STUDIA SOCJOLOGICZNE 2017, 3 (226)
ISSN 0039−3371

Remigiusz Żulicki
Uniwersytet Łódzki

POTENCJAŁ BIG DATA W BADANIACH SPOŁECZNYCH

Problematykę podjęto z powodu epistemologicznych „obietnic", pojawiających się wśród entuzjastów Big Data. Przedyskutowano możliwości wykorzystania Big Data jako metody czy techniki badań społecznych. Krytycznej ocenie poddano wspomniane „obietnice" oraz popularne w środowisku Big Data hasła o śmierci ekspertów. Wnioski dotyczą szans i zagrożeń poznawczych, szczególnie w naukach społecznych. Uznano, że stosowanie Big Data może być narzędziem poznania świata w różnych dziedzinach życia. Niezbędne jest jednak podejście sceptyczne. Dla socjologów poznawanie samego zjawiska przedstawiono jako cenne dla rozumienia społeczeństwa informacyjnego. Wskazano także możliwy kierunek przyszłych badań Big Data.

Główne pojęcia: Big Data; epistemologia; metody badań społecznych; społeczeństwo informacyjne.

Wstęp

Big Data to zjawisko definiowane jako układ złożony z: danych o określonych własnościach, metod przechowywania ich i przetwarzania, technik zaawansowanej analizy danych oraz potrzebnego środowiska i sprzętu informatycznego (Przanowski 2014). Rosnąca ilość danych i coraz większe zainteresowanie ich wykorzystaniem skłania nas do zajęcia się tą problematyką. Entuzjaści Big Data sądzą, że zjawisko to doprowadza do przełomu cywilizacyjnego porównywalnego z wynalezieniem Internetu, maszyny parowej czy druku (Cukier i Mayer-Schönberger 2014; Minelli, Chambers i Dhiraj 2013). Przełom ten zachodzić ma przede wszystkim w metodach i możliwościach poznania świata. Za Big Data stoją cztery epistemologiczne „obietnice": jakoby pozwala ująć całość problemu i zapewnić pełne wsparcie decyzjom; do uzyskania wartościowych wyników nie są potrzebne przedmiotowe teorie, ani stawianie hipotez; ponieważ dane mówią same za siebie, nieobciążone zbędną teorią, to wyniki analiz są znaczące i zgodne z prawdą o świecie; wyniki analiz, niezależnie od przedmiotu, może interpretować każdy posiadający rozeznanie w statystyce (Kitchin 2014: 4). Ponieważ te „obietnice" wydają się bardzo kontrowersyjne, przyjrzymy się tytułowemu

Katedra Socjologii Kultury, e-mail: remigiusz.zulicki@uni.lodz.pl

zjawisku jako potencjalnej metodzie oraz technice badań społecznych, a pośrednio uczynimy je samo przedmiotem zainteresowania nauk społecznych.

Jak sugerują entuzjaści, Big Data to „rewolucja, która zmieni nasze myślenie, życie i pracę" (Cukier i Mayer-Schönberger 2014). Temat uznajemy za szczególnie interesujący dla socjologów, także dlatego, że mają oni być przez tę zmianę poważnie dotknięci. Pojawiają się sugestie o końcu zapotrzebowania na ekspertów dziedzinowych, w tym oczywiście socjologów. Wystarczać ma analiza ogromnych zbiorów danych, której wyniki mówią, co się dzieje, a nie dlaczego się dzieje. Postaramy się spojrzeć na tytułowe zjawisko krytycznie, zarówno demaskując pustkę jego „obietnic", jak i ukazując potencjał zastosowania w pracy badaczy społecznych.

Czym jest Big Data?

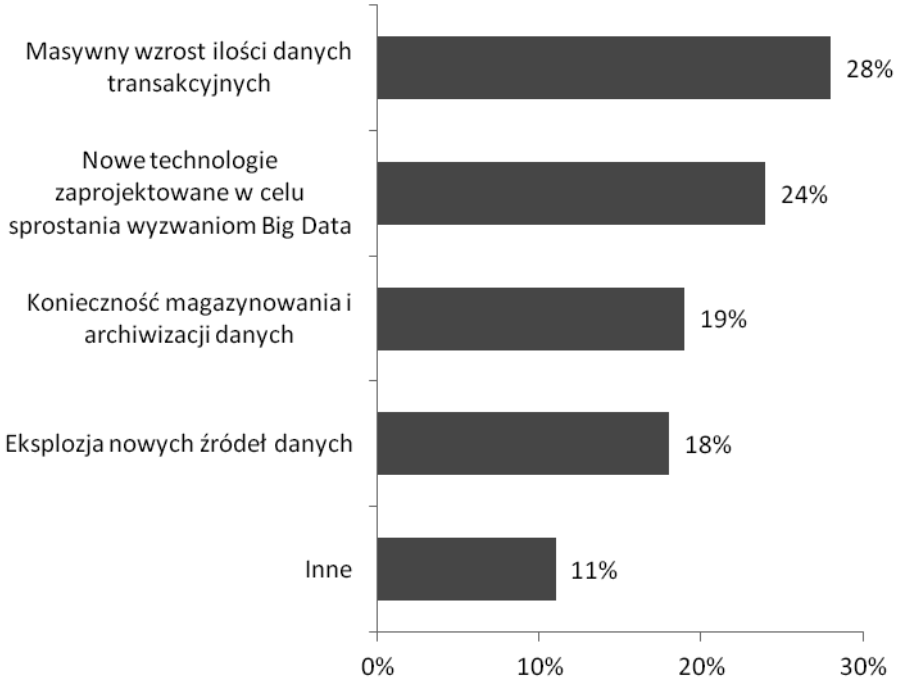
Termin i definicje

Termin „Big Data" prawdopodobnie powstał podczas rozmów w przerwie na lunch w firmie Silicon Graphics Inc. w połowie lat dziewięćdziesiątych ubiegłego wieku. Wydaje się, że w dużym stopniu przyczynił się do tego John R. Mashey, jeden z głównych pracujących tam informatyków. Pierwsze referencje akademickie w dziedzinie informatyki to zapewne praca S. M. Weissa i N. Indurkhya (*Predictive Data Mining. A practical guide* z 1998), w statystyce/ekonometrii zaś Francisa X. Diebolda (*Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting*, 2000 r.). W nocie serii Application Delivery Strategies z roku 2001 koncepcję Big Data znacząco rozwinął Douglas Laney z firmy doradczej Gartner. Powyższe fakty przytaczamy za jednym z „ojców chrzestnych" interesującego nas terminu (Diebold 2012). Francis Diebold wpadł na termin niejako przypadkiem, nie słysząc go wcześniej, a samo określenie uznał za „trafne, dźwięczne i intrygująco orwellowskie, szczególnie gdy zapisze się je z wielkich liter"[1] (tamże, s. 2).

Zjawisko Big Data definiowane (rozumiane?) jest różnorodnie. W przeprowadzonym w 2012 roku badaniu na grupie 154 osób – kierowników najwyższego szczebla międzynarodowych firm reprezentujących szeroką gamę branż – uzyskano pięć kategorii definicji Big Data, co przedstawiamy na rysunku 1.

[1] W przypadku przywołania prac anglojęzycznych cytaty podane są w tłumaczeniu autora artykułu.

Rysunek 1. Kategorie definiowania Big Data



Opracowanie własne na podstawie Gandomi i Haider 2015

Na rysunku 1. zaprezentowano kategorie, do jakich przyporządkowano swobodne wypowiedzi badanych kierowników. Częstość kategorii pokazano malejąco od góry do dołu. Brak jest wyraźnie wyróżniającej się kategorii odpowiedzi. Definicje dotyczyły różnych aspektów zjawiska. Koncentrowano się zarówno na tym, czym jest Big Data, jak i na tym, co i jak „robi" Big Data. Kategoria zawierająca 24% odpowiedzi w oryginale nazywa się „New technologies designed to adress the **3 Vs** challenges of Big Data" [wyróżnienie RŻ] (Gandomi i Haider 2015). To „3 Vs" jest powtarzającym się (Cukier i Mayer-Schönberger 2014; Laney 2001; Ohlhorst 2013; Soubra 2012; Berman 2013; Chen, Chiang i Storey 2012; Kwon, Lee i Shin 2014; TechAmerica 2012) sposobem definiowania Big Data przez wskazanie cech. Są to:

- Wielkość (*Volume*): rozmiar danych; bywa rozumiane jako zbiory większe niż 1 terabajt, a także takie, których nie da się przetwarzać za pomocą „tradycyjnych" narzędzi (Gandomi i Haider 2015).
- Różnorodność (*Variety*): zróżnicowanie, różne struktury, formaty i charakter danych (problem ten rozwijamy dalej).

- Szybkość (*Velocity*): szybkość pojawiania się nowych danych, ciągły ich napływ, co powoduje, że potrzebne są metody umożliwiające wydobywanie informacji z danych w czasie rzeczywistym (Gandomi i Haider 2015; Cukier i Mayer-Schönberger 2014).

Powyższe wymiary pozostają w zależności – zmiana jednego powoduje zmianę drugiego. Twórca definicji „3Vs", Laney zobrazował to jako trzy różne osie w trójwymiarowym układzie współrzędnych (2001, Figure 1). 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" (Dutcher 2014). Pojawiają się także propozycje kolejnych cech czy wymiarów Big Data, również określane jako „Vs". Są to (Gandomi i Haider 2015):

- 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 między innymi tym, że analizie poddawane są wszystkie dostępne dane tzn. cała populacja, nie zaś 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.

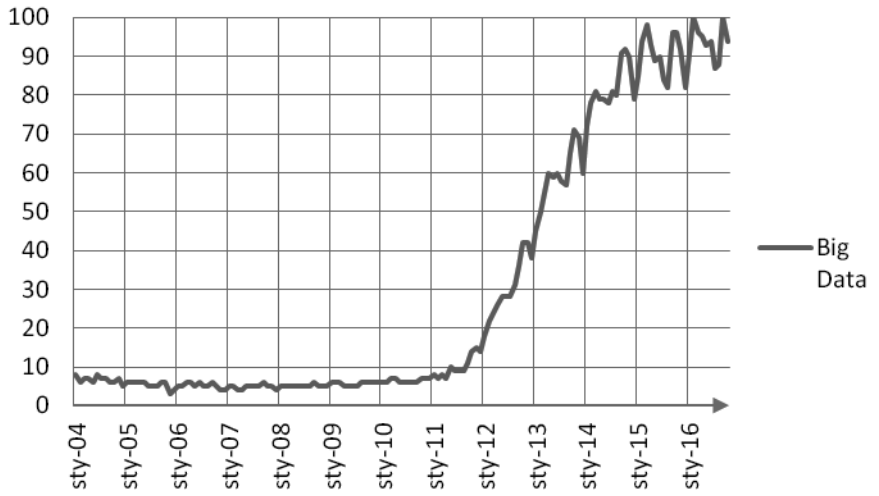
Sądzimy, że słuszna jest uwaga Karola Przanowskiego, który wskazuje, że powyżej przedstawione „Vs" dotyczą wyłącznie danych. Zdaniem tego autora na Big Data składają się jeszcze dwa obszary, więc „powinno się zatem definiować Big Data jako układ składający się z: danych opisanych własnościami 3Vs (5Vs), 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 związane z dużą liczbą klientów czy użytkowników" (Przanowski 2014: 13).

Big Data jako zjawisko społeczne

Niewątpliwie mamy do czynienia ze zjawiskiem społecznym – rosnącym zainteresowaniem Big Data, czymkolwiek „to" jest. Zmianę zainteresowania w poszukiwaniu informacji o Big Data na przestrzeni ostatniej dekady prezentujemy na rysunku 2.

[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).

Rysunek 2. Trend wyszukiwań zapytania „Big Data”



Opracowanie własne na podstawie Google Trends, https://www.google.pl/trends/explore?q=Big%20Data

Na rysunku 2. linia przedstawia względną częstotliwość wyszukiwania frazy „Big Data” (nie jest ważna wielkość liter) w wyszukiwarce Google na całym świecie. Można rozumieć to jako popularność wyszukiwania. Na osi X przedstawiono czas. Wykorzystano dane w interwale miesiąca, od stycznia 2004 do października 2016 roku, wyświetlono siatkę dla stycznia każdego roku. Skala od 1 do 100 na osi Y reprezentuje częstotliwość wyszukiwania, przy czym liczba wyszukań frazy jest dzielona przez liczbę wyszukań wszystkich fraz (czyli uwzględniono skalę korzystania z wyszukiwarki Google ogółem), a dodatkowo przedstawiona w liczbach względnych. Wartość 100 reprezentuje więc maksymalną zanotowaną dla danego okresu popularność – tu był to luty oraz wrzesień 2016 roku. Od początku 2012 roku do końca roku 2015 widać wyraźny trend wzrostowy. Popularność omawianej frazy wzrosła od 2004 do 2016 roku około dwanaście razy, a najwyższy wzrost nastąpił w okresie od maja 2011 do marca 2015 roku (około czternaście razy).

Powyższa analiza to także namiastka zastosowania Big Data. Skorzystaliśmy z dużej ilości danych behawioralnych, wygenerowanych niejako „przy okazji” przez użytkowników przeglądarki. Celem działania przeglądarki nie jest przecież prowadzenie badań społecznych, jednak korzystając z danych wprowadzanych i używanych w innym celu wnioskujemy o wzroście popularności pewnego zjawiska. A to wszystko za pomocą dostępnego online, bezpłatnego narzędzia Google Trends.

Big: dużo, ale nie tylko

Odnosimy wrażenie, że termin Big Data w dużym stopniu przykuwa uwagę dzięki słowu *big*. Samo *data* nie ma chyba zbyt wiele z intrygującej mocy tytułowego terminu, jest rzeczowe i na tym koniec. Tym samym, wzorem innych podejmujących temat autorów zaczniemy od opisów, jak wielkie są te „wielkie dane". Dane generują między innymi użytkownicy narzędzi dostarczanych przez technologicznych gigantów, jak Google i Facebook. Pierwsza z firm w 2012 roku 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 (Cukier i Mayer-Schönberger 2014). Ludzie korzystający z Internetu za pomocą różnego rodzaju serwisów, portali, aplikacji, na urządzeniach stacjonarnych i mobilnych są swoistymi generatorami danych. Pozostawiają cyfrowy ślad: sekwencje odwiedzanych stron internetowych, adresy IP, dane geolokalizacyjne GPS, dane pochodzące z sieci komórkowych, płatności elektronicznych i bezgotówkowych, wpisów na blogach, forach, portalach społecznościowych, wybory zakupowe w serwisach aukcyjnych i sklepach internetowych (Paharia 2014). Amerykańska firma doradcza IDC twierdzi, że „cyfrowy wszechświat" (*digital universe*) będzie rósł w tempie około 40% rocznie przez najbliższą dekadę. Oznacza to podwajanie ogólnego rozmiaru wszystkich światowych danych co dwa lata. W roku 2020 ilość cyfrowych bitów informacji ma niemal zrównać się z ilością gwiazd we wszechświecie fizycznym, osiągając rozmiar 44 zettabajtów (ZB^3), czyli 44 tryliony gigabajtów (GB). Poza podobnymi do wcześniej przywoływanych przykładami źródeł danych, jak np. social media, doradcy z IDC zwracają uwagę na tzw. Internet rzeczy (*Internet of Things*: IoT). Główną ideą IoT jest komunikacja pomiędzy przedmiotami bez udziału człowieka, co służyć ma monitorowaniu i zarządzaniu tymi przedmiotami. W roku 2014, z którego pochodzi raport IDC, liczba możliwych do komputeryzacji przedmiotów na całym świecie oszacowana została na 200 miliardów, z czego około 14 miliardów już było podłączonych do Internetu. Są to rozmaite przedmioty codziennego i „niecodziennego" użytku: zarówno samochody, jak i zabawki, samoloty odrzutowe, zmywarki czy obroże dla psów. Przedmioty te posiadać mają około 50 miliardów zbierających różnego rodzaje dane sensorów, przy czym w roku 2024 liczba ta ma wynosić około jednego tryliona (Turner 2014).

W powyższym akapicie oprócz epatowania wielkością chcieliśmy również pokazać, że Big Data to nie coś dotyczącego wąskiego grona przyrośniętych do komputerowych klawiatur informatyków. Właściwie każdy, kto choć raz w ostatniej dekadzie korzystał z Internetu na dowolnym urządzeniu, ma z Big

[3] Zetta oznacza 10^{21}, ale dla bajtów stosowany jest też mnożnik 1024, więc może oznaczać $1024^7 = 2^{70}$.

Data sporo wspólnego. Takie wrażenie, może nieco niepokojące, bo powodujące skojarzenia z byciem inwigilowanym, stara się naszym zdaniem wywrzeć wielu autorów piszących o Big Data (nie mówimy o tekstach technicznych). Zresztą orwellowskie skojarzenia miał już wspomniany powyżej „ojciec chrzestny" omawianego terminu Diebold. Czyżby zatem Big Data = Big Brother? Tego rodzaju sugestię (jak i subtelne odniesienie do innych wątków „dyskursu" o Big Data) sportretował Scott Adams w komiksie Dilbert, jednym z odcinków ukazującej się nieprzerwanie od 1989 roku serii satyrycznych historyjek z życia pewnej firmy i jej pracowników. Odcinek ten w całości prezentujemy na rysunku 3.

Rysunek 3. Big Data w komiksie Dilbert[4]

Źródło: http://dilbert.com/strip/2012-07-29

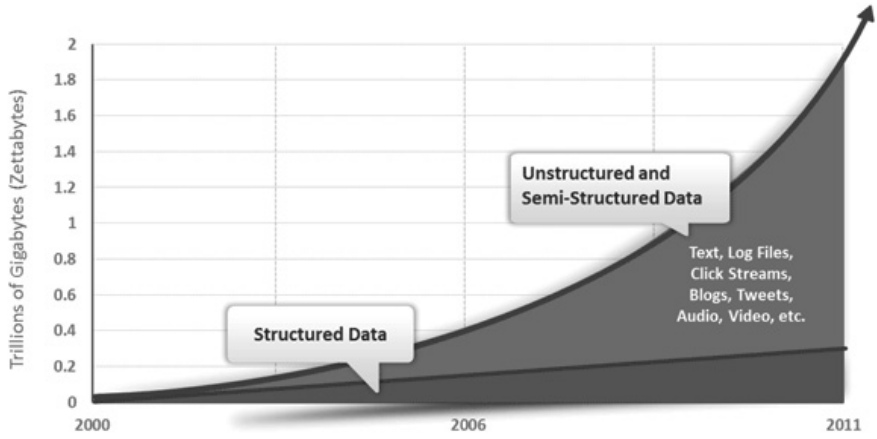
Pomijając na razie „dyskurs" o Big Data wracamy do kwestii technicznych. Co do samych danych, chodzi nie tylko o ich ilość czy wielkość, ale także charakter, rozumiany jako stopień ustrukturyzowania. Wpisy np. na portalu społecznościowym są przecież tekstami pisanymi swobodnie przez jego użytkowników, więc takie dane różnią się diametralnie od zapisów w bazie transakcji wykonanych kartą kredytową czy rejestrów pomiarów temperatury i ciśnienia w silniku samochodu F1, i to wcale nie dlatego, że dotyczą ludzi. Pierwsze z wymienionych rodzajów danych są nieustrukturyzowane, a dwa kolejne

[4] 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" (tłumaczenie własne).

– ustrukturyzowane. Charakter danych nieustrukturyzowanych mają więc te będące generowanym przez ludzi zapisem ich doświadczeń (jak teksty, fotografie, filmy). Ustrukturyzowane są dane pochodzące z zapisów zachowania ludzi czy innych obiektów (transakcje, kliknięcia, „polubienia", logowania itp.) oraz z czujników/sensorów określanych jako część IoT (UNECE 2014). O tym wymiarze klasyfikacji danych mówiliśmy przywołując definicje 3Vs – do charakteru danych odnosi się termin „różnorodność" (Variety).

Dla dodatkowej ilustracji na rysunku 4. prezentujemy przygotowaną przez autora jednego z blogów o Big Data wizualizację przyrostu wolumenu i charakteru danych. Na osi X przedstawiono czas, na osi Y rozmiar danych w zettabajtach, odcieniami szarości zaznaczono rodzaj danych (ustrukturyzowane/nieustrukturyzowane).

Rysunek 4. Przyrost ilości danych w podziale na ich charakter



Źródło: http://whatsbigdata.be/category/big-data-overview/

Choć rysunek 4. to właściwie wizualizacja intuicji autora bloga, jest ona co do zasady zgodna z innymi oszacowaniami (Cukier i Mayer-Schönberger 2014; Minelli i inni 2013). Ogólna ilość danych rośnie więc wykładniczo, ale dzieje się to głównie za sprawą coraz szybszego tempa przyrostu wolumenu danych nieustrukturyzowanych (ewentualnie „na wpół" ustrukturyzowanych). Ilość danych ustrukturyzowanych rośnie wolniej, raczej liniowo, a dane tego typu stanowią relatywnie coraz mniejszą część ogółu danych.

Na koniec zaznaczmy, co wyraźnie odróżnia dane w Big Data od danych pozyskanych w badaniach naukowych lub komercyjnych: są to zazwyczaj dane zbierane w celach innych niż analiza danych i zdobycie informacji/wiedzy. Oczywiście nie dotyczy to wszystkich rodzajów danych – przykładem

przeciwnym jest Internet rzeczy (IoT), którego jednym z założeń jest monitorowanie pracy urządzeń, czyli dane są zbierane w celach analitycznych. Jednak Big Data to raczej dane wykorzystywane powtórnie, zapisywane w celu wykonania jakiejś innej czynności bądź przy okazji jej wykonywania; mogą być chaotyczne, niekompletne i pozornie bezwartościowe. Tak zwane „dane resztkowe" (Cukier i Mayer- Schönberger 2014: 151) to np. wpisy w wyszukiwarce Google, z których już w tym artykule skorzystano dzięki Google Trends. Miliony użytkowników korzystają też nieświadomie z zastosowania danych resztkowych używając autouzupełniania i korekty w usłudze Gmail, Google Docs i Google Translate. Te wszystkie usługi zasilają modele statystyczne stworzone dzięki olbrzymiej ilości „googlowanych" nieprzerwanie, poprawnych i niepoprawnych zapytań. Zapytania te stanowią wciąż powiększającą się bazę danych najlepszego w historii systemu sprawdzania pisowni. Za każdym razem, gdy po popełnieniu literówki w wyszukiwanej frazie klikniemy podpowiadane „czy chodziło Ci o …?" trenujemy model. W odróżnieniu od danych zbieranych w badaniach – szczególnie społecznych – dane w Big Data to najczęściej dane niewymuszone, niedeklaratywne, behawioralne (będące zapisem zachowania) i w zasadzie – zastane, nie zaś wywołane.

Danetyzacja

Niewątpliwie danych jest obecnie bardzo, bardzo dużo. „Zwykłe pójście na spacer wygląda dziś zupełnie inaczej niż 15 lat temu. Niezależnie od tego, czy przemierzasz ulice wielkiego miasta, czy spacerujesz po lesie w wiejskiej okolicy, jeśli masz ze sobą przenośne urządzenie, towarzyszy Ci globalny tłum. […] Nowe informacje napływają tak szybko, że nieustannie próbujemy nadrobić zaległości. Im bardziej przyjmujemy taki styl życia, tym częściej dochodzę do wniosku, że mantra mojej przyjaciółki Marii naprawdę mogłaby stać się domyślnym powitaniem ery cyfrowej. Jak się masz? – Jestem zajęty, bardzo zajęty" (Powers 2014: 34–35). W cytowanym, bestsellerowym poradniku *Wyloguj się do życia* proponuje się różne strategie radzenia sobie z życiem w nadmiarze informacji wynikającym z ciągłego bycia online. Popularność książki może wskazywać na to, że ludzie żyjący w „połączonym świecie", wciąż rozpraszani i pobudzani przez „ekrany", jak William Powers nazywa wszystkie urządzenia elektroniczne z dostępem do Internetu (2014), czują potrzebę odpoczynku od strumienia informacji i szukają metod na radzenie sobie z nim. Problematykę podejmowali też inni autorzy, koncentrujący się między innymi na dysfunkcjach aparatu poznawczego czy wręcz uszkodzeniach mózgu, jak Manfred Spitzer w *Cyfrowej demencji*. Zwracamy jednak uwagę na to, że pojedynczy ludzie są nie tylko odbiorcami informacji, ale również ich twórcami. Nazwano ich swoistymi generatorami danych, a zjawisko polegające na zamianie różnych rodzajów ludzkiej aktywności w dane to danetyzacja (*datafication*) (Cukier i Mayer- Schönberger 2014).

Danetyzacja jest zatem „dążeniem do zawężania obszarów, które nie podlegają ewidencji" (Iwasiński 2016: 137). Łukasz Iwasiński przywołuje w tym kontekście słowa Lva Manovicha (2012: 335) który uważa, że efektem danetyzacji będzie zamiana świata w jedną wielką bazę danych. Danetyzacja nie musi jednak oznaczać kontrolowania bądź śledzenia kogokolwiek: ludzie sami chętnie generują dane o sobie, oraz na swoje potrzeby. Nie ma tu mowy o jakimkolwiek przymusie; być może co najwyżej niewielka świadomość użytkowników, co dzieje się z „ich" danymi. Istnieje swoisty ruch czy trend zwany *quantified self* (skwantyfikowany ja), polegający na dążeniu do nieustannego monitorowania parametrów swojego ciała, pracy czy właściwie dowolnych aktywności. Przykładowo polska grupa zarejestrowana na portalu społecznościowym pod nazwą „Quantified Self. Self knowledge through numbers" (QuantifiedSelfPoland) tak wyraziła swą misję: „U podstaw self-trackingu leży założenie o możliwości korygowania i usprawniania mechanizmów funkcjonowania ludzkiego organizmu i jego biologicznej, psychologicznej i społecznej aktywności". Chyba najpopularniejszą formą danetyzacji siebie są pomiary aktywności fizycznej: tętna, spalonych kalorii, intensywności treningu, przebytych kilometrów, prędkości biegu itp. Istnieją już urządzenia umożliwiające np. zapis i danetyzację fal mózgowych w trakcie snu (Cukier i Mayer-Schönberger 2014). Na dane zamieniane są zatem charakterystyki fizyczne, jak np. tętno w przypadku ludzi lub ciśnienie oleju w silniku samochodu wyścigowego pełnego sensorów, tzw. Internet rzeczy, ale nie tylko. Podobnie danetyzowane są charakterystyki społeczne i psychologiczne, jak postawy czy emocje, zamieniane na dane dzięki np. wpisom na portalach społecznościowych (o czym więcej przy okazji omawiania *text mining*).

Danetyzacja jest więc zamianą w dane tego wszystkiego, co nie jest zapisane w taki sposób. 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). Zdaniem badaczki komunikacji i nowych mediów José van Dijck, danetyzacja jest ideologią. Według niej wyznając danetyzację uznaje się, że dane są najważniejszym elementem dla zrozumienia rzeczywistości (Dijck 2014).

Big Data w praktyce

Amazon powstał w późnych latach dziewięćdziesiątych XX wieku jako jedna z pierwszych księgarni internetowych (Cukier i Mayer-Schönberger 2014). Od początku przyświecała mu idea zwiększania sprzedaży książek za pomocą polecania klientom nowych, interesujących pozycji. Owe polecenia były recenzjami, pisanymi przez zespół wysokiej klasy krytyków literackich i zamieszczanymi na Amazon.com. Pozytywne recenzje zespołu znacznie zwiększały sprzedaż konkretnych książek (Marcus 2004). Dążono jednak do spersonalizowania

rekomendacji i sięgnięto do zbieranych przez lata danych o transakcjach online. Początkowo badano wylosowaną, reprezentatywną próbę danych w celu wyłonienia kategorii klientów i dostosowania do nich rekomendacji (Cukier i Mayer-Schönberger 2014), szybko jednak okazało się, że rekomendacje te były bardzo powierzchowne i w efekcie wizyty w sklepie internetowym przypominały klientowi „zakupy w towarzystwie wiejskiego głupka" (Marcus 2004: 199). Dokonano więc dwóch poważnych zmian: zaczęto badać wszystkie zgromadzone dane, co obecnie zwane jest podejściem N = all, szukano zaś powiązań nie pomiędzy klientami, ale produktami. Nie stawiano hipotez, tylko badano relacje statystyczne pomiędzy wszystkimi książkami, niejako wbrew zdrowemu rozsądkowi i intuicji, nie przejmując się autorem, treścią czy gatunkiem literackim pozycji. Uzyskany algorytm działał w czasie rzeczywistym, wyświetlając klientowi rekomendacje na stronie internetowej sklepu.

Okazało się to metodą generującą sprzedaż wielokrotnie większą niż recenzje krytyków. Kalkulacja kosztów doprowadziła do decyzji o zwolnieniu wszystkich recenzentów i przejściu w całości na rekomendacje na podstawie badania danych. Podobno system ten generuje zakupy stanowiące ponad 30% zysków sprzedażowych Amazona, choć informacji firma nie potwierdziła oficjalnie (Cukier i Mayer-Schönberger 2014). Wielką zaletą systemu rekomendacji jest to, że nie ma żadnych przeszkód, by za jego pomocą rekomendować bardzo różnorodne produkty. Obecnie Amazon poza książkami sprzedaje filmy, sprzęt elektroniczny, sprzęty kuchenne, narzędzia do prac domowych i ogrodniczych, sprzęt sportowy, akcesoria i części samochodowe, ubrania, kosmetyki, jedzenie, alkohole, suplementy diety, biżuterię, zabawki i produkty dla dzieci, a także między innymi ręcznie wytwarzane meble, przestrzeń do magazynowania danych na serwerach, usługi hydraulika, elektryka, sprzątania, karty kredytowe, programy lojalnościowe czy różnego rodzaju przedmioty służące oprawie ślubów (amazon.com). Tego rodzaju systemy rekomendacji znalazły zastosowanie w bardzo wielu podobnych przedsiębiorstwach; np. w internetowej wypożyczalni filmów Netflix około 75% nowych zamówień generowanych jest dzięki podobnemu systemowi (Cukier i Mayer-Schönberger 2014).

Zwracamy uwagę, że systemy te wyłącznie identyfikują powiązania między produktami, nie wyjaśniając w żaden sposób przyczyn stojących za wyborami zakupowymi klientów. Jest to w sprzeczności ze stereotypowym przekonaniem o potrzebie poznania i zrozumienia potrzeb klienta jako podstawowych w skutecznej sprzedaży.

Google Flu Trends to rozwiązanie stworzone w 2009 roku, gdy rosła obawa przed epidemią grypy H1N1 (Cukier i Mayer-Schönberger 2014). Zadaniem było szacowanie rozprzestrzeniania się grypy w USA na podstawie wyszukiwań w Google. Zespół inżynierów Google pod kierunkiem Jeremy'ego Ginsburga stworzył algorytm statystyczny bazujący na korelacji pewnych zwrotów

wpisywanych w wyszukiwarkę z zachorowaniami na grypę na określonym obszarze (Ginsberg i inni 2009); wyrażenia pochodziły więc z olbrzymiej masy nieustrukturyzowanych danych generowanych przez internautów, a informacje o zachorowaniach to ustrukturyzowane, intencjonalnie zbierane dane różnych podmiotów świadczących usługi zdrowotne i rejestrujących pacjentów. Dane ustrukturyzowane zbierano na polecenie agencji rządowej ds. zapobiegania i zwalczania chorób zakaźnych. Zwracamy uwagę, że wyrażenia skorelowane nie zostały wybrane arbitralnie przez zdroworozsądkowe myślenie zespołu inżynierów, nie wybrali ich także eksperci ze świata nauk medycznych czy społecznych. Wyrażenia te, przy braku wstępnych założeń, zostały wyestymowane z danych na podstawie siły związku z liczbą zachorowań na określonym obszarze (Cukier i Mayer-Schönberger 2014). Badano dane z lat 2007–2008, z każdego dnia mając około trzech miliardów zapytań. Przetestowano następnie około 450 000 różnych modeli matematycznych i ostatecznie określono ten najlepiej dopasowany do danych historycznych. Uwzględniono 45 fraz wyszukiwania (Ginsberg i inni 2009). Narzędzie Google Flu Trends miało tę ogromną przewagę nad informacją agencji rządowej, że prognozowało rozprzestrzenianie się grypy praktycznie w czasie rzeczywistym, z opóźnieniem maksymalnie do jednego dnia, uwzględniając na bieżąco frazy „googlowane" w danej chwili przez miliony amerykańskich użytkowników. Informacje rządowe raportowały stan rzeczy z opóźnieniem około dwóch tygodni (Cukier i Mayer-Schönberger 2014).

Przykład ten był chyba pierwszym, dobrze znanym poza wąską branżą specjalistów zastosowaniem Big Data. Google Flu Trends, nazwane „the poster child[5] of big data" (Lazer i Kennedy 2015) jest jednak od niedawna dyżurnym przykładem słabych stron Big Data (Fung 2014, Lazer i inni 2014), ponieważ w roku 2013 prognoza rozmiaru szczytu sezonu grypy była różna od rzeczywistych danych o około 140% (Lazer i Kennedy 2015). Niemniej jednak inżynierów Google uznać wolno za swoistych pionierów, a ich projekt za przełomowy. W odpowiednim czasie i za pomocą wyjątkowych na owe czasy źródeł danych i podejścia do modelowania dostarczyli oni użytecznego narzędzia predykcji, które doceniły rządowe służby ochrony zdrowia (Dugas i inni 2012).

Zastosowania można by mnożyć: od przewidywania cen biletów lotniczych w celu najtańszego zakupu (Cukier i Mayer-Schönberger 2014), ocenę ryzyka kredytowego i optymalizację kosztów kredytu (Przanowski 2014), stawianie diagnozy na podstawie danych medycznych i genetycznych (Minelli, Chambers i Dhiraj 2013), po wskazanie posesji najbardziej zagrożonych pożarem czy wybór optymalnego czasu wymiany części w miejskich autobusach (Cukier

[5] *Poster child* oznacza typowy przykład, coś emblematycznego, utrwaloną w kulturze egzemplifikację.

i Mayer-Schönberger 2014). Wszystkie spełniają założenia Big Data zdefiniowanego przez Przanowskiego, czyli dane 3Vs (5Vs), specyficzną architekturę i sprzęt informatyczny, zaawansowane metody analityczne (Przanowski 2014). Niezależnie od dziedziny zastosowania, powtarzają się dwie cechy podejścia analitycznego: badanie populacji, nie tylko próby, w myśl zasady N = all, i zastąpienie wyjaśniania przyczyn wskazaniem korelacji. „Systemy rekomendacji Amazona znalazły wartościową korelację bez znajomości leżących u jej podstaw przyczyn. Wiedza **co**, a nie **dlaczego** jest wystarczająco dobra" (podkreślenie oryginalne) (Cukier i Mayer-Schönberger 2014: 76).

Big Data jako metoda/technika badań społecznych

Zamiast badań surveyowych?

„Po co słuchać ludzi, skoro wszystko o nich wiemy?" zapytali prelegenci szesnastego Kongresu Badaczy Rynku i Opinii (kongresbadaczy). Zbliżający się kryzys socjologii empirycznej ogłoszono już w 2007 roku (Savage i Burrows 2007). Skoro więc sami praktycy, a także naukowcy kwestionują sens prowadzenia badań społecznych (szczególnie ilościowych, gdzie dane zbierane są za pomocą technik wysokostandaryzowanych, co nazywać będziemy surveyami), nie powinno dziwić, że podobnie wyrazili się entuzjaści Big Data (Cukier i Mayer-Schönberger 2014). Stwierdzili, że nauki społeczne utraciły monopol w wyjaśnianiu danych empirycznych, specjaliści od teorii społecznych przestają być potrzebni, a „pasywne" zbieranie danych behawioralnych/niedeklaratywnych i podejście N = all pozwalają przezwyciężyć dobrze znane trudności realizacji badań kwestionariuszowych i doboru próby reprezentatywnej (Cukier i Mayer-Schönberger 2014).

O owej utracie monopolu napisali wspomniani Mike Savage i Roger Burrows (2007); ich zdaniem od czasów swoistego „zwycięstwa" badań surveyowych nad innymi metodami badań społecznych w latach czterdziestych XX wieku aż do lat dziewięćdziesiątych badacze surveyowi byli właściwie jedynymi, dostarczającymi informacji o populacjach ludzi. Rolę tę zaczęły przejmować początkowo badania marketingowe, później zwrócono uwagę na dane transakcyjne gromadzone przez firmy. Obecnie do głosu doszły analizy wielkich zbiorów danych z różnych źródeł, czyli tytułowe Big Data. Badania surveyowe autorzy postrzegają jako metodę/technikę osadzoną w kontekście historycznym (Savage i Burrows 2007): statystyczna koncepcja reprezentatywnej próby nie sięga dalej niż do początków XX wieku, a jej zastosowanie w badaniach społecznych – sondażu przedwyborczym Gallupa – przyniosło w latach trzydziestych widocznie lepsze rezultaty niż badanie na większej, ale obciążonej próbie (w której respondentami byli wyłącznie prenumeratorzy „Literary Digest").

Gallup stosował jednak próbę kwotową, co w 1948 roku doprowadziło do poważnej pomyłki i zwrotu ku próbom losowym (Babbie 2003). Ogólnie rzecz ujmując surveye były powszechnie uznawane za skuteczne i relatywnie tanie. Wraz z rozwojem technologii i swoistego, „kapitalistycznego" apetytu na wiedzę o społeczeństwie przeprowadzano ich coraz więcej i szybciej, co doprowadzić miało do coraz większych trudności z realizacją badania – obecnie ludzie postrzegają pytanie ich o zdanie raczej jako uciążliwość niż zaszczyt, a wskaźniki *response rate* są coraz niższe. Savage i Burrows (2007) zauważają też, że stosowane przez dziesięciolecia kategorie socjodemograficzne, określane przez dane metryczkowe, są w eksploracji danych z dużym powodzeniem zastępowane przez kategorie lokalizacyjne (np. kod pocztowy). Tego rodzaju szczegółowe określenie lokalizacji było przeważnie pomijane w surveyach.

Autorzy wskazują, że obecnie do zastosowań biznesowych korzystanie z eksploracji danych behawioralnych/niedeklaratywnych daje lepsze rezultaty niż surveye, głównie dzięki posiadaniu większej liczby bardziej szczegółowych danych, a co za tym idzie możliwości np. bardziej szczegółowej segmentacji klientów (Savage i Burrows 2007). Przy badaniach surveyowych na próbie reprezentatywnej, np. dla kraju, wyniki można uogólnić dla populacji kraju jako całości, ale wnioskowanie o wybranej podgrupie (np. o młodych, bezrobotnych mężczyznach z dużych miast) będzie nieuprawnione. Takie wnioskowanie na zasadzie przeskalowania obarczone byłoby niemożliwym do oszacowania błędem, bądź niemożliwe z powodu braku odpowiednich danych – braku takich respondentów, bądź braku zmiennych, jeżeli nie zadano odpowiednich pytań. Podejście Big Data i eksploracja danych (*data mining*) nie uprawnia do uogólniania wyników w sensie ekstrapolacji z próby na populację czy formułowania teorii. Nie jest to jednak celem w tym podejściu. Zasada pracy na pełnych zbiorach danych „N = all" zakłada, że analizujemy wszystkie dostępne dane i za ich pomocą tworzymy działający model. Widać to wyraźnie na opisanym już przykładzie systemu rekomendacji Amazona. Algorytm skutecznie „podpowiada" klientowi produkty i na tym koniec. Nie chodzi o odkrywanie ogólnych praw dotyczących wyborów zakupowych, a wyłącznie o rekomendowanie produktów jednego sprzedawcy jego klientom. Specyfikę tego podejścia widać też w metodach eksploracji danych. Dominuje podejście zwane uczeniem maszynowym (*machine learning*): w dużym uproszczeniu polega ono na wielokrotnym uczeniu modelu na zestawie danych i weryfikacji tego modelu na innym, nieznanym zestawie. Jest to jedna z form tzw. sztucznej inteligencji. Uczenie oznacza, że model poprawia swoje wyniki – np. trafność prognozy typu: klient spłaci/nie spłaci kredytu – przez dopasowywanie się do danych na podstawie doświadczenia, czyli kolejnych powtórzeń przebiegów przez dane. Algorytmy budowane są tu nierzadko na zasadzie tzw. czarnej skrzynki – nie wiadomo dokładnie, jakie zmiany w modelu wprowadza komputer przy kolejnych powtórzeniach, nie jest

znana także dokładna zależność między zmiennymi. Właściwie nie stosuje się p-value do oceny istotności modelu czy poszczególnych zmiennych w modelu. Pod uwagę brane jest dopasowanie modelu, a właściwie to, czy wyniki uzyskiwane na nieznanym zestawie danych są równie dobre, co na zestawie uczącym (Larose 2006).

Zwracamy więc uwagę, że choć Big Data i badania surveyowe można ogólnie nazwać ilościowymi metodami poznawania świata, to założenia stojące za metodami analizy danych są tu odmienne. Zdaniem Savage'a i Burrowsa (2007), surveye jako metoda/technika badań społecznych będą sukcesywnie tracić na znaczeniu, ale pozostaną ważne między innymi w badaniach wzdłuż czasu. Badaczy bazujących na surveyach przestrzegają przed byciem zepchniętym na margines i zachęcają do refleksji nad innymi metodami/technikami badań.

Na inny aspekt napięcia między pozyskiwaniem wiedzy z danych zbieranych aktywnie a pasywnie położono nacisk w przywołanym wystąpieniu na Kongresie Badaczy Rynku i Opinii. Autorzy zauważają, że choć zdroworozsądkowo zaufanie do danych niedeklaratywnych/behawioralnych (czyli pozyskanych pasywnie) może być np. u klienta agencji badawczej większe niż do danych deklaratywnych, to korzystanie wyłącznie z pierwszego rodzaju danych jest wysoce niewystarczające do formułowania użytecznych wniosków (Kongres Badaczy 2015). Uwaga zdaje się być bardzo słuszna. Nawet pomijając problem jakości czy sensowności danych zastanych, pomijając ważne zagadnienie ryzyka „odkrywania" korelacji pozornych, dane zastane mają tę wadę, że są… zastane. Nie otrzymamy więc niczego, co nie jest zebrane. Oczywiście istnieją metody tworzenia nowych, użytecznych zmiennych z tych istniejących czy przeciwnie, redukowania wielowymiarowości, czyli zmniejszania liczby zmiennych. Możliwe jest odkrywanie wzorów i zależności, których nie podejrzewano. Kłopotem staje się jednak weryfikowanie hipotez, szczególnie gdy nie mamy żadnych danych o interesującej nas zmiennej/zmiennych. Wydaje się, że w takim zadaniu najbardziej użyteczne jest wciąż skorzystanie z podejścia, które można nazwać tradycyjnym, wywodzącym się z nauk przyrodniczych badaniem naukowym: badania empirycznego na próbie losowej i statystycznego testowania hipotez opartego na p-value, z uwzględnieniem wielkości efektu i mocy testu. Co prawda o wadach takiego podejścia do badań społecznych socjologowie rozprawiają już od dziesięcioleci, jednak przywoływanie ich nie jest celem tego wywodu. Chcemy jedynie zaznaczyć, że choć eksploracja danych daje w niektórych zastosowaniach rezultaty „lepsze" niż surveye, to warto traktować te podejścia raczej jako komplementarne niż konkurencyjne. Podobnego zdania są polscy badacze marketingowi: autorzy wspominanej prelekcji (Kongres Badaczy 2015) czy wypowiadający się w rocznikach PTBRiO (Starzyński 2015; Lutostański 2015; Wójcik 2016; Mróz 2016; Maison 2016), choć w ich artykułach powtarzają się zdania o zagrożeniu branży badawczej.

Można by zatem postulować korzystanie z – ogólnie rzecz biorąc – Big Data jako kolejnej metody badań społecznych. Triangulacja metod i danych: „tradycyjnych" ilościowych, Big Data i jakościowych, ma naszym zdaniem duży potencjał w wieloaspektowym badaniu wybranego zjawiska. Savage i Burrows (2007: 895–896) uważają jednak, że to za mało; należy po pierwsze, zdać sobie sprawę z omawianej utraty monopolu na opis i wyjaśnianie świata społecznego, a w konsekwencji przemyśleć pole działania i rolę współczesnej socjologii.

Text mining

Jak zaznaczono wcześniej, szczególnie szybko rośnie ilość danych nieustrukturyzowanych. Jednym z typów danych o takim charakterze są teksty – np. wpisy na portalach społecznościowych, komentarze na forach internetowych, treści stron WWW, zdigitalizowane bądź wydawane wyłącznie w formie cyfrowej artykuły, książki, dokumenty – przeróżne teksty autorstwa zarówno profesjonalistów, jak i „zwykłych" użytkowników.

Eksploracja danych tekstowych (*text mining*) to komputerowa analiza tekstów, traktowanych jako dane. Używany jest także termin *computational text analysis* (O'Connor, Bamman i Smith 2011), co sugeruje analizę obliczeniową z użyciem komputera. Niekiedy wskazuje się, że jest to analiza zautomatyzowana, w odróżnieniu od tej ręcznej, w tym wspomaganej komputerowo analizy używanej w socjologicznych badaniach jakościowych z użyciem oprogramowania CAQDAS[6]. Analiza jakościowa, gdzie człowiek z pomocą komputera i np. programu NVivo koduje transkrypcje wywiadu, nie jest analizą *text mining*. *Text mining* to 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 częściowo z istniejących wcześniej metod eksploracji danych ustrukturyzowanych (*data mining*), częściowo z dziedziny „nauki" zwanej przetwarzaniem języka naturalnego (*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; Kao i Poteet red. 2010). Zastosowania *text mining* to między innymi takie zadania, jak: identyfikacja słów/zdań kluczowych, kategoryzacja dokumentów według wzorca albo bezwzorcowo (wyestymowanie kategorii), wykrywanie określonych treści w dokumentach, czy wspomniana jako ogólne zadanie *text mining*

[6] Computer Assisted Qualitative Data Analysis Software; por. podręcznik Jakuba Niedbalskiego (2013).

identyfikacja wzorów i powiązań między słowami/dokumentami (Dzieciątko i Spinczyk 2016).

Przykładem zastosowania *text mining* w badaniach społecznych jest stworzenie „kontekstualnego wykrywacza sarkazmu”[7] – modelu zbudowanego na podstawie wpisów na Twitterze (Bamman i Smith 2015). Celem modelu była kategoryzacja wpisów na sarkastyczne i niesarkastyczne. Zastosowano zestawy cech, charakteryzujących cztery różne składowe: samą treść wpisu (9 cech), autora wpisu (5 cech), odbiorców wpisu (trzy), reakcje na wpis i otoczenie wpisu (dwie). W wyniku użycia modelu, będącego odmianą regresji logistycznej, przy użyciu łącznie czterech zestawów cech uzyskano dopasowanie wyników do danych rzeczywistych na poziomie 85,1%. Autorzy zwracają uwagę, że zdecydowanie najważniejsze dla rozpoznawania sarkazmu okazały się cechy wpisu i autora wpisu (Bamman i Smith 2015). Zauważmy, że zastosowania *text mining* mogą być bardzo różnorodne. Bardzo popularnym narzędziem zbudowanym i stale ulepszanym dzięki eksploracji tekstów jest Tłumacz Google – ogólnie rzecz biorąc słownik działa na zasadzie oceny prawdopodobieństwa zastępowania słowa w jednym języku słowem z innego języka, czyli siły związku pomiędzy słowami. Funkcje, poza tłumaczeniem pojedynczych słów, to między innymi tłumaczenia maszynowe, rozpoznawanie języka tekstu i „odczytywanie na głos”. Eksplorując wielką ilość niedoskonałych tekstów z różnych źródeł Google zbudowało słownik uznawany za lepszy niż te mające za podstawę relatywnie małe zbiory profesjonalnych tłumaczeń (Cukier i Mayer-Schönberger 2014). Elementy *text mining* są ważną częścią słynnego projektu Google Flu Trends, a także – w najprostszej chyba formie – prezentowanej w tym artykule oceny popularności wyszukiwań terminu „Big Data”.

Oczywiście ogólne założenia metodologiczne Big Data – przywoływane wielokrotnie N = all i „co? zamiast dlaczego?” obowiązują również w analizach *text mining*. Tu również wykorzystuje się uczenie maszynowe, a nie „tradycyjne” podejście do badań ilościowych nastawione na reprezentatywną próbę losową i statystyczną weryfikację hipotez. Dodatkowo przy okazji omawiania eksploracji tekstów warto zauważyć jeszcze jedną cechę Big Data – akceptowanie chaosu na rzecz wielkości danych. Słownik Google jest najlepszy dlatego, że korzysta z największej, stale rosnącej bazy tekstów, które są niedoskonałe. Kenneth Cukier i Victor Mayer-Schönberger uważają, że: „Big Data, gdzie nacisk położony jest na złożone zbiory danych i brak uporządkowania, lepiej pomaga nam zbliżyć się do rzeczywistości, niż robi to uzależnienie od małych zbiorów [danych] i precyzji. […] Możemy ją [niejednoznaczność] zaakceptować, zakładając, że w zamian lepiej zrozumiemy rzeczywistość – tak jak w malarstwie impresjonistycznym, gdzie każde pociągnięcie pędzla oglądane z bliska wydaje się

[7] Oryg. „Contextualized Sarcasm Detection”.

bezcelowe, ale z oddali ukazuje się nam majestatyczny obraz" (Cukier i Mayer-Schönberger 2014: 71). Biznes docenia akceptowanie niedokładności również z uwagi na przyspieszenie analiz. Poniższy przykład nie dotyczy *text mining*, jednak przemawia do wyobraźni: dzięki zastosowaniu nieco mniej dokładnej, nierelacyjnej architektury baz danych firmie Visa – znanej z kart płatniczych – udało się skrócić czas przetwarzania danych dotyczących około 73 miliardów transakcji do 13 minut. Wcześniej zajmowało to cały miesiąc.

Potencjał *text mining* wydaje się zauważony przez przedstawicieli nauk społecznych i humanistycznych. Pojawia się pojęcie „cyfrowa humanistyka" (*digital humanities*). Już w 2004 roku powstał rodzaj podręcznika pod tytułem *A Companion to Digital Humanities* (Schreibman, Siemens i Unsworth 2004). Poruszono tam zarówno zagadnienia teoretyczne, jak i techniczne, wskazując na zastosowania eksploracji tekstów między innymi w lingwistyce, historii sztuki, filologii klasycznej czy archeologii. Nowszą tego typu pozycją jest np. *Text Mining. A Guidebook for the Social Sciences* (Ignatow i Mihalcea 2016). Poza publikacjami, za wyraz zainteresowania cyfrową humanistyką uznać wolno instytucjonalizowanie się grup uczonych, stosujących tego rodzaju podejście badawcze. Istnieje wpływowa, międzynarodowa wspólnota naukowców, nazywająca się akronimem DARIAH-EU – Digital Research Infrastructure for the Arts and Humanities. Jej członkami są badacze z 17 krajów europejskich. Organizację DARIAH-EU opisano jako infrastrukturę wspierającą badania i nauczanie metod humanistyki cyfrowej, w tym *text mining*. Zapewnia ona między innymi 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.eu). W Polsce działa konsorcjum DARIAH-PL, którego podstawowym celem było wprowadzenie Polski do europejskiej sieci DARIAH oraz pogłębienie i rozbudowanie współpracy ośrodków prowadzących projekty w zakresie humanistyki cyfrowej i dysponujących infrastrukturą w tym zakresie. Konsorcjum tworzy obecnie 18 uczelni, a jego liderem został Uniwersytet Warszawski. W 2015 roku doprowadzono do włączenia konsorcjum polskiego w strukturę europejską (dariah.pl). Na bardzo zbliżonym polu działa Laboratorium Cyfrowe Humanistyki Uniwersytetu Warszawskiego (LaCH UW). Laboratorium 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" (lach.edu.pl). Innym „wskaźnikiem" zainteresowania może być włączanie problematyki *text*

mining do programów nauczania na kierunkach społecznych. Przykładowo dla doktorantów kurs *Computational Text Analysis for Social Sciences* prowadzi King's College London (KCL); Barcelona Graduate School of Economics zaś proponuje przedmiot *Text Mining for Social Sciences* (Barcelona GSE). Na gruncie polskim warto odnotować zorganizowanie panelu poświęconego zastosowaniom Big Data na XVI Ogólnopolskim Zjeździe Socjologicznym. Sześć z ośmiu referatów dotyczyło *text mining* (PTS 2016).

Trudno oprzeć się wrażeniu, że za humanistyką cyfrową stoi nie tylko pragnienie poznania świata. Sądzimy, że motywacjami do sięgania po narzędzia informatyczne wśród badaczy społecznych i humanistów są także chęć „bycia na czasie", promocji swojej dyscypliny oraz zdobywania środków na działalność: „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" (lach.edu.pl). „Konsorcjum i poszczególne grupy robocze w jego ramach będą aktywnie zabiegać o projekty finansowane między innymi ze środków funduszy strukturalnych UE na lata 2014–2020 (w tym w ramach Programu Operacyjnego Polska Cyfrowa) oraz z Programu Ramowego Unii Europejskiej Horyzont 2020" (dariah.pl).

„[…] *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). Sądzimy, że eksploracja tekstów może znakomicie wzbogać warsztat pracy badaczy społecznych/humanistów, niekiedy czynić ich pracę bardziej efektywną, wspierać tworzenie teorii zarówno dzięki możliwości empirycznego sprawdzania hipotez, jak i heurystycznie wartościowej eksploracji danych bez założeń wstępnych. Nie zastąpi ona jednak wszystkich innych rodzajów analizy tekstów, ani nie wyeliminuje ekspertów dziedzinowych. Choć pojawiają się głosy, że nadciąga śmierć ekspertów.

Śmierć eksperta (czytaj: socjologa)?

Skoro danych jest tak wiele, są względnie łatwo dostępne za pomocą narzędzi informatycznych, a zdanetyzowanych jest wiele obszarów w różnych dziedzinach życia, to w celu poznania świata może wystarczy pozwolić „przemówić danym" (Cukier i Mayer-Schönberger 2014: 19, 35, 185)? Za pomocą zaawansowanych technik modelowania jest jakoby możliwe odkrywanie wiedzy z danych[8] – np. w postaci prognozy czy klasyfikacji – bez wiedzy dziedzinowej (substancjalnej)

[8] Nawiązujemy do tytułu popularnego podręcznika *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych* autorstwa Daniela T. Larose z 2006 roku.

o tym przedmiocie, którego dane dotyczą. Dane mają mówić same za siebie, czyli w myśl zasady „co? zamiast dlaczego?" interesujące – a także wystarczające do formułowania rekomendacji – są jedynie relacje między zmiennymi, nie zaś teorie te relacje wyjaśniające. Właściwie wiedza substancjalna jest więcej niż zbędna: ona przeszkadza. Opisując to zagadnienie Cukier i Mayer-Schönberger powołali się między innymi na zilustrowaną w filmie „Moneyball" historię drużyny baseballowej Oakland Athletics. Trener Billy Bean doprowadził tę, dotychczas bardzo 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 analizie danych. Podejmował bardzo niepopularne decyzje – zrezygnował zupełnie z pewnego bardzo 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: ponoć mówią oni, że „jakość przekładu rośnie za każdym razem, gdy z ich zespołu odejdzie jeden lingwista" (Cukier i Mayer-Schönberger 2014: 186).

Co ciekawe, omawiana tu eksploracyjna strategia analizy danych przypomina pewne podejścia metodologiczne w socjologii, np. Metodologię Teorii Ugruntowanej (MTU). W MTU istnieje koncepcja wyłaniania się czy odkrywania teorii z jakościowych danych empirycznych, postulat niestawiania hipotez i eliminowania lub co najmniej uświadamiania własnych założeń co do sfery badanej (Konecki 2000). Rezultatem projektu badawczego zrealizowanego zgodnie z MTU ma być teoria średniego zasięgu, a najważniejsze jest ugruntowanie tej teorii na materiałach zbieranych podczas badań terenowych – teoria jest odkrywana indukcyjnie, nie zaś dedukcyjnie. Sięgając w głąb historii badań społecznych zbliżone dyrektywy odnajdziemy tak u Bronisława Malinowskiego, jak i u przedstawicieli szkoły Chicago. W gruncie rzeczy takie podejście do pozyskania informacji czy wiedzy o świecie jest zbliżone do strategii stosowanej w Big Data. Najważniejsza cecha wspólna to podejście „najpierw dane". To dane są źródłem wiedzy, nie służą do weryfikacji wcześniej postawionych hipotez sformułowanych na gruncie teorii dziedzinowej. Sprowadzając obie strategie – Big Data i MTU – do skrajnego uproszczenia, pozwalamy sobie stwierdzić: nie należy znać się na przedmiocie badania, a znać dobrze narzędzie badawcze i metodę badawczą. Tym samym, jeżeli Big Data miałaby wyeliminować ekspertów dziedzinowych, to i tak potrzebuje ekspertów od

Big Data. Tacy już istnieją, zwani są oni *data scientists*. Zawód ten jest uznawany za jeden z najszybciej rozwijających się, nazywany „najbardziej seksownym zawodem świata", a zapotrzebowanie rynku pracy na takich specjalistów ma wciąż rosnąć (Davenport i Patil 2012). Jako ciekawostkę wskazujemy, że w Polsce w marcu 2017 roku ruszył pierwszy tzw. boot camp[9] data science, organizowany przez firmę Sages we współpracy z Politechniką Warszawską i PAN (kodolamacz.pl).

Mając zapewne na uwadze argumenty zbliżone do powyższych, już w 2008 roku redaktor naczelny magazynu „Wired" ogłosił śmierć ekspertów i zwycięstwo modeli matematycznych nad teorią, twierdząc, że: „tradycyjny proces odkryć naukowych – stawianie hipotez, które są testowane w realnym świecie z wykorzystaniem modelu przyczynowo-skutkowego – traci na znaczeniu i jest zastępowany analizą statystyczną korelacji, za którymi nie stoi żadna teoria" (Cukier i Mayer-Schönberger 2014: 99). Przykład Amazona, gdzie zastosowanie eksploracji danych i modelu do rekomendowania klientom książek przyczyniło się do zwolnienia dotychczas odpowiedzialnych za to krytyków literackich, jest emblematyczny – eksperci zwolnieni, *data scientist* pracuje dalej. Choć właściwie krytycy literaccy są ekspertami, ale nie są testującymi hipotezy naukowcami-empirykami, o których najpewniej mówił cytowany redaktor „Wired". Możemy też przedstawić niejako przeciwny przypadkowi Amazona dowód anegdotyczny: słynny projekt Google Flu Trends. Uznawany jest on za porażkę Big Data – jak wspominaliśmy w roku 2013 prognoza rozmiaru szczytu sezonu grypy była różna od rzeczywistych danych o około 140% (Lazer i Kennedy 2015). W odpowiedzi naukowcy posługujący się mniejszą liczbą bardziej precyzyjnych danych osiągnęli lepszą (bardziej trafną) prognozę (Fung 2014; Lazer i inni 2014). Pokazano także, że model Google przewidywał raczej zimę niż grypę (Lazer i inni 2014). Może w takim razie gruntowna wiedza dziedzinowa „wygrywa" z Big Data?

Szereg problemów dotyczących przede wszystkim prognozowania na podstawie wielkich zbiorów danych opisał 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). Ciekawym stwierdzeniem jest: „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

[9] Termin wszedł do branży IT z wojska i oznacza dosłownie obóz rekrutów; jest to intensywny, siedmiotygodniowy, praktyczny kurs zawodowy połączony ze wsparciem w wejściu na rynek pracy; w Polsce dotychczas działały bootcampy programistyczne.

zmieniła się tak bardzo od czasu wynalezienia Internetu, a nawet prasy drukarskiej. Większość danych to zwykły **szum**, podobnie jak większość wszechświata stanowi pusta przestrzeń" (podkreślenie oryginalne) (Silver 2014: 234–235). Autor wytyka błędy we wnioskowaniu nie tylko inżynierom Big Data, ale również naukowcom i innym analitykom – uważa, że przyczynami błędów są zarówno wypaczenia czy braki w wiedzy statystycznej oraz metodologicznej, jak i skłonności psychologiczne bądź czynniki motywacyjne (Silver 2014). Pogłoski o zwycięstwie analityki i analityków wielkich zbiorów danych nad całą resztą świata nauki należy zatem uznać za grubo przesadzone. Rezygnując z perspektywy konfliktu, sądzimy, że współpraca ekspertów dziedzinowych razem z *data scientists* byłaby najbardziej owocna w poznawaniu prawdy o świecie (jeśli taka prawda istnieje). Czy jednak tyczy się to wszystkich ekspertów? Czy naukowcy społeczni, a szczególnie socjologowie, mieliby coś do zaoferowania?

Sądzimy, że tak. Zaoferowali już kilkadziesiąt lat temu metaforę, która w dużej mierze zainspirowała fizyków i informatyków, zajmujących się problemami zbliżonymi do podejmowanych przez tzw. *data scientists*. Chodzi o metaforę sieci. Jak wskazuje Linton Freeman, pierwsze użycia tej metafory w filozofii sięgają XIII wieku, jednak za początki metodycznego stosowania metafory sieci w naukach społecznych uznaje on socjometrię Jacoba L. Moreno i Helen Jennings z lat trzydziestych ubiegłego stulecia (Freeman 2011: 26). Według Freemana wypracowano wtedy najważniejsze 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. Badania nastawione na analizę sieci kontynuowali między innymi Robert Merton czy Claude Lévi-Strauss. Później, w latach siedemdziesiątych, analizę sieci społecznych rozwinęła i ujednoliciła tzw. szkoła harvardzka Harrisona C. White'a. Dopiero w późnych latach dziewięćdziesiątych zagadnieniem zainteresowali się między innymi Albert-Laszlo Barábasi i Albert Réka (fizyk i biolożka). Od tamtego czasu, jak uważa Freeman, prace środowiska badaczy społecznych i fizyków przenikają się, jednak środowiska są odseparowane. Co ciekawe, badanie Barábasiego jest przywołane w pracy *Big Data. Rewolucja…*, gdzie autorzy uzasadniają 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). Barábasi w popularnonaukowej pracy *Linked: The*

New Science of Networks powołuje się zarówno na Leonarda Eulera, osiemnastowiecznego matematyka, który stworzył podstawy teorii grafów, jak i na Stanleya Milgrama i jego słynny opis sześciu stopni oddalenia poszczególnych ludzi w sieciach społecznych, nazwany problemem „małego świata" (Barábasi 2002). Oczywiście nie sugerujemy tu, że reprezentowana między innymi przez Barábasiego *New Science of Network* zawdzięcza wszelkie swoje dokonania naukom społecznym. Przecież socjometrycy z lat trzydziestych nie napisali pojęcia sieci społecznych na czystej tablicy: korzystali z zaplecza filozoficznego, aparatu matematycznego, a także z nauk medycznych – Moreno był psychiatrą (Freeman 2011). Z kolei Manuel Castells i jego *Społeczeństwo sieci* wydane po raz pierwszy w 1996 roku 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. A algorytm wyszukiwarki Google – PageRank – jest implementacją zainspirowanego przez socjologów i rozwiniętego przez fizyków pojęcia centralności sieci; w skrócie dotyczy ono tego, że węzły posiadające większą liczbę połączeń stają się istotnymi centrami (Freeman 2011).

Odwołując się do jeszcze innej metafory sieci w ujęciu Bruno Latoura (2013) chcemy podkreślić, że w wiedzotwórczym procesie budowania sieci wszyscy z wyżej wymienionych uczonych (i innych podmiotów) dokonywali translacji dokonań swoich poprzedników, tworząc tzw. „hybrydy" lub „quasi-obiekty". Należy to rozumieć bardzo prosto – inspiracje przebiegały wielokierunkowo, a pomysły i koncepcje ulegały różnego rodzaju przekształceniom, zniekształceniom i rozwinięciom. Tym samym uważamy, że pojawiające się w dyskursie o Big Data głosy dotyczące „śmierci ekspertów" w ogóle, czy konkretnie np. socjologów, są nową odsłoną starego sporu między dyscyplinami, który toczył się kilkanaście lat temu w ramach problematyki sieci społecznych. Każda ze stron będzie twierdzić, że to jej sposób poznania świata jest najlepszy, przy czym bez wątpienia poznanie świata jest tu co najmniej tak samo ważne, jak inne korzyści wynikające z uzyskania takiej „epistemologicznej przewagi". Dziś są to *data scientist*, ich Big Data i korzystający z wyników analiz menagerowie. Przywodzi to także na myśl spór wewnątrz socjologii, zapoczątkowany przez socjologię humanistyczną Williama Diltheya i Heinricha Rickerta w kontrze do pozytywizmu Augusta Comte'a. O ile socjologowie raczej zaakceptowali wielość perspektyw na gruncie swojej dyscypliny, o tyle niebywale trudniejsza, jeśli w ogóle możliwa, będzie zgoda w nieporównywalnie szerszym i bardziej różnorodnym gronie.

Chociaż ryzyko „wymarcia" socjologów czy ekspertów w ogóle uznajemy za bardzo małe, to szczególnie socjologom rośnie konkurencja. Niestety słuszna wydaje się teza o tym, że nauki społeczne utraciły monopol na badania społeczne i dostarczanie wiedzy o świecie społecznym. Nawet na gruncie polskim

znamienne jest to, że badanie przekazywania informacji między ludźmi za pomocą Twittera realizowane jest przez fizyków – projektem RENOIR – Reverse EngiNeering of sOcial Information pRocessing (renoirproject.eu) finansowanym w ramach programu EU Horyzont 2020 kieruje pracownia Fizyki w Ekonomii i Naukach Społecznych Wydziału Fizyki Politechniki Warszawskiej. Socjologowie – przynajmniej niektórzy – niewątpliwie zauważyli Big Data jako metodę badań, o czym wspominaliśmy w akapicie poświęconym *text mining*, ale przecież nie stworzyli tej metody, tylko ją adaptują. Na pewno jedną z barier jest poziom zaawansowania matematycznego i informatycznego Big Data, który naszym zdaniem może przerażać, szczególnie socjologów „jakościowych". Inną barierą zdaje się być charakter uzyskiwanych za pomocą Big Data rezultatów – temu, jaki charakter ma wiedza/informacje będące końcowym efektem analiz. Jak wskazaliśmy wcześniej („Zamiast badań surveyowych"), strategia metodologiczna w Big Data nie jest nastawiona ani na uogólnienia, ani na wyjaśnianie. Uzyskana wiedza/informacje muszą mieć przede wszystkim walor praktyczny; zaryzykujemy stwierdzenie, że muszą zwiększać zyski. Naszym zdaniem rezultaty Big Data nie mogą być, zgodnie z *lege artis*, uznane za wiedzę naukową (przynajmniej zgodnie z przytaczanym „tradycyjnym", a właściwie pozytywistycznym rozumieniem nauk empirycznych). Niewątpliwie jednak posługując się zarówno paradygmatem pozytywistycznym (strukturalnym), jak i humanistycznym (interpretatywnym) można korzystać z Big Data jako metody eksploracyjnej, heurystycznej czy pomocniczej w ramach stosowania triangulacji metod i technik badań społecznych.

Podobne zastosowanie Big Data w różnych dziedzinach nauki zaproponował Robert Kitchin. Tę potencjalną ścieżkę rozwoju nauki nazywa on „nauką opartą na danych"[10] – ma być to przeformułowanie sposobu jej uprawiania, w którym zmieszają się abdukcja, dedukcja i indukcja. Inną, naszym zdaniem niepokojącą, ścieżką rozwoju nauki może być empiryzm, czyli: „dane mówią same za siebie"[11] (Kitchin 2014: 10). Bardzo słuszna wydaje się krytyka Kitchina wobec takiego podejścia (2014: 5–6):

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

[10] Oryginalnie: „data-driven science" (tłumaczenie własne).

[11] Oryginalnie: „data can speak for themselves free of theory" (tłumaczenie własne).

- 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 niestawianie hipotez, i uzyskiwanie wartościowych informacji bez zadawania pytań – one zostały postawione wcześniej i „gdzie indziej" niż się wydaje.
- 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.
- Ignorowanie teorii substancjalnych, szczególnie w przypadku gdy badane są zachowania ludzi, prowadzi do bardzo ograniczonych wniosków, nie uwzględniając między innymi 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.

Wśród entuzjastów Big Data, szczególnie w zastosowaniach biznesowych, bez wątpienia dominuje empiryzm. Za ważne zadanie współczesnej socjologii uznajemy zatem krytyczne podejście do Big Data, demaskowanie uwodzicielskich „obietnic" o dostarczaniu obiektywnej prawdy o świecie oraz rozsądne korzystanie z wyników analiz tego rodzaju w działalności naukowej.

Podsumowanie

Po pierwsze, Big Data to zjawisko technologiczne. Dane o własnościach 3Vs (czy poszerzone 5Vs), źródła danych, sposoby przechowywania, przetwarzania i analizy danych rozpatrywane są nierzadko jako technologiczne osiągnięcia bądź problemy do rozwiązania.

Po drugie, to zjawisko ekonomiczne – nie bez przyczyny w kolejnych definicjach Big Data dodano następne V jak Value, czyli wartość rozumianą jako potencjał biznesowy, możliwości generowania zysków i przewagi konkurencyjnej dzięki informacjom wydobytym z danych. Firmy nie tylko zwyczajnie zarabiają na rezultatach Big Data – mówi się także o zmianie modelu biznesowego. Dawniej działy analityczne firm dostarczały raczej informacji zarządowi na zasadzie przekazywania głowie, co robią kończyny (Minelli i inni 2013, przedmowa). W erze Big Data firma ma być inteligentna, reagować natychmiast na różne sygnały z otoczenia: „będziemy tworzyć firmy bystrzejsze

i reagujące szybciej niż ludzie, którzy te firmy prowadzą" (Minelli i inni 2013, s. XVII).

Staraliśmy się przede wszystkim wskazać, że Big Data to zjawisko niejako epistemologiczne – w porównaniu z dominującym, szczególnie w naukach przyrodniczych podejściem pozytywistycznym, mamy tu raczej do czynienia z odmiennymi założeniami i metodami badań. Taki skrajny empiryzm, proponowany przez entuzjastów Big Data, wydaje się niepokojąco atrakcyjny – szczególnie dla biznesu. To powtarzane „co? zamiast dlaczego?" nieodmiennie kojarzy nam się z pochodzącym z memów internetowych hasłem „jeżeli coś jest głupie, ale działa, to nie jest głupie". Sądzimy, że tak pojęta pragmatyczność stwarza szereg zagrożeń, w tym etycznych: podobno opracowywany jest algorytm mający szacować prawdopodobieństwo popełnienia przestępstwa przez konkretną osobę – ma to na celu zatrzymywanie, a być może skazanie osoby, zanim (sic!) popełni niebezpieczny czyn (Cukier i Mayer-Schönberger 2014). Ograniczając się do zagadnień metodologicznych wymieńmy chociażby ryzyko „odkrywania" zależności pozornych, prowadzących do błędnych wniosków i decyzji. Temu oraz podobnym problemom poświęcona jest w całości wspomniana praca Nate'a Silvera *Sygnał i szum* (2014). Natknęliśmy się na nawiązującą do problemu zależności pozornych dyskusję w komentarzach pod artykułem zamieszczonym na portalu http://www.datasciencecentral.com. Jeden z wypowiadających się – Steve – prawdopodobnie praktyk *data science*, użył terminu „apofenia". Jego zdaniem: „Apofenia to dostrzeganie znaczących schematów bądź związków w danych przypadkowych czy bezsensownych. Ważna część pracy Data Scientist w erze Big Data to pomoc w odróżnianiu apofenii od znaczących zjawisk" (datasciencecentral.com). Taki głos rozumiemy jako przeciwny skrajnemu empiryzmowi. Jak wskazywaliśmy wcześniej, epistemologiczne „obietnice" Big Data są właściwie niemożliwe do spełnienia. Szczególnie, kiedy przedmiotem analiz są zachowania ludzi, stosowanie jedynie Big Data prowadzi do bardzo powierzchownych wyników.

Sądzimy, że świat nauki może z powodzeniem korzystać z potencjału Big Data, przy świadomości ograniczeń tego podejścia. W naukach społecznych wartościowe wydaje się włączanie elementów Big Data zarówno do analiz prowadzonych w paradygmacie pozytywistycznym (strukturalnym), jak i humanistycznym (interpretatywnym) w ramach triangulacji metod i technik. Sądzimy, że przedstawiciele nauk społecznych powinni rozwijać swoje kompetencje w posługiwaniu się Big Data i współpracować z *data scientists*, a także krytykować i podważać to podejście, ujawniać i dyskutować stojące za nim, głęboko ukryte założenia.

Big Data to bez wątpienia także zjawisko społeczne. Z perspektywy np. socjologii wiedzy można by spojrzeć na cały proces: od generowania danych przez ich przetwarzanie do wykorzystania i wdrożenia. W ramach socjologii kultury

czy socjologii organizacji również pojawia się szereg tematów: od strachu przed byciem inwigilowanym czy okradanym do władzy i zysku, jaki umożliwiają trafne prognozy. Jako zjawisko społeczne można również rozpatrywać ową szeroką rewolucję, jaką Big Data ma nieść. Materiałem do analizy jakościowej można by uczynić takie wypowiedzi, jak np. Drew Conwaya (firma Project Florida): „Big Data to ruch kulturowy, za pomocą którego kontynuujemy odkrywanie tego, jak ludzkość i świat oddziałują na siebie nawzajem", czy też Daniela Gillicka (firma Google): „Big Data reprezentuje zmianę kulturową, która polega na tym, że coraz więcej decyzji podejmowanych jest za pomocą algorytmów działających na podstawie udokumentowanych, niezmiennych dowodów" (Dutcher 2014). Analizie można by także poddać np. całą książkę wielokrotnie tu przywoływanej pary Cukier i Mayer-Schönberger. Ta wyjątkowo entuzjastyczna wobec Big Data publikacja stanowi potencjalnie bardzo wartościowy materiał do analizy dyskursu o tytułowym zjawisku.

Na gruncie polskim Big Data jako zjawisko społeczne zainteresowało głównie dwóch uczonych: wspomnianego już Łukasza Iwasińskiego oraz Kazimierza Krzysztofka. Iwasiński postrzega zjawisko – a właściwie jego część, danetyzację – jako społeczne zagrożenie na trzech płaszczyznach (Iwasiński 2016: 139):

- postępującej inwigilacji i utraty prywatności,
- kolonizowania przez rynek coraz drobniejszych elementów doświadczenia człowieka i świata społecznego,
- fetyszyzacji danych i reifikacji rzeczywistości społecznej.

O ile pierwsze dwa z wymienionych zagrożeń są dość typowe – można by je sprowadzić do Big Data = Big Brother i narzekania na „zły" kapitalizm – o tyle trzecie jest bardzo trafne. Autor słusznie zauważa, że zarówno dane, jak i algorytmy przetwarzające je w informacje są konstruowane społecznie, nie zaś obiektywnie prawdziwe. Tym samym należy zawsze korzystać z nich krytycznie. Zagrożeniem jest zatem traktowanie wyników Big Data jako ostatecznej wykładni prawdy. Problem ten zgłębił Krzysztofek (2012) – jego zdaniem ludzie przekazują władzę zbierania obserwacji i ich analizy technologiom z dwóch powodów. Otóż ludzki mózg nie wystarcza do wykonania tak dużych i złożonych analiz; co ważniejsze, zawierzenie technologiom jest wygodne, zwalnia z wysiłku poznawczego, i częściowo z odpowiedzialności za skutek decyzji podjętych na podstawie wyników Big Data. Krzysztofek w innym tekście (2015) stwierdził, że „W przeważającej większości współczesny człowiek nie rozumie technologii, jakimi się posługuje, są one dlań czarną skrzynką" (Krzysztofek 2015: 11). Płynące stąd zagrożenia to zarówno bezrobocie technologiczne, jak też rosnące nierówności społeczne. Coraz bardziej inteligentne maszyny – a sztuczna inteligencja jest niczym innym niż dostosowywaniem się poprzez ocenę prawdopodobieństwa uzyskiwaną z danych w czasie rzeczywistym – przejmują pracę

ludzi już nie tylko w produkcji, ale i usługach. Technologie wyprzedzają rozwój dużej części ludzkości, a przy tym „nie przejmują się" tym rozwojem. Cele rozwoju technologii nie są zdaniem Krzysztofka celami uwzględniającymi potrzeby i cechy człowieka, a nastawionymi na potrzeby i cechy układu człowiek–maszyna (Krzysztofek 2015). Może nawet maszyna–maszyna: na tym polega omawiany krótko Internet rzeczy. Co do nierówności, to mają one naturę intelektualną, tzn. niewielu będzie/jest ludzi, którzy będą/są „mądrzejsi" od maszyn, a wielu „głupszych", i w jakimś sensie podporządkowanych (Krzysztofek 2015).

Mamy wrażenie, że Big Data wciąż zyskuje na popularności i wzbudza zainteresowanie, ponieważ wyjątkowo „pasuje" do współczesnego świata: w tym zjawisku jak w soczewce koncentrują się pewne charakterystyczne cechy współczesności, określanej jako społeczeństwo ponowoczesne, postindustrialne, informacyjne, globalna wioska i neoliberalny kapitalizm. Te cechy to naszym zdaniem: utylitaryzm, technokracja, dehumanizacja, szybkość, elastyczność, nastawienie na zysk. Big Data jawi się zatem jako coś na kształt tajemniczego, potężnego i przenikliwego bytu, będącego połączeniem superkomputera, szpiega, szklanej kuli, wyroczni i kury znoszącej złote jajka. Czyż nie jest to spełnieniem marzeń? Kto z nas nie chce chociaż trochę zmniejszyć niepewności jutra? Uzyskać nieco więcej spokoju? A przy tym zarobić co nieco? Zjawisko Big Data można więc interpretować jako pozornie trafną odpowiedź na wyzwania ponowoczesności – przeładowanie informacjami, szybkość zmian, pragnienie konsumpcji, tonięcie w morzu możliwości; szeroko rozumianą „mozaikowość" i „płynność" życia, portretowaną wielokrotnie przez między innymi Zygmunta Baumana (2007). Powyższe rozważania to oczywiście tylko „wyobraźnia socjologiczna"; by zgłębić problematykę planujemy badania terenowe tytułowego zjawiska, inspirowane etnografią laboratorium czy szerzej Studiami nad Nauką i Techniką (Science and Technology Studies; STS) (por. Abriszewski 2010, 2012; Latour 2013). Sądzimy, że warto spojrzeć na zjawisko Big Data oczami jego praktyków.

Literatura

Abriszewski, Krzysztof. 2010. *Wszystko otwarte na nowo. Teoria Aktora-Sieci i filozofia kultury*. Toruń: Wydawnictwo Naukowe UMK.

Abriszewski, Krzysztof. 2012. *Poznanie, zbiorowość, polityka. Analiza Teorii Aktora-Sieci Bruno Latoura*. Kraków: TAiWPN Uniwersitas.

Amazon. 2017. https://www.amazon.com/gp/site-directory/ (dostęp 16.02.2017).

Babbie, Earl. 2003. *Badania społeczne w praktyce*. Tłum. W. Betkiewicz, M. Bucholc i P. Gadoms. Warszawa: WN PWN.

Barabási, Albert-László. 2002. *Linked: The New Science of Networks*. Cambridge, MA: Perseus Publishing.

Bamman, David i Noah A. Smith. 2015. *Contextualized Sarcasm Detection on Twitter*. „International AAAI Conference on Web and Social Media, North America" (04.2015.), http://www.aaai.org/ocs/index.ph0p/ICWSM/ICWSM15/paper/view/10538/10445.

Barcelona GSE. 2017. *Text Mining for Social Sciences*; http://www.barcelonagse.eu/tmp/pdf/Text_Mining_Social_Sciences.pdf (dostęp 22.02.2017).

Bauman, Zygmunt. 2007. *Płynne życie*. Tłum. T. Kunz. Kraków: Wydawnictwo Literackie.

Berman, J. Jules. 2013. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Waltham: Elsevier.

Castells, Manuel. 2013. *Społeczeństwo sieci*. Tłum. M. Marody, K. Pawluś i J. Stawiński. Warszawa: WN PWN.

Chen, Hsinchun, H.L. Roger Chiang i C. Veda Storey. 2012. *Business Intelligence and Analytics: From Big Data to Big Impast*. „MIS Quarterly" 36/4: 1165–1188.

Cukier, Kenneth i Victor Mayer-Schönberger. 2014. *Big Data. Rewolucja, która zmieni nasze myślenie, pracę i życie*. Tłum. M. Głatki. Warszawa: MT Biznes Ltd.

Data Science Central. 2012. http://www.datasciencecentral.com/forum/topics/the-3vs--that-define-big-data [wpis na forum], wypowiedź użytkownika o nicku „Steve" (06.07.2012 r., godzina 11:21) [dostęp 25.01.2017].

DARIAH-EU. 2017. http://www.dariah.eu/ (dostęp 27.02.2017).

Davenport, H. Thomas i D.J. Patil. 2012. *Data Scientist: The Sexiest Job of The 21st Century*. „Harvard Business Review" (10.2012), https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Diebold, X. Francis. 2012. *A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline,* (26.11.2012), http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf (dostęp 14.01.2017).

Dijck, van José. 2014. *Datafication, Dataism and Dataveillance: Big Data between scientific paradigm and ideology*. „Surveillance & Society" 12: 197–208.

Dugas, F. Anderea, Yu-Hsiang Hsieh, Scott R. Levin, Jesse M. Pines, Darren P. Mareiniss, Amir Mohareb, Charlotte A. Gaydos, Trish M. Perl i Richard E. Rothman. 2012. *Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics*. „Clinical Infection Diseases" 54/4: 463–469, doi:10.1093/cid/cir883.

Dutcher, Jennifer. 2014. *What is Big Data?,* Blog Data Science Berkley, https://datascience.berkeley.edu/what-is-big-data (dotęp 03.08.2014).

Dzieciątko, Mariusz i Dominik Spinczyk. 2016. *Text mining. Metody, narzędzia, zastosowania*. Warszawa: WN PWN.

Freeman, Linton. 2011. *The Development of Social Network Analysis –with an Emphasis on Recent Events*. W: J. Scott i P.J. Carrington (red.). *The Sage Handbook of Social Media Analysi*s. London: SAGE.

Fung, Kaiser. 2014. *Google Flu Trends' Failure Shows Good Data > Big Data.* „Harvard Business Review" (25.03.2014), https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data.

Gandomi, Amir i Murtaza Haider. 2015. *Beyond the Hype: Big Data Concepts, Methods, and Analytics*. „International Journal of Information Management" 35/2: 137–144.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski i Larry Brilliant. 2009. *Detecting Influenza Epidemics Using Search Engine Query Data*. „Nature" 457: 1012–1014, doi:10.1038/nature07634.

Google Trends, https://www.google.pl/trends/explore?q=Big%20Data

Ignatow, Gabe i Rada Mihalcea. 2016. *Text Mining. A Guidebook for the Social Sciences*. London: SAGE.

Iwasiński, Łukasz. 2016. *Społeczne zagrożenia danetyzacji rzeczywistości.* W: B. Sosińska-Kalata (red.). *Nauka o informacji w okresie zmian. Informatologia i humanistyka cyfrowa*. Warszawa: Wydawnictwo SBP, s. 135–146.

Kao, Anne i Stephen R. Poteet (red.). 2010. *Natural Language Processing and Text Mining.* Springer-Verlag London Limited.

KCS. *Computational Text Analysis for Social Sciences*; http://www.kcl.ac.uk/study/graduate-school/doctoral-training-centre/training/quantitative-training-competition.aspx.

Kitchin, Robert. 2014. *Big Data, New Epistemologies and Paradigm Shifts*. „Big Data and Society Sage" (Jun 2014), doi: 10.1177/2053951714528481.

Konecki, T. Krzysztof. 2000. *Studia z metodologii badań jakościowych. Teoria ugruntowana*. Warszawa: WN PWN.

Kongres Badaczy. 2015. http://www.kongresbadaczy.pl/2015/index.php/po-co-sluchac-ludzi-skoro-wszystko-o-nich-wiemy-o-sztuce-interpretacji-danych-niedeklaratywnych (dostęp 13.02.2017).

Krzysztofek, Kazimierz. 2012. *Big Data Society. Technologie samozapisu i samopokazu*. „Transformacje. Pismo interdyscyplinarne" 1–4(72–55): 223–257.

Krzysztofek, Kazimierz. 2015. *Technologie cyfrowe w dyskursach o przyszłości pracy.* „Studia Socjologiczne" 4: 5–31.

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: 387–394.

LaCH UW. 2017. http://lach.edu.pl/o-laboratorium/ (dostęp 04.02.2017).

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.

Larose, T. Daniel. 2006. *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*. Tłum. A. Wilbik. Warszawa: WN PWN.

Latour, Bruno. 2013. *Nadzieja Pandory. Esej o rzeczywistości w studiach nad nauką*. Tłum. K. Abriszewski, A. Derra, M. Smoczyński, M. Wróblewski i M. Zuber. Toruń: Wydawnictwo Naukowe UMK.

Lazer, David, Ryan Kennedy, Gary King i Alessandro Vespignani. 2014. *The Parable of Google Flu: Traps in Big Data Analysis*. „Science" 343/6167: 1203-1205, doi: 10.1126/science.1248506, http://nrs.harvard.edu/urn-3:HUL.InstRepos:12016836.

Lazer, David i Ryan Kennedy. 2015. *What We Can Learn From the Epic Failure of Google Flu Trends*. „Wired" (01.10.2015), https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.

Lohr, Steve. 2013. *The Origins of 'Big Data': An Etymological Detective Story*. Bits Blog of New York Times (01.02.2013), http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/?_r=0.

London KCLMS. 2015. *Computational Text Analysis for Social Sciences*; http://www.kcl.ac.uk/study/graduate-school/doctoral-training-centre/training/quantitative-training-competition.aspx (dostęp 22.02.2017).

Lutostański, Michał. 2015. *10 mln respondentów rocznie to nie jest mała grupa*. „Badania marketingowe. Rocznik Polskiego Towarzystwa Badaczy Rynku i Opinii – 2014/15 – edycja XIX", s. 58.

Maison, Dominika. 2016. *Czy badacze marketingowi będą jeszcze potrzebni?* „Badania marketingowe. Rocznik Polskiego Towarzystwa Badaczy Rynku i Opinii – 2015/16 – edycja XX", s. 32.

Manovich, Lev. 2011. *Język nowych mediów*. Tłum. P. Cypryański. Warszawa: Oficyna Wydawnicza Łośgraf.

Marcus, James. 2004. *Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut*. New York: The New Press.

Minelli, Michael, Michele Chambers i Ambiga Dhiraj. 2013. *Big Data, Big Analytics. Emerging Business Intelligence and Analytic Trends for Today's Businesses*. New Jersey: John Wiley & Sons Inc.

Mróz, Bogdan. 2016. *Zrozumieć duszę konsumenta*. „Badania marketingowe. Rocznik Polskiego Towarzystwa Badaczy Rynku i Opinii – 2015/16 – edycja XX", s. 5–9.

Niedbalski, Jakub. 2013. *Odkrywanie CAQDAS*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

O'Connor, Brendan, David Bamman i Noah A. Smith. 2011. *Computational Text Analysis for Social Science: Model Complexity and Assumptions*. Proc. of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds, http://people.cs.umass.edu/~wallach/workshops/nips2011css/papers/OConnor.pdf.

Ohlhorst, J. Frank. 2013. *Big Data Analytics: Turning Big Data into Big Money*. Wiley and SAS Business Series.

Paharia, Rajat. 2014. *Lojalność 3.0: jak zrewolucjonizować zaangażowanie klientów i pracowników dzięki big data i rywalizacji*. Tłum. D. Gasper. Warszawa: MT Biznes Ltd.

Powers, William. 2014. *Wyloguj się do życia*. Tłum. E. Kleszcz. Warszawa: Grupa Wydawnicza PWN.

Przanowski, Karol. 2014. *Credit Scoring w erze Big Data*. Warszawa: Oficyna Wydawnicza SGH.

PTS. 2016. *Program XVI Ogólnopolskiego Zjazdu Socjologicznego, Gdańsk 14-17 września 2016*, s. 44, http://16zjazdpts.pl (dostęp 08.02.2017).
Quantified Self Poland. 2017. https://pl-pl.facebook.com/QuantifiedSelfPoland/ (dostęp 21.02.2017).
RENOIR. 2017. http://www.renoirproject.eu/ (dostęp 17.02.2017).
Savage, Mike i Roger Burrows. 2007. *The Coming Crisis of Empirical Sociology*. „Sociology SAGE" 41(5): 885–899.
Schreibman, Susan, Ray Siemens i John Unsworth. 2004. *A Companion to Digital Humanities*. Oxford: Blackwell.
Silver, Nate. 2014. *Sygnał i szum. Sztuka prognozowania w erze technologii*. Tłum. M. Lipa. Gliwice: Wydawnictwo Helion.
Soubra, Diya. 2012. *The tree Vs that define big data,* Data Science Central, http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data (dostęp 28.01.2017).
Sumbul, Michel. 2014. *Big Data problematic* [wpis na blogu] http://whatsbigdata.be/category/big-data-overview/ (dostęp 30.01.2017).
Starzyński, Sebastian. 2015. *Big data zagrożenie czy szansa dla branży badawczej*. „Badania marketingowe. Rocznik Polskiego Towarzystwa Badaczy Rynku i Opinii – 2014/15 – edycja XIX", s. 34.
TechAmerica Foundation's Federal Big Data Commission. 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 (dostęp 26.01.2017).
Turner, Vernon. 2014. *The Digital Universe of Opportunities,* IDC (04.2014), http://www.emc.com/leadership/digital-universe/2014iview/index.htm (dostęp 19.01.2017).
UNECE. 2014. *How big is Big Data? Exploring the role of Big Data in Official Statistics*, http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307 (dostęp 14.01.2017).
Wójcik, Paweł. 2016. *Czy/kiedy podejście BIG DATA zastąpi badania marketingowe?* „Badania marketingowe. Rocznik Polskiego Towarzystwa Badaczy Rynku i Opinii – 2015/16 – edycja XX", s. 30.

The Potential of Big Data in Social Research

Summary

The problem was taken up by the epistemological promises emerging among the Big Data enthusiasts. There have been discussions about the use of Big Data as a method or a technique for social research. Also, the promises mentioned above and the common 'the end of experts' slogan were criticised. Conclusions concern cognitive opportunities and risks, especially in the social sciences. It was considered that Big Data could be known as a knowledge acquisition tool. However, a strong sceptical

approach is necessary. For sociologists, exploring the phenomenon itself is valuable for understanding the information society. The possible direction of future Big Data research is also indicated.

Key words: Big Data; research methodology; information society; epistemology.