędzia, no Power BI jest chyba najbardziej sensowny, więc yyy [krótka pauza] albo nawet jak mam po prostu jakiś problem to wizualizuję po prostu w Pythonie tak, typowo tak, ale jeżeli chcę coś szybko utworzyć to nie ma wstydu żadnego w używaniu Excela tak, no ale no po prostu Excel jest bardzo narażony na różnego rodzaju takie błędy, więc wystarczy że się formatowanie zmieni, przy czym przy dużym kopiowaniu, trzeba być po prostu ostrożnym przy korzystaniu z Excela. No ale tak naprawdę, no co można zrobić w Excelu, nie wiem czy można zrobić jakąś klasyfikację, regresję może można zrobić, ale [krótka pauza] Nie! Można zrobić regresję, można zrobić regresję chyba
> *B: Można narysować wykres z R^2, to na pewno*
> R: Tak, tak. No i po prostu ludzie się śmieją, dlatego że wielu ludzi po prostu nie wiem ucząc się dodatkowych funkcji, tak siebie postrzega. No i też jest takie, wydaje mi się w psychologii, jest takie prawo że ludzie którzy mają mniejszą wiedzę, w bardziej, mają tendencję do tego żeby ją, żeby, żeby postrzegać siebie jako lepszego, a ci którzy mają większą wiedzę mają tendencję żeby postrzegać siebie jako tego który nic nie wiem, więc eee
> *B: Mhm, mhm, masz rację słyszałem, słyszałem o czymś takim.*
> R: No jak Excel, jak ktoś mówi o Excelu no to znaczy że kruczę no nie bardzo, no to znaczy że to jest osoba która po prostu coś tam robi, może nawet ma dobre początki, no ale to nie jest droga po prostu, jeszcze powiedzmy średnią można obliczyć, tak? te okienkowe średnie policzyć, ale wykresy się topornie robi strasznie w Excelu, nic się nie zmieniło w tym, obecnie w tym temacie pomiędzy tymi wersjami, więc no [pauza] wiadomo tam jest też ograniczona ilość wierszy, no jakoś cały czas, mimo tych ciągłych updateów i tego nawet [krótka pauza] PowerQwery czy jak to się to nazywa, właśnie nie pamiętam, taki dodatek był zastosowany, rozszerzający dla większej ilości wierszy [pauza] To to, nieee no **to nie jest poważne narzędzie**. To jest dobre do takich wstępnych analiz, żeby sobie zobaczyć nawet tak jak wyglądają te dane, bo czasami nawet to pomaga, żeby zobaczyć po prostu sobie, eee żeby zobaczyć te pierwsze pięć wierszy, ale w sumie to można w każdej technologii to wyświetlić. {wywiad 12}

Excel dla wykonywania działania podstawowego w świecie DS jest więc postrzegany jako narzędzie nieodpowiednie, nieprzydatne, niewygodne, nieefektywne, mało sprawcze, nieciekawe (ponieważ od wielu lat względnie niezmienne), i ogółem bardzo ograniczone. **Excel to nie jest narzędzie techniczne, to narzędzie biznesowe**. W projektach DS jedynym akceptowanym w badanym świecie użyciem Excela jest komunikowanie się z biznesem, czy ogólnie rzecz biorąc z osobami nie-technicznymi. Istnieje książka do nauki podstaw ML, adresowana dla osób nie-technicznych, nie potrafiących programować, gdzie przykłady i ćwiczenia z modelowania wykonane są w arkuszach kalkulacyjnych. Książka przeznaczona jest np. dla wiceprezesów działu marketingu, analityków popytu, szefów start-upów w branży e-commerce, analityków biznesowych, marketerów internetowych (Foreman 2017:17). Status Excela w świecie DS jest zatem bliski do Power Pointa, czyli oprogramowania z tzw. pakietów biurowych typu MS Office, a nie do języków Python lub R.

Zakończenie

Założone cele naszej rozprawy zostały zrealizowane. Dokonaliśmy obszernego opisu etnograficznego polskiego społecznego świata data science, co stanowiło cel główny. Na podstawie rekonstrukcji tworzonych przez uczestników społecznego świata DS definicji działania podstawowego zaproponowaliśmy własne ujęcie, realizując pierwszy z celów szczegółowych rozprawy. Działaniem podstawowym uczestników badanego świata jest, według nas, pisanie kodu do przetwarzania, analizy i modelowania danych.

Przedstawiliśmy nasze rozumienie roli technologii, umożliwiających wykonanie działania podstawowego w odniesieniu do zachodzących w społecznym świecie procesów, m.in. wyznaczania granic społecznego świata, legitymizacji i zaświadczania o autentyczności oraz segmentacji – profesjonalizacji. W ten sposób zrealizowany został drugi cel szczegółowy rozprawy. W odniesieniu do technologii, granice społecznego świata DS ogólnie wyznacza wykonywanie działania podstawowego w językach programowania: przede wszystkim w Pythonie i w mniejszym stopniu w R. Bardziej szczegółowo, ale pozostając w sferze języka Python, można wskazać na szereg narzędzi i metod, uznawanych w badanym świecie za właściwe. Jest to np. Anaconda – dystrybucja Pythona do zastosowań DS, a wraz z nią notatnik Jupyter i repozytorium conda. Do przetwarzania, analizy danych są to m.in. pakiety pandas, NumPy, matplotlib, a także narzędzie Apache Spark, czyli zunifikowaną technologię do rozproszonego przetwarzania danych. Do modelowania metodami tzw. tradycyjnego uczenia maszynowego służy najczęściej pakiet scikit-learn. Głównie do metod uczenia głębokiego, za właściwe uznawane są pakiety Tensor Flow (TF), PyTorch oraz narzędzie Keras, zwane nakładką, ułatwiającą korzystanie z TF. Zidentyfikowaliśmy także dyskursywne odgraniczanie się od obszarów bliskich DS poprzez wskazywanie na odmienne wartości. W ten sposób np. obszar IT określany jest jako nudny (bardziej inżynieryjny), zaś DS jako ciekawy (bardziej eksperymentalny); nauka akademicka uznawana jest za „sztukę dla sztuki”, zaś DS za sprawcze. Zaświadczanie o autentyczności odbywa się w badanym świecie, rzecz jasna, poprzez odwołanie do używania właściwych narzędzi i metod. Zidentyfikowaliśmy także technologie przemilczane, a niezbędne do pełnego uczestnictwa w świecie DS – np. podstawy języków SQL i bash. Inne praktyki zaświadczania o autentyczności to powołanie się na doświadczenie w „rozwiązywaniu realnych problemów” i cechy osobiste, np. bycie zainteresowanym matematyką lub informatyką od dziecka. Ogólnie zaświadczanie o

autentyczności uczestnictwa dokonywane jest poprzez kategorię bycia osobą techniczną. Proponujemy także postrzeganie autentycznego uczestnika świata społecznego jako kogoś, kto uosabia wartości tego świata. Opisaliśmy także praktyki prowadzące do bycia uznanym za uczestnika silnie osadzonego, osiągającego poziom maestrii. Jest to np. samodzielne budowanie narzędzi do DS oraz agnostycyzm technologiczny i metodologiczny. Wskazaliśmy ogólne legitymizacje świata DS, jak narracja o jego powstaniu. Data science powstało, ponieważ z uwagi na gwałtownie rosnącą ilość dostępnych danych cyfrowych i rozwój mocy obliczeniowej komputerów, przy jednoczesnym spadku ich kosztów (zarówno danych, jak i mocy) możliwe, a zarazem opłacalne, stało się rozwiązywanie realnych problemów z wykorzystaniem metod uczenia maszynowego. Opisaliśmy tworzące się subświaty / segmenty świata DS, głównie w kategoriach profesjonalizacji i zmiany zakresu wybranych aspektów działania podstawowego. Wyróżniliśmy inżynierię danych (*data engineering*), badania nad uczeniem maszynowym (*ML research*), inżynierię uczenia maszynowego (*ML engineering*) i analitykę danych. Działania podstawowe, jak i właściwe technologie są dla nich inne, niż dla świata DS.

Rozpoznaliśmy i opisaliśmy wartości uczestników świata DS – efektywność, sprawczość, samorozwój, samodzielność, ciekawość, swobodę i racjonalność, realizując trzeci z celów szczegółowych naszej pracy. Za centralną wartość badanego świata uznaliśmy efektywność. To wartości merytokratyczne, charakterystyczne dla późnonowoczesnego neoliberalizmu i wywodzące się z jednej strony z tzw. światopoglądu technokratycznego, z drugiej zaś z etosu naukowca akademickiego wg. R. K. Mertona. Pokazaliśmy, jak w sposób wartościujący, w tych siedmiu wyróżnionych kategoriach aksjonormatywnych, oceniane są rozmaite działania uczestników świata DS. W kategoriach wartości oceniane i budowane są także technologie używane w świecie DS. Widzimy preferencje i wybory technologiczne jako wyraz wartości. W dużej mierze w kategoriach wartości opisaliśmy także zachodzące w świecie DS procesy, a szczególnie areny proponujemy postrzegać jako spory o wartości.

Określiliśmy i opisaliśmy główne areny sporów dla społecznego świata DS, realizując czwarty cel szczegółowy rozprawy. Jako szeroką arenę widzimy modelowanie, zaś model ML potraktowaliśmy jako obiekt graniczny, wokół którego toczy się spór o władzę, zasoby, wartości, o przetrwanie i rozwój każdego z zaangażowanych światów. W tę arenę zaangażowane są społeczne światy DS, biznesu, akademii, IT, mediów, polityki i prawa. Dostęp do obiektu granicznego (szczególnie, jeżeli zawęzimy jego ujęcie do produkcyjnie wdrażanych modeli ML) mają światy techniczne, czyli DS i IT, oraz te światy nie-techniczne,

które dysponują władzą nad wpływowymi aktorami na arenie. Jest to świat biznesu (aktor – pieniądze) i świat prawa (regulacje prawne). Podkreślamy rolę gigantów technologicznych, jak Google, Facebook, Amazon, we wspieraniu rozwoju ważnych na arenie technologii, także tych „wolnych" (*open source*). Dość znaczący na arenie jest też świat polityki, jako najbardziej sprawczy w zakresie kształtowania regulacji prawnych, ale nie ich egzekwowania, oraz obszar nauk ścisłych i przyrodniczych, jako subświat nauki akademickiej, kształtujący metody ML, publikujący teksty i pakiety dotyczące modelowania. Niemniej te światy nie mają bezpośredniego dostępu do obiektu granicznego. Jeszcze mniej wpływowe na arenie są światy mediów oraz obszar nauk społecznych i humanistycznych jako subświat nauki akademickiej, do którego sami przynależymy. Podkreślamy różnice w dyskursach zaangażowanych w arenę światów. Światy DS, IT oraz subświat nauk ścisłych i przyrodniczych traktują obiekt graniczny w kategorii „uczenia maszynowego". Świat prawa posługuje się zarówno tą kategorią, jak i „sztuczną inteligencją". Pozostałe światy posługują się głównie nie-technicznym pojęciem „sztucznej inteligencji", stanowiącym opakowanie dla modeli ML (i wielu innych technologii). Te światy pozostają pod wpływem praktyk celowo stosowanych przez biznesowych rzeczników świata DS i IT (głównie marketingu i HR), częściowo także pod wpływem popkulturowych ujęć AI, które można traktować jako tło dla areny. Twierdzimy, że uczestnicy świata DS postrzegają realizowane przez siebie projekty, których małą częścią jest budowanie modeli ML, jako majsterkowanie, podczas gdy z perspektywy nie-technicznej zarówno ich praca jak i samo działanie modeli ML jest tak przedstawiane jak i widziane w kategorii magii. Stąd popularny w świecie DS żart: jeżeli to jest napisanie w Pythonie, to prawdopodobnie uczenie maszynowe; jeżeli jest napisanie w PowerPointcie, prawdopodobnie to AI. Zwracamy uwagę także na to, że zarówno użytkownicy końcowi systemów bazujących na modelach ML, osoby wykonujące nisko płatne i niewykwalifikowane zadania (tzw. klikacze, mogą np. etykietować dane źródłowe na potrzeby trenowania modeli nadzorowanego ML) jak i środowisko naturalne, obciążane energochłonną infrastrukturą obliczeniową (a w wybranych przypadkach marketingowo zwaną „chmurą") to aktorzy słabo obecni na arenie, traktowani nierzadko przedmiotowo, niczym zasoby do zużywania.

Jako nieco węższą widzimy arenę dotyczącą sytuacji kobiet w badanym świecie społecznym. W Polsce nie zauważamy, aby problematyzowano sytuacje innych mniejszości. Kobiety stanowią 10% – 20% osób uczestniczących w badanym świecie, co oznacza, że kiedy w zespole DS pracuje kobieta, to niemal zawsze jest to jedyna kobieta wśród kilku mężczyzn.

Opisaliśmy spór o wartości pomiędzy dwiema pozycjami w świecie DS – pozycją „doskonałej merytokracji" a pozycją „różnorodności". Pierwsza z nich zakłada doskonałą merytokrację, w której nie liczą się takie cechy jak płeć, a tylko i wyłącznie kompetencje techniczne i metodologiczne konkretnych osób. Druga, bardziej poprawna politycznie i wyraźniej widoczna pozycja „różnorodności", afirmuje włączanie do komercyjnego DS osoby reprezentujące mniejszości. Takie zespoły mają budować lepsze produkty i realizować lepsze projekty, bo wypracowują rozwiązania jakoby właśnie bardziej różnorodne, nie obciążone (in vivo „zbiasowane") w kierunku perspektywy młodych, białych mężczyzn – grupy ilościowo bezwzględnie dominującej.

Zaprezentowaliśmy także arenę dotyczącą używanych w świecie DS narzędzi, szczególnie spór o konkurencyjność, komplementarność i rozłączność zastosowań języków Python i R. Przedstawiliśmy różnice historyczne w rozwoju tych języków i dzisiejsze różnice w ich zastosowaniach. Ogólnie język Python jest obecnie w badanym świecie traktowany jako pierwszy wybór, jest językiem ilościowo znacznie bardziej popularnym – w Polsce może być on używany w około 87% firm zajmujących się projektami DS. W około 38% jest to język R, zaś w co najmniej 25% wykorzystywane są oba języki. Ze względu na cechy samego języka, jak i dostępne pakiety, Python uznawany jest za lepszy do modelowania metodami ML (szczególnie do uczenia głębokiego), zaś R do statystyki i analizy danych. Mniej więcej od 2018 r. Python uznawany jest, przynajmniej w komercyjnej sferze działalności badanego świata, za domyślny język do wykonywania działania podstawowego. Język R należałoby poznać czy chociażby potrafić się posłużyć nim jako drugim, w konkretnych zastosowaniach np. analitycznych czy statystycznych, posługując się już swobodnie Pythonem i stając się silnie osadzonym uczestnikiem świata DS. Spór zatem stracił na znaczeniu, a w ogóle przesadna koncentracja na jednej tylko technologii jest w badanym świecie postrzegana jako coś, co cechuje początkujących uczestników. Ci silnie osadzeni charakteryzują się raczej agnostycyzmem technicznymi i metodologicznym, czyli dość swobodnie dobierają narzędzia i metody do rozwiązywanego problemu. Opisywana arena jednoznacznie definiuje uczestników świata DS jako osoby wykonujące działanie podstawowe za pomocą kodu. Narzędzia analityczne bazujące na interfejsach graficznych, jak np. Excel, są uznawane za nie-techniczne, niewłaściwe do działania podstawowego DS.

Zrealizowaliśmy piąty cel szczegółowy rozprawy, czyli ujęcie pełnej sytuacji badania. Łącznie składają się na nie rozdziały: 2. poświęcony przeglądowi literatury akademickiej, gdyż naukowców traktujemy jako zaangażowanych w arenę; 3. i 4., gdyż nasze badania wraz

z ich ramą teoretyczną i metodologią stanowią wyraz właśnie takiego zaangażowania, a te rozdziały pokazują usytuowanie badacza (RŻ) w odniesieniu do świata DS; i na końcu rozdziały 5. i 6., jako prezentacja wyników naszych badań i analiz.

Musimy powiedzieć o tym, co zostało przemilczane w naszej pracy (*sites of silence,* por. Clarke 2005:85). Są to dwa zagadnienia. Po pierwsze, szczegóły dotyczące metod stosowanych w DS. Po drugie, procesy wychodzenia uczestników z badanego, społecznego świata. Pierwsze zagadnienie pominęliśmy świadomie, ponieważ zagłębienie się w szczegóły matematyczne stosowanych metod byłoby dla nas niewspółmiernie pracochłonne w stosunku do efektu w postaci głębszego zrozumienia badanego świata, a więc i założonych celów tej rozprawy. W pracy pozostaliśmy przy doraźnych, raczej intuicyjnych wyjaśnieniach np. czym różni się ML nadzorowane od nienadzorowanego, co to jest propagacja wsteczna itp. Zdecydowaliśmy się zagłębić bardziej szczegółowo w technologie świata DS jako narzędzia, a nie jako metody, bo naszym zdaniem narzędzia są znacznie silniej obecne w dyskursie tego świata. Z uwagi na uzyskany przez nas, właśnie głównie przez pryzmat narzędzi, wgląd w elementy i procesy w społecznym świecie DS ta decyzja wydaje nam się właściwa. Drugie zagadnienie, wychodzenia ze świata DS, pominęliśmy nieświadomie. W toku własnych badań i analiz nie spotkaliśmy się z takim procesem, ani nie wpadliśmy na pomysł, aby go poszukać. Możemy powiedzieć jedynie o wychodzeniu ze świata w kontekście procesów segmentacji i profesjonalizacji, np. przechodzenia z definiowania się jako data scientista do inżyniera uczenia maszynowego (*ML engineer*). Nie znamy jednak praktyk wychodzenia „na zewnątrz" badanego świata, a bez wątpienia takie istnieją.

Ufamy, że nasza praca może przyczynić się do zmian w szerokim dyskursie nie-technicznym, dotyczącym tego, co związane z badanym przez nas społecznym światem DS, a nazywane np. chmurą, big data, sztuczną inteligencją. Proponujemy cztery stwierdzenia, stanowiące końcowe podsumowanie naszych wniosków:

- „data scientist" to nie programista, a ktoś, kto pisze kod do przetwarzania, analizy i modelowania danych;
- dane i obliczenia w „chmurze" to dane i obliczenia na cudzym komputerze, a rzędy takich komputerów – serwerów w betonowych halach nie mają w sobie nic zwiewnego, potrzebują stałego zasilania energią elektryczną, chłodzenia wodą i ludzkiej obsługi;

- „big data” to nieco przestarzałe określenie technologii rozproszonego i równoległego składowania i przetwarzania danych, jak np. Apache Hadoop i Spark, a nie inwigilacja dużych grup ludzi ani nie coś, co odmieni oblicze nauki i przyniesie śmierć ekspertom dziedzinowym;
- „sztuczna inteligencja” / „AI” nie istnieje – istnieją różnego rodzaju automaty, zaś sami twórcy komercyjnie stosowanych modeli uczenia maszynowego, czyli jakoby najbardziej inteligentnych, bo wyszukujących prawidłowości w zbiorach danych cyfrowych części systemów zwanych AI, uznają „sztuczną inteligencję” przede wszystkim za termin marketingowy, mający oczarowywać klientów.

Co do konkluzji dla socjologii, chcemy, aby nasza praca przyczyniła się do zwiększenia zainteresowania badaczy społecznych nowoczesnymi technologiami i ich wpływem na świat życia ludzi i środowisko naturalne. Z satysfakcją zauważamy, że sami dość wcześnie w Polsce zajęliśmy się tą tematyką, czyli obszarem DS. Właściwie pięć lat przed słusznie postulowanym przez K. Koneckiego podjęciem tych problemów „prawie w każdym badaniu socjologicznym” (Konecki 2020:196–97). Zapraszamy badaczy do jakościowego zanurzenia się w dyskurs techniczny, wyjścia poza czerpany z mediów, popkultury i komunikatów marketingowych żargon nie-techniczny. Życzymy powodzenia w przezwyciężeniu pokusy powtarzania tradycyjnej krytyki nowoczesnych technologii w duchu orwellowskim, huxleyowskim czy postpanoptycznym, bez własnych badań empirycznych. Zachęcamy badaczy społecznych do bliskiego obcowania z technologiami w ramach badań jakościowych, szczególnie za pomocą własnoręcznego pisania kodu w językach programowania. Namawiamy do etnograficznego obcowania z grupami ludzi odpowiedzialnymi za tworzenie i wdrażanie technologii na wszystkich jej poziomach. Cenne poznawczo wydaje się nam także włączanie do analiz kwestii ekonomicznych, jak np. wysokość i sposoby wynagradzania twórców technologii, popyt i podaż na ich rynku pracy, sposoby finansowania i koszty projektów, czy nawet ceny akcji spółek.

Socjologia od początku swego istnienia jest nauką o nowoczesności. Już A. Comte zauważył, że kapitalizm, jako reżim ekonomiczny, jest tym, co ukształtowało nowoczesność. Obecnie, na koniec drugiej dekady XXI w., nie zmieniło się wiele. Późną nowoczesność – w rozumieniu A. Giddensa, S. Lasha i U. Becka – kształtuje neoliberalny kapitalizm. Stosowanie nawet najbardziej zaawansowanych technologicznie i metodologicznie automatyzacji, zwanych potocznie sztuczną inteligencją, nie jest zatem żadną rewolucją. Nawet, kiedy wzrośnie skala stosowania tego rodzaju technologii i gdyby zmieniła się

struktura zawodowa rynku pracy. To jest i zapewne będzie nieomal taki sam, tryumfujący kapitalizm. Wciąż napędzany pozytywizmem, choć obecnie mówi się o napędzaniu biznesu danymi (*data-driven*), a jeszcze niedawno o rewolucji big data. Stąd mogłoby się wydawać, że zasadniczo nic nie zagraża przetrwaniu i dalszemu rozwojowi opisanego przez nas społecznego świata DS w Polsce – dopóki będzie on zatrudniony w służbie kapitalizmu. Z naszych badań wynika, iż uczestnicy polskiego świata DS widzą swój świat, szczególnie w kategoriach branży i rynku na usługi DS, jako obszar na wczesnym etapie rozwoju. Polscy klienci są postrzegani jako powoli i nieufnie wchodzący na rynek, a ulokowane w Polsce zespoły DS czy freelancerzy pracują w znakomitej większości dla klientów, wewnętrznych jak i zewnętrznych, z Europy Zachodniej i USA. Uczestnicy spodziewają się zatem dalszego rozwoju i stabilizacji świata DS, w wielu wymiarach (np. edukacji, technologii, prawa, rynku). Spodziewają się także spadku entuzjazmu wobec „sztucznej inteligencji" i przesunięcia postrzegania ich pracy w stronę IT, w stronę technologii tak mało magicznych, jak strony internetowe.

Źródła potencjalnej „rewolucji", czyli zmiany społecznej, upatrujemy gdzie indziej. Pandemia COVID-19 pokazała, że sprawczość racjonalna i sprawczość pieniądza jest zawodna. Wirus, czyli coś, co nie posiada nawet DNA, nie poddaje się przecież łatwo temu porządkowi. Można rozumieć to szeroko, my powiedzmy jedynie, że wiele systemów opartych na modelach uczenia maszynowego przestało działać zgodnie z oczekiwaniem, bowiem dane wejściowe przestały przypominać te, na jakich trenowano modele przed przełomem 2019/2020 r. Być może zmieni to kierunek rozwoju DS, o ile neoliberalny, scjentystyczny, efektywnościowy paradygmaty działania tego świata będzie w wymiarze globalnym tracił na znaczeniu. Być może obserwujemy właśnie początek końca neoliberalizmu, bazującego na egoizmie i konkurencji, na rzecz systemu społecznego opartego na współpracy.

Aneks

Spis badań i analiz własnych

Tabela 3: Spis wywiadów swobodnych – badania własne

Nr. wywiadu	Termin realizacji wywiadu	Długość wywiadu [minuty]
1	21.10.2016	28
2	21.06.2018	47
3	4.07.2018	46
4	9.07.2018	45
5	10.07.2018	46
6	25.07.2018	62
7	25.07.2018	55
8	26.09.2018	52
9	26.09.2018	74
10	1.10.2018	57
11	5.10.2018	74
12	30.10.2018	51
13	30.10.2018	57
14	4.01.2019	40
15	5.01.2019	67
16	9.01.2019	55
17	13.01.2019	69
18	4.03.2019	68
19	9.03.2019	106
20	3.04.2019	68
21	24.04.2019	135

Nr. wywiadu	Termin realizacji wywiadu	Długość wywiadu [minuty]
22	21.05.2019	97
23	27.05.2019	88
24	27.05.2019	82
25	28.05.2019	108
26	30.05.2019	167

Źródło: opracowanie własne

Tabela 4: Spis obserwacji uczestniczących – badania własne

Nr. obserwacji	Termin realizacji obserwacji	Długość obserwacji [godziny]	Rodzaj obserwacji	Obserwacja online
1	5.10.2016 – 5.05.2017	10	kurs	-
2	7.11.2016	2	wykład	-
3	3.12.2016 – 20.05.2017	40	kurs	tak
4	12.12.2016	1	meetup	-
5	22.02.2017	1	meetup	-
6	28.03.2017	1,5	meetup	-
7	24.04.2017	2	warsztaty	tak
8	2.06.2017 – 11.03.2019	12	kurs	tak
9	27. - 29.09.2017	22	konferencja / warsztaty	-
10	19. - 21.10.2017	2	konferencja	-
11	16.11.2017	1	wykład	-
12	12.12.2017	6	konferencja	-
13	13.12.2017	2	meetup	-
14	15.12.2017	2	konferencja	-

Nr. obserwacji	Termin realizacji obserwacji	Długość obserwacji [godziny]	Rodzaj obserwacji	Obserwacja online
15	15.01.2018	1,5	meetup	-
16	9.02.2018	45	kurs	tak
17	2. - 29.03.2018	1	inne	tak
18	2.03.2018	1,5	konferencja	-
19	6.04.2018	3	wykład	-
20	17.04.2018	1	inne	tak
21	17.04.2018	1	inne	tak
22	18.04.2018	8	konferencja	-
23	25.04.2018	2	konferencja	-
24	26.04.2018	6	warsztaty	-
25	21.06.2018	2	meetup	-
26	2. - 5.07.2018	28	konferencja	-
27	27.08.2018	0,5	inne	tak
28	16.10.2018	1,5	meetup	-
29	30.10.2018	1,5	meetup	-
30	18.11.2018	6	warsztaty	-
31	19. - 22.11.2018	4	kurs	tak
32	6.12.2018	1	kurs	tak
33	11.12.2018	1,5	meetup	-
34	14.01.2019	2	kurs	tak
35	11. - 15.02.2019	10	kurs	tak
36	12.02.2019	2	meetup	-
37	27.02.2019	7	konferencja	-
38	4.04.2019	2	meetup	-
39	8. - 12.04.2019	10	kurs	tak
40	13.04.2019	7	konferencja	-

Nr. obserwacji	Termin realizacji obserwacji	Długość obserwacji [godziny]	Rodzaj obserwacji	Obserwacja online
41	2. - 15.05.2019	5	kurs	tak
42	8.05.2019	1,5	meetup	-
43	13. - 17.05.2019	10	kurs	tak
44	16.05.2019	5	warsztaty	-
45	18. - 31.05.2019	5	kurs	tak
46	14.06.2019	8	konferencja	-

Źródło: opracowanie własne

W analizach ilościowych i wizualizacji danych wykorzystano następujące pakiety dla języka GNU R:

- cowplot (Wilke 2019);
- ggrepel (Slowikowski 2019);
- jsonlite (Ooms 2014);
- lubridate (Grolemund i Wickham 2011);
- maptools (Bivand i Lewin-Koh 2019);
- PerformanceAnalytics (Peterson i Carl 2020);
- RColorBrewer (Neuwirth 2014);
- readODS (Schutten i in. 2020);
- rgeos (Bivand i Rundel 2019);
- rmarkdown (Allaire i in. 2019; Xie i in. 2018);
- scales (Wickham 2018d);
- sf (Pebesma 2018);
- tidyverse (Wickham i in. 2019).

Kod w języku R, umożliwiający powtórzenie całości analiz i wizualizacji, wraz ze wszystkimi plikami wejściowymi i wyjściowymi dostępny jest publicznie na koncie GitHub Remigiusza Żulickiego:

- dane o meetupach data science w Polsce – https://github.com/zremek/meetup-harvesting;

- dane z sondaży „środowiskowych” data science – https://github.com/zremek/survey-polish-data-science;
- dane z Google Trends – https://github.com/zremek/google-trends-ds-ml-ai;
- wizualizacja wyników analizy jakościowej (mapa pozycyjna) – https://github.com/zremek/qualitative-maps.

Indeks rysunków

Indeks tabel

Bibliografia

48forward. 2019. „Curt Simon Harlinghausen”. https://48forward.com/greencircle/curt-simon-harlinghausen/ (dostęp 27.06.2020).

Abadi, Martín i in. 2016. „TensorFlow: A system for large-scale machine learning”. W *12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah: USENIX.

Abriszewski, Krzysztof. 2010. *Wszystko otwarte na nowo. Teoria Aktora-Sieci i filozofia kultury*. Toruń: Wydawnictwo Naukowe UMK.

———. 2012. *Poznanie, zbiorowość, polityka. Analiza Teorii Aktora-Sieci Bruno Latoura*. Kraków: TAiWPN UNIVERSITAS.

ActiveState. 2018. *Python: A Lingua Franca*. https://www.activestate.com/wp-content/uploads/2018/10/Python-Lingua-Franca-Whitepaper-2018_1.pdf.

Adamczyk, Florian i in. 2018. *Automat do skaryfikacji żołędzi wraz z identyfikacją zmian chorobowych*. Poznań: Przemysłowy Instytut Maszyn Rolniczych.

Afeltowicz, Łukasz, i Krzysztof Pietrowicz. 2008. „Koniec socjologii, jaka znamy, czyli o maszynach społecznych i inżynierii socjologicznej”. *Studia Socjologiczne* 3(190): 43–73.

Agarwal, Rahul. 2019. „A Layman guide to moving from Keras to Pytorch”. *[wpis na blogu 06.01.2019]*. https://mlwhiz.com/blog/2019/01/06/pytorch_keras_conversion/ (dostęp 2.05.2019).

AI Fund. 2018. „AI Fund”. https://aifund.ai/ (dostęp 21.11.2018).

AI Now Institute. 2018. „The AI Now Institute”. https://ainowinstitute.org/ (dostęp 14.06.2018).

Albright, Jonathan. 2018. „The Graph API: Key Points in the Facebook and Cambridge Analytica Debacle”. *Tow Center*. https://medium.com/tow-center/the-graph-api-key-points-in-the-facebook-and-cambridge-analytica-debacle-b69fe692d747 (dostęp 7.02.2019).

Aldiabat, Khaldoun M., i Carole Lynne Le Navenec. 2018. „Data saturation: The mysterious step in grounded theory methodology”. *Qualitative Report* 23(1): 245–61.

Alekseichenko, Vladimir. 2017. „Praktyczne uczenie maszynowe dla programistów 2.0”. *DataWorkshop*. http://www.dataworkshop.eu/ (dostęp 26.09.2017).

———. 2018. *Przegląd Strategii Rozwoju Sztucznej Inteligencji na Świecie*. Warszawa. https://www.digitalpoland.org/assets/publications/przegląd-strategii-rozwoju-sztucznej-inteligencji-na-swiecie/przeglad-strategii-rozwoju-ai-digitalpoland-report.pdf.

———. 2018. „Sztuczna inteligencja, cyberbezpieczeństwa i hackerzy”. *18.06.2018 Biznes Myśli*. http://biznesmysli.pl/sztuczna-inteligencja-cyberbezpieczenstwa-i-hackerzy/ (dostęp 20.06.2018).

———. 2019a. „10 mitów o sztucznej inteligencji”. *Biznes Myśli*. http://biznesmysli.pl/10-mitow-o-sztucznej-inteligencji/ (dostęp 29.04.2019).

———. 2019b. „10 właściwych pytań przy wdrażaniu uczenia maszynowego”. *Biznes Myśli*. https://biznesmysli.pl/10-wlasciwych-pytan-przy-wdrazaniu-uczenia-maszynowego/ (dostęp 30.04.2019).

———. 2019c. „Drony zmieniają branże ubezpieczeń, budowlaną i inne”. *Biznes Myśli*. https://biznesmysli.pl/drony-zmieniaja-branze-ubezpieczen-budowlana-i-inne/ (dostęp 9.07.2019).

Allaire, J. 2012. „RStudio: integrated development environment for R”. W *The R User Conference, useR! 2011*, University of Warwick, 14. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.651.1157&rep=rep1&type=pdf#page=14.

Allaire, J J i in. 2019. „rmarkdown: Dynamic Documents for R”. https://rmarkdown.rstudio.com.

Allaire, J J, Kevin Ushey, i Yuan Tang. 2018. „reticulate: Interface to «Python»”. https://cran.r-project.org/package=reticulate.

Allo, Patrick i in. 2016. „The ethics of algorithms: Mapping the debate”. *Big Data & Society* 3(2). https://journals.sagepub.com/doi/10.1177/2053951716679679.

Amatriain, Xavier, i Justin Basilico. 2012. „Netflix Recommendations: Beyond the 5 stars (Part 1)”. *The Netflix Tech Blog*. https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429 (dostęp 9.07.2019).

Amazon. 2019a. „Amazon Go”. https://www.amazon.com/b?node=16008589011 (dostęp 17.04.2019).

———. 2019b. „AWS Marketplace: Machine Learning; Artificial Intelligence”. https://aws.amazon.com/marketplace/solutions/machinelearning/ (dostęp 2.08.2019).

Ameisen, Emmanuel. 2018. „How to deliver on Machine Learning projects: A guide to the ML Engineering Loop”. *Insight Data*. https://blog.insightdatascience.com/how-to-deliver-on-machine-learning-projects-c8d82ce642b0 (dostęp 25.10.2018).

Amie Ismail, Nur, Wardah Zainal Abidin, Nur Amie Ismail, i Wardah Zainal Abidin. 2016. „Data Scientist Skills”. *IOSR Journal of Mobile Computing & Application* 03(04): 52–61. http://iosrjournals.org/iosr-jmca/papers/Vol3-Issue4/G03045261.pdf.

Amrhein, Valentin, Sander Greenland, i Blake McShane. 2019. „Scientists rise up against statistical significance". *Nature* 567(7748): 305–7. http://www.nature.com/articles/d41586-019-00857-9.

Anaconda. 2018. „Distribution - Anaconda". https://www.anaconda.com/distribution/ (dostęp 3.12.2018).

———. 2019a. „Anaconda Distribution Starter Guide". *Anaconda*. https://docs.anaconda.com/_downloads/9ee215ff15fde24bf01791d719084950/Anaconda-Starter-Guide.pdf.

———. 2019b. „Anaconda Enterprise for Data Scientists". https://www.anaconda.com/enterprise/ (dostęp 22.07.2019).

———. „NumFOCUS - Anaconda". https://www.anaconda.com/numfocus/ (dostęp 3.12.2018).

Anand, Rajaraman, i D Ullman Jeffrey. 2011. 2011-01-03]. Http://Infolab. Stanford *Mining of massive datasets*. http://scholar.google.com/scholar?q=related:o4rvxS-5BtwJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5.

Andersen, Nina Blom. 2014. „"Dioxins are the easiest topic to mention": Resident activists' construction of knowledge about low-level exposure to toxic chemicals". *Public Understanding of Science* 25(3): 303–16. https://doi.org/10.1177/0963662514552600.

Anderson, Ashton, Jon Kleinberg, i Sendhil Mullainathan. 2016. „Assessing Human Error Against a Benchmark of Perfection". W *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, New York, New York, USA: ACM Press, 705–14. http://arxiv.org/abs/1606.04956%0Ahttp://dx.doi.org/10.1145/2939672.2939803.

Anderson, Chris. 2008. „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *2008.06.23 Wired*. https://www.wired.com/2008/06/pb-theory/ (dostęp 14.11.2017).

Anderson, Ken, Dawn Nafus, Tye Rattenbury, i Ryan Aipperspach. 2009. „Numbers Have Qualities Too: Experiences with Ethno-Mining". *Ethnographic Praxis in Industry Conference Proceedings* 2009(1): 123–40. http://doi.wiley.com/10.1111/j.1559-8918.2009.tb00133.x.

Andrus, Calvin, Jon Cook, i Suresh Sood. 2017. *Data Science: An Introduction*. Wikibooks. https://en.wikibooks.org/wiki/Data_Science: _An_Introduction.

Angrosino, Michael. 2010. *Badania etnograficzne i obserwacje*. Warszawa: Wydawnictwo Naukowe PWN.

Anguelov, Dragomir i in. 2005. „The Correlated Correspondence Algorithm for Unsupervised Registration of Nonrigid Surfaces". *Advances in Neural Information Processing Systems* 17: 33–40. http://robotics.stanford.edu/~drago/cc/video.mp4.

Anguelov, Dragomir, Daphne Koller, Evan Parker, i Sebastian Thrun. 2004. „Detecting and modeling doors with mobile robots". W *International Conference on Robotic & Automation*, 3777-3784 Vol.4. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1308857%5Cnhttp://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1308857.

Annechino, Rachelle. 2016. „Interviewing for Introverts". *20.01.2016 Ethnography Matters*. https://medium.com/ethnography-matters/interviewing-for-introverts-cdd4adfab547 (dostęp 7.05.2018).

Anscombe, F. J. 1973. „Graphs in Statistical Analysis Graphs in Statistical Analysis *". *The American Statistician* 27(1): 17–21.

Apramian, Tavis, Sayra Cristancho, Chris Watling, i Lorelei Lingard. 2017. „(Re)Grounding grounded theory: a close reading of theory in four schools". *Qualitative Research* 17(4): 359–76.

Arakelyan, Sophia. 2017. „Tech giants are using open source frameworks to dominate the AI community". *Venture Beat*. https://venturebeat.com/2017/11/29/tech-giants-are-using-open-source-frameworks-to-dominate-the-ai-community/ (dostęp 17.07.2019).

Arena, Michael J., Alex Pentland, i David Price. 2010. „Honest Signals - Hard Measures for Social Behavior". *Organization Development Journal* 28(3): 11–20. http://www.questia.com/library/journal/1P3-2127733261/honest-signals-hard-measures-for-social-behavior.

Ariely, Dan. 2013. „@danariely: Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone..." *2013.01.06 Twitter*. https://twitter.com/danariely/status/287952257926971392 (dostęp 14.02.2018).

Aron, Jacob. 2018. „Stop being the product". *New Scientist* 237(3172): 24–25. http://linkinghub.elsevier.com/retrieve/pii/S0262407918306110.

Ashcraft, Mark H. 2002. „Math anxiety: Personal, educational, and cognitive consequences". *Current Directions in Psychological Science* 11(5): 181–85.

Associated Press. 2015. „Big Brother is watching: how China is compiling computer ratings on all its citizens". *South China Morning Post*. https://www.scmp.com/news/china/policies-politics/article/1882533/big-brother-watching-how-china-compiling-computer (dostęp 21.01.2019).

Avram, Abel. 2013. „Docker: Automated and Consistent Software Deployments". *InfoQ*. https://www.infoq.com/news/2013/03/Docker/ (dostęp 23.01.2019).

Awad, Edmond i in. 2020. „Drivers are blamed more than their automated cars when both make mistakes". *Nature Human Behaviour* 4(2): 134–43. http://www.nature.com/articles/s41562-019-0762-8.

Azam, Anum. 2014. „The First Rule of Data Science". *27.04.2014 Berkeley Science Review*. http://berkeleysciencereview.com/article/first-rule-data-science/ (dostęp 26.01.2018).

Azevedo, Ana, i Manuel Filipe Santos. 2008. „KDD, SEMMA and CRISP-DM: a parallel overview". *IADIS European Conference Data Mining* (January): 182–85. http://recipp.ipp.pt/handle/10400.22/136.

Baath, Rasmus. 2012. „The State of Naming Conventions in R". *The R Journal* 4(2): 74–75. http://journal.r-project.org/archive/2012-2/RJournal_2012-2_Baaaath.pdf.

Babbie, Earl. 2006. *Badania społeczne w praktyce*. Warszawa: Wydawnictwo Naukowe PWN.

Badawy, Adam, i Emilio Ferrara. 2018. „The rise of Jihadist propaganda on social networks". *Journal of Computational Social Science* 1(2): 453–70. https://doi.org/10.1007/s42001-018-0015-z.

Bagiński, Konrad. 2018. „Nasze dane mogą być groźniejsze od karabinów. Ale pozwalają też rozwikłać największe problemy świata. Rozmowa z Piotrem Migdałem". *3.04.2018 INN:Poland*. http://innpoland.pl/142337,Dane-moga-byc-tak-samo-grozne-jak-rakiety-Ale-pozwalaja-tez-rozwiklac-problemy-nierozwiazywalne (dostęp 20.04.2018).

Baitu, N. T. 2014. „What is a data scientist? 14 definitions of a data scientist!" *14.07.2014 Big Data made simple. A Crayon Data resource*.

Bamman, David, i Noah A Smith. 2015. „Contextualized Sarcasm Detection on Twitter". W *International AAAI Conference on Web and Social Media*, Association for the Advancement of Artificial Intelligence. http://www.aaai.org/ocs/index.ph0p/ICWSM/ICWSM15/paper/view/10538/10445.

Bank Zachodni WBK. 2018. „bankITup - Dane rządzą!" https://challengerocket.com/pl/bankitup2018 (dostęp 22.03.2018).

Bannister, Benjamin. 2017. „SEO Secrets: Reverse-Engineering Google's Algorithm". https://medium.freecodecamp.org/seo-secrets-reverse-engineering-googles-algorithm-92fad4f5a39 (dostęp 11.02.2019).

Barabási, Albert-lászló. 2002. *Linked: The New Science of Networks*. Cambridge, Massachussets: Perseus Publishing.

Barlow, Mike. 2013. *The Culture of Big Data*. O'Reilly. https://www.wandisco.com/assets/bltc29a3ca0d2cb7aad/Oreilly_the_culture_of_big_data.pdf.

Barnet, David. 2017. „The robots are coming – but will they really take all our jobs?" *30.11.2017 Independent*. http://www.independent.co.uk/news/science/robots-are-coming-but-will-they-take-our-jobs-uk-artificial-intelligence-doctor-who-a8080501.html (dostęp 14.02.2018).

Barrio, Pablo J., Daniel G. Goldstein, i Jake M. Hofman. 2016. „Improving Comprehension of Numbers in the News". W *Proceedings of the 2016 CHI Conference on Human Factors*

in Computing Systems - CHI '16, New York, New York, USA: ACM Press, 2729–39. http://dl.acm.org/citation.cfm?doid=2858036.2858510.

Barry, Dwight. 2016. *Business Intelligence with R. From Acquiring Data to Pattern Exploration*. Leanpub. https://leanpub.com/businessintelligencewithr.

Batorski, Dominik. 2004. „Sieci społeczne: Charakterystyka, uwarunkowania i konsekwencje struktur relacji społecznych na przykładzie komunikacji internetowej". Uniwersytet Warszawski.

———. 2018. „Facebook. Wąż, który pożera własny ogon". *Magazyn Opinii Pismo*.

Batorski, Dominik, i Krzysztof Olechnicki. 2007. „Wprowadzenie do socjologii Internetu". *Studia Socjologiczne* 3(186): 5–14.

Bauman, Zygmunt, i David Lyon. 2013. *Płynna inwigilacja. Rozmowy*. Kraków: Wydawnictwo Literackie.

Baurska, Hanna i in. 2014. „Monocytic differentiation induced by side-chain modified analogs of vitamin D in ex vivo cells from patients with acute myeloid leukemia". *Leukemia Research* 38(5): 638–47. https://linkinghub.elsevier.com/retrieve/pii/S014521261400071X (dostęp 5.12.2018).

Bayer, Michael. 2012. „SQLAlchemy". W *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*, red. Amy Brown i Greg Wilson. aosabook.org. http://aosabook.org/en/sqlalchemy.html.

Becker, Howard. 1974. „Art as Collective Action". *American Sociological Review* 39(6): 767–76.

———. 1986. *Doing Things Together*. Evanston IL: Northwestern University Press.

Beede, David N. i in. 2011. „Women in STEM: A Gender Gap to Innovation". *SSRN Electronic Journal*. http://www.ssrn.com/abstract=1964782.

Benthall, Sebastian. 2018. „The politics of AI ethics is a seductive diversion from fixing our broken capitalist system". *[wpis na blogu] Digifesto*. https://digifesto.com/2018/12/18/the-politics-of-ai-ethics-is-a-seductive-diversion-from-fixing-our-broken-capitalist-system/ (dostęp 8.02.2019).

Berkeley Institute of Data Science. 2013. „Launch of the Berkeley Institute for Data Science". https://bids.berkeley.edu/resources/videos/launch-berkeley-institute-data-science (dostęp 26.05.2018).

Berman, Jules J. 2013. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Waltham: Elsevier.

Bhardwaj, Prachi. 2018. „#DeleteFacebook is trending due to Facebook's poor handling of Cambridge Analytica". *Business Insider*.

https://www.businessinsider.com/deletefacebook-facebook-cambridge-analytica-fiasco-twitter-2018-3?IR=T (dostęp 5.02.2019).

Bhatt, Niraj. 2013. „Big Data, NoSQL and MapReduce". *Architect's Blog*. https://nirajrules.wordpress.com/2013/05/ (dostęp 13.02.2018).

Białko, Michał. 2005. *Sztuczna inteligencja i elementy hybrydowych systemów ekspertowych*. Koszalin: Wydawnictwo Uczelniane Politechniki Koszalińskiej.

Biecek, Przemysław. 2014. *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*. Warszawa: Wydawnictwo Naukowe PWN.

———. 2015a. *Pogromcy Danych. Przetwarzanie danych w programie R*. Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego. http://pogromcydanych.icm.edu.pl/ (dostęp 28.12.2016).

———. 2015b. *Pogromcy Danych. Wizualizacja oraz modelowanie danych*. Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego. http://pogromcydanych.icm.edu.pl/.

———. 2015c. „PogromcyDanych: PogromcyDanych / DataCrunchers is the Masive Online Open Course that Brings R and Statistics to the People". https://cran.r-project.org/package=PogromcyDanych.

———. 2016. *Odkrywać! Ujawniać! Objaśniać! Zbiór esejów o sztuce przedstawiania danych*. Warszawa: Fundacja Naukowa SmarterPoland.pl. http://www.biecek.pl/Eseje/index.html.

———. 2017a. *Przewodnik po pakiecie R*. Wrocław: Oficyna Wydawnicza GiS.

———. 2017b. „Studia doktoranckie z data science". *SmarterPoland.pl*. http://smarterpoland.pl/index.php/2017/05/studia-doktoranckie-z-data-science/ (dostęp 10.12.2017).

———. 2018a. „Amazing adventures of Beta nad Bit!" http://betabit.wiki/ (dostęp 6.12.2018).

———. 2018b. „Ceteris Paribus Plots – a new DALEX companion". *SmarterPoland [01.06.2018 wpis na blogu]*. http://smarterpoland.pl/index.php/2018/06/ceteris-paribus-plots-a-new-dalex-companion/ (dostęp 2.06.2018).

———. 2018c. „CV - Przemysław Biecek". http://biecek.pl/CV/ (dostęp 11.08.2018).

———. 2018d. „DALEX: Explainers for Complex Predictive Models in R". *Journal of Machine Learning Research* 19(84): 1–5. https://arxiv.org/abs/1806.08915.

———. 2018e. „Który z nich zostanie najgorszym wykresem 2018?" *SmarterPoland.pl*. http://smarterpoland.pl/index.php/2018/12/najgorszy-wykres-2018/ (dostęp 3.04.2019).

———. 2018f. „RODO + DALEX, kilka słów o moim referacie na DSS”. *SmarterPoland [30.05.2018 wpis na blogu]*. http://smarterpoland.pl/index.php/2018/05/rodo-dalex-kilka-slow-o-moim-referacie-na-dss/ (dostęp 2.06.2018).

———. 2019. „MDP: Model Development Process v. 0.1”. *GitHub 2019.06.19*. https://github.com/ModelOriented/DrWhy/blob/master/images/ModelDevelopmentProcess.pdf (dostęp 3.07.2019).

———. 2020. „XAI Stories”. *github.io*. https://pbiecek.github.io/xai_stories/ (dostęp 5.08.2020).

Biecek, Przemysław, Ewa Baranowska, i Piotr Sobczyk. 2019. *Wykresy unplugged*. Warszawa: Fundacja Naukowa SmarterPoland.pl.

Biecek, Przemysław, i Tomasz Burzykowski. 2020. *Explanatory Model Analysis. Explore, Explain, and Examine Predictive Models. With examples in R and Python*. CRC Press. https://pbiecek.github.io/ema/ (dostęp 8.09.2020).

Biecek, Przemysław, i S. Cebrat. 2008. „Why Y chromosome is shorter and women live longer?” *The European Physical Journal B* 65(1): 149–53. http://www.springerlink.com/index/10.1140/epjb/e2008-00325-4 (dostęp 5.12.2018).

Biecek, Przemysław, i Marcin Kosinski. 2017. „archivist : An R Package for Managing, Recording and Restoring Data Analysis Results”. *Journal of Statistical Software* 82(11). http://www.jstatsoft.org/v82/i11/ (dostęp 5.12.2018).

Biecek, Przemysław, i Krzysztof Trajkowski. 2011. *Na przełaj przez Data Mining*. http://www.biecek.pl/NaPrzelajPrzezDataMining/.

Biedrzycki, Norbert. 2017a. „Budzą przerażenie, ale czy są uzasadnione? Oto 4 mity o sztucznej inteligencji”. *25.06.2017 Business Insider*. https://businessinsider.com.pl/technologie/nowe-technologie/mity-o-sztucznej-inteligencji/cbmd8hb (dostęp 22.03.2018).

———. 2017b. „Machine learning. Komputery nie są już niemowlętami”. *27.04.2017 [wpis na blogu]*. https://norbertbiedrzycki.pl/machine-learning-komputery-nie-sa-juz-niemowletami/ (dostęp 22.03.2018).

Bielecka-Prus, Joanna. 2014. „Po co nam autoetnografia? Krytyczna analiza autoetnografii jako metody badawczej”. *Przegląd Socjologii Jakościowej* X(3): 76–95.

Big Data Borat. 2013. „@BigDataBorat: Data science is statistics on Mac”. *2013.08.27 Twitter*. https://twitter.com/bigdataborat/status/372350993255518208 (dostęp 19.02.2018).

Bivand, Roger, i Nicholas Lewin-Koh. 2019. „maptools: Tools for Handling Spatial Objects”. https://cran.r-project.org/package=maptools.

Bivand, Roger, i Colin Rundel. 2019. „rgeos: Interface to Geometry Engine - Open Source ('GEOS')". https://cran.r-project.org/package=rgeos.

Błaszczak, Anita. 2016. „Specjaliści big data będą wkrótce na wagę złota". *Rzeczpospolita*. http://archiwum.rp.pl/artykul/1318468-Specjalisci-big-data-beda-wkrotce-na-wage-zlota.html.

Blei, David M., Andrew Ng, i Michael I. Jordan. 2003. „Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3: 993–1022.

Blumenkrantz, Deena. 2018. „Slack Workspaces for Data Science". *[13.08.2018 wpis na blogu]*. https://medium.com/deena-does-data-science/all-the-slack-workspaces-for-data-science-323380abf8ba (dostęp 5.09.2019).

Blumer, Herbert. 2007. *Interakcjonizm symboliczny. Perspektywa i metoda*. Kraków: Zakład Wydawniczy NOMOS.

Bobriakov, Igor. 2017. „Top 15 Python Libraries for Data Science in 2017". *2017.05.09 Medium*. https://medium.com/activewizards-machine-learning-company/top-15-python-libraries-for-data-science-in-in-2017-ab61b4f9b4a7 (dostęp 3.02.2018).

Bogdan, Małgorzata i in. 2008. „Extending the Modified Bayesian Information Criterion (mBIC) to Dense Markers and Multiple Interval Mapping". *Biometrics* 64(4): 1162–69. http://doi.wiley.com/10.1111/j.1541-0420.2008.00989.x (dostęp 5.12.2018).

Bomba, Radosław. 2015. „«Simowie» na wspak. Gra «This War of Mine» w perspektywie retoryki proceduralnej". *Wielogłos* 3(25): 87–95.

———. 2017. „Bazodanowe interfejsy. Projektowanie interakcji z dużymi zasobami danych kulturowych". *Kultura Popularna* 4(50): 64–73. http://kulturapopularna-online.pl/gicid/01.3001.0010.0044.

Bond, Robert M i in. 2012. „A 61-million-person experiment in social influence and political mobilization". *Nature* 489: 295. https://doi.org/10.1038/nature11421.

Bornakke, Tobias, i Brian L Due. 2018. „Big–Thick Blending: A method for mixing analytical insights from big and thick data sources". *Big Data & Society* 5(1): 1–16. http://journals.sagepub.com/doi/10.1177/2053951718765026.

Borowiecki, Łukasz, i Piotr Mieczkowski. 2019. *Map of the Polish AI*. Warszawa. https://www.digitalpoland.org/assets/publications/mapa-polskiego-ai/map-of-the-polish-ai-2019-edition-i-report.pdf.

Borowik, Magdalena, Leszek Maśniak, Robert Kroplewski, i Hubert Romaniec. 2018. *Przemysł+. Gospodarka oparta o dane*. https://www.gov.pl/cyfryzacja/gospodarka-oparta-o-dane-przemysl-.

Botsman, Rachel. 2017. „Big data meets Big Brother as China moves to rate its citizens". *Wired 21.10.2017*. http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion (dostęp 18.06.2018).

Bourdieu, Pierre. 1977. Cambridge Studies in Social and Cultural Anthropologypology *Outline of a Theory of Practice*. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/outline-of-a-theory-of-practice/193A11572779B478F5BAA3E3028827D8.

———. 1986. „The forms of capital". W *Handbook of Theory and Research for the Sociology of Education*, red. John G Richardson. Westport, CT: Greenwood, 241–58.

Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. Cambridge, Massachussets: MIT Press.

Boyd, Danah i in. 2017. „Ten simple rules for responsible big data research". *PLoS Computational Biology* 13(3): 1–10.

———. 2017. „Toward Accountability: Data, Fairness, Algorithms, Consequences". *12.04.2017 Points: Data & Society [wpis na blogu]*. https://points.datasociety.net/toward-accountability-6096e38878f0 (dostęp 30.04.2018).

———. „what's in a name? - danah michele boyd". http://www.danah.org/name.html (dostęp 11.08.2018).

Boyd, Danah, i Kate Crawford. 2011. „Six Provocations for Big Data". *SSRN Electronic Journal*: 1–17. http://www.ssrn.com/abstract=1926431.

———. 2012. „CRITICAL QUESTIONS FOR BIG DATA". *Information, Communication & Society* 15(5): 662–79. http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878.

Brackenbury, Will i in. 2018. „Draining the Data Swamp". W *Proceedings of the Workshop on Human-In-the-Loop Data Analytics - HILDA'18*, New York, New York, USA: ACM Press, 1–7. http://dl.acm.org/citation.cfm?doid=3209900.3209911.

Breda, Thomas, i Clotilde Napp. 2019. „Girls' comparative advantage in reading can largely explain the gender gap in math-related fields". *Proceedings of the National Academy of Sciences* 116(31): 15435–40. http://www.pnas.org/lookup/doi/10.1073/pnas.1905779116.

Breiman, Leo. 2001. „Statistical Modeling: The Two Cultures". *Statistical Science* 16(3): 199–231.

Bridges, Judith. 2017. „Gendering metapragmatics in online discourse: «Mansplaining man gonna mansplain…»". *Discourse, Context & Media* 20: 94–102. https://linkinghub.elsevier.com/retrieve/pii/S2211695817300910.

Brinkmeier, Michael. 2006. „PageRank revisited”. *ACM Transactions on Internet Technology* 6(3): 282–301. http://portal.acm.org/citation.cfm?doid=1151087.1151090.

BrodieG. 2018. „r - data.table vs dplyr: can one do something well the other can’t or does poorly? - Stack Overflow”. *[wątek na Stack Overflow]*. https://stackoverflow.com/questions/21435339/data-table-vs-dplyr-can-one-do-something-well-the-other-cant-or-does-poorly/27840349#27840349 (dostęp 5.12.2018).

Bródka, Piotr, Anna Chmiel, Matteo Magnani, i Giancarlo Ragozini. 2018. „Quantifying layer similarity in multiplex networks: a systematic study”. *Royal Society Open Science* 5(8): 171747. http://rsos.royalsocietypublishing.org/lookup/doi/10.1098/rsos.171747.

Brooks, Hannah. 2014. „Interviews with Data Scientists”. *Data Science Weekly* 1(April).

Brosz, Maciej, Grzegorz Bryda, i Piotr Siuda. 2017. „Od redaktorów: Big Data i CAQDAS a procedury badawcze w polu socjologii jakościowej”. *Przegląd Socjologii Jakościowej* 13(2): 6–23. http://link.springer.com/10.1007/978-81-322-2494-5.

Bryan, Jennifer. 2018. „Excuse Me, Do You Have a Moment to Talk About Version Control?” *The American Statistician* 72(1): 20–27. https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1399928.

Bryan, Jennifer, Jim Hester, David Robinson, i Hadley Wickham. 2019. „reprex: Prepare Reproducible Example Code via the Clipboard”. https://cran.r-project.org/package=reprex.

Bryan, Jennifer, i Hadley Wickham. 2017. „Data Science: A Three Ring Circus or a Big Tent?” *Journal of Computational and Graphical Statistics* 26(4).

Bryan, Jenny. 2018. „Happy Git and GitHub for the useR”. http://happygitwithr.com/ (dostęp 6.02.2018).

Brynjolfsson, Erik, Sinan Aral, i Marshall Van Alstyne. 2007. „Information, Technology and Information Worker Productivity: Task Level Evidence”. *NBER Working Paper No. w13172*. https://ssrn.com/abstract=993403.

Brynjolfsson, Erik, i Andrew McAfee. 2015. *Wyścig z maszynami. Jak rewolucja cyfrowa napędza innowacje, zwiększa wydajność i w nieodwracalny sposób zmienia rynek pracy*. Warszawa: Kurhaus Publishing.

Brzezińska, Ana, i Aleksandra Przegalińska. 2015. „Oko w oko z androidem. BINA48: Jestem jak gąbka. Pochłaniam każdą wiedzę, z którą się stykam”. *19.09.2015 Gazeta.pl*. http://weekend.gazeta.pl/weekend/1,152121,18816606,oko-w-oko-z-androidem-bina48-jestem-jak-gabka-pochlaniam.html (dostęp 1.08.2018).

Brzeziński, Tomasz. 2018. „Lenistwo matką wynalazków, czyli dlaczego nie gram w Kaggle. Wystąpienie na konferencji Data Workshop Club Conf”. *YouTube*. https://www.youtube.com/watch?v=Y5i05jBhQE8&feature= (dostęp 14.09.2019).

Bugalski, Piotr. 2019. „Kurs Raspberry Pi – #8 – praca w konsoli, podstawy Linuksa". https://forbot.pl/blog/kurs-raspberry-pi-praca-w-konsoli-podstawy-linuksa-id23911 (dostęp 21.10.2019).

Bugnion, Pascal. 2016. *Scala for Data Science*. Birmingham - Mumbai: Packt Publishing. http://proquest.safaribooksonline.com.ezproxy.lib.vt.edu/9781785281372.

Burda, Katarzyna. 2018. „Uczłowieczanie komputera". *Newsweek Polska*: 64–67.

Burkart, Patrick, i Jonas Andersson Schwarz. 2014. „Post-Privacy and Ideology". W *Media, Surveillance and Identity: Social Perspectives*, Digital Formations, New York: Peter Lang Publishing Group, 218–37.

Burrows, John F. 1996. „Numbering the streaks of the tulip? Reflections on a Challenge to the Use of Statistical Methods in Computational Stylistics". *Computing in the Humanities Working Papers1* A.2. http://projects.chass.utoronto.ca/chwp/burrows/#ednote.

Burski, Jacek. 2014. „Relacja badacz–narzędzie – analiza konsekwencji użycia narzędzi komputerowych w analizie danych jakościowych na przykładzie QDA Miner". W *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, red. Jakub Niedbalski. Łódź: Wydawnictwo Uniwersytetu Łódzkiego. http://hdl.handle.net/11089/24534.

Burtch, Linda. 2014. „Tell Your Kids to Be Data Scientists, Not Doctors". *Wired*. https://www.wired.com/insights/2014/06/tell-kids-data-scientists-doctors/ (dostęp 14.02.2018).

———. 2018. *The Burtch Works Study Salaries of Data Scientists*. https://www.burtchworks.com/wp-content/uploads/2018/05/Burtch-Works-Study_DS-2018.pdf.

Cadwalladr, Carole. 2017. „The great British Brexit robbery: how our democracy was hijacked". *The Guardian*. https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy (dostęp 3.10.2019).

———. 2018a. „'I made Steve Bannon's psychological warfare tool': meet the data war whistleblower". *The Guardian*. https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump (dostęp 5.10.2018).

———. 2018b. „The Brexit whistleblower: 'Did Vote Leave use me? Was I naive'". *The Guardian*. https://www.theguardian.com/uk-news/2018/mar/24/brexit-whistleblower-shahmir-sanni-interview-vote-leave-cambridge-analytica (dostęp 5.02.2019).

Cadwalladr, Carole, i Emma Graham-Harrison. 2018a. „Facebook and Cambridge Analytica face mounting pressure over data scandal". *19.03.2018 The Guardian*. https://www.theguardian.com/news/2018/mar/18/cambridge-analytica-and-facebook-accused-of-misleading-mps-over-data-breach (dostęp 22.03.2018).

———. 2018b. „Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach". *The Guardian*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (dostęp 5.02.2019).

Cadwalladr, Carole, i Mark Townsend. 2018. „Revealed: the ties that bind Vote Leave's data firm to controversial Cambridge Analytica". *The Guardian*. https://www.theguardian.com/uk-news/2018/mar/24/aggregateiq-data-firm-link-raises-leave-group-questions (dostęp 5.02.2019).

Cain, M, i J Finch. 1981. „Towards a Rehabilitation of Data". W *Practice and Progress: British Sociology 1950-1980*, London: George Allen and Unwin.

Van Camp, Jeffrey. 2019. „8 Best Smart Speakers in 2019: Alexa, Google Assistant, Siri, Cortana". *Wired*. https://www.wired.com/story/best-smart-speakers/ (dostęp 10.07.2019).

Campbell, Heidi A., i Antonio C. Pastina. 2010. „How the iphone became divine: New media, religion and the intertextual circulation of meaning". *New Media and Society* 12(7): 1191–1207.

Canton, James. 2016. „From Big Data to Artificial Intelligence: The Next Digital Disruption". *07.05.2016 The Huffington Post*. https://www.huffingtonpost.com/james-canton/from-big-data-to-artifici_b_10817892.html (dostęp 12.02.2018).

Cao, Longbing. 2017a. „Data Science: A Comperhensive Overview". *ACM Computing Surveys* 50(3): 1–42. http://dl.acm.org/citation.cfm?doid=3101309.3076253.

———. 2017b. „Data science: challenges and directions". *Communications of the ACM* 60(8): 59–68. http://dl.acm.org/ft_gateway.cfm?id=3015456&ftid=1895823&dwn=1&CFID=777607742&CFTOKEN=94761055.

Carey, James W. 2002. „Cultural studies and symbolic interactionism: Notes in critique and tribute to Norman Denzin". *Studies in Symbolic Interaction* (25): 199–209. https://www.emeraldinsight.com/10.1016/S0163-2396(02)80047-3.

Carneiro, Tiago i in. 2018. „Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications". *IEEE Access* 6: 61677–85.

Caro, Daniel H., i Przemysław Biecek. 2017. „intsvy : An R Package for Analyzing International Large-Scale Assessment Data". *Journal of Statistical Software* 81(7). http://www.jstatsoft.org/v81/i07/ (dostęp 5.12.2018).

Carr, Nicholas. 2013. „To Save Everything Click Here: Technology, Solutionism and the Urge to Fix Problems That Don't Exist". *NATURE* 495(7439): 45.

Casas, Pablo. 2018. *Data Science Live Book: An intuitive and practical approach to data analysis, data preparation and machine learning, suitable for all ages!* Pablo Adrian Casas. https://livebook.datascienceheroes.com/.

Cass, Stephen. 2016. „Linux At 25 Qa With Linus Torvalds". *IEEE Spectrum*. https://spectrum.ieee.org/computing/software/linux-at-25-qa-with-linus-torvalds (dostęp 24.10.2019).

Cath, Corinne i in. 2018. „Artificial Intelligence and the 'Good Society': the US, EU, and UK approach". *Science and Engineering Ethics* 24(2).

Centre for the New Economy and Society. 2018. *The Future of Jobs Report 2018 Insight Report Centre for the New Economy and Society*. Geneva. http://reports.weforum.org/future-of-jobs-2018/preface/?doing_wp_cron=1540064059.9713280200958251953125.

Cerf, Vint. 2007. „An Information Avalanche". *IEEE Computer* 40, nr 1: 104–5. http://doi.ieeecomputersociety.org/10.1109/MC.2007.5.

Ceri, Stefano. 2018. „On the role of statistics in the era of big data: A computer science perspective". *Statistics & Probability Letters* 136: 68–72. http://linkinghub.elsevier.com/retrieve/pii/S0167715218300646 (dostęp 1.03.2018).

ChallengeRocket.com. 2018. „Warsaw AI Hackathon at Google Campus Warsaw". https://challengerocket.com/warsaw-artificial-intelligence-hackathon-google-campus-warsaw (dostęp 25.04.2018).

Chang, Emily. 2018. *Brotopia: Breaking up the boys' club of Silicon Valley*. New York: Portfolio/Penguin.

Chang, Fay i in. 2006. „Bigtable: A Distributed Storage System for Structured Data". W *7th {USENIX} Symposium on Operating Systems Design and Implementation (OSDI)*, 205–18.

Chang, Ray M., Robert J. Kauffman, i YoungOk Kwon. 2014. „Understanding the paradigm shift to computational social science in the presence of big data". *Decision Support Systems* 63: 67–80. https://www.sciencedirect.com/science/article/pii/S0167923613002212 (dostęp 21.12.2018).

Chang, Robert. 2015. „Doing Data Science at Twitter. A reflection of my two year Journey so far. Sample size N = 1". *Medium.com*. https://medium.com/@rchang/my-two-year-journey-as-a-data-scientist-at-twitter-f0c13298aee6 (dostęp 29.08.2018).

Chang, Winston, Javier Luraschi, i Timothy Mastny. 2019. „profvis: Interactive Visualizations for Profiling R Code". https://cran.r-project.org/package=profvis.

Charmaz, Kathy. 2009. *Teoria ugruntowana. Praktyczny przewodnik po analizie jakościowej*. Warszawa: Wydawnictwo Naukowe PWN.

Charmaz, Kathy, i Adele E. Clarke, red. 2013. *Grounded Theory and Situational Analysis*. London: Sage.

Chemaly, Soraya, i Catherine Buni. 2016. „The secret rules of the internet". *The Verge*. https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech (dostęp 5.01.2019).

Chen, Catherine, i Haoqiang Jiang. 2018. „Important Skills for Data Scientists in China: Two Delphi Studies". *Journal of Computer Information Systems* 00(00): 1–10. https://doi.org/10.1080/08874417.2018.1472047.

Chen, Hao. 2010. „Comparative Study of C, C++, C# and Java Programming Languages". VAASAN AMMATTIKORKEAKOULU UNIVERSITY OF APPLIED SCIENCES. https://www.theseus.fi/bitstream/handle/10024/16995/Chen_Hao.pdf.

Chen, Hsinchun, Roger H L Chiang, i Veda C. Storey. 2012. „Business Intelligence and Analytics: From Big Data to Big Impact". *Management Information Systems Quarterly* 36(4): 1165–88. http://aisel.aisnet.org/misq/vol36/iss4/16.

Chen, Min, Shiwen Mao, i Yunhao Liu. 2014. „Big data: A survey". *Mobile Networks and Applications* 19(2): 171–209.

Chen, Tianqi, i Carlos Guestrin. 2016. „XGBoost". W *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, New York, New York, USA: ACM Press, 785–94. http://dl.acm.org/citation.cfm?doid=2939672.2939785.

Chibana, Nayomi. 2017. „5 Data Storytelling Tips for Improving Your Charts and Graphs". *Visual Learning Center by Visme*. http://blog.visme.co/data-storytelling-tips/ (dostęp 6.09.2018).

Chojnowski, Maciej. 2020. „Biecek: SI? Musimy coś sobie wyjaśnić - Sztuczna Inteligencja". https://www.sztucznainteligencja.org.pl/biecek-si-musimy-cos-sobie-wyjasnic/ (dostęp 26.04.2020).

Chollet, Francois. 2015. „Keras, now running on TensorFlow". *The Keras Blog*. https://blog.keras.io/keras-now-running-on-tensorflow.html (dostęp 17.07.2019).

Chomsky, Noam. 2013. „The Corporatization of the University". *2013.07.12 Rackham Auditorium, University of Michigan*. https://www.youtube.com/watch?reload=9&v=66w9Lf9yuZA&feature=player_detailpage#t=3210 (dostęp 9.02.2019).

Choudhury, Tanzeem, Alex Pentl, i Alex Pentland. 2002. „The Sociometer: A Wearable Device for Understanding Human Networks". https://dam-prod.media.mit.edu/x/files/tech-reports/TR-554.pdf.

Cicilline, David N. 2019. *Text - H.R.5 - 116th Congress (2019-2020): Equality Act*. 116th Congress (2019-2020): https://www.congress.gov/bill/116th-congress/house-bill/5/text (dostęp 5.04.2019).

Ciechanowski, Leon, Aleksandra Przegalińska, Mikołaj Magnuski, i Peter Gloor. 2018. „In the shades of the uncanny valley: An experimental study of human–chatbot interaction".

Future Generation Computer Systems. https://www.sciencedirect.com/science/article/pii/S0167739X17312268 (dostęp 16.04.2018).

Clark, Liat. 2012. „Google's Artificial Brain Learns to Find Cat Videos". *2012.06.26 Wired*. https://www.wired.com/2012/06/google-x-neural-network/ (dostęp 21.01.2018).

Clark, Stuart. 2014. „Artificial intelligence could spell end of human race – Stephen Hawking". *02.12.2014 The Guardian*. https://www.theguardian.com/science/2014/dec/02/stephen-hawking-intel-communication-system-astrophysicist-software-predictive-text-type (dostęp 14.02.2018).

Clarke, Adele E. 1991. „Social Words / Arenas Theory as Organizational Theory". W *Social Organization and Social Process. Essays in Honor of Anselm Strauss*, New York: Aldine de Gruyter, 119–58.

———. 1997. „A Social Worlds Research Adventure: The Case of Reproductive Science". W *Grounded Theory in Practice*, red. Anselm L Strauss i Juliet Corbin. Thousand Oaks, CA: SAGE Publications, 63–94.

———. 2003. „Situational Analyses: Grounded Theory Mapping After the Postmodern Turn". *Symbolic Interaction* 26(4): 553–76.

———. 2005. *Situational Analysis. Grounded Theory After the Postmodern Turn*. London: Sage.

———. 2015. „From Grounded Theory to Situational Analysis. What's New? Why? How?" W *Situational Analysis in Practice. Mapping Research with Grounded Theory*, red. Adele E Clarke, Carrie Friese, i Rachel S. Washburn. Walnut Creek: Left Coast Press Inc., 84–118.

Clarke, Adele E., i Monica J. Casper. 1996. „From Simple Technology to Complex Arena: Classification of Pap Smears, 1917-90". *Medical Anthropology Quarterly* 10(4): 601–23. http://doi.wiley.com/10.1525/maq.1996.10.4.02a00120.

Clarke, Adele E., i Carrie Friese. 2007. „Situational Analysis: Going Beyond Traditional Grounded Theory". W *Handbook of Grounded Theory*, red. Kathy Charmaz i Anthony Bryant. London: Sage, 694–743.

Clarke, Adele E., Carrie Friese, i Rachel S. Washburn. 2015. „Introducing Situational Analysis". W *Situational Analysis in Practice. Mapping Research with Grounded Theory*, red. Adele E. Clarke, Carrie Friese, i Rachel S. Washburn. Walnut Creek: Left Coast Press Inc., 11–75.

———. 2017. *Situational Analysis: Grounded Theory After the Interpretive Turn*. Los Angeles: Sage.

Clarke, Adele E., i Susan Leigh Star. 2008. „The Social Worlds Framework: A Theory/Method Package". W *The Handbook of Science and Technology Studies*, red. J.

Hackett, Edward, Olga Amsterdamska, Michael Lynch, i Judy Wajcman. Cambridge, London: The MIT Press, 113–58. http://doi.wiley.com/10.1002/9780470377994.ch6.

Clarke, Peter. 2012. „Google neural network teaches itself to identify cats". *2012.06.27 EE Times*. https://www.eetimes.com/document.asp?doc_id=1266579 (dostęp 21.01.2018).

Class Central. 2018. „Data Science | Free Online Courses MOOCs". https://www.class-central.com/subject/data-science (dostęp 21.02.2018).

Cleveland, William S. 2001. „Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics". *International Statistical Review* 69(1): 21–26. http://doi.wiley.com/10.1111/j.1751-5823.2001.tb00477.x.

Codd, Edgar Frank. 1970. „A relational model of data for large shared data banks". *Communications of the ACM* 13(6): 377–87. http://portal.acm.org/citation.cfm?doid=357980.358007.

Collobert, Ronan, Clement Farabet, Koray Kavukcuoglu, i Soumith Chintala. 2019. „Torch". http://torch.ch/ (dostęp 15.07.2019).

Colmerauer, Alain, i Philippe Roussel. 1993. „The Birth of Prolog". *SIGPLAN Not.* 28(3): 37–52. http://doi.acm.org/10.1145/155360.155362.

Conway, Drew. 2010. „The Data Science Venn Diagram". *2010.08.30*. http://www.dataists.com/2010/09/the-data-science-venn-diagram/ (dostęp 18.02.2018).

———. 2014. „Data science through the lens of social science". *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*: 1520–1520. http://dl.acm.org/citation.cfm?doid=2623330.2630824.

Cook, Gary i in. 2017. *Clicking Clean: Who Is Winning the Race To Build a Green Internet?* https://storage.googleapis.com/planet4-international-stateless/2017/01/35f0ac1a-clickclean2016-hires.pdf.

Courtland, Rachel. 2018. „Bias detectives: the researchers striving to make algorithms fair". *Nature* 558(7710): 357–60. http://www.nature.com/articles/d41586-018-05469-3.

Couture-Beil, Alex. 2018. „rjson: JSON for R". https://cran.r-project.org/package=rjson.

Crain, Matthew. 2018. „The limits of transparency: Data brokers and commodification". *New Media and Society* 20(1).

Crawford, Kate. 2016. „Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics". *Science Technology and Human Values* 41(1): 77–92.

———. 2018. *AI Now Report 2018*. New York. https://ainowinstitute.org/AI_Now_2018_Report.pdf.

Crawford, Kate, Sarah Myers West, i Meredith Whittaker. 2019. *Discriminating Systems: Gender, Race and Power in AI*. New York. https://ainowinstitute.org/discriminatingsystems.pdf.

Creemers, Rogier. 2015. „Planning Outline for the Construction of a Social Credit System (dostęp 2014-2020)". *China Copyright and Media [wpis na blogu]*. https://chinacopyrightandmedia.wordpress.com/2014/06/14/planning-outline-for-the-construction-of-a-social-credit-system-2014-2020/ (dostęp 21.01.2019).

CrowdFlower. 2017. *2017 Data Scientist Report*. https://visit.crowdflower.com/rs/416-ZBE-142/images/data-scientist-report-dec.pdf?mkt_tok=eyJpIjoiWXpNek5EQmtNalJsTkdWayIsInQiOiJPb29MV2JJdU81alRhbGh6OUVWcmt2UWpibXJ3cG5pSlFrNUxlVUdwT2hna1VOOU5Gd2tMU3ZEWmhoTVVmVHRXNWFhMFM4eTI1dDJwbWRJczVoTVlnRjFkQjl4ekNmT.

CS Department Toronto University. „Geoffrey E. Hinton - Home Page". http://www.cs.toronto.edu/~hinton/ (dostęp 29.03.2018).

Cukier, Kenneth, i Victor Mayer-Schönberger. 2014. *Big Data. Rewolucja, która zmieni nasze myślenie, pracę i życie*. Warszawa: MT Biznes Ltd.

Culkin, John M. 1967. „A Shoolman's Guide to Marshall McLuhan". *The Saturday Review*: 51–53; 70–72. http://www.unz.com/print/SaturdayRev-1967mar18-00051/Contents/.

Ćwiklak, Dariusz. 2017. „Wielki Brat szepcze do ciebie". *Newsweek Polska*: 73–75.

Czapska, Martyna. 2018. „RODO a sztuczna inteligencja". *Lex Robotica*. http://lexrobotica.pl/2018/05/25/rodo-a-sztuczna-inteligencja/ (dostęp 10.09.2018).

Czarnocka-Cieciura, Marta, i Piotr Migdał. 2015. „TagOverflow". *GitHub*. https://github.com/stared/tagoverflow (dostęp 9.12.2018).

Czarnowski, Ireneusz, Krzysztof Krawiec, Jacek Mańdziuk, i Jerzy Stefanowski. 2018. *Raport z pierwszego Zjazdu Polskiego Porozumienia na Rzecz Rozwoju Sztucznej Inteligencji*. Poznań. https://pp-rai.cs.put.poznan.pl/pp-rai-2018-raport.pdf.

Czubkowska, Sylwia. 2018. „Dane, kłamstwa i wybory: Facebook wybiera Ci prezydenta". *Gazeta Wyborcza*.

Dalton, Craig M, Linnet Taylor, i Jim Thatcher. 2016. „Critical Data Studies: A dialog on data and space". *Big Data & Society* 3(1). https://journals.sagepub.com/doi/10.1177/2053951716648346.

Dar, Pranav. 2018. „Python or R? Hadley Wickham and Wes McKinney are Building Platform Independent Libraries!" *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2018/05/python-and-r-are-joining-hands-to-eliminate-platform-dependency/ (dostęp 9.06.2018).

DARIAH. 2018. „DARIAH - Main". http://www.dariah.eu/ (dostęp 24.01.2018).

Darmach, Krystian. 2017. „Autoetnografia 2.0”. *Kultura i Społeczeństwo* 3(LXI): 87–101.

Data & Society. 2018. „Data & Society research institute”. https://datasociety.net/ (dostęp 30.04.2018).

Data Science Association. 2020. „Data Science Code of Professional Conduct”. http://datascienceassn.org/code-of-conduct.html (dostęp 2.08.2020).

Data Science Warsaw. 2018a. „Data Science Summit”. http://dssconf.pl/ (dostęp 30.03.2018).

———. 2018b. „Data Science Warsaw (Warszawa, Polska) | Meetup”. https://www.meetup.com/pl-PL/Data-Science-Warsaw/ (dostęp 19.11.2018).

DataCamp. 2018. „DataCamp Scholarship for Women and Gender Minorities Application Form 2018”. https://docs.google.com/forms/d/e/1FAIpQLScPBQuNDqucnpQoMUhFAKcpFSXYskb_zWr5k8Uy3pPB6o0Uag/viewform (dostęp 10.09.2018).

Davenport, Thomas H., i D. J. Patil. 2012. „Data scientist: The sexiest job of the 21st century”. *Harvard Business Review*.

Davidson, James i in. 2010. „The YouTube video recommendation system”. W *Proceedings of the fourth ACM conference on Recommender systems - RecSys ’10*, New York, New York, USA: ACM Press, 293. http://portal.acm.org/citation.cfm?doid=1864708.1864770.

Davies, Harry. 2015. „Ted Cruz campaign using firm that harvested data on millions of unwitting Facebook users”. *The Guardian*. https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data (dostęp 1.02.2019).

———. 2018. „Facebook told me it would act swiftly on data misuse – in 2015”. *26.03.2018 The Guardian*. https://www.theguardian.com/commentisfree/2018/mar/26/facebook-data-misuse-cambridge-analytica (dostęp 5.04.2018).

Day, Matt. 2018. „Amazon Go cashierless convenience store opens to the public in Seattle”. *The Seattle Times*. https://www.seattletimes.com/business/amazon/amazon-go-cashierless-convenience-store-opening-to-the-public/ (dostęp 10.07.2019).

DB-Engines. 2019. „DB-Engines Ranking - popularity ranking of database management systems”. https://db-engines.com/en/ranking (dostęp 13.08.2019).

Dean, Jeff. 2019. „Deep Learning to Solve Challenging Problems (Google I/O’19)”. *YouTube*. https://www.youtube.com/watch?v=rP8CGyDbxBY (dostęp 11.06.2019).

Dean, Jeffrey, i Sanjay Ghemawat. 2008. „MapReduce: Simplified Data Processing on Large Clusters”. *Communications of the ACM* 51(1): 107–13. http://portal.acm.org/citation.cfm?doid=1327452.1327492.

Delapenha, Lauren. 2017. „42 Essential Quotes by Data Science Thought Leaders". https://www.kdnuggets.com/2017/05/42-essential-quotes-data-science-thought-leaders.html (dostęp 6.02.2018).

Denyer, Simon. 2016. „China's plan to organize its society relies on 'big data' to rate everyone". *The Washington Post*. https://www.washingtonpost.com/world/asia_pacific/chinas-plan-to-organize-its-whole-society-around-big-data-a-rating-for-everyone/2016/10/20/1cd0dd9c-9516-11e6-ae9d-0030ac1899cd_story.html.

Devlin, Josh. 2018. „Want a Job in Data? Learn This." *17.01.2018 Dataquest*. https://www.dataquest.io/blog/why-sql-is-the-most-important-language-to-learn/ (dostęp 20.04.2018).

Devlin, Kate. 2018. *Turned On. Science, Sex and Robots*. Bloomsbury Sigma.

DeZyre. 2015. „Data Science Programming: Python vs R". https://www.dezyre.com/article/data-science-programming-python-vs-r/128 (dostęp 29.09.2018).

Diamond, Jeremy, i Diana Bash. 2018. „Who is Brad Parscale?" *02.03.2018 CNN*. https://edition.cnn.com/2018/02/27/politics/who-is-brad-parscale/index.html (dostęp 18.06.2018).

Diaz-Bone, Rainer. 2013. 14 Forum Qualitative Sozialforschung *Situationsanalyse - Strauss meets Foucault?* http://www.qualitative-research.net/index.php/fqs/article/view/1928/3466.

Diebel, James, Sebastian Thrun, i James Diebel and Sebastian Thrun. 2005. „An application of markov random fields to range sensing". W *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 291–98. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.6296&rep=rep1&type=pdf.

Diebold, Francis X. 2012. *A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline*. http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf.

Dijck van, José. 2014. „Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology". *Surveillance and Society* 12(2): 197–208.

dimlakgorkehgz. 2019. „GUIs to save from typing R code". *AlternativeTo.net*. https://alternativeto.net/list/2063/guis-to-save-from-typing-r-code/ (dostęp 11.02.2019).

Dixon, James. 2010. „Pentaho, Hadoop, and Data Lakes". *[wpis na blogu]*. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (dostęp 14.08.2019).

Dobrzynska, Ewelina, Joanna Rymaszewska, Przemysław Biecek, i Andrzej Kiejna. 2016. „Do Mental Health Outpatient Services Meet Users' Needs? Trial to Identify Factors Associated with Higher Needs for Care". *Community Mental Health Journal* 52(4): 472–78. http://link.springer.com/10.1007/s10597-015-9923-z (dostęp 5.12.2018).

Dodge, David. 2018. „5 Reasons Python Programming Is Perfect for Kids". *{coda}kid*. https://codakid.com/5-reasons-python-programming-is-perfect-for-kids/ (dostęp 20.07.2019).

Domańska, Ewa. 2007. „Zwrot performatywny we współczenej humanistyce". *Teksty Drugie* 5: 48–61. http://tekstydrugie.pl/wp-content/uploads/2016/06/57a60b0dab998f282822a6b305dfb23e.pdf.

Donizy, Piotr i in. 2018. „Nuclear pseudoinclusions in melanoma cells: prognostic fact or artifact? The possible role of Golgi phosphoprotein 3 overexpression in nuclear pseudoinclusions generation". *Pathology International* 68(2): 117–22. http://doi.wiley.com/10.1111/pin.12629 (dostęp 5.12.2018).

Donoho, David. 2015. „50 Years of Data Science". *R Software*: 41.

Dopierała, Renata. 2013. *Prywatność w perspektywie zmiany społecznej*. Kraków: Zakład Wydawniczy NOMOS.

Doran, Derek. 2018. „Data Scientist". W *Encyclopedia of Big Data*, Cham: Springer International Publishing, 1–4. http://link.springer.com/10.1007/978-3-319-32001-4_61-1.

Downey, Allen. 2017. „Programming as a Way of Thinking". *Scientific American Blog Network*. https://blogs.scientificamerican.com/guest-blog/programming-as-a-way-of-thinking/ (dostęp 29.07.2019).

Drabas, Tomasz, Denny Lee, i Holden Karau. 2017. *Learning PySpark*. Birmingham, UK: Packt Publishing. http://han3.lib.uni.lodz.pl/han/ebsco/search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1477650&lang=pl&site=eds-live.

Draper, Nora. 2017. „Fail Fast: The Value of Studying Unsuccessful Technology Companies". *Media Industries Journal* 4(1). http://hdl.handle.net/2027/spo.15031809.0004.101.

Drejewicz, Szymon. 2017. „Podcast Data Science po polsku". https://soundcloud.com/szymon-drejewicz-184766256 (dostęp 20.06.2017).

Drozd-Sokołowska, Joanna i in. 2018. „Atypical chronic myeloid leukaemia: A case of an orphan disease-A multicenter report by the Polish Adult Leukemia Group". *Hematological Oncology* 36(3): 570–75. http://doi.wiley.com/10.1002/hon.2501 (dostęp 5.12.2018).

Druga, Stefania, Randi Williams, Hae Won Park, i Cynthia Breazeal. 2018. „How Smart Are the Smart Toys?: Children and Parents' Agent Interaction and Intelligence Attribution". *Proceedings of the 17th ACM Conference on Interaction Design and Children*: 231–40. http://doi.acm.org/10.1145/3202185.3202741.

Drury, Benjamin J., John Oliver Siy, i Sapna Cheryan. 2011. „When Do Female Role Models Benefit Women? The Importance of Differentiating Recruitment From Retention in STEM". *Psychological Inquiry* 22(4): 265–69. http://www.tandfonline.com/doi/abs/10.1080/1047840X.2011.620935.

Dunn, Jeff. 2016. „Siri vs. Google Assistant vs. Alexa vs. Cortana: Which AI is best?" *Business Insider*. https://www.businessinsider.com/siri-vs-google-assistant-cortana-alexa-2016-11?IR=T#how-do-i-say-where-is-the-library-in-spanish-38 (dostęp 10.07.2019).

Dunn, Jeffrey. 2016. „Introducing FBLearner Flow: Facebook's AI backbone". *Facebook Code*. https://code.fb.com/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/ (dostęp 9.07.2019).

Dutcher, Jennifer. 2014. „What is Big Data?" *[wpis na blogu] Data Science Berkley*. https://datascience.berkeley.edu/what-is-big-data/ (dostęp 9.12.2016).

Dwoskin, Elizabeth, Drew Harwell, i Craig Timberg. 2018. „Facebook had a closer relationship than it disclosed with the academic it called a liar". *The Washington Post*. https://www.washingtonpost.com/business/economy/facebook-had-a-closer-relationship-than-it-disclosed-with-the-academic-it-called-a-liar/2018/03/22/ca0570cc-2df9-11e8-8688-e053ba58f1e4_story.html (dostęp 6.02.2019).

van Dyk, David i in. 2015. „ASA Statement on the Role of Statistics in Data Science". *AMSTAT News*. http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/ (dostęp 7.02.2018).

Dzieciątko, Mariusz, i Dominik Spinczyk. 2016. *Text mining. Metodyka, narzędzia, zastosowania*. Warszawa: Wydawnictwo Naukowe PWN.

Eagle, Nathan. 2005. „Machine Perception and Learning of Complex Social Systems". Massachusetts Institute of Technology.

Eagle, Nathan, i Kate Greene. 2014. *Reality Mining: Using Big Data to Engineer a Better World*. Cambridge, London: The MIT Press. http://www.jstor.org/stable/j.ctt9qf8q3.

Eagle, Nathan, i Alex Pentland. 2006. „Reality mining: sensing complex social systems". *Journal Personal and Ubiquitous Computing* 10(4): 255–68.

Eaton, John W. 2019. „GNU Octave". https://www.gnu.org/software/octave/ (dostęp 15.07.2019).

Economic Graph Team. 2017. *LinkedIn's 2017 U.S. Emerging Jobs Report*. https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report.

Eder, Maciej. 2013. „Mind your corpus: systematic errors in authorship attribution". *Literary and Linguistic Computing* 28(4): 603–14. http://dx.doi.org/10.1093/llc/fqt039.

———. 2014. „Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii". *Teksty Drugie* 2: 90–105.

———. 2015. „Does size matter? Authorship attribution, small samples, big problem". *Digital Scholarship in the Humanities* 30(2): 167–82. http://dx.doi.org/10.1093/llc/fqt066.

———. 2017. „Visualization in stylometry: Cluster analysis using networks". *Digital Scholarship in the Humanities* 32(1): 50–64. http://dx.doi.org/10.1093/llc/fqv061.

Eder, Maciej, Jan Rybicki, i Mike Kestemont. 2016. „Stylometry with R: A Package for Computational Text Analysis". *The R Journal* 8(1): 107–21. https://doi.org/10.32614/RJ-2016-007.

Elish, M. C., i Danah Boyd. 2018. „Situating methods in the magic of Big Data and AI". *Communication Monographs* 85(1): 57–80. https://www.tandfonline.com/doi/full/10.1080/03637751.2017.1375130.

Elliott, Larry. 2018. „Robots will take our jobs. We'd better plan now, before it's too late". *The Guardian*. https://www.theguardian.com/commentisfree/2018/feb/01/robots-take-our-jobs-amazon-go-seattle (dostęp 14.02.2018).

Esh. 2016. „Commoditizing Music Machine Learning: Services | Labs". *Spotify Labs*. https://labs.spotify.com/2016/08/07/commodity-music-ml-services/ (dostęp 9.07.2019).

Exxact. 2019. „NVIDIA Data Science Workstations". https://www.exxactcorp.com/NVIDIA-Data-Science-Workstations (dostęp 12.09.2019).

Facebook. 2018. „Yann LeCun – Facebook Research". https://research.fb.com/people/lecun-yann/ (dostęp 23.08.2018).

———. 2019. „About". https://www.facebook.com/pg/facebook/about/ (dostęp 28.01.2019).

Faraway, Julian J., i Nicole H. Augustin. 2018. „When small data beats big data". *Statistics and Probability Letters*.

FAT ML. 2019. „Fairness, Accountability, and Transparency in Machine Learning". http://www.fatml.org/ (dostęp 8.02.2019).

Feinberg, Donald. 2017. „Data Lake, Big Data, NoSQL - The Good, The Bad and The Ugly". *Gartner blogs*. https://blogs.gartner.com/donald-feinberg/2017/07/29/oh-terms-use-technology-need-new-hype/ (dostęp 6.08.2019).

Ferate, Pat. 2016. „meetup-api · PyPI". *PyPI*. https://pypi.org/project/meetup-api/ (dostęp 2.06.2019).

Ferretti, Federico. 2019. „Mapping do-it-yourself science". *Life Sciences, Society and Policy* 15(1): 1. https://lsspjournal.biomedcentral.com/articles/10.1186/s40504-018-0090-1.

Ferrucci, David i in. 2010. „Building Watson: An Overview of the DeepQA Project". *AI Magazine* 31(3): 59. https://aaai.org/ojs/index.php/aimagazine/article/view/2303.

Feuerstein, Steven. 1996. *Advanced Oracle PL/SQL Programming with Packages*. red. Debby Russell. Sebastopol, CA, USA: O'Reilly Associates, Inc.

Fierro, Miguel, Mathew Salvaris, i Tao Wu. 2017. „Lessons Learned From Benchmarking Fast Machine Learning Algorithms". *Microsoft Machine Learning Blog*. https://blogs.technet.microsoft.com/machinelearning/2017/07/25/lessons-learned-benchmarking-fast-machine-learning-algorithms/ (dostęp 15.11.2019).

Fjeld, Jessica i in. 2020. Research Publication No. 2020-1 *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Cambridge, Massachussets. https://cyber.harvard.edu/publication/2020/principled-a.

Fleishman, G. 2019. *Take Control of Slack*. alt concepts incorporated.

Flores, Anthony W., Christopher T. Lowenkamp, Alexander M. Holsinger, i Edward J. Latessa. 2006. „Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity". *Journal of Criminal Justice* 34(5): 523–29.

Floridi, Luciano. 2006. „The Logic of Being Informed". *Logique et Analyse* 49: 433–60. citeulike-article-id:8287843%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.7807.

———. 2012. „Big Data and Their Epistemological Challenge". *Philosophy & Technology* 25(4): 435–37. http://link.springer.com/10.1007/s13347-012-0093-4.

———. 2014a. „Right to Be Forgotten – A Diary of the Google Advisory Council Tour". *Philosophy of Information*. http://www.philosophyofinformation.net/right-to-be-forgotten-a-diary-of-the-google-advisory-council-tour/ (dostęp 14.09.2018).

———. 2014b. *The 4th revolution: how the infosphere is reshaping human reality*. Oxford: OUP.

———. 2015. „«The Right to Be Forgotten»: a Philosophical View". *Annual Review of Law and Ethics* 1(May 2014): 1–18.

———. 2018. „Soft Ethics and the Governance of the Digital". *Philosophy and Technology* 31(1).

———. 2019a. „What the Near Future of Artificial Intelligence Could Be".

———. 2019b. „Why I have accepted the invitation to...” *2019.04.04 Facebook [post]*. https://www.facebook.com/lucianofloridioxford/posts/784927045241088 (dostęp 5.04.2019).

Floridi, Luciano, i Mariarosaria Taddeo. 2016. „What is data ethics?” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374(2083).

Ford, Heather. 2014. „Big Data and Small: Collaborations between ethnographers and data scientists”. *Big Data & Society* 1(2). http://journals.sagepub.com/doi/10.1177/2053951714544337.

———. 2016. „What does it mean to be a “participant observer” in a place like Wikipedia?” *19.01.2016 Ethnography Matters*. https://medium.com/ethnography-matters/what-does-it-mean-to-be-a-participant-observer-in-a-place-like-wikipedia-89d6727276ba (dostęp 7.05.2018).

Foreman, John W. 2017. *Mistrz analizy danych. Od danych do wiedzy*. Gliwice: Helion.

Formulated.by. 2018. „15 Data Science Slack Communities to Join”. *Towards Data Science*. https://towardsdatascience.com/15-data-science-slack-communities-to-join-8fac301bd6ce (dostęp 5.09.2019).

Fowler, James H. 2006. „Connecting the Congress: A Study of Cosponsorship Networks”. *Political Analysis* (dostęp 14): 456–87.

Fowler, James H, i Nicholas Christakis. 2008a. „Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study”. *British Medical Journal* 337(a2338). https://www.bmj.com/content/bmj/337/bmj.a2338.full.pdf.

———. 2008b. „The Collective Dynamics of Smoking in a Large Social Network”. *New England Journal of Medicine* 358(21): 2249–58. http://nrs.harvard.edu/urn-3:HUL.InstRepos:3710308.

Fox, John. 2009. „Aspects of the Social Organization and Trajectory of the R Project”. *The R Journal* 1(December): 5–13. http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Fox.pdf.

Fox, John, i Allison Leanage. 2016. „R and the Journal of Statistical Software”. *Journal of Statistical Software* 73(2). http://www.jstatsoft.org/v73/i02/.

Franceschi-Bicchierai, Lorenzo. 2018. „Why We’re Not Calling the Cambridge Analytica Story a «Data Breach»”. *Motherboard*. https://motherboard.vice.com/en_us/article/3kjzvk/facebook-cambridge-analytica-not-a-data-breach (dostęp 5.02.2019).

Frankowski, Jacek. 2018. „Dane: przerażająca inwigilacja vs nowe możliwości". *8.03.2018*. https://www.linkedin.com/pulse/dane-przerażająca-inwigilacja-vs-nowe-możliwości-jacek-frankowski/ (dostęp 22.03.2018).

Freeman, Linton C. 2014. „The Development of Social Network Analysis – with an Emphasis on Recent Events". W *The SAGE Handbook of Social Network Analysis*, London: SAGE Publications Ltd, 26–39. http://sk.sagepub.com/reference/the-sage-handbook-of-social-network-analysis.

Frenken, Koen, i Juliet Schor. 2017. „Putting the sharing economy into perspective". *Environmental Innovation and Societal Transitions* 23: 3–10. http://dx.doi.org/10.1016/j.eist.2017.01.003.

Frey, Carl Benedikt, i Michael A. Osborne. 2017. „The future of employment: How susceptible are jobs to computerisation?" *Technological Forecasting and Social Change* 114(1): 254–80. https://linkinghub.elsevier.com/retrieve/pii/S0040162516302244.

Friedman, Batya, i Helen Nissenbaum. 1996. „Bias in computer systems". *ACM Transactions on Information Systems* 14(3): 330–47. http://portal.acm.org/citation.cfm?doid=230538.230561.

Fuller, Abby. 2019. „On Soft Talks and Being Technical - talk at Monktoberfest 2019". *RedMonk Tech Events*. https://www.youtube.com/watch?reload=9&v=cDgi4oaoIto (dostęp 30.04.2020).

Fundacja Panoptykon. 2019. „Prawo do wyjaśnienia w pytaniach i odpowiedziach". https://panoptykon.org/prawo-do-wyjasnienia (dostęp 5.09.2019).

Fung, Kaiser. 2014. „Junk Charts". https://junkcharts.typepad.com/ (dostęp 13.03.2018).

Future of Life Institute. 2017. „AI Principles". https://futureoflife.org/ai-principles/ (dostęp 7.02.2018).

Gągolewski, Marek. 2016. *Programowanie w języku R Analiza Danych. Obliczenia. Symulacje*. Warszawa: Wydawnictwo Naukowe PWN.

Gągolewski, Marek, Maciej Bartoszuk, i Anna Cena. 2016. *Przetwarzanie i analiza danych w języku Python*. Warszawa: Wydawnictwo Naukowe PWN.

Gallopoulos, E., E. Houstis, i J.R. Rice. 1994. „Computer as thinker/doer: problem-solving environments for computational science". *IEEE Computational Science and Engineering* 1(2): 11–23. http://ieeexplore.ieee.org/document/326669/.

Gandomi, Amir, i Murtaza Haider. 2015. „Beyond the hype: Big data concepts, methods, and analytics". *International Journal of Information Management*. http://www.mendeley.com/research/beyond-hype-big-data-concepts-methods-analytics.

Gardner, Dan. 2012. „Introduction". W *The human face of big data*, red. Rick. Smolan i Jennifer Erwitt. Sausalito, CA: Against All Odds Productions, 14–15.

Garner, Bennett. 2018. „Why I Code in Python". *Medium*. https://medium.com/@BennettGarner/why-i-code-in-python-a1e4012eb859 (dostęp 24.07.2019).

Geertz, Clifford. 1983. *Local Knowledge: Further essays in interpretive anthropology*. New York: Basic Books.

Gershon, Ilana. 2011. „Neoliberal Agency". *Current Anthropology* 52(4): 537–55. https://www.journals.uchicago.edu/doi/10.1086/660866.

Gerson, Elihu M. 1983. „Scientific Work and Social Worlds". *Knowledge: Creation, Diffusion, Utilization* 4(3): 357–77.

Gersz, Aleksandra. 2018. „Wyciek danych z Facebooka wpłynął na wybory w Polsce?" *Dziennik Łódzki*.

Ghosh, Shona. 2018. „The data scientist behind the Cambridge Analytica scandal did paid consultancy work for Facebook and has close ties to staff". *Business Insider*. https://www.businessinsider.com.au/aleksandr-kogan-did-paid-work-for-facebook-has-ties-to-staff-2018-4 (dostęp 1.02.2019).

Gierałowska, Urszula. 2013. „Inwestycja w fundusz hedgingowy jako przykład instrumentu alternatywnego". *Zeszyty Naukowe Uniwersytetu Szczecińskiego* (768): 127–42.

Ginsberg, Jeremy i in. 2009. „Detecting influenza epidemics using search engine query data". *Nature* 457(7232): 1012–14. http://www.nature.com/articles/nature07634.

Gitelman, Lisa, i Virginia Jackson. 2013. „Introduction". W *„Raw Data" is an oxymoron*, red. Lisa Gitelman. Cambridge, London: The MIT Press.

Glaser, Barney G, i Anselm L Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine.

Główny Urząd Statystyczny. 2018. „ShowMeData. Hackathon GUS". https://hackathon.stat.gov.pl/.

———. 2019. „Pytania i zamówienia / Jak zamówić dane". https://stat.gov.pl/pytania-i-zamowienia/jak-zamowic-dane/ (dostęp 6.08.2019).

Gogoi, Namrata. 2019. „How to Use AI Camera Features on Any Android Phone". *Guiding Tech*. https://www.guidingtech.com/ai-camera-features-android-phone/ (dostęp 10.07.2019).

Gogołek, Włodzimierz. 2016. „Rafinacja dużej skali zasobów sieciowych - Big Data. Dziennikarskie źródło informacji". *Ekonomika i Organizacja Przedsiębiorstwa* 7(798): 12–18.

Gold, R. L. 1958. „Roles in Sociological Field Observations". *Social Forces* 36(3): 217–23. https://academic.oup.com/sf/article-lookup/doi/10.2307/2573808.

Goldstein, Ira. 2019. „What ! No GUI ? – Teaching A Text Based Command Line Oriented Introduction to Computer Science Course". *Information Systems Education Journal* 17(February): 40–48.

Gontarz, Andrzej. 2019. *Rynek pracy Od BigData i AI do BI*. Warszawa. https://bigdatatechwarsaw.eu/report-job-market-for-data-professionals-from-big-data-and-ai-to-bi/.

Goode, Lauren. 2018. „How Google's Eerie Robot Phone Calls Hint at AI's Future". *08.05.2018 Wired*. https://www.wired.com/story/google-duplex-phone-calls-ai-future/ (dostęp 9.05.2018).

Goodfellow, Ian, Yoshua Bengio, i Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org/.

Google. 2012. „Google's R Style Guide". https://google.github.io/styleguide/Rguide.xml (dostęp 5.01.2018).

———. 2018. „Peter Norvig – Google AI". https://ai.google/research/people/author205 (dostęp 22.05.2018).

———. 2019a. „Colaboratory – FAQ". https://research.google.com/colaboratory/faq.html (dostęp 23.07.2019).

———. 2019b. „Informacje". https://www.google.com/about/ (dostęp 28.01.2019).

Google Trends. 2020a. „Big data, Machine learning, Data science, Artificial intelligence - Explore - Google Trends". https://trends.google.com/trends/explore?date=2008-01-01 2020-09-30&q=%2Fm%2F0bs2j8q,%2Fm%2F01hyh_,%2Fm%2F0jt3_q3,%2Fm%2F0mkz&hl=en-US (dostęp 5.10.2020).

———. 2020b. „Big data, Machine learning, Data science - Explore - Google Trends". https://trends.google.com/trends/explore?date=2008-01-01 2020-09-30&q=%2Fm%2F0bs2j8q,%2Fm%2F01hyh_,%2Fm%2F0jt3_q3&hl=en-US (dostęp 5.10.2020).

Gorecki, Jan, i Jiaming Yuan. 2019. „Database-like ops benchmark". https://h2oai.github.io/db-benchmark/index.html (dostęp 6.05.2019).

Gostkiewicz, Michał. 2017. „Chiny podłączą 1,3 miliarda ludzi i wielkie firmy do Matrixa. Prawdziwego. Czerwonej pigułki nie będzie". *gazeta.pl 16.11.2017*. http://weekend.gazeta.pl/weekend/1,152121,22621623,chiny-podlacza-1-3-miliarda-ludzi-i-wielkie-firmy-do-matrixa.html (dostęp 18.06.2018).

Grace, Katja i in. 2017. „When Will AI Exceed Human Performance? Evidence from AI Experts". 1–21. http://arxiv.org/abs/1705.08807.

Granville, Vincent. 2014. „16 analytic disciplines compared to data science". *2014.07.24 Data Science Central*. https://www.datasciencecentral.com/profiles/blogs/17-analytic-disciplines-compared (dostęp 3.01.2017).

Greene, Daniel, Anna Lauren Hoffman, i Luke Stark. 2019. „Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning". W *Hawaii International Conference on System Sciences (HICSS)*, Maui. http://dmgreene.net/wp-content/uploads/2018/09/Greene-Hoffman-Stark-Better-Nicer-Clearer-Fairer-HICSS-Final-Submission.pdf.

Grewal, Paul. 2018. „Suspending Cambridge Analytica and SCL Group From Facebook". *Facebook Newsroom*. https://newsroom.fb.com/news/2018/03/suspending-cambridge-analytica/ (dostęp 5.02.2019).

Groen, Albin. 2019. „Maximizing use of the terminal". https://medium.com/@albingroen/maximizing-use-of-the-terminal-9b7b12ab5dd2 (dostęp 11.09.2019).

Grolemund, Garrett, i Hadley Wickham. 2011. „Dates and Times Made Easy with {lubridate}". *Journal of Statistical Software* 40(3): 1–25. http://www.jstatsoft.org/v40/i03/.

Grommé, Francisca, Evelyn Ruppert, i Baki Cakici. 2018. „Data Scientists: A new faction of the transnational field of statistics". W *Ethnography for a data-saturated world*, red. Hannah Knox i Dawn Nafus. Manchester: Manchester University Press, 33–61. https://research.gold.ac.uk/20522/1/Grommé et al 2017 Data Science prepub dist copy.pdf.

Grothendieck, G. 2017. „sqldf: Manipulate R Data Frames Using SQL". https://cran.r-project.org/package=sqldf.

Grush, Loren. 2015. „Google engineer apologizes after Photos app tags two black people as gorillas". *1.07.2015 The Verge*. https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas (dostęp 26.02.2018).

Gualtieri, Mike i in. 2018. *The Forrester Wave™: Multimodal Predictive Analytics And Machine Learning Solutions, Q3 2018*. https://reprints.forrester.com/#/assets/2/1451/RES141374/reports (dostęp 14.01.2019).

Gulipalli, Pradeep. 2019. „The Pareto Principle for Data Scientists". *KDnuggets*. https://www.kdnuggets.com/2019/03/pareto-principle-data-scientists.html (dostęp 15.04.2019).

Guo, Philip J. 2018. „Non-Native English Speakers Learning Computer Programming". W *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, New York, New York, USA: ACM Press, 1–14. http://dl.acm.org/citation.cfm?doid=3173574.3173970.

Gutierrez, Sebastian. 2014. *Data Scientists at Work: Sexy Scientists Wrangling Data And Begetting New Industries*. Berkeley, CA: Appres.

Guzella, Thiago S., i Walmir M. Caminhas. 2009. „A review of machine learning approaches to Spam filtering". *Expert Systems with Applications* 36(7): 10206–22. http://dx.doi.org/10.1016/j.eswa.2009.02.037.

Haggerty, Kevin, i Richard V. Ericson. 2000. 51 The British Journal of Sociology *The Surveillant Assemblage*.

Hałas, Elżbieta. 2006. *Interakcjonizm symboliczny*. Warszawa: Wydawnictwo Naukowe PWN.

Hale, Jeff. 2018. „The Most in Demand Skills for Data Scientists". *Towards Data Science*. https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-4a4a8db896db (dostęp 25.10.2018).

Halevy, Alon, Peter Norvig, i Fernando Pereira. 2009. „The Unreasonable Effectiveness of Data". *IEEE Intelligent Systems* 24(2): 8–12. http://ieeexplore.ieee.org/document/4804817/.

Halford, Susan, i Mike Savage. 2017. „Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research". *Sociology* 51(6): 1132–48. http://journals.sagepub.com/doi/10.1177/0038038517698639.

Hamideh. 2015. „tools - Do data scientists use Excel? - Data Science Stack Exchange". https://datascience.stackexchange.com/questions/5443/do-data-scientists-use-excel (dostęp 4.09.2018).

Hansen, Steven. 2017. „How Big Data Is Empowering AI and Machine Learning?" *hackernoon.com*. https://hackernoon.com/how-big-data-is-empowering-ai-and-machine-learning-4e93a1004c8f (dostęp 13.02.2018).

Harari, Yuval Noah. 2017. *Homo Deus: A Brief History of Tomorrow*. London: Vintage.

———. 2018. *21 lekcji na 21 wiek*. Kraków: Wydawnictwo Literackie.

Haratyk, Karol, Kamila Biały, i Marcin Gońda. 2017. „Biographical meanings of work: the case of a Polish freelancer". *Przegląd Socjologii Jakościowej* XIII(4): 136–59. http://www.qualitativesociologyreview.org/PL/Volume40/PSJ_13_4_Bialy_Gonda_Haratyk.pdf.

Harper, Norma Wynn, i C. J. Daane. 1998. „Causes and Reduction of Math Anxiety in Preservice Elementary Teachers". *Action in Teacher Education* 19(4): 29–38. http://www.tandfonline.com/doi/abs/10.1080/01626620.1998.10462889.

Harris, Harlan, D., Patric Murphy, Sean, i Marck Vaisman. 2013. *Analyzing The Analyzers: An Introspective Survey of Data Scientists and Their Work*. 1. wyd. Beijing, Cambridge: O'Reilly.

Harris, Jonathan. 2012. „Data Driven". W *The Human Face of Big Data*, red. Rick Smolan i Jennifer Erwitt. Sausalito, CA: Against All Odds Production, 200–203.

Harrison, John. 2019. „RSelenium: R Bindings for «Selenium WebDriver»". https://cran.r-project.org/package=RSelenium.

Hatton, Celia. 2015. „China «social credit»: Beijing sets up huge system". *BBC News*. https://www.bbc.com/news/world-asia-china-34592186 (dostęp 21.01.2019).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, i Jian Sun. 2015. „Deep Residual Learning for Image Recognition". https://arxiv.org/abs/1512.03385 (dostęp 2.08.2019).

Henderson, Peter i in. 2017. „Deep Reinforcement Learning that Matters". http://arxiv.org/abs/1709.06560.

Henke, Nicolaus, Jordan Levine, i Paul Mcinerney. 2018. „You Don't Have to Be a Data Scientist to Fill This Must- Have Analytics Role". *05.02.2018 Harvard Business Review*. https://hbr.org/2018/02/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role (dostęp 20.08.2018).

Henry, Lionel, i Hadley Wickham. 2018. „purrr: Functional Programming Tools". https://cran.r-project.org/package=purrr.

Hester, Jim, i Hadley Wickham. 2018. „odbc: Connect to ODBC Compatible Databases (using the DBI Interface)". https://cran.r-project.org/package=odbc.

Hill, Kashmir. 2012. „How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did". *16.02.2012 Forbes*. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#6d733aa66668 (dostęp 2.05.2018).

Hockey, Susan. 2004. „The History of Humanities Computing". W *A Companion to Digital Humanities*, red. Susan Schreibman, Ray Siemens, i John Unsworth. Oxford: Blackwell, 3–19. http://digitalhumanities.org/companion/.

Hockney, Ashley, i Hilary Green. 2018. „How Cambridge Analytica Used Machine Learning to Mine Facebook Data". *29.03.2018 Codeacademy*. http://news.codecademy.com/cambridge-analytica-machine-learning-facebook-data/ (dostęp 6.04.2018).

Hoffmann, Anna Lauren, Nicholas Proferes, i Michael Zimmer. 2018. „"Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users". *New Media and Society* 20(1): 199–218.

Hofman, Jake M., Amit Sharma, i Duncan J. Watts. 2017. „Prediction and explanation in social systems". *Science* 355(6324): 486–88. http://link.springer.com/10.1007/s10764-012-9616-1.

Hosni, Hykel, i Angelo Vulpiani. 2018. „Data science and the art of modelling". *Lettera Matematica* 6(2): 121–29. http://dx.doi.org/10.1007/s40329-018-0225-5.

Hsu, Wendy F. 2014. „Digital Ethnography Toward Augmented Empiricism: A New Methodological Framework". *Journal of Digital Humanities* 3(1). http://journalofdigitalhumanities.org/3-1/digital-ethnography-toward-augmented-empiricism-by-wendy-hsu/.

Hu, Jane C. 2018. „From ISIS to NIPS: How organizations handle unfortunate acronyms — Quartz". *Quartz*. https://qz.com/1437829/from-isis-to-nips-how-organizations-handle-unfortunate-acronyms/ (dostęp 12.02.2019).

Huang, Ronggui. 2018. „RQDA: R-based Qualitative Data Analysis". http://rqda.r-forge.r-project.org.

Huet, Ellen. 2016. „The Humans Hiding Behind the Chatbots". *Bloomberg*. https://www.bloomberg.com/news/articles/2016-04-18/the-humans-hiding-behind-the-chatbots (dostęp 6.01.2019).

Hughes, Phil. 1999. „An Interview with Guido van Rossum". *Linux Journal*. http://web.archive.org/web/20160310004241/http://www.linuxjournal.com/article/5028 (dostęp 16.07.2019).

Hunter, John D. 2007. „Matplotlib: A 2D Graphics Environment". *Computing in Science & Engineering* 9(3): 90–95. http://ieeexplore.ieee.org/document/4160265/.

Hunter, John D., i Michael Droettboom. 2012. „matplotlib". W *The Architecture of Open Source Applications, Volume II: Structure, Scale, and a Few More Fearless Hacks*, red. Amy Brown i Greg Wilson. Lulu.com. http://aosabook.org/en/matplotlib.html.

Hyndman, Rob. 2014. „Am I a data scientist?" *2014.12.09 [wpis na blogu]*. https://robjhyndman.com/hyndsight/am-i-a-data-scientist/ (dostęp 5.01.2018).

Ierusalimschy, Roberto, Waldemar Celes, i Luiz Henrique de Figueiredo. 2019. „Lua". https://www.lua.org/ (dostęp 15.07.2019).

Ihaka, Ross, i Robert Gentleman. 1996. „R: A Language for Data Analysis and Graphics". *Journal of Computational and Graphical Statistics* 5(3): 299. https://www.jstor.org/stable/1390807?origin=crossref.

Inbar, Ohad, Noam Tractinsky, i Joachim Meyer. 2007. „Minimalism in information visualization". W *Proceedings of the 14th European conference on Cognitive ergonomics invent! explore! - ECCE '07*, New York, New York, USA: ACM Press, 185. http://portal.acm.org/citation.cfm?doid=1362550.1362587.

Introna, Lucas, i David Wood. 2002. „Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems". *Surveillance & Society* 2(2/3): 177–98. https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/3373.

Isaak, Jim, i Mina J Hanna. 2018. „User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection". *Computer* 51(8): 56–59. https://ieeexplore.ieee.org/document/8436400/.

Ismay, Chester, i Albert Y Kim. 2018. *An Introduction to Statistical and Data Sciences via R*. ModernDive. https://moderndive.com/index.html.

IT Central Station. 2019. *Data Science Platforms: Buyer's Guide and Reviews*. https://www.itcentralstation.com/landing/report-data-science-platforms.

Iwasiński, Łukasz. 2016. „Społeczne zagrożenia danetyzacji rzeczywistości". W *Nauka o informacji w okresie zmian. Informatologia i humanistyka cyfrowa*, red. Barbara Sosińska-Kalata. Warszawa: Wydawnictwo SBP, 135–46.

———. 2017. „Przyczynek do rozważań nad suwerennością konsumenta w epoce danetyzacji i big data". *Kultura - Historia - Globalizacja* 21: 119–33.

Jackson, Michelle. 2001. „Meritocracy, Education and Occupational Attainment: What Do Employers Really See as Merit?" *Working Paper, Department of Sociology, University of Oxford.* 3: 1–24. https://www.sociology.ox.ac.uk/materials/papers/2001-03.pdf.

Jacobs, Adam. 2009. „The Pathologies of Big Data". *Communications of the ACM* 52(8): 36. http://portal.acm.org/citation.cfm?doid=1536616.1536632.

Jacyno, Małgorzata. 2007. *Kultura indywidualizmu*. Warszawa: Wydawnictwo Naukowe PWN.

Jain, Varsha. 2018. „Luxury: Not for Consumption but Developing Extended Digital Self". *Journal of Human Values* 24(1): 25–38.

James, Manyika i in. 2011. McKinsey Global Institute *Big data: The next frontier for innovation, competition, and productivity*.

Jarvis, Jeremy. 2014. „@jeremyjarvis: A data scientist is a statistician who lives in San Fransisco". *2014.01.30 Twitter*. https://twitter.com/jeremyjarvis/status/428848527226437632 (dostęp 7.12.2017).

Jemielniak, Dariusz. 2018. „Socjologia 2.0: o potrzebie łączenia big data z etnografią cyfrową, wyzwaniach jakościowej socjologii cyfrowej i systematyzacji pojęć". *Studia Socjologiczne* 2(229): 7–29.

———. 2019. *Socjologia internetu*. Warszawa: Scholar.

Jifa, Gu, i Zhang Lingling. 2014. „Data, DIKW, Big Data and Data Science". *Procedia Computer Science* 31: 814–21. https://www.sciencedirect.com/science/article/pii/S1877050914005092 (dostęp 1.03.2018).

Jodkowski, Kazimierz. 1990. *Wspólnoty uczonych, paradygmaty i rewolucje naukowe. Realizm. Racjonalność. Relatywizm*. Lublin: Wydawnictwo UMCS.

Johnson, Bobbie. 2010. „Privacy no longer a social norm, says Facebook founder". *11.01.2010 The Guardian*.

https://www.theguardian.com/technology/2010/jan/11/facebook-privacy (dostęp 10.04.2018).

Joler, Vladan, i Kate Crawford. 2018. „Anatomy of an AI system". https://anatomyof.ai/img/ai-anatomy-publication.pdf.

Jolly, Eshin. 2018. „Pymer4: Connecting R and Python for Linear Mixed Modeling". *Journal of Open Source Software* 3(31): 862. http://joss.theoj.org/papers/10.21105/joss.00862.

Jordaan, Yolanda, i Gené Van Heerden. 2017. „Online privacy-related predictors of Facebook usage intensity". *Computers in Human Behavior* 70: 90–96. http://dx.doi.org/10.1016/j.chb.2016.12.048.

Jozuka, Emiko. 2016. „A Japanese AI Almost Won a Literary Prize". *VICE*. https://www.vice.com/en_us/article/wnxnjn/a-japanese-ai-almost-won-a-literary-prize (dostęp 10.07.2019).

Jóźwiak, Michał. 2019. „Grzechomaty i księża-roboty, czyli o… Kościele przyszłości?" *21.08.2019 misyjne.pl*. https://misyjne.pl/grzechomaty-i-ksieza-roboty-czyli-o-kosciele-przyszlosci/ (dostęp 26.08.2020).

Julia, Angwin, Jeff Larson, Surya Mattu, i Lauren Kirchner. 2016. „Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." *23.05.2016 ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (dostęp 30.04.2018).

Junco, Pablo Ruiz. 2017. „Data Scientist Personas: What Skills Do They Have and How Much Do They Make?" *Glassdoor Economic Research*. https://www.glassdoor.com/research/data-scientist-personas/ (dostęp 25.10.2018).

Jung, Alexander. 2019. „imgaug Documentation. Release 0.2.9". https://buildmedia.readthedocs.org/media/pdf/imgaug/latest/imgaug.pdf.

Juza, Marta. 2016. „Internet w życiu społecznym - nadzieje, obawy, krytyka". *Studia Socjologiczne* 1(220): 199–221.

Kacperczyk, Anna. 2007. „Badacz i jego poszukiwania w świetle «Analizy Sytuacyjnej» Adele E. Clarke". *Przegląd Socjologii Jakościowej* 3(2): 5–32.

———. 2011. „Obiekt graniczny (Boundary object)". W *Słownik socjologii jakościowej*, red. Krzysztof T. Konecki i Piotr Chomczyński. Warszawa: Diffin, 192.

———. 2014. „Autoetnografia – technika, metoda, nowy paradygmat? O metodologicznym statusie autoetnografii". *Przegląd Socjologii Jakościowej* 10(3): 32–74.

———. 2016. *Społeczne światy. Teoria - empiria - metody badań: na przykładzie społecznego świata wspinaczki*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

———. 2017. „Rozum czy emocje? O odmianach autoetnografii oraz epistemologicznych przepaściach i pomostach między nimi". *Kultura i Społeczeństwo* 3(LXI): 127–48.

Kaczorowska-Spychalska, Dominika. 2019. „UŁ komentuje: Sztuczna inteligencja i biznes". https://www.uni.lodz.pl/aktualnosc/szczegoly/ul-komentuje-sztuczna-inteligencja-i-biznes-zwiazek-prawie-doskonaly (dostęp 4.07.2019).

Kaggle. 2017. „2017: The State of Data Science & Machine Learning". https://www.kaggle.com/surveys/2017 (dostęp 27.03.2018).

———. 2018. „Kaggle Days". https://www.kaggledays.com/ (dostęp dostęp 25.04.2018).

Kallenberg, Michiel i in. 2016. „Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring". *IEEE Transactions on Medical Imaging* 35(5): 1322–31. http://ieeexplore.ieee.org/document/7412749/.

Kane, Michael J., John W. Emerson, i Stephen Weston. 2013. „Scalable strategies for computing with massive data". *Journal of Statistical Software* 55(14): 1–19.

Kapur, Manu, i Nikol Rummel. 2012. „Productive failure in learning from generation and invention activities". *Instructional Science* 40(4): 645–50. http://link.springer.com/10.1007/s11251-012-9235-4.

karupakalas. 2018. „Answer to: tools - IDE alternatives for R programming (RStudio, IntelliJ IDEA, Eclipse, Visual Studio)"; *Data Science Stack Exchange*. https://datascience.stackexchange.com/a/28853 (dostęp 11.07.2019).

Kędzierski, Robert. 2018. „Niepokorni nie mogą nawet jeździć pociągiem. Pierwsze ofiary chińskiego systemu oceny obywateli". *gazeta.pl 02.06.2018*. http://next.gazeta.pl/next/7,156830,23474436,niepokorni-nie-moga-nawet-jezdzic-pociagiem-pierwsze-ofiary.html (dostęp 18.06.2018).

Keras. 2019. „Backend - Keras Documentation". https://keras.io/backend/ (dostęp 17.07.2019).

Kerzel, Urlich. 2017. „Big Data vs. Artificial Intelligence". *[wpis na blogu] BlueYoned*.

Ketkar, Nikhil. 2017. „Introduction to PyTorch". W *Deep Learning with Python*, Berkeley, CA: Apress, 195–208. http://link.springer.com/10.1007/978-1-4842-2766-4_12.

King, John, i Roger Magoulas. 2016. *Data Science Salary Survey*.

King, Thomas C., Nikita Aggarwal, Mariarosaria Taddeo, i Luciano Floridi. 2019. „Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". *Science and Engineering Ethics*: 1–32. https://doi.org/10.1007/s11948-018-00081-0.

Kitchin, Rob. 2014. „Big Data, new epistemologies and paradigm shifts". *Big Data & Society* 1(1): 1–12. http://journals.sagepub.com/doi/10.1177/2053951714528481.

Kitchin, Robert, i Mark Carrigan. 2014. „Big data should complement small data, not replace them". 1–8. http://blogs.lse.ac.uk/impactofsocialsciences/2014/06/27/series-philosophy-of-data-science-rob-kitchin/ (dostęp 23.04.2018).

Kling, Rob, i Elihu M. Gerson. 1978. „Patterns Of Segmentation And Intersection In The Computing World". *Symbolic Interaction* 1(2): 24–43. http://doi.wiley.com/10.1525/si.1978.1.2.24.

Knapik, Rozalia. 2018. *Sztuczny Bóg. Wizerunki technologicznej Osobliwości w (pop)kulturze*. Poznań: Instytut Kultury Popularnej.

Knight, Will. 2017. „Andrew Ng Is Leaving Baidu in Search of a Big New AI Mission". *2017.03.22 MIT Technology Review*. https://www.technologyreview.com/s/603897/andrew-ng-is-leaving-baidu-in-search-of-a-big-new-ai-mission/ (dostęp 1.04.2018).

Knox, Hannah, i Dawn Nafus, red. 2018. *Ethnography for a data-saturated world*. Manchester: Manchester University Press.

Koch, Therese. 2018. „NIPS AI Conference to Continue Laughing about Nipples at the Expense of Women in Tech". https://medium.com/@therese.koch1/nips-ai-conference-to-continue-laughing-about-nipples-at-the-expense-of-women-in-tech-8c0fa74b1ec4 (dostęp 12.02.2019).

Kodołamacz.pl, i Centrum Zarządzania Innowacjami i Transferem Technologii Politechniki Warszawskiej. 2017. „Zawody Przyszłości #1". http://kodolamacz.pl/zawodyprzyszlosci/ (dostęp 11.06.2017).

Koehrsen, Will. 2018. „Python and Slack: A Natural Match". *Towards Data Science*. https://towardsdatascience.com/python-and-slack-a-natural-match-60b136883d4d (dostęp 4.09.2019).

Koloch, Grzegorz i in. 2017. *Intensywność wykorzystania danych w gospodarce a jej rozwój – analiza diagnostyczna Wprowadzenie*. https://mc.bip.gov.pl/rok-2017/analiza-diagnostyczna-intesywnosc-wykorzystania-danych-w-gospodarce-a-jej-rozwoj.html.

Komisja Europejska. 2020. „White Paper on Artificial Intelligence - A European approach to excellence and trust". *COM(2020) 65 final*. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Kondratiuk, Ilona i in. 2017. „GSK-3β and MMP-9 Cooperate in the Control of Dendritic Spine Morphology". *Molecular Neurobiology* 54(1): 200–211. http://link.springer.com/10.1007/s12035-015-9625-0 (dostęp 5.12.2018).

Konecki, Krzysztof. 2020. „Uwagi na temat tego , co jest postrzegane jako ważne i nieważne w socjologii". *Przegląd Socjologii Jakościowej* XVI(2): 188–207.

Konecki, Krzysztof T. 2000. *Studia z metodologii badań jakościowych. Teoria ugruntowana*. Warszawa: Wydawnictwo Naukowe PWN.

———. 2010. „W stronę socjologii jakościowej: badanie kultur, subkultur i światów społecznych”. W *Kultury, subkultury i światy społeczne w badaniach jakościowych*, red. Jacek Leoński i Magdalena Fiternicka-Gorzko. Szczecin: Volumina.pl, 1–23.

Kopf, Dan. 2015. „Hadley Wickham, the Man Who Revolutionized R”. *Priceonomics*. https://priceonomics.com/hadley-wickham-the-man-who-revolutionized-r/ (dostęp 5.01.2018).

Koro-Ljungberg, Mirka. 2013. „«Data» As Vital Illusion”. *Cultural Studies - Critical Methodologies* 13(4): 274–78.

Kosinski, M., D. Stillwell, i T. Graepel. 2013. „Private traits and attributes are predictable from digital records of human behavior”. *Proceedings of the National Academy of Sciences* 110(15): 5802–5. http://www.pnas.org/cgi/doi/10.1073/pnas.1218772110.

Kotuła, Sebastian Dawid. 2014. *Wstęp do Open Source*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Kozinets, Robert V. 2003. „The Field behind the Screen: Using Netnography for Marketing Research in Online Communities”. *Journal of Marketing Research* 39(1): 61–72.

Kozyrkov, Cassie. 2018. „Top 10 roles in AI and data science”. *hackernoon.com*. https://medium.com/hackernoon/top-10-roles-for-your-data-science-team-e7f05d90d961 (dostęp 23.10.2018).

Krawczyk, Stanisław, i Piotr Migdał. 2011. „Zespół Aspergera, nauki ścisłe i kultura nerdów”. *Rocznik Kognitywistyczny* (V): 93–101.

Krawiec, Krzysztof. 2003. *Uczenie maszynowe i sieci neuronowe*. Poznań: Wydawnictwo Politechniki Poznańskiej.

Kretek-Kamińska, Agnieszka. 2016. „Strategie metodologiczno-badawcze realizowane w socjologicznych rozprawach doktorskich w latach 1993-2013”. Uniwersytet Łódzki.

Kriegel, Alex, i Boris M Trukhnov. 2011. *Discovering SQL: A Hands-On Guide for Beginners*. Indianapolis, Ind: Wrox.

Kromme, Jeroen. 2017. „Python & R vs. SPSS & SAS”. *18.03.2017 The Analytics Lab*. http://www.theanalyticslab.nl/2017/03/18/python-r-vs-spss-sas/ (dostęp 26.10.2018).

Kroplewski, Robert i in. 2019. *Polityka Rozwoju Sztucznej Inteligencji w Polsce na lata 2019-2027*. Warszawa. https://www.gov.pl/attachment/0aa51cd5-b934-4bcb-8660-bfecb20ea2a9.

Krzemiński, Ireneusz. 1986. *Symboliczny interakcjonizm i socjologia*. Warszawa: Państwowe Wydawnictwo Naukowe.

Krzysztofek, Kazimierz. 2010. „«Fragteracja» złożonych systemów społecznych - kilka pytań i hipotez badawczych”. *Studia i Prace Uniwersytetu Ekonomicznego w Krakowie* 13:

30–47. https://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ekon-element-000169260839.

———. 2011. „W stronę maszyn społecznych. Jaka będzie socjologia, której nie znamy?" *Studia Socjologiczne* 2(201): 123–45.

———. 2012. „Big Data Society. Technologie samozapisu i samopokazu: ku humanistyce cyfrowej". *Transformacje* 1–4(72–75): 223–57.

———. 2015. „Technologie cyfrowe w dyskursach o przyszłości pracy". *Studia Socjologiczne* 4(219): 50–31.

Kuhn, Thomas S. 2001. *Struktura rewolucji naukowych*. Warszawa: Fundacja Aletheia.

Kulesza, Kamil. 2018a. „Dyktatura big data. Kto pierwszy okiełzna dane - demokratyczna Europa czy autorytarne Chiny?" *Gazeta Wyborcza*. http://wyborcza.pl/Jutronauci/7,160052,22973430,prywatnosc-kocham-i-rozumiem.html.

———. 2018b. „Prywatność kocham i rozumiem". *Gazeta Wyborcza*.

Kulisiewicz, Marcin, Przemysław Kazienko, Boleslaw K Szymanski, i Radosław Michalski. 2018. „Entropy Measures of Human Communication Dynamics". *Scientific Reports* 8(1): 15697. https://doi.org/10.1038/s41598-018-32571-3.

Kuncewicz, Łukasz. 2019a. „Lukasz Kuncewicz na LinkedIn: „Data Science job interview – how the questions will change in 5 years?" *LinkedIn*. https://www.linkedin.com/feed/update/urn:li:activity:6556607403457155074 (dostęp 31.07.2019).

———. 2019b. „The I dare you list for Data Scientists". *LinkedIn*. https://www.linkedin.com/posts/kuncewicz_the-i-dare-you-list-for-data-scientists-activity-6568229477464313856-hD3g/ (dostęp 1.09.2019).

———. 2019c. „You're ready to become a Data Scientist if…". *LinkedIn*. https://www.linkedin.com/posts/kuncewicz_youre-ready-to-become-a-data-scientist-if-activity-6566757953188311040-Ie3C (dostęp 22.09.2019).

Kupilik, Matthew, i Frank Witmer. 2018. „Spatio-temporal violent event prediction using Gaussian process regression". *Journal of Computational Social Science* 1(2): 437–51. http://link.springer.com/10.1007/s42001-018-0024-y.

Kurczewska, Joanna. 1997. *Technokraci i ich świat społeczny*. Warszawa: Wydawnictwo Iinstytutu Filozofii i Socjologii PAN.

Kurzweil, Ray. 2016. *Nadchodzi Osobliwość. Kiedy człowiek przekroczy granice biologii*. Warszawa: Kurhaus Publishing.

Kurzweil, Ray, i Brian Strope. 2017. „Efficient Smart Reply, now for Gmail". *Google AI Blog*. https://ai.googleblog.com/2017/05/efficient-smart-reply-now-for-gmail.html (dostęp 9.07.2019).

Kuźba, Michał, i Przemysław Biecek. 2020. „What Would You Ask the Machine Learning Model? Identification of User Needs for Model Explanations Based on Human-Model Conversations". http://arxiv.org/abs/2002.05674.

Kvale, Steinar. 2004. *InterViews. Wprowadzenie do jakościowego wywiadu badawczego*. Białystok: Trans Humana.

———. 2010. *Prowadzenie wywiadów*. Warszawa: Wydawnictwo Naukowe PWN.

Kwiatkowska, Agnieszka. 2017. „«Hańba w Sejmie» – zastosowanie modeli generatywnych do analizy debat parlamentarnych". *Przegląd Socjologii Jakościowej* 13(2): 82–109.

Kwon, Ohbyung, Namyeon Lee, i Bongsik Shin. 2014. „Data quality management, data usage experience and acquisition intention of big data analytics". *International Journal of Information Management* 34(3): 387–94. https://linkinghub.elsevier.com/retrieve/pii/S0268401214000127.

Laboratorium Cyfrowe Humanistyki UW. 2017. „O Laboratorium". http://lach.edu.pl/o-laboratorium/ (dostęp 2.04.2017).

Landing AI. 2018. „Landing AI". https://landing.ai/ (dostęp 21.11.2018).

Laney, Douglas. 2001. „3-D data management: Controlling data volume, velocity and variety". *Application delivery strategies META Group Inc.* http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (dostęp 14.03.2016).

Langville, Amy N., i Carl D. Meyer. 2004. „Deeper inside Pagerank". *Internet Mathematics* 1(3): 335–80.

Lanthier, Mark. 2011. „An Introduction to Computer Science and Problem Solving". W *COMP 1405*, red. Mark Lanthier. Ottawa: Carleton University, 1–38. https://people.scs.carleton.ca/~lanthier/teaching/ProcessingNotes/COMP1405_Ch1_IntroductionToComputerScience.pdf.

Lapowsky, Issie. 2016. „The Man Behind Trump's Bid to Finally Take Digital Seriously". *08.19.2016 Wired*. https://www.wired.com/2016/08/man-behind-trumps-bid-finally-take-digital-seriously/ (dostęp 18.06.2018).

———. 2017. „What Did Cambridge Analytica Really Do for Trump's Campaign?" *Wired*. https://www.wired.com/story/what-did-cambridge-analytica-really-do-for-trumps-campaign/ (dostęp 7.02.2019).

———. 2019. „How Cambridge Analytica Sparked the Great Privacy Awakening". *Wired 2019.03.17*. https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/ (dostęp 17.03.2019).

Lardinois, Frederic. 2015. „As Kubernetes Hits 1.0, Google Donates Technology To Newly Formed Cloud Native Computing Foundation". *TechCrunch*.

https://techcrunch.com/2015/07/21/as-kubernetes-hits-1-0-google-donates-technology-to-newly-formed-cloud-native-computing-foundation-with-ibm-intel-twitter-and-others/ (dostęp 5.05.2019).

Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, New Jersey: John Wiley & Sons.

———. 2012. *Metody i modele eksploracji danych*. Warszawa: Wydawnictwo Naukowe PWN.

Larson, Jeff, Surya Mattu, Lauren Kirchner, i Angwin Julia. 2016. „How We Analyzed the COMPAS Recidivism Algorithm". *23.05.2016 ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (dostęp 30.04.2018).

Lasek, Mirosława. 2002. *Data Mining. Zastosowania w analizach i ocenach klientów bankowych*. Warszawa: Oficyna Wydawnicza „Zarządzanie i Finanse".

———. 2007. *Metody Data Mining w analizowaniu i prognozowaniu kondycji ekonomicznej przedsiębiorstw : zastosowania SAS Enterprise Miner*. Warszawa: Centrum Doradztwa i Informacji Difin.

Lassiter, Luke Eric. 2005. *The Chicago Guide to Collaborative Ethnography*. Chicago, London: The University of Chicago Press.

Latour, Bruno. 2009. „Tarde's idea of quantification". W *The Social After Gabriel Tarde: Debates and Assessments*, red. Mattei Candea. London: Routledge, 145–62. http://www.bruno-latour.fr/node/144.

———. 2010. *Splatając na nowo to, co społeczne. Wprowadzenie do Teorii Aktora-Sieci*. Kraków: TAiWPN UNIVERSITAS.

———. 2013. „Technologia jako utrwalone społeczeństwo". *AVANT. Pismo awangardy filozoficzno-naukowej* 4(1): 17–48.

Lazer, D., R. Kennedy, G. King, i A. Vespignani. 2014. „The Parable of Google Flu: Traps in Big Data Analysis". *Science* 343(6176): 1203–5. http://www.sciencemag.org/cgi/doi/10.1126/science.1248506.

Lazer, David i in. 2009. „Computational Social Science". *Science* 323(February): 721–23.

Le, Quoc V. i in. 2011. „Building high-level features using large scale unsupervised learning". http://arxiv.org/abs/1112.6209.

LearnDataSci. 2018. „Top Data Science Online Courses in 2018". https://www.learndatasci.com/best-data-science-online-courses/ (dostęp 21.02.2018).

LeCun, Y. i in. 1989. „Backpropagation Applied to Handwritten Zip Code Recognition". *Neural Computation* 1(4): 541–51. http://www.mitpressjournals.org/doi/10.1162/neco.1989.1.4.541.

LeCun, Y., L. Bottou, Y. Bengio, i P. Haffner. 1998. „Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86(11): 2278–2324. http://ieeexplore.ieee.org/document/726791/.

LeCun, Yann. „MNIST Demos on Yann LeCun's website". http://yann.lecun.com/exdb/lenet/index.html (dostęp 20.02.2018a).

———. „Yann LeCun's Biography". http://yann.lecun.com/ex/bio.html (dostęp 12.09.2018b).

Lee, Adrian. 2016. „Geoffrey Hinton, the «godfather» of deep learning, on AlphaGo". https://www.macleans.ca/society/science/the-meaning-of-alphago-the-ai-program-that-beat-a-go-champ/ (dostęp 23.11.2018).

Lee, Honglak, Peter Pham, Yan Largman, i Andrew Ng. 2009. „Unsupervised feature learning for audio classification using convolutional deep belief networks". W *Advances in Neural Information Processing Systems 22*, red. Y Bengio i in. Curran Associates, Inc., 1096–1104. http://papers.nips.cc/paper/3674-unsupervised-feature-learning-for-audio-classification-using-convolutional-deep-belief-networks.pdf.

Leek, J. T., i R. D. Peng. 2015. „What is the question?" *Science* 347(6228): 1314–15. https://www.sciencemag.org/lookup/doi/10.1126/science.aaa6146.

Leek, Jeff. 2015. *The Elements of Data Analytic Style*. Leanpub. https://leanpub.com/datastyle/.

Leek, Jeffrey. 2016. „How to be a Modern Scientist". http://leanpub.com/modernscientist.

Lerman, Rachel, i Matt O'Brien. 2019. „Google's privacy push gets a mixed reception". *Washington Times*. https://www.washingtontimes.com/news/2019/may/8/googles-privacy-promises-dont-sway-many-experts/ (dostęp 20.05.2019).

Leviathan, Yaniv. 2018. „Google Duplex: An AI System for Accomplishing Real World Tasks Over the Phone". *08.05.2018 Google AI Blog*. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html (dostęp 9.05.2018).

Levin, Sam. 2019. „Google scraps AI ethics council after backlash: «Back to the drawing board»". *2019.04.05 The Guardian*. https://www.theguardian.com/technology/2019/apr/04/google-ai-ethics-council-backlash (dostęp 5.04.2019).

Levy, Steven. 2019. „Cambridge Analytica, Whistle-Blowers, and Tech's Dark Appeal". *Wired*. https://www.wired.com/story/cambridge-analytica-whistle-blowers-and-techs-dark-appeal/ (dostęp 15.10.2019).

Lewis-Kraus, Gideon. 2017. „Budowniczowie wieży Googel [przedruk z "The New York Times" 18.12.2016, tłum. Marcin Orliński]". *Przekrój* 2(3557): 151–62.

Li, Jun. 2008. „Ethical Challenges in Participant Observation : A Reflection on Ethnographic Fieldwork". *The Qualitative Report* 13(1): 100–115. http://www.nova.edu/ssss/QR/QR13-1/li.pdf.

Li, Michael, i Paul Paczuski. 2017. *Ranked: 15 Python packages for Data Science*. http://blog.thedataincubator.com/wp-content/uploads/2017/04/Ranked-15-Python-Packages-for-Data-Science.pdf.

Lin, Yu-Wei. 2011. „A qualitative enquiry into OpenStreetMap making". *New Review of Hypermedia and Multimedia* 17(1): 53–71. https://doi.org/10.1080/13614568.2011.552647.

Linden, Gregory D, Jennifer A Jacobi, i Eric A Benson. 2001. „Collaborative recommendations using item-to-item similarity mappings". http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=/netahtml/PTO/srchnum.htm&r=1&f=G&l=50&s1=6,266,649.PN.&OS=PN/6,266,649&RS=PN/6,266,649.

Lohr, Steve. 2009. „For Today's Graduate, Just One Word: Statistics". *The New York Times*. http://www.nytimes.com/2009/08/06/technology/06stats.html (dostęp 15.02.2018).

López, Gustavo, Luis Quesada, i Luis A Guerrero. 2018. „Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces BT - Advances in Human Factors and Systems Interaction". W *International Conference on Applied Human Factors and Ergonomics*, red. Isabel L Nunes. Cham: Springer International Publishing, 241–50.

Lorenz, Chris. 2012. „If you're so smart, why are you under surveillance? Universities, neoliberalism, and new public management". *Critical Inquiry* 38(3): 599–629.

Loukides, Mike. 2010. „What is data science?" *O'Reilly Media*. https://www.oreilly.com/ideas/what-is-data-science (dostęp 8.09.2016).

Lowrie, Ian. 2016. „Caring for Computers: How Russian Data Scientists Refashion Their Laptops". *Anthropology Now* 8(2): 25–33.

———. 2017. „Algorithmic rationality: Epistemology and efficiency in the data sciences". *Big Data & Society* 4(1): 1–13. http://journals.sagepub.com/doi/10.1177/2053951717700925.

———. 2018. „Becoming a real data scientist. Expertise, flexibility and lifelong learning". W *Ethnography for a data-saturated world*, red. Hannah Knox i Dawn Nafus. Manchester: Manchester University Press, 62–81. http://www.manchesterhive.com/view/9781526127600/9781526127600.00010.xml.

Lulek, Izabela. 2013. „Rozwój rynku venture capital/private equity w Polsce". *Copernican Journal of Finance & Accounting* 2(1): 119–37. http://wydawnictwoumk.pl/czasopisma/index.php/CJFA/article/view/1802.

Lutyński, Jan. 1974. „Uwagi na temat typologii wywiadów". *[maszynopis]*.

Lyotard, Jean-François. 1997. *Kondycja ponowoczesna: raport o stanie wiedzy*. Warszawa: Fundacja Aletheia.

Mac, Ryan. 2018. „Cambridge Analytica Data Scientist Aleksandr Kogan Wants You To Know He's Not A Russian Spy". *BuzzFeedNews*. https://www.buzzfeednews.com/article/ryanmac/facebook-cambridge-analytica-aleksandr-kogan-not-spy (dostęp 1.02.2019).

Maciąg, Rafał. 2014. „Sieć i Społeczeństwo Sieci – zarys rozwoju". *Zarządzanie Mediami* 2(4): 157–67.

Majek, Karol. 2020. „Polskie blogi o sztucznej inteligencji i analizie danych". *Deep Drive PL*. https://deepdrive.pl/polskie-blogi-o-sztucznej-inteligencji-i-analizie-danych/ (dostęp 20.04.2020).

Mallan, Kerry Margaret, Parlo Singh, i Natasha Giardina. 2010. „The challenges of participatory research with 'tech-savvy' youth". *Journal of Youth Studies* 13(2): 255–72. http://www.tandfonline.com/doi/full/10.1080/13676260903295059.

Mandl, Ulrike, Adriaan Dierx, i Fabienne Ilzkovitz. 2008. *The effectiveness and efficiency of public spending*. https://ideas.repec.org/p/euf/ecopap/0301.html.

Mandziejewski, Adam. 2018. „Na ile pozwolimy sztucznej inteligencji: wywiad z Luciano Floridi". *12.09.2018 Przekrój*. https://przekroj.pl/nauka/na-ile-pozwolimy-sztucznej-inteligencji-adam-mandziejewski (dostęp 14.09.2018).

Manhart, Klaus. 1996. „Artificial Intelligence Modelling: Data Driven and Theory Driven Approaches". *Social Science Micro Simulation* (dostęp September 2007): 416–31.

Mannes, John. 2017. „Geofferey Hinton was briefly a Google intern in 2012 because of bureaucracy". *TechCrunch*. https://techcrunch.com/2017/09/14/geoffrey-hinton-was-briefly-a-google-intern-in-2012-because-of-bureaucracy/ (dostęp 23.11.2018).

Manovich, Lev. 2012a. *Język nowych mediów*. Warszawa: Oficyna Wydawnicza Łośgraf.

———. 2012b. „Trending: The Promises and the Challenges of Big Social Data". W *Debates in the Digital Humanities*, red. Matthew K Gold. Minneapolis: University of Minnesota Press, 460–75. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.

———. 2018. *AI Aesthetics*. Moskwa: Strelka Press. http://manovich.net/index.php/projects/ai-aesthetics.

Manyika, James i in. 2017. McKinsey Global Institute *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*.

Marcus, George E. 1995. „Ethnography in / of the World System : The Emergence of Multi-Sited Ethnography". *Annual Review of Anthropology* (24): 95–117.

Marczuk, Piotr, Piotr Mieczkowski, Leonardo Calini, i Bartosz Paszcza. 2019. *Iloraz sztucznej inteligencji. Potencjał AI w polskiej gospodarce*. Warszawa. https://www.digitalpoland.org/assets/publications/iloraz-sztucznej-inteligencji/iloraz-sztucznej-inteligencji-edycja-2-2019.pdf.

Marr, Bernard. 2016. „What Is The Difference Between Artificial Intelligence And Machine Learning?" *Forbes*. https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#23890af2742b (dostęp 14.02.2018).

Martin, Vivian B. 2006. „The Postmodern Turn: Shall Classic Grounded Theory Take That Detour? A Review Essay". *The Grounded Theory Review* 5(2/3): 119–29. http://groundedtheoryreview.com/wp-content/uploads/2012/06/GT-Review-vol5-no2-3-final-printers-version4.pdf.

Maslow, Abraham H. 1966. *The Psychology of Science. A Reconnaissance*. South Bend, Indiana: Gateway Editions.

Mason, Hilary. 2010. „A Taxonomy of Data Science". *25.09.2010 dataists. Fresher than seeing your model doesn't have heteroscedastic errors*. http://www.dataists.com/2010/09/a-taxonomy-of-data-science/ (dostęp 30.07.2018).

masterindatascience. 2018. „Top 23 Schools with Data Science Master's Programs". https://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/ (dostęp 21.02.2018).

Matloff, Norman. 2017. „Norm Matloff's answer to Will Python take over R? - Quora". *25.12.2017 Quora*. https://www.quora.com/Will-Python-take-over-R/answer/Norm-Matloff (dostęp 8.09.2018).

———. 2020. „TidyverseSkeptic: An opinionated view of the Tidyverse «dialect» of the R language." *GitHub*. https://github.com/matloff/TidyverseSkeptic#readme (dostęp 23.10.2020).

Matplotlib. 2018. „Matplotlib: Python plotting — Matplotlib 3.0.2 documentation". https://matplotlib.org/index.html (dostęp 16.09.2018).

Matuszewski, Paweł. 2018. *Cyberplemiona: analiza zachowań użytkowników Facebooka w trakcie kampanii parlamentarnej*. Warszawa: Wydawnictwo Naukowe PWN.

Mayo Clinic. „Alumni - Mayo Clinic Research". https://www.mayo.edu/research/labs/advanced-medical-imaging-technology/about/alumni (dostęp 3.12.2018).

Mccarthy, John. 1960. „Recursive Functions of Symbolic Expressions and Their Computation by Machine, Part I". *Communications of the ACM* (April): 1–34. http://www-

formal.stanford.edu/jmc/recursive.pdf%5Cnpapers2://publication/uuid/F449381A-D60A-4567-B758-C5F74E0F45ED.

McCarthy, John, Marvin L Minsky, Nathaniel Rochester, i Claude E Shannon. 1955. „Darmouth AI Project Proposal". http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

McClure, Sean. 2019. „Step-by-Step Guide to Creating R and Python Libraries (in JupyterLab)". *Towards Data Science*. https://towardsdatascience.com/step-by-step-guide-to-creating-r-and-python-libraries-e81bbea87911 (dostęp 16.07.2019).

McCulloch, Gretchen. 2019. „Coding Is for Everyone—as Long as You Speak English". *Wired*. https://www.wired.com/story/coding-is-for-everyoneas-long-as-you-speak-english/ (dostęp 10.09.2019).

McKinney, Wes. 2010. „Data Structures for Statistical Computing in Python". W *Proceedings of the 9th Python in Science Conference*, red. Stéfan van der Walt i Jarrod Millman. 51–56.

———. 2011. „pandas: a Foundational Python Library for Data Analysis and Statistics". W *PyHPC 2011: Python for High Performance and Scientific Computing*,.

———. 2012. *Python for Data Analysis. Data Wrangling With Pandas, NumPy, And IPython*. Sebastopol: O'Reilly.

———. 2018a. „About - Wes McKinney". http://wesmckinney.com/pages/about.html (dostęp 4.12.2018).

———. 2018b. *Python w analizie danych. Przetwarzanie danych za pomocą pakietów Pandas i NumPy oraz środowiska IPython*. Gliwice: Helion.

Meetup. 2019. „Meetup API - Documentation". https://www.meetup.com/pl-PL/meetup_api/docs/ (dostęp 2.06.2019).

Melton, Jim. 1996. „SQL language summary". *ACM Computing Surveys* 28(1): 141–43.

Merity, Stephen. 2017. „Bias is not just in our datasets, it's in our conferences and community". *2017.12.11 [wpis na blogu]*. https://smerity.com/articles/2017/bias_not_just_in_datasets.html (dostęp 10.02.2019).

Merritt, Jonathan. 2018. „Sztuczna inteligencja zagrożeniem dla chrześcijaństwa?" *Miesięcznik Znak* (762): 52–57. https://www.miesiecznik.znak.com.pl/sztuczna-inteligencja-zagrozeniem-dla-chrzescijanstwa/.

Metcalf, Jacob, i Kate Crawford. 2016. „Where are human subjects in Big Data research? The emerging ethics divide". *Big Data & Society* 3(1). http://journals.sagepub.com/doi/10.1177/2053951716650211.

MI^2. 2018. „MI^2. Do Things That Matter". http://mi2.mini.pw.edu.pl/ (dostęp 6.09.2018).

Michel, Jean Baptiste i in. 2011. „Quantitative analysis of culture using millions of digitized books". *Science* 331(6014): 176–82.

Microsoft. 2019a. „Microsoft Azure Marketplace". https://azuremarketplace.microsoft.com/en-us/marketplace/ (dostęp 2.08.2019).

———. 2019b. „Microsoft R Open: The Enhanced R Distribution". *MRAN*. https://mran.microsoft.com/open (dostęp 19.07.2019).

Microsoft Azure. 2019. „Azure Databricks". https://azure.microsoft.com/pl-pl/services/databricks/ (dostęp 3.06.2019).

Microsoft Research. 2019a. „People: danah boyd". https://www.microsoft.com/en-us/research/people/dmb/ (dostęp 11.02.2019).

———. 2019b. „People: Kate Crawford". https://www.microsoft.com/en-us/research/people/kate/#!publications (dostęp 11.02.2019).

Migdał, Piotr. 2014. „Symmetries and self-similarity of many-body wavefunctions". http://arxiv.org/abs/1412.6796 (dostęp 23.07.2019).

———. 2016. „From Science to Data Science, a Comprehensive Guide for Transition". *KDnuggets*. https://www.kdnuggets.com/2016/04/data-science-comprehensive-guide-transition.html (dostęp 3.02.2018).

———. 2017. „After PyData Warsaw 2017". *[wpis na blogu]*. https://p.migdal.pl/2017/11/15/after-pydata-warsaw-2017.html (dostęp 1.12.2017).

———. 2018. „Keras or PyTorch as your first deep learning framework - deepsense.ai". *deepsense.ai*. https://deepsense.ai/keras-or-pytorch/ (dostęp 17.07.2019).

Miller, Justin J. 2013. „Graph database applications and concepts with Neo4j". *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA* 2324: 36.

Miller, Tim. 2019. „Explanation in artificial intelligence: Insights from the social sciences". *Artificial Intelligence* 267: 1–38. https://www.sciencedirect.com/science/article/abs/pii/S0004370218305988 (dostęp 21.12.2018).

Miloslavskaya, Natalia, i Alexander Tolstoy. 2016. „Big Data, Fast Data and Data Lake Concepts". *Procedia Computer Science* 88: 300–305. http://dx.doi.org/10.1016/j.procs.2016.07.439.

Minelli, Michael, Ambiga Dhiraj, i Michele Chambers. 2013. *Big Data, Big Analytics. Emerging business intelligence and analytic trends for today's businesses*. New Jersey: John Wiley & Sons.

Ministerstwo Cyfryzacji RP. 2018a. „RODO: Informator". https://www.gov.pl/cyfryzacja/rodo-informator.

———. 2018b. „Założenia do strategii AI w Polsce". https://www.gov.pl/documents/31305/436699/Założenia_do_strategii_AI_w_Polsce_-_raport.pdf/a03eb166-0ce5-e53c-52a4-3bfb903edf0a.

Misal, Disha. 2018. „PyTorch vs Keras: Who Suits You The Best". *Analytics India Magazine*. https://www.analyticsindiamag.com/pytorch-vs-keras-who-suits-you-the-best/ (dostęp 19.02.2019).

Mohamed, Shakir, Marie Therese Png, i William Isaac. 2020. „Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence". Philosophy & Technology 33(4): 659–84. http://link.springer.com/10.1007/s13347-020-00405-8.

Molnar, Christoph. 2018. „iml: An R package for Interpretable Machine Learning". *Journal of Open Source Software* 3(26): 786. http://joss.theoj.org/papers/10.21105/joss.00786.

Moravčík, Matej i in. 2017. „DeepStack: Expert-level artificial intelligence in heads-up no-limit poker". *Science* 356(6337): 508–13. http://www.sciencemag.org/lookup/doi/10.1126/science.aam6960.

Morrison, Daniel R. 2016. „Mapping to Make Sense of Messy Worlds". *Symbolic Interaction* 39(3): 519–21.

Mortimer, Steven. 2018. „Most Starred R Packages on GitHub". *2018.06.18*. https://stevenmortimer.com/most-starred-r-packages-on-github/ (dostęp 11.07.2018).

Morzy, Mikołaj. 2014. „Big Data: fakty, mity, obietnice i zagrożenia". *29.12.2014 [wpis na blogu]*. https://dataminingalapolonaise.wordpress.com/2014/12/29/big-data-fakty-mity-obietnice-i-zagrozenia/ (dostęp 14.08.2018).

Morzy, Tadeusz. 2013. *Eksploracja danych. Metody i algorytmy*. Warszawa: Wydawnictwo Naukowe PWN.

Mróz, Maciej. 2020. „Nowy renesans, czyli SI po ludzku - Sztuczna Inteligencja". *06.03.2020 Sztuczna Inteligencja*. https://www.sztucznainteligencja.org.pl/nowy-renesans-czyli-si-po-ludzku/ (dostęp 10.03.2020).

Muenchen, Robert A. 2014. „Why R is Hard to Learn". *r4stats*. http://r4stats.com/articles/why-r-is-hard-to-learn/ (dostęp 21.03.2018).

———. 2019. „The Popularity of Data Science Software". *r4stats*. http://r4stats.com/articles/popularity/ (dostęp 1.07.2018).

Müller, Kirill, i Lorenz Walthert. 2020. „styler: Non-Invasive Pretty Printing of R Code". https://cran.r-project.org/package=styler.

Müller, Kirill, i Hadley Wickham. 2018. „tibble: Simple Data Frames". https://cran.r-project.org/package=tibble.

Munroe, Randall. 2010. „xkcd: Convincing". *xkcd. A Webcomic of Romance, Sarcasm, Math and Language*. https://www.xkcd.com/833/ (dostęp 29.09.2017).

Murphy, Samantha. 2014. „Facebook Changes Its «Move Fast and Break Things» Motto". *Mashable*. https://mashable.com/2014/04/30/facebooks-new-mantra-move-fast-with-stability/ (dostęp 14.01.2018).

Muûlu, Kıvanç i in. 2012. „Speculative analysis of integrated development environment recommendations". *ACM SIGPLAN Notices* 47(10): 669–82. http://dl.acm.org/citation.cfm?doid=2398857.2384665.

Naur, Peter. 1974. *Concise Survey of Computer Methods*. Lund: Studentlitteratur. http://www.naur.com/Conc.Surv.html.

Nauta, Gerhard Jan. 2008. „As You Can See: Applying Visual Collaborative Filtering to Works of Art". *Digital Humanities Quarterly* 1(2). http://www.digitalhumanities.org/dhq/vol/2/1/000019/000019.html.

Neff, Gina, Anissa Tanweer, Brittany Fiore-Gartland, i Laura Osburn. 2017. „Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science". *Big Data* 5(2): 85–97.

Netflix. 2019. „How Netflix's Recommendations System Works". https://help.netflix.com/en/node/100639 (dostęp 9.07.2019).

NeurIPS. 2018. „NeurIPS Code Of Conduct". https://neurips.cc/public/CodeOfConduct (dostęp 12.02.2019).

Neuwirth, Erich. 2014. „RColorBrewer: ColorBrewer Palettes". https://cran.r-project.org/package=RColorBrewer.

Newton, Casey. 2019. „The Trauma Floor. The secret lives of Facebook moderators in America". *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona (dostęp 26.02.2019).

Ng, Andrew. 2012. „Machine Learning". *Coursera*. https://www.coursera.org/learn/machine-learning (dostęp 18.01.2017).

———. 2017. „AI is the new electricity". W *AI Frontiers. Applied Deep Learning*, Santa Clara. https://nov2017.aifrontiers.com/#speakers.

———. 2018a. „About - Andrew Ng". https://www.andrewng.org/about/ (dostęp 21.01.2018).

———. 2018b. *Machine Learning Yearning - Draft Version*. deeplearning.ai. https://d2wvfoqc9gyqzf.cloudfront.net/content/uploads/2018/09/Ng-MLY01-13.pdf.

———. 2018c. „Yann LeCun Interview - Foundations of Convolutional Neural Networks”. *Coursera*. https://www.coursera.org/lecture/convolutional-neural-networks/yann-lecun-interview-4PnfT (dostęp 23.10.2018).

———. „Andrew Ng’s Home page”. http://ai.stanford.edu/~ang/originalHomepage.html (dostęp 21.01.2018).

Ng, Andrew, i Jennifer Widom. *Origins of the Modern MOOC (xMOOC)*. http://www.robotics.stanford.edu/~ang/papers/mooc14-OriginsOfModernMOOC.pdf.

Ng, Andrew, Alice X Zheng, i Michael I Jordan. 2001. „Stable Algorithms for Link Analysis”. W *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’01*, New York, New York, USA: ACM Press, 258–66. http://portal.acm.org/citation.cfm?doid=383952.384003.

Nguyen, Clinton. 2016. „China’s social credit score is like a «Black Mirror» episode”. *Business Insider*. https://www.businessinsider.com/china-social-credit-score-like-black-mirror-2016-10 (dostęp 21.01.2019).

Nicolaus, Henke i in. 2016. McKinsey Global Institute *The age of analytics: Competing in a data-driven world*. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

Niedbalski, Jakub. 2013. *Odkrywanie CAQDAS. Wybrane bezpłatne programy komputerowe wspomagające analizę danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

———, red. 2014. *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Nissenbaum, Helen. 2001. „How Computer Systems Embody Values”. *IEEE Computer* (120): 118–19. https://nissenbaum.tech.cornell.edu/papers/embodyvalues.pdf.

Northpointe Inc. 2011. „Practitioners Guide to COMPAS”.

Norvig, Peter. 1987. „A Unified Theory of Inference for Text Understanding”. University of California, Berkeley. http://www.eecs.berkeley.edu/Pubs/TechRpts/1987/5995.html.

———. 1992. *Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

———. 2001. „Teach Yourself Programming in Ten Years”. http://norvig.com/21-days.html (dostęp 13.04.2018).

———. 2012. „Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning”. *Significance* 9(4): 30–33. http://doi.wiley.com/10.1111/j.1740-9713.2012.00590.x (dostęp 22.11.2018).

———. 2018. „Peter Norvig - Resume”.

Norvig, Peter, i Sebastian Thrun. 2012. „Intro to Artificial Intelligence”. *Udacity*. https://eu.udacity.com/course/intro-to-artificial-intelligence--cs271 (dostęp 22.06.2018).

Nowacki, Filip. 2014. „Marketing 4.0 - nowa koncepcja w obliczu przemian współczesnego konsumenta”. *Marketing i Rynek* 6(June): 11–19. https://www.researchgate.net/profile/Filip_Nowacki/publication/278006393_Marketing_40_-_nowa_koncepcja_w_obliczu_przemian_wspolczesnego_konsumenta/links/557824ce08aeb6d8c01dd680.pdf.

Nowosad, Jakub. 2019. *Elementarz programisty. Wstęp do programowania używając R*. Poznań: Space A. https://nowosad.github.io/elp/elementarz-programisty-nowosad-2019.pdf.

NumFOCUS. 2018. „NumFOCUS: Open Code = Better Science”. https://numfocus.org/ (dostęp 3.12.2018).

Nunns, James. 2017. „How Python rose to the top of the data science world”. *Computer Business Review*. https://www.cbronline.com/big-data/analytics/python-rose-top-data-science-world/ (dostęp 2.10.2018).

Nurczyk, Ewelina, i Barbara Ramza. 2017. „Zawody przyszłości: Data Scientist”. *Kariera w Finansach i Bankowości 2017/2018*: 26–30. https://www.karierawfinansach.pl/przewodnik/2017/zawody-przyszlosci/data-scientist.

NYU Center for Data Science. 2013. „Yann LeCun Appointed Director of NYU Center for Data Science”. *2013.02.18*. https://cds.nyu.edu/yann-lecun-appointed-director-of-nyu-center-for-data-science/ (dostęp 19.09.2018).

O’Connor, Brendan O, David Bamman, i Noah a Smith. 2011. „Computational Text Analysis for Social Science : Model Assumptions and Complexity”. *Second Workshop on Computational Social Science andWisdom of the Crowds (NIPS 2011)*: 1–8. http://people.cs.umass.edu/~wallach/workshops/nips2011css/papers/OConnor.pdf.

O’Neil, Cathy. 2013. O’Reilly Media *On Being a Data Skeptic*. Sebastopol: O’Reilly.

———. 2017a. *Broń matematycznej zagłady. Jak algorytmy zwiększają nierówności i zagrażają demokracji*. Warszawa: Wydawnictwo Naukowe PWN.

———. 2017b. „The Ivory Tower Can’t Keep Ignoring Tech”. *The New York Times*. https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html (dostęp 11.02.2019).

O’Neil, Cathy, i Rachel Schutt. 2015. *Badanie danych: raport z pierwszej linii działań*. Gliwice: Helion.

Obem, Anna. 2018. „Nowa afera, stare wyzwania". *23.03.2018 Fundacja Panoptykon*. https://cyfrowa-wyprawka.org/aktualnosci/nowa-afera-stare-wyzwania (dostęp 24.04.2018).

———. 2020. „Co możesz zrobić po obejrzeniu Social Dilemma (poza wyrzuceniem telefonu)?" *Fundacja Panoptykon*. https://panoptykon.org/spoleczny-dylemat (dostęp 12.10.2020).

OECD. 2015. „Does math make you anxious?" *PISA in Focus* 02: 1–4. https://www.oecd-ilibrary.org/docserver/5js6b2579tnx-en.pdf?expires=1571755126&id=id&accname=guest&checksum=05935AB2BB78220F9D61C443782798C7.

Ogilvy. 2018. „The Dress For Respect". https://www.ogilvy.com/work/the-dress-for-respect/ (dostęp 30.11.2018).

Ohlhorst, Frank J. 2013. *Big Data Analytics: Turning Big Data into Big Money*. New York: Wiley.

Oliphant, Travis E. 2006. *A guide to NumPy*. USA: Trelgol Publishing. https://archive.org/details/NumPyBook/.

———. 2007. „Python for Scientific Computing". *Computing in Science & Engineering* 9(3): 10–20. http://ieeexplore.ieee.org/document/4160250/.

Oliphant, Travis E, Armando Manduca, Richard L Ehman, i James F Greenleaf. 2001. „Complex-Valued Stiffness Reconstruction for Magnetic Differential Equation". *Magnetic Resonance in Medicine* 45(September 2000): 299–310.

Olsen, Mark. 1993. „Signs, symbols and discourses: A new direction for computer-aided literature studies". *Computers and the Humanities* 27(5–6): 309–14. http://link.springer.com/10.1007/BF01829380.

Olszewski, Adrian. 2017. „Adrian Olszewski's answer to Will Python take over R? - Quora". *04.06.2017 Quora*. https://www.quora.com/Will-Python-take-over-R/answer/Adrian-Olszewski-1 (dostęp 8.09.2018).

———. 2018. „Adrian Olszewski's answer to Why do so many statisticians not want to become data scientists? Why are they not interested in Big Data? - Quora". *16.06.2018 Quora*. https://www.quora.com/Why-do-so-many-statisticians-not-want-to-become-data-scientists-Why-are-they-not-interested-in-Big-Data/answer/Adrian-Olszewski-1 (dostęp 8.09.2018).

OLX. 2018. „Data Ninja - Fight with real-world machine learning problems". https://dataninja.olx.pl/ (dostęp 3.01.2019).

Onalytica. 2017. *Big Data 2017: Top 100 Influencers And Brands*. http://www.onalytica.com/wp-content/uploads/2017/05/Onalytica-Big-Data-Top-100-Influencers-and-Brands.pdf.

———. 2018. *Data Science: Top 100 Influencers, Brands & Publications*. http://www.onalytica.com/wp-content/uploads/2018/04/Onalytica-Data-Science-Top-100-Influencers-Brands-and-Publications.pdf.

Ooms, Jeroen. 2014. „The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects". http://arxiv.org/abs/1403.2805.

van den Oord, Aaron, Sander Dieleman, i Benjamin Schrauwen. 2013. „Deep content-based music recommendation". W *Advances in Neural Information Processing Systems 26*, red. C J C Burges i in. Curran Associates, Inc., 2643–51. http://papers.nips.cc/paper/5004-deep-content-based-music-recommendation.pdf.

Orsini, Lauren. 2014. „Why Python Makes A Great First Programming Language". *ReadWrite*. https://readwrite.com/2014/07/08/what-makes-python-easy-to-learn/ (dostęp 25.07.2019).

Osowski, Stanisław. 2006. *Sieci neuronowe do przetwarzania informacji*. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej.

———. 2013. *Metody i narzędzia eksploracji danych*. Warszawa: Wydawnictwo BTC.

Ozimek, Adam. 2017. „The Paradox Of Robots Taking All Our Jobs". *Forbes*. https://www.forbes.com/sites/modeledbehavior/2017/10/28/the-paradox-of-robots-taking-all-our-jobs/#1f8e16501274 (dostęp 14.02.2018).

Ożóg, Maciej. 2009. „Transgresje panoptykonu. Nadzór w dobie technologii cyfrowych". *Kultura Współczesna* 2(60): 14–30.

Page, Lawrence, Sergey Brin, Rajeev Motwani, i Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/.

Paharia, Rajat. 2014. *Lojalność 3.0: jak zrewolucjonizować zaangażowanie klientów i pracowników dzięki big data i rywalizacji*. Warszawa: MT Biznes Ltd.

Paliszkiewicz, Joanna. 2018. „Kreowanie w mediach społecznościowych wizerunku kandydata do pracy na przykładzie portalu LinkedIn". *Zarządzanie Zasobami Ludzkimi* 5(124): 79–91.

Paluch, Robert i in. 2018. „Fast and accurate detection of spread source in large complex networks". *Scientific Reports* 8(1): 2508. https://doi.org/10.1038/s41598-018-20546-3.

Paprocki, Wojciech. 2016. „CYFRYZACJA GOSPODARKI I SPOŁECZEŃSTWA – SZANSE I WYZWANIA DLA SEKTORÓW INFRASTRUKTURALNYCH". W *Cyfryzacja gospodarki i społeczeństwa*, red. Wojciech Paprocki, Jerzy Grajewski, i Jana Pieriegud. Gdańsk: Europejski Kongres Finansowy, 39–57. https://efcongress.com/sites/default/files/publikacja_ekf_2016_cyfryzacja_gospodarki_i_spoeczestwa.pdf.

Pascnau, Razovan, Victoria Patraucean, i Doina Precup. 2018. *Transylvanian Machine Learning Summer School (TMLSS). Summary of the first edition*. https://drive.google.com/file/d/1Bcuzv9MM-U3CG_QLA63zzGB1E3zfjUgQ/view.

Pasquale, Frank. 2018. „Odd Numbers". *Real Life*. https://reallifemag.com/odd-numbers/ (dostęp 8.02.2019).

Paszke, Adam i in. 2017. „Automatic differentiation in PyTorch". W *NIPS-W*, https://openreview.net/pdf?id=BJJsrmfCZ.

Patrik, Linda E. 2007. „Encoding for Endangered Tibetan Texts". *Digital Humanities Quarterly* 1(1). http://www.digitalhumanities.org/dhq/vol/1/1/000004/000004.html.

Paulson, Linda Dailey. 2007. „Developers Shift to Dynamic Programming Languages". *Computer* 40(2). http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=4085614.

Pavlicek, Antonin, Frantisek Sudzina, i Ludmila Malinova. 2017. „Impact of Gender and Personality Traits (Bfi-10) on Tech Savviness". *Idimt-2017 - Digitalization in Management, Society and Economy* 46: 195–99.

Pebesma, Edzer. 2018. „Simple Features for R: Standardized Support for Spatial Vector Data". *The R Journal* 10(1): 439–46. https://doi.org/10.32614/RJ-2018-009.

Pedregosa, F i in. 2011. „Scikit-learn: Machine Learning in {P}ython". *Journal of Machine Learning Research* 12: 2825–30.

Pedregosa, Fabian. 2015. „Feature extraction and supervised learning on fMRI : from practice to theory". Université Pierre et Marie Curie. https://tel.archives-ouvertes.fr/tel-01100921.

———. 2018. „About me". http://fa.bianp.net/pages/about.html (dostęp 4.12.2018).

Pedregosa, Fabian, Francis Bach, i Alexandre Gramfort. 2014. „On the Consistency of Ordinal Regression Methods". *Journal of Machine Learning Research* 18.

Pedregosa, Fabian, Rémi Leblond, i Simon Lacoste-Julien. 2017. „Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization". *Advances in neural information processing systems*.

Peng, R. D. 2011a. „Reproducible Research in Computational Science". *Science* 334(6060): 1226–27.

———. 2011b. „Reproducible Research in Computational Science". *Science* 334(6060): 1226–27. http://www.sciencemag.org/cgi/doi/10.1126/science.1213847.

Peng, Roger D. 2016a. *Exploratory Data Analysis with R*. Leanpub. http://leanpub.com/exdata.

———. 2016b. *Report Writing for Data Science in R*. Leanpub. https://leanpub.com/reportwriting.

Pentland, Alex. 2009. „Reality Mining of Mobile Communications: Toward a New Deal on Data". W *The Global Information Technology Report, World Economic Forum*, red. INSEAD. 75–80. http://hd.media.mit.edu/wef_globalit.pdf.

Pentland, Alex, i Tracy Heibeck. 2008. *Honest Signals. How They Shape Our World*. Cambridge: The MIT Press.

Perez, Fernando, i Brian E Granger. 2015. „Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science". 1–24. http://archive.ipython.org/JupyterGrantNarrative-2015.pdf.

Pérez, Fernando, i Brian E Granger. 2007. „{IP}ython: a System for Interactive Scientific Computing". *Computing in Science and Engineering* 9(3): 21–29. https://ipython.org.

Peruzzi, Antonio, Fabiana Zollo, Walter Quattrociocchi, i Antonio Scala. 2018. „How News May Affect Markets' Complex Structure: The Case of Cambridge Analytica". *Entropy* 20(10): 765. http://www.mdpi.com/1099-4300/20/10/765.

Peters, Gjalt-Jorn Ygram. 2019. „rock: Reproducible Open Coding Kit". https://cran.r-project.org/package=rock.

Peters, Isabella. 2009. *Folksonomies*. Berlin, New York: Walter de Gruyter.

Peterson, Brian G, i Peter Carl. 2020. „PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis". https://cran.r-project.org/package=PerformanceAnalytics.

Petrov, Dmitry. 2017. „How A Data Scientist Can Improve His Productivity - Data Version Control". *[wpis na blogu 15.05.2017]*. https://blog.dataversioncontrol.com/how-a-data-scientist-can-improve-his-productivity-730425ba4aa0 (dostęp 9.08.2019).

Phandroid. 2019. „Google IO 2019 Duplex for the web". *YouTube*. https://www.youtube.com/watch?v=2KlEBw02J5s (dostęp 15.05.2019).

Piatetsky, Gregory. 2017a. „Optimism about AI improving society is high, but drops with experience developing AI systems". *KDnuggets*. https://www.kdnuggets.com/2017/07/optimism-ai-impact-experience.html (dostęp 8.10.2018).

———. 2017b. „Python overtakes R, becomes the leader in Data Science, Machine Learning platforms". *08.2017 KdNuggets*. https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html (dostęp 3.08.2018).

———. 2018a. „Data Scientist – best job in America, 3 years in a row". https://www.kdnuggets.com/2018/01/glassdoor-data-scientist-best-job-america-3years.html (dostęp 25.02.2018).

———. 2018b. „Here are the most popular Python IDEs / Editors”. *KDnuggets*. https://www.kdnuggets.com/2018/12/most-popular-python-ide-editor.html (dostęp 17.07.2019).

———. 2018c. „Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis”. *KDnuggets*. https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html/2 (dostęp 25.11.2018).

———. 2018d. „The 6 components of Open-Source Data Science/ Machine Learning Ecosystem; Did Python declare victory over R?” *KDnuggets*. https://www.kdnuggets.com/2018/06/ecosystem-data-science-python-victory.html (dostęp 25.03.2019).

Piątkowska, Krystyna. 2017. „Kwartet Anscombe’a, Datasaurus - czyli po co w ogóle rysować?” *28.06.2017*. https://www.statystyczny.pl/kwartet-anscombe-datasaurus/ (dostęp 5.09.2018).

Pietrowiak, Kamil. 2014. „Etnografia oparta na współpracy. Założenia, możliwości, ograniczenia”. *Przegląd Socjologii Jakościowej* X(4): 18-.

Pink, Sarah, Minna Ruckenstein, Robert Willim, i Melisa Duque. 2018. „Broken data: Conceptualising data in an emerging world”. *Big Data & Society* 5(1). http://journals.sagepub.com/doi/10.1177/2053951717753228.

Piotrowski, Andrzej. 1998. *Ład interakcji. Studia z socjologii interpretatywnej*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Polanyi, Michael. 2005. *Personal Knowledge. Towards a Post-Critical Philosophy*. London: Routledge.

Politechnika Gdańska. 2019. „Inżynieria danych - data science”. https://ftims.pg.edu.pl/studia-podyplomowe-ftims/inzynieria-danych-data-science (dostęp 6.01.2019).

Polska Agencja Prasowa. 2019. „Centrum Badań nad Sztuczną Inteligencją ruszyło na UŁ”. *Nauka w Polsce*. http://naukawpolsce.pap.pl/aktualnosci/news%2C32456%2Ccentrum-badan-nad-sztuczna-inteligencja-ruszylo-na-ul.html (dostęp 1.02.2019).

Polskie Towarzystwo Badaczy Rynku i Opinii. 2017. *Data driven decisions. Przewodnik po źródłach wiedzy w marketingu*. http://datadrivendecisions.pl/ddd_guide.pdf.

Pontificia Accademia per la Vita. 2020a. „renAIssance - Rome Call for AI Ethics”. http://www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html (dostęp 10.03.2020).

———. 2020b. *Rome Call for AI Ethics*. Watykan. http://www.academyforlife.va/content/dam/pav/documenti pdf/2020/CALL 28 febbraio/AI Rome Call x firma_DEF_DEF_.pdf.

Portmess, Lisa, i Sara Tower. 2015. „Data barns, ambient intelligence and cloud computing: the tacit epistemology and linguistic representation of Big Data". *Ethics and Information Technology* 17(1): 1–9. http://link.springer.com/10.1007/s10676-014-9357-2.

Powell, Gavin, i Carol McCullough-Dieter. 2005. *Oracle SQL: Jumpstart with Examples*. Amsterdam: Digital Press.

Power, Daniel J. 2002. „What is the true story about using data mining to identify a relation between sales of beer and diapers?" *DSS News* 3(23). http://www.dssresources.com/newsletters/66.php (dostęp 8.07.2019).

Powers, David M W. 2007. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Adelaide. http://david.wardpowers.info/BM/Evaluation_SIETR.pdf.

Prendki, Jennifer. 2018. „Before you launch your machine learning model, start with an MVP". *VentureBeat*. https://venturebeat.com/2018/11/24/before-you-launch-your-machine-learning-model-start-with-an-mvp/ (dostęp 30.03.2019).

Press, Gil. 2018. „The Brute Force Of IBM Deep Blue And Google DeepMind". *Forbes*. https://www.forbes.com/sites/gilpress/2018/02/07/the-brute-force-of-deep-blue-and-deep-learning/#741dad3249e3 (dostęp 10.07.2019).

Princeton Alumni Weekly. 2012. „John D. Hunter '90". https://paw.princeton.edu/memorial/john-d-hunter-'90 (dostęp 4.10.2018).

Prokulski, Łukasz. 2017. „Ankieta – wyniki (i jak je podsumować w R)" *[wpis na blogu]*. https://blog.prokulski.science/index.php/2017/10/17/ankieta-wyniki/ (dostęp 14.11.2017).

———. 2019. „Analiza ofert pracy. Ile jest pracy dla ludzi zajmujących się analizą danych i machine learningiem?" *[wpis na blogu]*. https://blog.prokulski.science/index.php/2019/02/18/analiza-ofert-pracy/#more-2093 (dostęp 12.06.2019).

Provost, Foster, i Tom Fawcett. 2013. „Data Science and its Relationship to Big Data and Data-Driven Decision Making". *Big Data* 1(1): 51–59. http://www.mendeley.com/research/data-science-relationship-big-data-datadriven-decision-making-2.

Przanowski, Karol. 2014. *Credit Scoring w erze Big Data*. Warszawa: Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie.

Przegalińska, Aleksandra. 2013. „Fenomenologia istot wirtualnych". Uniwersytet Warszawski. https://depotuw.ceon.pl/bitstream/handle/item/463/Fenomenologia istot wirtualnych-1.pdf.

———. 2014. „Nawroty bylejakości". *dwutygodnik.com* 10(144). http://www.dwutygodnik.com/artykul/5472-nawroty-bylejakosci.html.

———. 2015. „Jeśli nie neoluddyzm, to co?” *Res Publica Nova* 3(221).

———. 2016a. „Ciało i umysł - wstęp kuratorski”. W *Interfejsy, kody, symbole. Przyszłość komunikowania*, red. Ewa Drygalska. Wrocław: Miasto Przyszłości / Laboratorium Wrocław, 11–19.

———. 2016b. *Istoty wirtualne. Jak fenomenologia zmieniała sztuczną inteligencję*. Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych UNIVERSITAS.

———. 2016c. „Konflikt generacyjny wśród botów: między ELIZĄ, Cleverbotem a Tay”. W *Live'Bot*, red. Violka Kuś. Bydgoszcz: E-naukowiec, 10–14. http://e-naukowiec.eu/live-b0t-monografia-pod-redakcja-_violki-kus/.

Przegalińska, Aleksandra, Leon Ciechanowski, Mikołaj Magnuski, i Peter Gloor. 2018. „Muse Headband: Measuring Tool or a Collaborative Gadget?” W *Collaborative Innovation Networks: Building Adaptive and Resilient Organizations*, red. Francesca Grippa i in. Cham: Springer International Publishing, 93–101. https://doi.org/10.1007/978-3-319-74295-3_8.

Przybyłowska, Ilona. 1978. „Wywiad swobodny ze standaryzowaną listą poszukiwanych informacji i możliwości jego zastosowania w badaniach socjologicznych”. *Przegląd Socjologiczny* 30: 53–63.

Puschmann, Cornelius, i Jean Burgess. 2014. „Metaphors of big data”. *International Journal of Comunication* 8: 1690–1709.

PwC. 2015. „What's next for the 2017 data science and analytics job market?”

PyData. 2018. „PyData Warsaw 2018”. https://pydata.org/warsaw2018/ (dostęp 14.08.2018).

PyLadies. 2019. „About: PyLadies”. https://www.pyladies.com/about/ (dostęp 12.02.2019).

Pyszczuk, Artur. 2011. *Programowanie w Bashu, czyli jak pisać skrypty w Linuksie*. https://www.arturpyszczuk.pl/files/bash/bash.pdf.

Python Software Foundation. 2019. „Search results · PyPI”. https://pypi.org/search/?c=Intended+Audience+%3A%3A+Science%2FResearch&o=-created&q=&page=1 (dostęp 16.07.2019).

R-Ladies. 2019. „About us – R-Ladies Global”. https://rladies.org/about-us/ (dostęp 12.02.2019).

R Foundation. 2019. „CRAN - Contributed Packages”. https://cran.r-project.org/web/packages/ (dostęp 16.07.2019).

Radomski, Andrzej, i Radosław Bomba. 2013. *Zwrot cyfrowy w humanistyce. Internet, nowe media, kultura 2.0*. http://www.sbc.org.pl/dlibra/doccontent?id=77488.

Rajpurkar, Pranav i in. 2018. „Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists" red. Aziz Sheikh. *PLOS Medicine* 15(11): e1002686. http://dx.plos.org/10.1371/journal.pmed.1002686.

Raschka, Sebastian. 2018. *Python. Uczenie maszynowe*. Gliwice: Helion.

Rath, Matthias. 2018. „Data Science – die neue Leitwissenschaft ?" W *Weltkulturatlas. Kultur in Zeit der Globalisierung. Daten, Geschichten, Grafiken*, red. Thomas Knubben, Erich Schols, i Uli Braun. Stuttgart: av Edition, 21–37. https://www.researchgate.net/publication/329786919_Data_Science_-_die_neue_Leitwissenschaft/references.

Ratner, Alexander i in. 2017. „Snorkel: Rapid Training Data Creation with Weak Supervision". http://arxiv.org/abs/1711.10160.

Raymond, Eric S. 2001. „The cathedral and the bazaar". W *In The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*, red. Eric S Raymond. Sebastopol: O'Reilly, 16–64.

Realbotix. 2018. „Harmony AI Sex Robot". https://realbotix.com/ (dostęp 22.01.2018).

Reitz, Kenneth. 2018. *Python Guide Documentation Release 0.0.1*. https://buildmedia.readthedocs.org/media/pdf/python-guide/latest/python-guide.pdf.

RENOIR. 2018. „RENOIR Project". http://renoirproject.eu/ (dostęp 21.02.2018).

Reshama, Shaikh. 2018. „Why Women Are Flourishing In R Community But Lagging In Python". *github.io*. https://reshamas.github.io/why-women-are-flourishing-in-r-community-but-lagging-in-python/ (dostęp 12.12.2018).

Rexer Analytics. 2018. „Data Science Survey". http://www.rexeranalytics.com/data-science-survey.html (dostęp 1.02.2018).

Ribeiro, Marco Tulio, Sameer Singh, i Carlos Guestrin. 2016. „«Why Should I Trust You?» Explaining the Predictions of Any Classifier". W *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, 1135–44. http://aclweb.org/anthology/N16-3020.

Richardson, Rashida, Jason Schultz, i Kate Crawford. 2019. „Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice". *New York University Law Review Online* February: 192–233. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423.

Rinker, Tyler W. 2020. „{qdap}: {Q}uantitative Discourse Analysis Package". http://github.com/trinker/qdap.

Robertson, Adi. 2019. „The 5 biggest announcements from Facebook's F8 developer conference keynote". *The Verge*.

https://www.theverge.com/2019/4/30/18524068/facebook-f8-2019-keynote-highlights-summary-news-feed-messenger-instagram-oculus (dostęp 20.05.2019).

Rodak, Olga. 2017. „Twitter jako przedmiot badań socjologicznych i źródło danych społecznych: perspektywa konstruktywistyczna". *Studia Socjologiczne* 3(226): 209–36.

Rogers, Richard. 2018. „Aestheticizing Google critique: A 20-year retrospective". *Big Data & Society* 5(1). http://journals.sagepub.com/doi/10.1177/2053951718768626.

Rogozhnikov, Alex. 2016. „Jupyter (IPython) notebooks features". *[wpis na blogu]*. https://arogozhnikov.github.io/2016/09/10/jupyter-features.html (dostęp 22.07.2019).

Ross, Casey, i Ike Swetlitz. 2018. „IBM's Watson recommended «unsafe and incorrect» cancer treatments, internal documents show". *2018.07.25 STAT*. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (dostęp 2.01.2019).

van Rossum, Guido. 1997. „Comparing Python to Other Languages". https://www.python.org/doc/essays/comparisons/ (dostęp 16.07.2019).

Rothenberger, Lea, Benjamin Fabian, i Elmar Arunov. 2019. „Relevance of Ethical Guidelines for Artificial Intelligence - a Survey and Evaluation". *European Conference on Information Systems* (May): 0–11.

RStudio. 2018a. „About - RStudio". https://www.rstudio.com/about/ (dostęp 26.07.2018).

———. 2018b. „RStudio - RStudio Products". https://www.rstudio.com/products/rstudio/ (dostęp 27.11.2018).

RStudio Team. 2016. „RStudio: Integrated Development Environment for R". http://www.rstudio.com/.

———. 2017. „Cheat Sheet: RStudio IDE". https://www.rstudio.org/links/ide_cheat_sheet.

Rudzki, Piotr J, Przemysław Biecek, i Michał Kaza. 2017. „Comprehensive graphical presentation of data from incurred sample reanalysis". *Bioanalysis* 9(12): 947–56. http://www.future-science.com/doi/10.4155/bio-2017-0038 (dostęp 5.12.2018).

Rumelhart, David E., Geoffrey E. Hinton, i Ronald J. Williams. 1986. „Learning representations by back-propagating errors". *Nature* 323(6088): 533–36. http://www.nature.com/doifinder/10.1038/323533a0.

Russel, Stuart J, i Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach (3rd ed.)*. New Jersey: Prentice Hall. http://web.cecs.pdx.edu/~mperkows/CLASS_479/2017_ZZ_00/02__GOOD_Russel=Norvig=Artificial Intelligence A Modern Approach (3rd Edition).pdf.

Rutenberg, Jim. 2013. „Data You Can Believe In". *The New York Times*. https://www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html (dostęp 16.05.2018).

Rybicki, Jan, i Maciej Eder. 2011. „Deeper Delta across genres and languages: do we really need the most frequent words?" *Literary and Linguistic Computing* 26(3): 315–21. http://dx.doi.org/10.1093/llc/fqr031.

Ryciak, Norbert. 2019. „Danetyka, czyli o polskim tłumaczeniu Data Science". *Kodołamacz.pl - Zostań programistą w 7 tygodni!* https://www.kodolamacz.pl/blog/data-science-po-polsku/ (dostęp 22.01.2020).

Sack, Harald. 2018. „Joseph Weizenbaum and his famous Eliza". *SciHi Blog*. http://scihi.org/joseph-weizenbaum-eliza/ (dostęp 11.06.2019).

Sadowski, Jathan. 2019. „When data is capital: Datafication, accumulation, and extraction". *Big Data & Society* 6(1). https://journals.sagepub.com/doi/10.1177/2053951718820549

Sanchez-Ocana, Alejandro Suarez. 2013. *Tajemnice Google'a. Wielki Brat ery informatycznej*. Warszawa: Wydawnictwo Bellona.

Sato, Kaz. 2016. „How a Japanese cucumber farmer is using deep learning and TensorFlow". *Google Cloud Blog*. https://cloud.google.com/blog/products/gcp/how-a-japanese-cucumber-farmer-is-using-deep-learning-and-tensorflow (dostęp 14.06.2019).

Satyaseel, Harshit. 2018. „Most Popular Python Libraries for Data Science in 2018". *2018.09.30 Technotification*. https://www.technotification.com/2018/09/popular-python-libraries-data-science.html (dostęp 1.10.2018).

Sauer, Gerard. 2017. „A Murder Case Tests Alexa's Devotion to Your Privacy". *Wired*. https://www.wired.com/2017/02/murder-case-tests-alexas-devotion-privacy/ (dostęp 30.06.2019).

Savage, Mike, i Roger Burrows. 2007. „The Coming Crisis of Empirical Sociology". *Sociology* 41(5): 885–99. http://journals.sagepub.com/doi/10.1177/0038038507080443.

Saxena, A, M Sun, i Andrew Ng. 2009. „Make3D: Learning 3D Scene Structure from a Single Still Image". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(5): 824–40.

Schmidhuber, Jürgen. 2015. „Deep learning in neural networks: An overview". *Neural Networks* 61: 85–117. https://linkinghub.elsevier.com/retrieve/pii/S0893608014002135.

Schmidt, Eric, Jonathan Rosenberg, i Alan Eagle. 2014. *Jak działa Google*. Kraków: Insignis Media.

Schoenfeld, Alan H. 1992. „Learning to Think Mathematically: Problem Solving, Metacognition, And Sense Making in Mathematics". W *Handbook of Research on Mathematics Teaching and Learning*, red. D Grouws. New York: Macmillan Publishers Limited, 334–70.

Schulze, Elizabeth. 2019. „40% of A.I. start-ups in Europe have almost nothing to do with A.I., research finds". *CNBC*. https://www.cnbc.com/2019/03/06/40-percent-of-ai-start-ups-in-europe-not-related-to-ai-mmc-report.html (dostęp 2.04.2019).

Schutten, Gerrit-Jan i in. 2020. „readODS: Read and Write ODS Files". https://cran.r-project.org/package=readODS.

Schwartz, Barry. 2016. „How Google uses machine learning in its search algorithms". *Search Engine Land*. https://searchengineland.com/google-uses-machine-learning-search-algorithms-261158 (dostęp 9.07.2019).

Schwartz, Mattathias. 2017. „Facebook User Data Harvested by Cambridge Analytica". *The Intercept*. https://theintercept.com/2017/03/30/facebook-failed-to-protect-30-million-users-from-having-their-data-harvested-by-trump-campaign-affiliate/ (dostęp 1.02.2019).

scikit-learn. 2018. „About us — scikit-learn 0.20.1 documentation". https://scikit-learn.org/stable/about.html#citing-scikit-learn (dostęp 4.07.2018).

SciPy.org. „HomePage - SciPy.org". *2018*. https://scipy.org/index.html (dostęp 13.08.2018).

Scriptol.com. 2006. „List of Hello World Programs in 200 Programming Languages". https://www.scriptol.com/programming/hello-world.php (dostęp 15.07.2019).

Sekercioglu, Eser. 2018. „Eser Sekercioglu's answer to For data science, which will die first: R or Python? - Quora". *27.08.2018 Quora*. https://www.quora.com/For-data-science-which-will-die-first-R-or-Python/answer/Eser-Sekercioglu (dostęp 8.09.2018).

Sharif, Bonita, i Jonathan I. Maletic. 2010. „An Eye Tracking Study on camelCase and under_score Identifier Styles". W *2010 IEEE 18th International Conference on Program Comprehension*, IEEE, 196–205. http://ieeexplore.ieee.org/document/5521745/.

Sharp Sight. 2016. *Data Science Crash Course*. http://www.sharpsightlabs.com/.

———. 2018. „R vs Python ... which to learn for data science". *13.08.2018*. https://www.sharpsightlabs.com/blog/r-vs-python/ (dostęp 14.08.2018).

Shaw, Zed A. 2014. *Learn Python the Hard Way: a very simple introduction to the terrifyingly beautiful world of computers and code*. Third. Donnelley: Addison-Wesley.

Shead, Sam. 2018. „«NIPS» AI Conference Changes Name, Kind Of". *Forbes*. https://www.forbes.com/sites/samshead/2018/11/21/nips-ai-conference-changes-name-kind-of/#8020bc73bc15 (dostęp 10.02.2019).

Shiab, Nael. 2015. „On the Ethics of Web Scraping and Data Journalism". *Global Investigative Journalism Network*. https://gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/ (dostęp 5.02.2018).

Shibutani, Tamotsu. 1955. „Reference Groups as Perspectives". *American Journal of Sociology* 60(6): 562–69. https://doi.org/10.1086/221630.

Silaparasetty, Vinita. 2018. „Python vs R for Machine Learning". W *International Conference on Security*, Bangalore: Garden City University, 1–19. https://www.academia.edu/36911486/PYTHON_VS_R_FOR_MACHINE_LEARNING.

Silge, Julia, i David Robinson. 2018. *Text Mining with R: A Tidy Approach*. O'Reilly. https://www.tidytextmining.com/.

Silver, David i in. 2016. „Mastering the game of Go with deep neural networks and tree search". *Nature* 529(7587).

———. 2017. „Mastering the game of Go without human knowledge". *Nature* 550(7676): 354–59. https://doi.org/10.1038/nature24270.

———. 2018. „A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". *Science* 362(6419): 1140–44. http://www.sciencemag.org/lookup/doi/10.1126/science.aar6404.

Silver, Nate. 2014. *Sygnał i szum*. Gliwice: Helion.

Silverman, David. 2010. *Prowadzenie badań jakościowych*. Warszawa: Wydawnictwo Naukowe PWN.

Simonite, Tom. 2018. „AI Researchers Fight Over Four Letters: NIPS". *Wired*. https://www.wired.com/story/ai-researchers-fight-over-four-letters-nips/ (dostęp 12.02.2019).

Singleton, Keith. 2018. „MeetUp API (article) - DataCamp". *DataCamp*. https://www.datacamp.com/community/tutorials/meetup-api-data-json (dostęp 2.06.2019).

Singularity University. 2018. „Peter Norvig - Mentor - Singularity University". https://su.org/mentors/peter-norvig/ (dostęp 1.08.2018).

Skeet, Jon. 2010. „Writing the perfect question". *[wpis na blogu 29.08.2010]*. https://codeblog.jonskeet.uk/2010/08/29/writing-the-perfect-question/ (dostęp 9.09.2019).

Skirpan, Michael, i Micha Gorelick. 2017. „The Authority of «Fair» in Machine Learning". http://arxiv.org/abs/1706.09976.

Skura, Małgorzata, i Michał Lisicki. 2018. *Gen liczby: jak dzieci uczą się matematyki?* Warszawa: Wydawnictwo Mamania.

Słomczyński, Krzysztof. 2017. „GitHub - WhyR2017/meetup-harvesting". *GitHub*. https://github.com/WhyR2017/meetup-harvesting (dostęp 2.08.2017).

Slowikowski, Kamil. 2019. „ggrepel: Automatically Position Non-Overlapping Text Labels with «ggplot2»". https://cran.r-project.org/package=ggrepel.

Smart, Francis. 2013. „Export R Results Tables to Excel – Please don't kick me out of your club". *R-bloggers*. https://www.r-bloggers.com/2013/08/export-r-results-tables-to-excel-please-dont-kick-me-out-of-your-club/ (dostęp 28.10.2017).

Smith, Gavin JD. 2018. „Data doxa: The affective consequences of data practices". *Big Data & Society* 5(1). http://journals.sagepub.com/doi/10.1177/2053951717751551.

Socha, Beata. 2018. „Are You Human?" *Warsaw Business Journal*: 48–51.

Solon, Olivia. 2018. „Cambridge Analytica whistleblower says Bannon wanted to suppress voters". *The Guardian*. https://www.theguardian.com/uk-news/2018/may/16/steve-bannon-cambridge-analytica-whistleblower-suppress-voters-testimony (dostęp 5.02.2019).

Somers, James. 2017. „Is AI Riding a One-Trick Pony?" *MIT Technology Review*. https://www.technologyreview.com/s/608911/is-ai-riding-a-one-trick-pony/ (dostęp 23.11.2018).

———. 2018. „The Scientific Paper Is Obsolete. Here's What's Next". *The Atlantic*. https://www.theatlantic.com/science/archive/2018/04/the-scientific-paper-is-obsolete/556676/ (dostęp 12.08.2018).

Sommerville, Ian i in. 1993. „Integrating ethnography into the requirements engineering process". W *[1993] Proceedings of the IEEE International Symposium on Requirements Engineering*, IEEE Comput. Soc. Press, 165–73. http://ieeexplore.ieee.org/document/324821/.

Sopyła, Krzystof. 2019. „Dlaczego porzuciłem Tensorflow na rzecz Pytorch". *About Data*. https://ksopyla.com/machine-learning/pytorch-vs-tensorflow/ (dostęp 19.03.2019).

Sopyła, Krzysztof. 2017. „Podsumowanie ankiety Data Science Polska 2017". *23.09.2017 About Data*. https://ksopyla.com/data-science/podsumowanie-ankiety-data-science-polska-2017/ (dostęp 10.10.2017).

Sorensen, Chris. 2017. „How U of T's «godfather» of deep learning is reimagining AI". *University of Toronto*. https://www.utoronto.ca/news/how-u-t-s-godfather-deep-learning-reimagining-ai (dostęp 23.11.2018).

Sotrender. 2018. „Sotrender. No-bullshit analytics. Analyze and optimize your marketing over social media." https://www.sotrender.com/pl/team/ (dostęp 1.12.2018).

Soubra, Diya. 2012. *The three Vs that define big data*. http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data.

Spector, Alfred, Peter Norvig, i Slav Petrov. 2012. „Google's hybrid approach to research". *Communications of the ACM* 55(7): 34. http://dl.acm.org/citation.cfm?doid=2209249.2209262.

Spectre, Aleksandr i in. 2015. „On wealth and the diversity of friendships: High social class people around the world have fewer international friends". *Personality and Individual Differences* 87: 224–29. http://dx.doi.org/10.1016/j.paid.2015.07.040.

Spinellis, Diomidis. 2005. „Version control systems". *IEEE Software* 22(5): 108–9.

Springer. 2018. „Journal of computational social science".

de St. Germain, James H. 2008. „Problem Solving". *School of Computing University of Utah - Jim's Computer Science Topics Area*. https://www.cs.utah.edu/~germain/PPS/Topics/problem_solving.html (dostęp 27.01.2019).

Stallman, Richard. 2016. „FLOSS i FOSS". *gnu.org*. https://www.gnu.org/philosophy/floss-and-foss.html (dostęp 12.07.2019).

Stanford University. 2018. „Sebastian Thrun – Personal". http://robots.stanford.edu/personal.html (dostęp 23.04.2018).

Staniak, Mateusz, i Przemyslaw Biecek. 2019. „The Landscape of R Packages for Automated Exploratory Data Analysis". http://arxiv.org/abs/1904.02101.

Staniak, Mateusz, i Przemysław Biecek. 2018. „Explanations of model predictions with live and breakDown packages". http://arxiv.org/abs/1804.01955.

Star, Susan Leigh, i James R. Griesemer. 1989. „Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39". *Social Studies of Science* 19(3): 387–420. http://journals.sagepub.com/doi/10.1177/030631289019003001.

StatCounter. 2019. „Desktop Operating System Market Share Worldwide". https://gs.statcounter.com/os-market-share/desktop/ (dostęp 18.09.2019).

Statt, Nick. 2014. „Zuckerberg: «Move fast and break things» isn't how Facebook operates anymore". *CNET*. https://www.cnet.com/news/zuckerberg-move-fast-and-break-things-isnt-how-we-operate-anymore/ (dostęp 14.01.2018).

Stavens, David, Gabriel Hoffmann, i Sebastian Thrun. 2007. „Online speed adaptation using supervised learning for high-speed, off-road autonomous driving". W *IJCAI International Joint Conference on Artificial Intelligence*, 2218–24.

Stevens, Marthe, Rik Wehrens, i Antoinette de Bont. 2018. „Conceptualizations of Big Data and their epistemological claims in healthcare: A discourse analysis". *Big Data & Society* 5(2). http://journals.sagepub.com/doi/10.1177/2053951718816727.

Stiegler, Bernard. 2010. „Manifest 2010". *Ars Industrialis*. http://arsindustrialis.org/manifest2010pl (dostęp 5.04.2018).

———. 2017. *Wstrząsy. Głupota i wiedza w XXI wieku*. Warszawa: Wydawnictwo Naukowe PWN.

Strauss, Anselm L. 1978. „A Social World Perspective". *Studies in Symbolic Interaction* 1: 119–28.

———. 1982. „Social Worlds and Legitimation Processes". *Studies in Symbolic Interaction* 4: 171–90.

———. 1984. „Social Worlds and Their Segmentation Processes". W *Studies in Symbolic Interaction*, red. Norman Denzin. Greenwitch CT: JAI Press, 123–39.

———. 1993. *Continual Permutations of Action*. New York: Aldine de Gruyter.

Strauss, Anselm L, i Juliet Corbin. 1990. *Basics of Qualitative Research. Grounded Theory Procedures and Techniques*. Newbury Park−London−New Delhi: Sage.

Striphas, Ted. 2015. „Algorithmic culture". *European Journal of Cultural Studies* 18(4–5): 395–412. http://journals.sagepub.com/doi/10.1177/1367549415577392.

Strong, Alexus. 2018. „Why I'm Learning Python in 2018". *12.01.2018 Codeacademy*.

Sullivan, Danny. 2016. „RIP Google PageRank score: A retrospective on how it ruined the web". *Search Engine Land*. https://searchengineland.com/rip-google-pagerank-retrospective-244286 (dostęp 11.02.2019).

Sumbul, Michel. 2014. „Big Data problematic". *What's Big Data? [wpis na blogu]*. https://whatsbigdata.be/category/big-data-overview/ (dostęp 30.01.2017).

Surma, Jerzy. 2017. *Cyfryzacja życia w erze big data. Człowiek, biznes, państwo*. Warszawa: Wydawnictwo Naukowe PWN.

Surowiecki, James. 2017. „Chill: Robots Won't Take All Our Jobs". *Wired*. https://www.wired.com/2017/08/robots-will-not-take-your-job/ (dostęp 14.02.2018).

switchup. 2018. „2018 Best Data Science Bootcamps". https://www.switchup.org/rankings/best-data-science-bootcamps (dostęp 21.02.2018).

Symeonidis, Iraklis, Pagona Tsormpatzoudi, i Bart Preneel. 2015. „Collateral damage of Facebook Apps: an enhanced privacy scoring model". *IACR Cryptology ePrint Archive*: 1–18. http://dblp.uni-trier.de/db/journals/iacr/iacr2015.html#SymeonidsBTP15.

System Wspomagania Analiz i Decyzji GUS. 2019. „Miasta największe pod względem liczby ludności". http://swaid.stat.gov.pl/Dashboards/Miasta największe pod względem liczby ludności.aspx (dostęp 11.06.2019).

Szacki, Jerzy. 2002. *Historia myśli socjologicznej*. Warszawa: Wydawnictwo Naukowe PWN.

Szahaj, Andrzej. 2004. *Zniewalająca moc kultury. Artykuł i szkice z filozofii kultury, poznania i polityki*. Toruń: Wydawnictwo UMK.

Szczurek, Ewa i in. 2011. „Deregulation upon DNA damage revealed by joint analysis of context-specific perturbation data". *BMC Bioinformatics* 12(1): 249.

http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-249 (dostęp 5.12.2018).

Szeliga, Marcin. 2017. *Data science i uczenie maszynowe*. Warszawa: Wydawnictwo Naukowe PWN.

Szkoła Główna Handlowa. 2018. „Metody ilościowe w ekonomii i systemy informacyjne". http://oferta.sgh.waw.pl/pl/studialicencjackie/kierunki-pl/misi/Strony/default.aspx (dostęp 11.10.2018).

Szpunar, Magdalena. 2005. „Antropomorfizm wcielony - komputer w roli osoby". W *Komputer - przyjaciel czy wróg?*, red. Agnieszka Szewczyk. Szczecin: Uniwersytet Szczeciński, 114–21.

———. 2006. „Technofobia versus technofilia - technologia i jej miejsce we współczesnym świecie". *Problemy społeczne w grze politycznej*: 370–84.

———. 2008. „Internet a kultura daru. Fenomen ruchu open source". W *Idee i Myśliciele. Filozoficzne i kulturoznawcze rozważania o duchowości i komunikowaniu*, red. Ignacy S Fiut. Kraków: Wydawnictwa AGH, 99–112. http://magdalenaszpunar.com/_publikacje/2008/opensource.pdf.

———. 2012. *Nowe-Stare Medium.* Warszawa: Wydawnictwo IFiS PAN.

———. 2013. „Wokół koncepcji gatekeepingu. Od gatekeepingu tradycyjnego do technologicznego". W *Idee i Myśliciele: medialne i społeczne aspekty filozofii*, red. Ignacy S Fiut. Kraków: Wydawnictwa AGH, 56–61.

———. 2015. „Sieć ukryta a sieć widzialna. O zasobach WWW nieindeksowanych przez wyszukiwarki". *Przegląd Kulturoznawczy* 1(1): 44–55.

———. 2016a. „Humanistyka cyfrowa a socjologia cyfrowa. Nowy paradygmat badań naukowych". *Zarządzanie w kulturze* 17(4): 355–69.

———. 2016b. *Kultura cyfrowego narcyzmu*. Kraków: Wydawnictwa AGH.

———. 2018a. „Kultura algorytmów". *Zarządzanie w Kulturze* 19(1): 1–10. http://www.ejournals.eu/Zarzadzanie-w-Kulturze/2018/19-1-2018/art/11556/.

———. 2018b. „Kultura lęku (nie tylko) technologicznego". *Kultura Współczesna* 2(101): 114–23.

Szreder, Mirosław. 2015a. „Big data wyzwaniem dla człowieka i statystyki". *Wiadomości Statystyczne* 8(651): 1–11.

———. 2015b. „Czy statystyka pozwala lepiej zrozumieć świat?" *Polityka*. https://www.polityka.pl/tygodnikpolityka/nauka/1618675,1,czy-statystyka-pozwala-lepiej-zrozumiec-swiat.read (dostęp 10.02.2018).

———. 2016. „Złudzenie Big Data”. *Tygodnik Powszechny* 10(3478): 50–51. https://www.tygodnikpowszechny.pl/zludzenie-big-data-32574 (dostęp 10.02.2018).

———. 2018a. „O algorytmach Big Data (na marginesie książki Cathy O’Neil pt. Broń matematycznej zagłady. Jak algorytmy zwiększają nierówności i zagrażają demokracji, Wydawnictwo Naukowe PWN, 2017)”. *Wiadomości Statystyczne* 2(681): 78–81.

———. 2018b. „Populacja małych frakcji”. *Polityka*. https://www.polityka.pl/tygodnikpolityka/nauka/1734106,1,wazki-1-proc.read (dostęp 10.02.2018).

Sztompka, Piotr. 2007. *Socjologia. Analiza społeczneństwa*. Kraków: Wydawnictwo Znak.

Szumakowicz, Eugeniusz. 2000. „Sztuczna inteligencja - problem czy pseudoproblem?” W *Granice sztucznej inteligencji. Eseje i studia*, red. Eugeniusz Szumakowicz. Kraków: Wydawnictwo Politechniki Krakowskiej, 11–42.

Szupiluk, Ryszard. 2013. *Dekompozycje wielowymiarowe w agregacji predykcyjnych modeli data mining*. Warszawa: Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie.

Szymielewicz, Katarzyna. 2017. „Algorytm zwycięstwa”. *Fundacja Panoptykon*. https://panoptykon.org/wiadomosc/algorytm-zwyciestwa (dostęp 9.04.2018).

Szymielewicz, Katarzyna, i Anna Obem. 2020. „Sztuczna inteligencja non-fiction”. *Fundacja Panoptykon*. https://panoptykon.org/sztuczna-inteligencja-non-fiction (dostęp 25.07.2020).

Szynglarewicz, Bartlomiej i in. 2017. „Epithelial-mesenchymal transition inducer Snail1 and invasive potential of intraductal breast cancer”. *Journal of Surgical Oncology* 116(6): 696–705. http://doi.wiley.com/10.1002/jso.24708 (dostęp 5.12.2018).

Taddeo, Mariarosaria, i Luciano Floridi. 2018a. „How AI can be a force for good”: *Science* 361(6404): 751–52.

———. 2018b. „Regulate artificial intelligence to avert cyber arms race comment”. *Nature* 556(7701): 296–98.

Tadeusiewicz, Ryszard. 1989. *W stronę uśmiechniętych maszyn: spacer pograniczem biologii i techniki*. Warszawa: Wydawnictwa „Alfa”.

———. 1998. *Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami*. Warszawa: Akademicka Oficyna Wydawnicza.

———. 2007. *Odkrywanie właściwości sieci neuronowych przy użyciu programów w języku C#*. Kraków: Polska Akademia Umiejętności.

———. 2015a. *Poznaj domowe urządzenia*. Kraków: Dział Informacji i Promocji Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie.

———. 2015b. *Poznaj prawa przyrody*. Kraków: Dział Informacji i Promocji Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie.

———. 2018. „Dossier w pigułce - Strona Prof. Tadeusiewicza". https://www.uci.agh.edu.pl/uczelnia/tad/dossier.php (dostęp 5.12.2018).

Tadeusiewicz, Ryszard, i Piotr Chrząstowski. 1994. *Informatyka? Ależ to bardzo proste!* Warszawa: Wydawnictwa Naukowo-Techniczne. http://otworzksiazke.pl/ksiazka/informatyka/.

Tadeusiewicz, Ryszard, i Mariusz Flasiński. 1991. *Rozpoznawanie obrazów*. Warszawa: Państwowe Wydawnictwo Naukowe. http://otworzksiazke.pl/ksiazka/rozpoznawanie_obrazow/.

Tadeusiewicz, Ryszard, i Zbigniew Mikrut. 1978. „Identyfikacja i modelowanie neuronu na maszynie cyfrowej". *Archiwum Automatyki i Telemechaniki* 3(23): 345–56.

Tarkowski, Alek, Bartosz Paszcza, i Natalia Mileszyk. 2020. „«AlgoPolska»: 11 postulatów, których realizacja pozwoli uniknąć mrocznych wizji rodem z «Black Mirror»". *Klub Jagielloński i Fundacja Centrum Cyfrowe*. https://klubjagiellonski.pl/2019/12/09/algopolska-11-tez-ktorych-realizacja-pozwoli-uniknac-nam-mrocznych-wizji-rodem-z-black-mirror/ (dostęp 5.01.2020).

Tatman, Rachael. 2019. „Six steps to more professional data science code". *Kaggle*. https://www.kaggle.com/rtatman/six-steps-to-more-professional-data-science-code (dostęp 29.09.2019).

Taylor, David. 2016. „Battle of the Data Science Venn Diagrams". https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html4. (dostęp 14.12.2017).

Taylor, Linnet, i Lina Dencik. 2020. „Constructing Commercial Data Ethics". *Technology and Regulation*. https://techreg.org/index.php/techreg/article/view/35.

TechAmerica. 2012. *Demystifying big data: A practical guide to transforming the business of Government*. http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf.

Teddlie, Charles, i Abbas Tashakkori. 2009. *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Los Angeles: SAGE Publications.

The Neural Information Processing Systems Foundation Board of Trustees. 2018. „From the Board: Changing our Acronym". https://nips.cc/Conferences/2018/News (dostęp 12.02.2019).

The pandas project. 2018. „pandas: Python Data Analysis Library". http://pandas.pydata.org/about.html (dostęp 4.12.2018).

Thieme, Nick. 2018. „R generation”. *Significance* 15(4): 14–19. http://doi.wiley.com/10.1111/j.1740-9713.2018.01169.x.

Thomas, Suzanne L, Dawn Nafus, i Jamie Sherman. 2018. „Algorithms as fetish: Faith and possibility in algorithmic work”. *Big Data & Society* 5(1): 1–11. http://journals.sagepub.com/doi/10.1177/2053951717751552.

Tocci, Jason. 2009. Publicly Accessable Penn Dissertations „Geek cultures: Media and identity in the digital age”. University of Pennsylvania. http://repository.upenn.edu/dissertations/AAI3395723/.

Tomanek, Krzysztof, i Grzegorz Bryda. 2015. „Odkrywanie postaw dydaktyków zawartych w komentarzach studenckich. Analiza treści z zastosowaniem słownika klasyfikacyjnego”. *Przegląd Socjologiczny* 64(4): 51–81. http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-1272e7e6-acfb-4795-95ea-7f775015936d.

———. 2017. „Metodyka dla analizy treści w projektach stosujących techniki text mining i rozwiązania CAQDAS piątej generacji”. *Przegląd Socjologii Jakościowej* 13(2): 128–43.

Tompa, Frank W. 1996. „An Overview of Waterloo’s Database Software for the OED”. *Computing in the Humanities Working Papers)* B.13. http://projects.chass.utoronto.ca/chwp/tompa/.

Törnberg, Petter, i Anton Törnberg. 2018. „The limits of computation: A philosophical critique of contemporary Big Data research”. *Big Data & Society* 5(2): 1–12. http://journals.sagepub.com/doi/10.1177/2053951718811843.

Torres, J C. 2018. „The future of smartphone cameras is AI”. *SlashGear*. https://www.slashgear.com/the-future-of-smartphone-cameras-is-ai-15530789/ (dostęp 10.07.2019).

Troszczyński, Marek, i Aleksander Wawer. 2017. „Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych”. *Przegląd Socjologii Jakościowej* 13(2): 62–80.

Trzpiot, Grażyna. 2017. „Rozumienie Data Science”. W *Statystyka a Data Science*, Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, 6–30.

Tufekci, Zeynep. 2014. „Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls”. W *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 1–10. http://arxiv.org/abs/1403.7400.

———. 2015. „Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency”. *Telecomm & High Tech* (203): 203–18.

———. 2018. „@zeynep: Facebook’s defense that Cambridge Analytica harvesting...” *17.03.2018 Twitter*. https://twitter.com/zeynep/status/975067185889427456 (dostęp 22.03.2018).

Tufte, Edward R. 1983. *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.

———. 1990. *Envisioning Information*. Cheshire: Graphics Press.

Turilli, Matteo, i Luciano Floridi. 2009. „The ethics of information transparency". *Ethics and Information Technology* 11(2): 105–12.

Turing, Alan M. 1950. „Computing Machinery and Intelligence". *Mind* LIX(236): 433–60. https://academic.oup.com/mind/article-lookup/doi/10.1093/mind/LIX.236.433.

Turner, Andrew. 2019. „Using Data from Git and GitHub in Ethnographies of Software Development". W , 35–49. http://link.springer.com/10.1007/978-3-030-17152-0_4.

Turner, Anna, Marcin W Zieliński, i Kazimierz M. Słomczyński. 2018. „Google big data: charakterystyka i zastosowanie w naukach społecznych". *Studia Socjologiczne* 4(231): 49–71.

Turner, Jonathan H. 2004. *Struktura teorii socjologicznej*. Warszawa: Wydawnictwo Naukowe PWN.

Udacity. 2018. „Udacity - Our Story". https://www.udacity.com/us (dostęp 23.04.2018).

Umemoto, Daigo, i Nobuyasu Ito. 2018. „Power-law distribution in an urban traffic flow simulation". *Journal of Computational Social Science* 1(2): 493–500. http://link.springer.com/10.1007/s42001-018-0028-7.

UNECE. 2014. *How big is Big Data? Exploring the role of Big Data in Official Statistics*. http://www1.unece.org/stat/platform/display/bigdata/How+big+is+Big+Data.

Uniwersytet Warszawski. 2019. „Quantitative psychology and economics". https://www.wne.uw.edu.pl/en/candidates/doctoral-program-quantitative-psychology-and-economics/ (dostęp 15.07.2019).

Unruh, David R. 1980. „The Nature of Social Worlds". *The Pacific Sociological Review* 23(3): 271–96. http://journals.sagepub.com/doi/10.2307/1388823.

UPIAN. 2015. *Do Not Track*. France, Canada. https://donottrack-doc.com/.

Uri, Therese. 2015. „The Strengths and Limitations fo Using Situational Analysis Grounded Theory as Research Methodology". *Journal of Ethnographic & Qualitative Research* 10(1): 135–51.

Vaidhyanathan, Siva. 2018. *Antisocial media. Jak Facebook oddala nas od siebie i zagraża demokracji.* Warszawa: Grupa Wydawnicza Foksal.

Vanderplas, Jake. 2016. „Conda: Myths and Misconceptions". *[wpis na blogu Pythonic Perambulations]*. https://jakevdp.github.io/blog/2016/08/25/conda-myths-and-misconceptions/ (dostęp 22.07.2019).

Vasconcelos, Ana. 2007. „The use of Grounded Theory and of Arenas/Social Worlds Theory in Discourse Studies: A case study on the discursive adaptation of information systems". *Electronic Journal of Business Research Methods* 2(5): 125–36.

Vazquez, Favio. 2017. „Data version control with DVC. What do the authors have to say? Interview with Dmitry Petrov". *Towards Data Science*. https://towardsdatascience.com/data-version-control-with-dvc-what-do-the-authors-have-to-say-3c3b10f27ee (dostęp 3.08.2018).

Vedder, Anton. 1999. „KDD: The challenge to individualism". *Ethics and Information Technology* 1: 275–81.

Viaene, Stijn. 2013. „Data scientists aren't domain experts". *IT Professional* 15(6): 12–17.

Vicknair, Chad i in. 2010. „A comparison of a graph database and a relational database - A Data Provenance Perspective". *Proceedings of the 48th Annual Southeast Regional Conference on ACM SE 10*: 1. http://portal.acm.org/citation.cfm?doid=1900008.1900067.

Vinsel, Lee. 2017. „You Have Been Waiting Your Whole Life for Me: Cathy O'Neil Misrepresents Reality". https://medium.com/@sts_news/you-have-been-waiting-your-whole-life-for-me-cathy-oneil-misrepresents-reality-706d5b0b3fbb (dostęp 11.02.2019).

Vopson, Melvin M. 2020. „The information catastrophe". *AIP Advances* 10(8): 085014. http://aip.scitation.org/doi/10.1063/5.0019941.

Votta, Fabio. 2018. „@favstats: Our @hadleywickham, who art at RStudio, Hallowed be thy name. Thy functions run, thy unit tests..." *2018.09.23 Twitter*. https://twitter.com/favstats/status/1043836510880116738 (dostęp 23.01.2019).

Wachter, Sandra, Brent Mittelstadt, i Luciano Floridi. 2017. „Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". *International Data Privacy Law* 7(2): 76–99. https://academic.oup.com/idpl/article-lookup/doi/10.1093/idpl/ipx005.

Wagner, Ben. 2018. „Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?" W *Being Profiling: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen*, red. Emre Bayamlioglu, Irina Baraliuc, Liisa Albertha Wilhelmina Janssens, i Mirelle Hildebrandt. Amsterdam: Amsterdam University Press, 84–90.

Wainer, Howard. 1984. „How to Display Data Badly". *The American Statistician* 38(2): 137–47. https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1984.10483186.

Walne Zgromadzenie Delegatów Polskiego Towarzystwa Socjologicznego. 2012. „Kodeks etyki socjologa". http://pts.org.pl/wp-content/uploads/2016/04/kodeks.pdf.

Wang, Tricia. 2013. „Why Big Data Needs Thick Data". *13.03.2013 Ethnography Matters*. https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7 (dostęp 7.05.2018).

———. 2016. „The human insights missing from big data”. *[prezentacja TEDxCambridge]*. https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data#t-9755 (dostęp 7.05.2018).

Wang, Yilun, i Michal Kosinski. 2018. „Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.” *Journal of Personality and Social Psychology* 114(2): 246–57. http://doi.apa.org/getdoi.cfm?doi=10.1037/pspa0000098.

Warsaw.AI. 2020. „Przemysław Biecek”. https://warsaw.ai/speaker/przemyslaw-biecek/ (dostęp 5.08.2020).

Wąsowski, Michał. 2018a. „Facebook zablokował dostęp do platformy twórcom aplikacji i wyłączył ważną funkcję reklamową”. *29.03.2018 Business Insider Polska*.

———. 2018b. „Mark Zuckerberg pokazał «prawdziwą twarz» na przesłuchaniach w Kongresie USA”. *12.04.2018 Bussines Insider Polska*. https://businessinsider.com.pl/firmy/strategie/jak-poradzil-sobie-mark-zuckerberg-na-przesluchaniu-senatu-usa/b1yjyfg (dostęp 9.05.2018).

Watson, Samuel. 2018. „Samuel S. Watson’s answer to Who will eventually win the Python versus R debate in data science? - Quora”. *12.08.2018 Quora*. https://www.quora.com/Who-will-eventually-win-the-Python-versus-R-debate-in-data-science/answer/Samuel-S-Watson-1 (dostęp 8.09.2018).

Watson, Sara M. 2015. „Data is the New «__»”. *dismagazine*. http://dismagazine.com/discussion/73298/sara-m-watson-metaphors-of-big-data/ (dostęp 23.04.2018).

WDI18. 2018. „Warszawskie Dni Informatyki 2018: Data Science by Data Science Warsaw”. https://warszawskiedniinformatyki.pl/ (dostęp 11.05.2018).

Weber, Max. 1989. *Polityka jako zawód i powołanie*. Kraków: Wydawnictwo Znak.

Wellman, Barry. 2001. „Computer Networks As Social Networks”. *Science* 293: 2031–34. http://science.sciencemag.org/content/293/5537/2031/tab-pdf.

White, John. 2012. „Julia, I Love You”. *[wpis na blogu]*. http://www.johnmyleswhite.com/notebook/2012/03/31/julia-i-love-you/ (dostęp 15.07.2019).

White, Tom. 2012. *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O’Reilly.

———. 2015. *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O’Reilly.

Whitehorn, Mark. 2006. „The parable of the beer and diapers”. *The Register*. https://www.theregister.co.uk/2006/08/15/beer_diapers/ (dostęp 8.07.2019).

Why R? Foundation. 2018. „Why R?” http://whyr.pl/ (dostęp 11.05.2018).

Wickham, Hadley. 2010. „A Layered Grammar of Graphics”. *Journal of Computational and Graphical Statistics* 19(1): 3–28. http://www.tandfonline.com/doi/abs/10.1198/jcgs.2009.07098.

———. 2014a. *Advanced R*. New York: Chapman and Hall/CRC Press. http://adv-r.had.co.nz/Introduction.html.

———. 2014b. „Data science: how is it different to statistics ? « IMS Bulletin”. *2014.08.04 IMS Bulletin Online*. http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics / (dostęp 3.09.2018).

———. 2014c. „Tidy Data”. *Journal of Statistical Software* 59(10): 1–23. http://www.jstatsoft.org/v59/i10/.

———. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

———. 2017. „tidyverse: Easily Install and Load the «Tidyverse»”. https://cran.r-project.org/package=tidyverse.

———. 2018a. „forcats: Tools for Working with Categorical Variables (Factors)”. https://cran.r-project.org/package=forcats.

———. 2018b. „Hadley Wickham - About”. http://hadley.nz/ (dostęp 16.01.2018).

———. 2018c. „profr: An Alternative Display for Profiling Information”. https://cran.r-project.org/package=profr.

———. 2018d. „scales: Scale Functions for Visualization”. https://cran.r-project.org/package=scales.

———. 2018e. „stringr: Simple, Consistent Wrappers for Common String Operations”. https://cran.r-project.org/package=stringr.

———. 2018f. „You can’t do data science in a GUI”. *Association for Computing Machinery*. https://www.youtube.com/watch?v=cpbtcsGE0OA (dostęp 27.11.2018).

———. 2019a. „Letter To A Young Maintainer - talk at Monktoberfest 2019”. *RedMonk Tech Events*. https://www.youtube.com/watch?v=1K7u5hkciLI (dostęp 30.04.2020).

———. 2019b. „rvest: Easily Harvest (Scrape) Web Pages”. https://cran.r-project.org/package=rvest.

———. 2019. „Welcome to the Tidyverse”. *Journal of Open Source Software* 4(43): 1686.

Wickham, Hadley, Romain François, Lionel Henry, i Kirill Müller. 2018. „dplyr: A Grammar of Data Manipulation”. https://cran.r-project.org/package=dplyr.

Wickham, Hadley, i Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. First. O’Reilly. http://r4ds.had.co.nz/.

———. 2018. *Język R. Kompletny zestaw narzędzi dla analityków danych.* Gliwice: Helion.

Wickham, Hadley, i Lionel Henry. 2018. „tidyr: Easily Tidy Data with «spread()» and «gather()» Functions". https://cran.r-project.org/package=tidyr.

Wickham, Hadley, Jim Hester, i Romain François. 2017. „readr: Read Rectangular Text Data". https://cran.r-project.org/package=readr.

Więcko, Konrad, Krzysztof Słomczyński, i Przemysław Biecek. 2016. „Data analysis of pracuj.pl". http://52.31.27.158/ml/ (dostęp 21.02.2018).

Wiggers, Kylie. 2019. „How to keep Alexa, Cortana, Siri, Google Assistant, and Bixby from recording you". *VentureBeat*. https://venturebeat.com/2019/04/16/how-to-prevent-alexa-cortana-siri-google-assistant-and-bixby-from-recording-you/ (dostęp 10.07.2019).

Wikipedia. „Data scientist". https://pl.wikipedia.org/wiki/Data_scientist (dostęp 13.02.2018).

Wilke, Claus O. 2019. „cowplot: Streamlined Plot Theme and Plot Annotations for «ggplot2»". https://cran.r-project.org/package=cowplot.

Williams, Alex. 2017. „Will Robots Take Our Children's Jobs?" *The New York Times*. https://www.nytimes.com/2017/12/11/style/robots-jobs-children.html (dostęp 14.02.2018).

Williams, L., R.R. Kessler, W. Cunningham, i R. Jeffries. 2000. „Strengthening the case for pair programming". *IEEE Software* 17(4): 19–25. http://ieeexplore.ieee.org/document/854064/.

Williams, Randi i in. 2018. „My doll says it's ok: a study of children's conformity to a talking doll". *Proceedings of the 17th ACM Conference on Interaction Design and Children*: 625–31.

Williams, Wendy M., i Stephen J. Ceci. 2015. „National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track". *Proceedings of the National Academy of Sciences* 112(17): 5360–65. http://www.pnas.org/lookup/doi/10.1073/pnas.1418878112.

Wills, Josh. 2012. „@josh_wills: Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician." *2012.05.03 Twitter*. https://twitter.com/josh_wills/status/198093512149958656 (dostęp 7.12.2017).

WiMLDS. 2018. „Home - Women in Machine Learning & Data Science". http://wimlds.org/ (dostęp 26.09.2018).

Winner, Langdon. 1980. „Do Artifacts Have Politics?" *Daedalus* 109(1): 121–36. http://www.jstor.org/stable/20024652.

Witkowska, Dorota. 2002. *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe*. Warszawa: Wydawnictwo C.H.Beck.

Wolff, Mark. 2007. „Reading Potential: The Oulipo and the Meaning of Algorithms". *Digital Humanities Quarterly* 1(1).

Wollowski, Michael i in. 2016. „A Survey of Current Practice and Teaching of AI". *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*: 4119–24.

Wolpert, David H., i William G. Macready. 1997. „No free lunch theorems for optimization". *IEEE Transactions on Evolutionary Computation* 1(1): 67–82.

Wong, Gillian, i Josh Chin. 2016. „China's New Tool for Social Control: A Credit Rating for Everything". *The Wall Street Journal*. https://www.wsj.com/articles/chinas-new-tool-for-social-control-a-credit-rating-for-everything-1480351590 (dostęp 21.01.2019).

Wong, Julia Carrie. 2018. „Mark Zuckerberg apologises for Facebook's «mistakes» over Cambridge Analytica". *The Guardian*. https://www.theguardian.com/technology/2018/mar/21/mark-zuckerberg-response-facebook-cambridge-analytica (dostęp 5.02.2019).

Woodgate, Rob. 2019. „What Is Slack, and Why Do People Love It?" *How-To Geek*. https://www.howtogeek.com/428046/what-is-slack-and-why-do-people-love-it/ (dostęp 4.09.2019).

Wray, Christina C. 2016. „Staying in the Know: Tools You Can Use to Keep Up With Your Subject Area". *Collection Management* 41(3): 182–86. http://www.tandfonline.com/doi/full/10.1080/01462679.2016.1196628.

Wrenn, Mary V. 2015. „Agency and neoliberalism". *Cambridge Journal of Economics* 39(5): 1231–43. https://academic.oup.com/cje/article-lookup/doi/10.1093/cje/beu047.

Wu, Jeff. 1997. „Statistics = Data Science?" www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf.

Wu, Yonghui i in. 2016. „Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". http://arxiv.org/abs/1609.08144.

Wydział Fizyki Technicznej i Matematyki Stosowanej PŁ. 2018a. „Kurs: na FTIMS I stopień". http://ftims.p.lodz.pl/course/view.php?id=11#MMADB (dostęp 6.04.2018).

———. 2018b. „Kurs: Specjalność Inżynieria Oprogramowania i Analiza Danych". http://ftims.p.lodz.pl/course/view.php?id=83 (dostęp 6.04.2019).

Wydział Informatyki Politechniki Białostockiej. 2019. „Data Science - Studia podyplomowe WI PB". http://podyplomowe.wi.pb.edu.pl/data-science/ (dostęp 6.01.2019).

Wydział Matematyki i Informatyki UŁ. 2016. „Analiza danych i data mining". http://analizadanych.math.uni.lodz.pl/ (dostęp 18.08.2016).

———. 2018. „Nowy kierunek – Analiza danych". http://www.math.uni.lodz.pl/analiza-danych/ (dostęp 6.06.2018).

Wydział Matematyki i Nauk Informacyjnych PW. 2019. „Inżynieria i Analiza Danych". https://ww2.mini.pw.edu.pl/studia/inzynierskie-i-licencjackie/inzynieria-i-analiza-danych/ (dostęp 5.01.2019).

Xie, Yihui. 2016. Bookdown: Authoring Books and Technical Documents with R Markdown *Bookdown: Authoring books and technical documents with R Markdown*.

Xie, Yihui, J J Allaire, i Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida, Florida: Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Amber Thomas, i Alison P Hill. 2019. *blogdown: Creating Websites with R Markdown*. https://bookdown.org/yihui/blogdown/ (dostęp 2.07.2019).

Yair, Gad. 2007. „Meritocracy". W *The Blackwell Encyclopedia of Sociology*, red. George Ritzer. Oxford, UK: John Wiley & Sons, Ltd. http://doi.wiley.com/10.1002/9781405165518.wbeosm082.

Yau, Nathan. 2009. „Rise of the Data Scientist". *FlowingData*. https://flowingdata.com/2009/06/04/rise-of-the-data-scientist/ (dostęp 12.12.2017).

Yegulalp, Serdar. 2017. „Facebook brings GPU-powered machine learning to Python | InfoWorld". *InfoWOrld*. https://www.infoworld.com/article/3159120/facebook-brings-gpu-powered-machine-learning-to-python.html (dostęp 17.07.2019).

Youtie, Jan, Alan L. Porter, i Ying Huang. 2016. „Early social science research about Big Data". *Science and Public Policy* 44(1): 1–10. https://academic.oup.com/spp/article-lookup/doi/10.1093/scipol/scw021.

Zagórna, Anna. 2020. „Unia wypuściła «Białą księgę sztucznej inteligencji»". *19.02.2020 Sztuczna Inteligencja*. https://www.sztucznainteligencja.org.pl/unia-wypuscila-biala-ksiege-sztucznej-inteligencji/ (dostęp 23.02.2020).

Zaharia, Matei i in. 2016. „Apache Spark: A Unified Engine for Big Data Processing". *Communications of the ACM* 59(11): 56–65. http://dl.acm.org/citation.cfm?doid=3013530.2934664.

Zaród, Marcin. 2018. „Aktorzy-sieci w kolektywach hakerskich w Polsce". Uniwersytet Warszawski. https://depotuw.ceon.pl/handle/item/2909.

Zawistowska, Alicja. 2013. „«Płeć matematyki». Zróżnicowania osiągnięć ze względu na płeć wśród uzdolnionych uczniów". *Studia Socjologiczne* 3(210): 75–95.

Zeileis, Achim i in. 2016. „fortunes: R Fortunes". https://cran.r-project.org/package=fortunes.

Zespół Fundacji Panoptykon. 2016. *Zabawki wielkiego brata*. Warszawa. https://panoptykon.org/sites/default/files/Panoptykon_ZabawkiWielkiegoBrata_przewodnik.pdf.

———. 2017. „Mikrosekunda w sieci”. *19.05.2017*. https://panoptykon.org/biblio/infografiki/mikrosekunda-w-sieci-wersja-pelna (dostęp 18.08.2018).

Zgiep, Łukasz. 2014. „Sharing economy jako ekonomia przyszłości”. *Myśl Ekonomiczna i Polityczna* 4(47): 193–205. https://mysl.lazarski.pl/fileadmin/user_upload/dokumenty/czasopisma/mysl-ekonomiczna-polityczna/2014/MEiP_4_9_2014_Zgiep.pdf.

Zhang, Amy X, Michael Muller, i Dakuo Wang. 2020. „How do Data Science Workers Collaborate? Roles, Workflows, and Tools”. http://arxiv.org/abs/2001.06684.

Zhao, Shuai i in. 2018. „Packaging and Sharing Machine Learning Models via the Acumos AI Open Platform”. W *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 841–46. https://ieeexplore.ieee.org/document/8614160/.

Zuboff, Shoshana. 2015. „Big other: Surveillance capitalism and the prospects of an information civilization”. *Journal of Information Technology* 30(1): 75–89.

———. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

Zunger, Yonatan. 2018. „Computer science faces an ethics crisis. The Cambridge Analytica scandal proves it”. *The Boston Globe*. https://www.bostonglobe.com/ideas/2018/03/22/computer-science-faces-ethics-crisis-the-cambridge-analytica-scandal-proves/IzaXxl2BsYBtwM4nxezgcP/story.html (dostęp 1.02.2019).

Zybertowicz, Andrzej i in. 2015. *Samobójstwo Oświecenia? Jak neuronauka i nowe technologie pustoszą ludzki świat*. Kraków: Wydawnictwo Kasper.

Żulicki, Remigiusz. 2016. „Big Data - nowa wiedza?” W *Postęp cywilizacyjny - stan obecny i perspektywy*, red. Monika Maciąg i Filip Polakowski. Lublin: Wydawnictwo Naukowe TYGIEL, 19–31. http://bc.wydawnictwo-tygiel.pl/public/assets/130/Postęp cywilizacyjny - stan obecny i perspektywy v.2.3.pdf.

———. 2017. „Potencjał Big Data w badaniach społecznych”. *Studia Socjologiczne* 3(226): 176–207. http://www.studiasocjologiczne.pl/pliki/Studia_Socjologiczne_2017_nr3_str175_207.pdf

———. 2019. „Pułapki myślowe data-driven. Krytyka (nie tylko) metodologiczna”. *Marketing i Rynek* 8(XXVI): 3–14.

Żulicki, Remigiusz, i Michał Żytomirski. 2020. „Próba wypracowania metodologii pomiaru baniek filtrujących w wyszukiwarce Google”. *Zarządzanie Mediami* 8(3): 243–57. http://www.ejournals.eu/ZM/2020/3-2020/art/16420/.