Leonard Dajerling
Uniwersytet im. Adama Mickiewicza w Poznaniu
DOI: 10.14746/r.2016.2.4

Era Big Data. Czyli jak postęp technologiczny i metodologiczny wpływa na wybory prezydenckie?

Kampanie wyborcze w Stanach Zjednoczonych od zawsze wzbudzały emocje. Nie od dziś jest oczywistym, że marketing polityczny jest kluczem do serc wyborców. Wyjątkiem nie są ostatnie wybory prezydenckie w 2016 r., kiedy Kandydat Partii Republikańskiej ekstrawagancki milioner Donald Trump, jako niedoświadczony polityk, kontra Hilary Clinton doświadczona polityk żona byłego prezydenta USA. Niespodziewany, jak na pierwszą chwilę wynik wzbudza potrzebę zadania pytania o to co mogło, bądź co stało za wygraną D. Trumpa.

Na to pytanie odpowiedzi dostarcza Internet, a dokładnie termin *Big Data* tj. duże zbiory danych. Postęp a co za tym idzie rozwój technologii od zawsze wpływał na kształtowanie środowiska życia człowieka, w tym politycznego. W historii ludzkich wynalazków niewątpliwie jako przełomowy moment można wpisać moment wynalezienia Internetu. Wraz z rozwojem technologii komputerowej, Internet upowszechnił się w życiu większości ludzi na świecie. Owe upowszechnienie utożsamiane jest z trzecią rewolucją przemysłową (IT, automatyzacja produkcji itd.). Obecnie coraz częściej można się spotkać z twierdzeniami, że jako ludzkość stoimy na przełomie czwartej rewolucji technologicznej, nazywanej często *cyber-physical*, która opisuje rzeczywistość jako swoistą integrację człowieka z Internetem czy maszyną.

Przedstawiony powyżej rozwój generuje ogromne ilości informacji. Przyjmuje się, że ilość informacji, jaka wygenerowana zostanie w najbliższych 48 godzinach (Maciołek, 2017), to więcej niż cywilizacja stworzyła od początku swojego istnienia. Coraz częściej kwestia analizy dużych ilości informacji wykracza poza granice nauk ścisłych odnajdując miejsce w ramach nauk humanistycznych, dając możliwości wykorzystywania tych technik np. w wyborach prezydenckich. Taka tendencja zmusza do refleksji nad obecnym stanem rozwoju człowieka tj. czy jest przygotowany na zmiany technologiczne, czy potrafi skutecznie przeciwdziałać niekontrolowanemu zbieraniu informacji.

Celem niniejszego artykułu jest próba opisu oraz wyjaśnienia terminu *Big Data* oraz możliwości zastosowania technik analizy danych podczas kampanii wyborczych. W pierwszej kolejności autor przytacza historię rozwoju Internetu w świecie, następnie opisuje kolejne trzy rewolucje przemysłowe, uznając że zagadnienia te stanowią istotną część w celu zrozumienia fenomenu przyrostu informacji oraz, jak proces ten może kształtować się w przyszłości. Przechodząc do czwartej rewolucji przemysłowej, autor skupia się na głównym zagadnieniu, jakim jest analiza danych przy użyciu *Big Data*. Analizując ową tendencję, dokonane zostały próby rekonstrukcji oraz określenie czym *Big Data* jest i gdzie tkwi różnica w stosunku do klasycznej analizy danych. W głównej, końcowej części artykułu autor szczególną uwagę poświęca na kwestie związane z możliwościami wykorzystania analizowanej metody w celach politycznych, na przykładzie kampanii prezydenckiej D. Trumpa. W ostatniej części autor wskazuje na możliwe scenariusze wykorzystania *Big Data*, które wykraczają poza analizy typowo politologiczne.

Rozwój internetu a ilość informacji

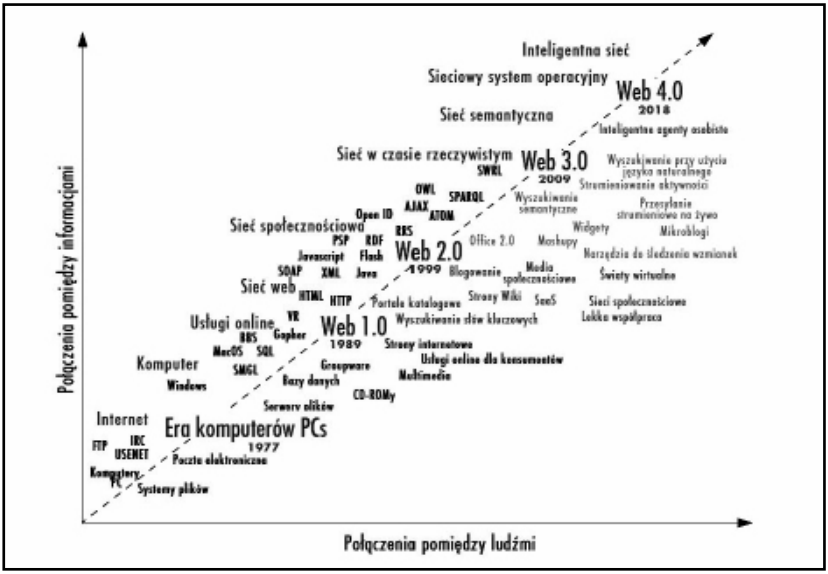
Źródłem większości przełomowych technologii czy wynalazków jest sektor wojskowy. Nie inaczej było w przypadku Internetu, kiedy w dniu 4 października 1957 r. Związek Radziecki wystrzelił pierwszego sztucznego satelitę w kosmos (Sputnik 1), w odpowiedzi narodziła się koncepcją stworzenia systemu komunikacji dla armii Stanów Zjednoczonych. Paul Baran amerykański informatyk polskiego pochodzenia pracujący w firmie Rand Corporation opracował w 1962 r. na zlecenie amerykańskich sił zbrojnych projekt działania sieci, która mogłaby połączyć Pentagon, Departament Obrony i bazę pod górą Cheyenne (Krysiak, 2005, s. 169). Wówczas jednak projekt miał niewiele wspólnego z dzisiejszym Internetem i był stosunkowo trudny w obsłudze.

Dalszym, milowym krokiem w historii Internetu, był projekt A*rpanet* (ang. *Advanced Research Projects Agency Network*) mającym na celu połączenie uniwersytetów w USA. W 1969 r. rozpoczęto uruchamianie pierwszego węzła sieci, na amerykańskich uczelniach m.in. na Uniwersytecie Kalifornijskim w Los Angeles. Powodem tych działań była limitowana liczba komputerów ze względów na ówczesne koszty oraz konstrukcję (Huurdeman, 2003, s. 584). Ponieważ było ich niewiele, a liczne uniwersytety chciały z nich korzystać, wystąpiono z propozycją, aby połączyć ośrodki akademickie siecią, która pozwalałaby na zdalne korzystanie z dostępnego sprzętu komputerowe-

go. Dynamiczny rozwój sieci ARPANET w latach 1969-1977 pozwala wywnioskować potencjał tkwiący w komunikacji elektronicznej.

Dzięki rozwiązaniu World Wide Web (www) opracowanego w 1991 r. przez m.in. Tima Berners-Lee udało się przyłączyć technologię hipertekstu do Internetu, co stało się podstawą nowego typu komunikacji sieciowej (McPherson, 2009, s. 5). Jednak dopiero wraz z wynalezieniem technologii Web 2.0, zmieniło się postrzeganie serwisów internetowych, które dotychczas działały jednokierunkowo (Kaznowski, 2008, s. 28). Od tej pory postrzegane są, jako nowa infrastruktura globalna oparta na interakcjach pomiędzy użytkownikami, gdzie każdy ma możliwość współtworzenia oraz przetwarzania treści zamieszczonych na portalach internetowych (Schemat nr 1).

Schemat 1 Ewolucje technologii sieciowych



Źródło: www.absta.pl.

W ciągu ostatnich lat technologie medialne dojrzewały w ramach codziennych praktyk społecznych. Jako bilans tychże praktyk mamy dziś w pełni wykształtowane serwisy społecznościowe oparte na wymianie informacji w czasie rzeczywistym. Użytkownicy mają możliwość nadawania przekazów w formie video z telefonów komórkowych (smartfony) na żywo. Interesującym zjawiskiem jest umieszczanie w serwisach internetowych typu Facebook ogłoszeń o prace. Możli-

wości, jakich obecnie dostarczają wspomniane technologie z pewnością jest więcej, a jedyne co ogranicza tworzenie nowych funkcji czy ulepszeń, jest pomysł oraz możliwości programistów. Opisany rozwój serwisów internetowych opartych na wymianie informacji w czasie rzeczywistym uważa się już za technologię Web 3.0 (Schemat nr 2). Kolejnym krokiem jest technologia Web 4.0, która określana jest mianem sieci inteligentnej bądź Internetem Rzeczy, gdzie urządzenia np. lodówka czy mikrofalówka połączona do Internetu generuje informację mające na celu poinformowanie np. o dacie przydatności produktu itd.

Określenie Web 4.0 jest połączone z koncepcją „czwartej rewolucji przemysłowej". Rewolucje przemysłowe przybliżają możliwość innego spojrzenia na rozwoju nauki, gdzie głównym przedmiotem rozważań jest technologia. Nazewnictwo „Rewolucje Przemysłowe" proces zawdzięcza temu, że ich wpływ na funkcjonowanie modelu gospodarki było i jest znaczne. Za pierwszą rewolucje przyjmuje tzw. wiek pary, kiedy podczas produkcji wykorzystywano energię pozyskiwaną z wody oraz pracy. Za drugą rewolucję przemysłową uznaje się tzw. wiek elektryczności, kiedy do produkcji dóbr wykorzystywano energie elektryczną. Następnie za trzecią rewolucję przemysłową zwaną inaczej Wiekiem Komputerów, uznaje się wynalezienie w 1969 r. programowalnego układu logicznego. Coraz częściej spotkać się można z twierdzeniami, że żyjemy w czasach czwartej rewolucji przemysłowej (Schwab, 2016). Cechą charakterystyczną ostatniej rewolucji jest stopniowy zanik umownej bariery łączącej człowieka z maszyną oraz wzajemnym wykorzystywaniu automatyzacji, przetwarzania, wymiany danych i technik wytwórczych (Schwab, 2016). W definicji czwarta rewolucja przemysłowa jest zbiorczym terminem dla technik i zasad funkcjonowania organizacji łańcucha wartości łącznie stosujących lub używających systemów cyber-fizycznych, Internetu Rzeczy i przetwarzania (analizy) chmurowego (Hermann, Pentek, Otto, 2015, s. 6-9). Owa synergia przedstawia się w codziennej obecności w życiu człowieka urządzeń z połączeniem do Internetu, a także z tzw. Internetem Rzeczy. Uznaje się, że kwintesencją czwartej rewolucji przemysłowej będzie wynalezienie sztucznej inteligencji, co szacuje się przy uznaniu teorii Ray'a Kurzweil, na 2045 r., Wspomniana teoria, której twórcą jest Ray Kurzweil wskazuje, że można przewidzieć wynalezienie sztucznej inteligencji na podstawie ekstrapolacji dotychczasowych trendów rozwoju techniki i uogólnienia prawa

Moore'a[1] na inne dziedziny technologii. Według Kurzweila, przełomy w rozwoju technologicznym zdarzają się w coraz krótszych odstępach (postęp wykładniczy) czasu, które zmaleją do minimalnych wartości około 2045 r. (Kurzweil, 2005, s. 122).

Tabela 1 Rewolucje przemysłowe

Rewolucja Przemysłowa	Umowny początek	Nazwa (umowna)	Charakterystyczne wynalazki
I	1784	Wiek pary	Produkcja przy zastosowaniu energii z wody i pary
II	1870	Wiek elektryczności	Produkcja przy zastosowaniu energii elektrycznej
III	1969	Wiek komputerów	Programowalny układ logiczny
IV	???	Cyber-physical System	Sztuczna inteligencja

Źródło: opracowanie własne.

Wpływ informacji na metodologie nauk

Nierozłącznym elementem przedstawionych powyżej zjawisk, jest informacja. To właśnie informacja wraz z rozwojem cywilizacyjnym dynamicznie nabiera na znaczeniu. Studiując metodologię bądź historie nauki można natknąć się na podstawowe rozróżnienie nauk, tj. nauki humanistyczne oraz przyrodnicze (Heller, 2011, s. 22). Nauki humanistycznie, przyjmuje się jako wartościujące a więc subiektywne. Natomiast nauki przyrodnicze za obiektywne, ponieważ dostarczają jednostek mierzalnych, których brak jest w naukach humanistycznych (autor za nauki humanistyczne przyjmuje również społeczne) (Heller, 2011, 22-23). Metody wykorzystywane w nauce w ogóle mieszają się tworząc przy tym nowe podejścia, wnioski bądź rekomendacje. Najczęściej wyróżnia się trzy podejścia: interdyscyplinarne, multidyscyplinarne oraz transdyscyplinarne (Konieczny, 2016, s. 443).

[1] Prawo Moore'a – prawo empiryczne, wynikające z obserwacji, że ekonomicznie optymalna liczba tranzystorów w układzie scalonym zwiększa się w kolejnych latach zgodnie z trendem wykładniczym (podwaja się w niemal równych odcinkach czasu). G. Moore w 1965 r. zaobserwował podwajanie się liczby tranzystorów co ok. 18 miesięcy. Liczba ta była następnie korygowana i obecnie przyjmuje się, że liczba tranzystorów w mikroprocesorach od wielu lat podwaja się co ok. 24 miesiące. Na zasadzie analogii, prawo Moore'a stosuje się też do wielu innych parametrów sprzętu komputerowego, np. pojemności dysków twardych czy wielkości pamięci operacyjnej (Moore, 1965).

Przykładem mogą służyć: bio-informatyka, astrofizyka itd. Również w naukach humanistycznych bazować można na danych wyrażanych w sposób matematyczny (numeryczny). Próby takiego podejścia są znane już od dawna. Uznanie, że stosowanie metod ilościowych daje możliwość zdobywania obiektywnej wiedzy (uznawanej za prawde) wiedzie do szkoły Platońskiej (zainspirowani pitagorejskim podejściem) wizji nauki, kiedy Platon uznawał, że prawdziwe istnienie przysługuje tylko światu idei, zaś świat fizyczny jest jedynie jego cieniem (patrz więcej: Dembiński, 2013). Przykładem, obrazującym różne podejścia może być Naukach o Stosunkach Międzynarodowych (NSM). Jednym z podejść do analizy NSM, jest dobata (inter)paradygmatyczna, gdzie występuje tzw. rewolucja behawioralna (Czaputowicz, 2008, s. 91-99). W ramach podejścia behawiorystów to struktura i przestrzeń zyskują w porównaniu do czasu i kontekstu (Vaughan-Williams, 2005, s. 115-116), czyniąc tym samym to podejście częścią nauk nomotetycznych, jak tych mających na celu odkrywanie praw (Grobler, 2006, s. 250). Innym przykładem obrazującym wykorzystanie metod ilościowych był program Korelaty Wojny (gromadzenie informacji oraz przetwarzanie w celu przewidzenia wystąpienia konfliktu), jednak i w tym przypadku nie można mówić o sukcesie ze względu na zmiany jakie zachodziły w badanej strukturze oraz ograniczonym dostępie do danych i możliwości ich przetwarzania (Sułek, 2010, s. 112). Podobny problem zauważalny jest w Naukach o Polityce (dalej: NOP), gdzie narracje dotyczące metodologii ukształtowały dwie przeciwstawne tendencje. Pierwsza, znajduje wyraz w upodobnianiu NOP do nauk przyrodniczych, gdzie szczególny nacisk położony jest na obiektywny opis oraz prawa. Drugi natomiast, kształtuje NOP w kierunku normatywnym opartym na wartościach i etyce. Oznacza to, że w nauce – szeroko pojętej humanistyce – można wskazać stały, niezmienny problem, który dzieli podejścia na te oparte na analityce oraz wartościowaniu, gdzie źródłem jest głównie interpretacja historii.

Przenosząc znaczący wzrost ilości informacji do problemu podwójnej narracji występującej w naukach w kontekście mieszania metodologii MRM (ang. *Mixed Research Methodology*), należy jedynie zwrócić uwagę na zmiany, jakie być może wywoła *Big Data* w świecie nauki. Oznacza to, że analizowanie obecnie dostępnych danych daje możliwość prawidłowego wnioskowania, które już nie tylko prowadzi do generalnych praw probabilistycznych, ale również do ustalania indywidualnych osobliwości, tak jak to jest w przypadku nauk idiograficznych.

Społeczeństwo stworzyło potężne komputery, większość ludzi w państwach rozwiniętych posiada telefon komórkowy (smartfon) z połączeniem do Internetu. Dziś zjawisko gromadzenia tych wszystkich informacji określamy z języka angielskiego, jako *Big Data*, czyli potężne zbiory informacji o dużej objętości, różnorodności i zmienności. W związku z tym, że prawie każdy pozostawia w sieci informacje na swój temat, świadomie bądź nieświadomie, rodzi się pytanie o możliwości wykorzystania tych informacji. Jednak, aby odpowiednio podejść do analizy kampanii wyborczej D. Trumpa, należy w pierwszej kolejności usystematyzować oraz wyjaśnić pojęcie *Big Data*.

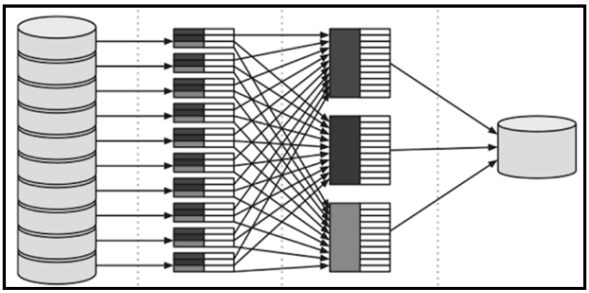
Czym Jest Big Data?

Na pytania, *„(...) jaki kolor nadwozia powinno się wybrać, kupując używany samochód, jeżeli chce się zwiększyć prawdopodobieństwo nabycia samochodu w dobrym stanie? W jaki sposób władze mogą zlokalizować najbardziej niebezpieczne studzienki w Nowym Jorku, żeby zapobiec ich eksplozjom? Jak według firmy Google będzie rozprzestrzeniała się epidemia wirusa grypy H1N1?"* (Mayer-Schonberger, Cukier, 2013, s. 1-5). Jaki przekaz należy kierować i o której godzinie do wyborców, aby wygrać wybory prezydenckie w USA? Na te i inne pytania odpowiada termin *Big Data*.

Co było początkiem ery *Big Data*? Czy można jednoznacznie wskazać okres, wynalazek lub technikę, która przyczyniła się do omawianego zjawiska? Otóż tak, przyjmuje się, że genezą jest decyzja firmy Google o utworzeniu nowego sposobu przechowywania danych ze względu na niskie możliwości ówczesnych serwerowni oraz wysokich kosztów przechowywania danych. Pomysł o stworzeniu nowej metody przechowywania danych pozwolił odnieść sukces na skale globalną wspomnianej firmie. Dokładnie w połowie lat 90. XX wieku powstały pierwsze zarysy koncepcji tego, co dzisiaj nazywa się *Big Data*. Firma Google, aby podnieść jakość świadczonych usług postanowiła zaindeksować treść stron www potrzebowała niewyobrażalnych, jak na tamte czasy, zasobów i przestrzeni serwerowej, a na drodze stały koszty oraz możliwości techniczne serwerów. W wyniku tej decyzji powstał GFS (ang. *Google File System*), który wykorzystywał koncepcję przetwarzania i przechowywania danych w systemie rozproszonym DFS (ang. *Distributed File System*) oraz skuteczny mechanizm dostępu do danych, czyli Map-Reduce (Ghemawat, Gobioff, Leung, 2003, s. 1-4). Dzięki zastosowanej metodzie osiągnięto znaczną redukcje

kosztów, zwiększono przestrzeń dyskową czy możliwość równoległych operacji przetwarzania danych.

Schemat 2 Uproszczony model DFS



Źródło: Opracowanie własne.

Kolejnym krokiem, ku dzisiejszej *Big Data*, była decyzja organizacji Apache o powołaniu projektu stworzenia analogicznego rozwiązania, jednak objętego licencją typu open source. Decyzja ta była odpowiedzią na nowa metodę zastosowaną przed firmę Google. Dziś rozwiązanie to jest jednym z najczęściej spotykanym sygnowanym logo Apache. Hadoop, ponieważ tak nazywa się rozwiązanie organizacji Apache, jest systemem rozproszonego przechowywania i przetwarzania plików. Sposób działania opiera się rozpraszaniu danych na wielu serwerach i dzielenie na bloki, które zostają rozdystrybuowane pomiędzy węzłami. Dzięki zastosowaniu takiej metody dane maksymalnie optymalizują się w kwestii przechowywania. Metadane, inaczej tzw. mapa, pozwalająca uzyskać dostęp do określonego fragmentu pliku, przechowywane są w pamięci operacyjnej serwera NameNode. Natomiast system rozpraszania pików opracowany przez organizację Apache nazywa się rozproszonym systemem plików HDFS (ang. *Hadoop Distributed File System*) (White, 2012, s. 15-17).

Chcąc dokładnie przeanalizować *Big Data* należy, w pierwszej kolejności zrozumieć czym to pojęcie jest, tj. dokonać próby definicji. Już na samym początku napotkać można na poważny problem. Otóż, uznając termin *Big Data*, jako nowe zjawisko przetwarzania danych, należy wskazać na podstawowe różnice między tzw. starymi metodami przetwarzania a nowymi. Termin *Big Data* pierwszy raz został użyty na początku lat 2000, kiedy analityk Doug Laney przedstawił rozpowszechnioną już dziś definicję, według której dane masowe to tzw. 3V: volume (czyli ilość), velocity (czyli szybkość) i variety (czyli różnorodność) (Douglas, 2001, s. 1-4).

- ilość – zbieranie dane z różnorodnych źródeł: transakcje biznesowe, media społecznościowe, dane z sensorów, dane wymieniane między urządzeniami. W przeszłości przechowywanie tych danych stanowiło problem, ale obecnie nowe technologie (takie jak np. Hadoop) znacznie to ułatwiły.
- szybkość – dane powstają i są dostarczane niezwykle szybko i muszą być obsługiwane z odpowiednim reżimem czasowym. Czas analizy danych uznaje się zbliżony do rzeczywistego.
- różnorodność – dane ściągane są w różnych formatach, od ustrukturyzowanych, numerycznych danych w tradycyjnych bazach danych do niestrukturalnych dokumentów tekstowych, email, video, audio, danych znaczników magazynowych lub transakcji finansowych.

Obecnie wskazuje się na model 4V, w ramach którego wyróżnia się również weryfikację posiadanych danych. Której celem jest wyciąganie wniosków. Natomiast największą zaletą Big Data jest możliwość przetwarzania danych nieustrukturyzowanych (m.in. obrazy cyfrowe, pliki wideo, pliki audio, posty z portali społecznościowych, poczta e-mail, pliki programu Word, arkusze kalkulacyjne, pliki PDF) (Chen, Mao, Liu, 2014, 1-39).

Tabela 2. Przetwarzanie danych w Big Data a tradycyjne metody

Charakterystyka	Big Data	Tradycyjne metody
Przetwarzanie danych ustrukturyzowanych	+	+
Przetwarzanie danych nieustrukturyzowanych	+	+
Szybkość przetwarzania w czasie rzeczywistym	+	-
Niskie koszty przetwarzania danych	+	-
Brak usystematyzowanej struktury	+	-
Oszczędność miejsca na serwerach	+	-

Źródło: Opracowanie własne.

Po przeanalizowaniu powyższych informacji można wskazać kilka zmiennych, które mają kluczową rolę w kwestii odróżnienia *Big Data* od tradycyjnych metod zbierania i analizy danych. Do najważniejszych z nich należą: możliwość przetwarzania w czasie rzeczywistym (bądź zbliżonym do rzeczywistego), brak usystematyzowanej struktury co związane jest z dzieleniem i przechowywaniem danych w wielu bazach oraz najważniejsza, to niskie koszty obsługi oraz oszczędność miejsca na serwerach.

Podsumowując czym jest *Big Data*, można śmiało stwierdzić, że jest to system na bazie którego możliwe jest pobieranie, przechowywanie bardzo dużych ilości danych zarówno ustrukturyzowanych, jaki i nieustrukturyzowanych (np. informacje na portalach społecznościowych), dający możliwość szybkiej analizy informacji w czasie zbliżonym do rzeczywistego, która prowadzi do odkrywczych związków, które dotychczas wydawały się nieoczywiste.

Nasuwa się zatem pytanie o możliwości wykorzystania tego potencjału przez np. polityków w celu kształtowania polityki nastawionej na realizację zrównoważonego rozwoju bądź dochodzeniu do władzy. W dalszej części artykuł autor analizuje wykorzystanie Big Data podczas kampanii prezydenckiej D. Trumpa.

Big Data w służbie politykom

Przeglądając portale internetowe w celu uzyskania informacji dotyczącej wyboru D. Trumpa na prezydenta, często można natrafić na nagłówki typu: *„Donald Trump, kandydat Partii Republikańskiej na prezydenta USA, a także biznesmen i telewizyjny celebryta bez politycznego dorobku i doświadczenia, niespodziewanie wygrał wybory prezydenckie, pokonując byłą sekretarz stanu demokratkę Hillary Clinton"* (Portal tvn24). To co łączy większość komentarzy, to zamiennie używane stwierdzenie „niespodziewanie". Jednak, gdy przyjrzeć się bliżej kulisom kampanii wyborczej można natrafić na informację dotyczące firmy Cambridge Analytica (Hunter, 2016). Sama firma, jak informuje na swojej stronie internetowej wyznaje konkretną filozofię: *„W Cambridge Analytica rozumiemy, że każdy klient, przypadek, czy kampania jest wyjątkowa. Dlatego pomożemy Ci połączyć się z każdym członkiem grupy docelowej na poziomie indywidualnym, w taki sposób, aby angażować, informować i kierować do działania. Łączymy 25-letnie doświadczenie w zakresie badań zachowań, pionierskiej analizy danych oraz najnowocześniejszej technologii […]*" (Cambridge Analytica).

Skąd zainteresowanie kandydata na prezydenta zachowaniem wyborców, skoro można je określić w typowych sondażach? Otóż 9 września 2016 r. odbyła się konferencja, tuż przed wyborami w Stanach Zjednoczonych. Prelegentem był Alexander Nix, prezes firmy Cambridge Analytica, który stwierdził, że jego firma jest w stanie określić osobowość każdego dorosłego w kraju (Hunter, 2016). Dzięki użyciu metody OCEAN, która obejmuje pięć czynników: neurotyczność, ekstrawersję, otwartość na doświadczenie, ugodowość i sumienność.

Pytanie skąd pobierała dane jest stosunkowo prosta, kupowała dane z list wyborców, prenumerat czasopism, danych medycznych, wypisów z ksiąg wieczystych oraz portali społecznościowych. Po przeanalizowaniu uzyskanych danych metodą OCEAN stało się możliwe określenie indywidualnych cech wyborców do których D. Trump mógł skierować hasła wyborcze. Głównym kanałem był Facebook, jak stwierdził dyrektor reklamy Komitetu Wyborczego Republikanów Gary Coby: *„Facebook okazał się skutecznym narzędziem dla zespołu Trumpa"* (Lapowsky, 2016). Dzięki temu portalowi stało się możliwe kierowanie różnorodnych komunikatów, które w trzecim dniu debaty prezydenckiej osiągnęły 170 000 wariacji, równocześnie portal dał możliwość analizy zachowania wyborców na komunikaty dzięki czemu można było dokonywać odpowiednich korekt treści czy formy (Lapowsky, 2016).

Jak się okazuje nie tylko osiągnięcia Cambridge Analytica przy użyciu metody OCEAN do analizy danych miały wpływ na wybory. Swój wkład miał również polski naukowiec Michał Kosiński pracujący obecnie na Uniwersytecie Stanforda. Jak można przeczytać w wywiadzie dla gazeta.pl: *„Do polskiego naukowca z uniwersytetu Stanforda dzwonią firmy i ludzie, proponując mu pracę. Wszystko dlatego, że jego odkrycie może się przysłużyć nie tylko zdobyciu wiedzy, ale także pieniędzy i władzy"* (Gostkiewicz, 2017). Polski naukowiec opracował metodę analizy dużych ilości informacji, dzięki której można określić *„preferencje seksualne (u mężczyzn skutecznie w 88 proc. przypadków), wygląd, zainteresowania, poziom inteligencji, pochodzenie etniczne i kolor skóry (u Amerykanów skutecznie w 95 proc. przypadków), wyznanie, poziom zadowolenia z życia, uzależnienia, wiek, płeć oraz poglądy społeczne, religijne i polityczne (te ostatnie w USA w 85 proc. przypadków) z dokładnością do jednej osoby"* (Gostkiewicz, 2017). Idąc dalej, można się dowiedzieć, że u podstaw badania M. Kosińskiego legła metoda OEAN. W 2014 r. M. Kosiński otrzymał propozycję doradztwa firmie Strategic Communications Laboratories w stworzeniu modeli do analizy 10 milionów profili Amerykanów na Facebooku, jak sam zaznaczył „odmówił, bo zaniepokoiło go, że firma miała specjalizować się we wpływaniu na wybory" (Gostkiewicz, 2017). Dziennikarze „Das Magazin", powołując się na posiadane dokumenty, piszą, że *„Strategic Communications Laboratories okazała się potem spółką-matką Cambridge Analytica"* (Bamford, 2016), a metoda M. Kosińskiego została poznana poprzez tego samego naukowca, który proponował mu transakcję w Strategic Communica-

tions Laboratories. Z łatwością wywnioskować możne, że w kampanii D. Trumpa znalazła zastosowanie metoda badań identyczna albo przynajmniej bardzo podobna do tej stworzonej przez M. Kosińskiego (Gostkiewicz, 2017).

Na czym dokładniej polegało badanie M. Kosińskiego? Jak można się dowiedzieć z artykułu pt. Prywatne cechy i atrybuty są przewidywalne z rejestrów cyfrowych ludzkich zachowań (ang. *Private traits and attributes are predictable from digital records of human behawior*). Po pierwsze, istnieją znaczące psychologicznie połączenia między osobowościami użytkowników, ich preferencjami dotyczącymi stron internetowych i cechami profili na Facebooku (Kosinski, Stillwella, Graepelb, 2013). Po drugie, osobowość konkretnego człowieka może zostać określona na podstawie cech jego profilu na Facebooku, i że maszyna robi to lepiej niż człowiek (Youyoua, Kosinski, Stillwella, 2014). A to wszystko dzięki zastosowaniu odpowiednich algorytmów oraz tzn. uczeniu maszynowemu (ang. *machine learining*) w analizie Big Data w celu wykrycia nieznanych prawidłowości w danych i możliwości przewidywania (Kosinski, Wang, Lakkaraju, Leskovec, 2016). Problemem nie jest również dostęp do danych. W swoim artykule M. Kosiński zauważa, że korzystał z dostępnych informacji za pośrednictwem portalu myPersonality project (http://mypersonality.org), które następnie podzielił na trzy próbki. Popularna aplikacja działająca w ramach portalu Facebook umożliwiała uczestnikom przeprowadzenie testu psychologicznego oraz uzyskanie gotowego raportu. W przeprowadzonym badaniu średni wiek uczestników wyniósł 24,1 lat, kobiety stanowiły 61,1% a mężczyźni 38,9% badanych (Youyou, Schwartz, Stillwell, Kosinski, 2016, s. 1). Pierwsza próbka została wykorzystana do budowy modelu oceny osobowości oparta na tzw. lubię to (Likes). Zawierała 295,320 informacji uczestników, którzy wypełnili kwestionariusze osobowości i posiadali co najmniej 20 polubień na swoim profilu. Druga próbka została wykorzystana do opracowania modelu oceny osobowości opartych na języku. Zawierała 59,547 uczestników, którzy wypełnili kwestionariusze osobowości i napisał co najmniej 500 słów we wszystkich aktualizacjach statusu na profilu. Trzecia próbka została wykorzystana do zbadania istnienia podobieństwa osobowości osób przebywających w związkach oraz przyjaciółmi. Zawierała ona 247,773 jednostek tworzących łącznie 5,042 heteroseksualnych związków oraz 138,553 dwóch przyjaciół. Osoby w związkach zostały zidentyfikowane przy użyciu informacji profilowej Facebook „stan cywilny". Przyjacielskie połączenia zostały zidentyfikowane przy

użyciu listy znajomych Facebook (Youyou, Schwartz, Stillwell, Kosinski, 2016, s. 4). Zastosowanie takiego podziału pozwoliło na zidentyfikowanie stopnia podobieństwa osobowości osób będących w związkach oraz między będącymi znajomymi, co bezpośrednio łączy się z możliwościami zastosowania metody do wyborów np. prezydenckich.

Powyższy przykład jasno wskazuje możliwości, jakimi obecnie mogą dysponować politycy w celu zidentyfikowania oraz dotarcia nawet do pojedynczego wyborcy. Identyfikacja ludzkich potrzeb, zainteresowań itd., daje możliwości dotychczas nieznane. Analizując zjawisko nasnuwają się kolejne pytania dotyczące kwestii etycznych. Jak bronią się szefowie firm konsultingowych czy badacze doradzający politykom? Najczęściej odwojują się do racjonalnego wyboru każdego wyborcy, bo przecież na samym końcu to każdy indywidualnie podejmuje decyzję. Jednak jedno jest pewne, każdy nowy kandydat na prezydenta USA będzie zmuszony wykorzystać potęgę jaką daje analiza danych Big Data, jeśli chce zwyciężyć kolejne wybory prezydenckie.

Podsumowanie

Obecnie większość (szczeólnie w Państwach rozwiniętych) ludzi posiada indywidualny Tracker – telefon komórkowy oraz dostęp do Internetu[2]. Codziennie pozostawiane są duże ilości informacji, przy dużych zbiorach nieustrukturyzowanych danych rośnie prawdopodobieństwo określenia prawidłowego profilu psychologicznego oraz naszych zainteresowań. Wnioski płynące z próby definicji analizowanego zjawiska pokazują, że próba budowy możliwych scenariuszy jest niezwykle trudna, ponieważ brak jest empirycznych przesłanek dotyczących wyborów czy polityki w ogóle na bazie których można byłoby wnioskować. Jedyne dane na których dokonać można wstępnych analiz dotyczą wykorzystania tejże technologii w celach komercyjnych. Przykładem może służyć firma Google, która stworzyła projekt *Google Flu Trends* w celu ustalania możliwość rozwoju oraz rozprzestrzeniania się chorób np. wirusa grypy na podstawie sposobu, w jaki chorzy szu-

[2] Zgodnie z raportem 2016 Digital w skali globalnej, to: 3,42 mld użytkowników internetu, czyli 46% globalnej penetracji, 2,31 mld użytkowników mediów społecznych – 31% globalnej penetracji, 3,79 mld unikalnych użytkowników telefonów komórkowych, co stanowi 51% globalnej penetracji, 1,97 mld użytkowników mobilnych mediów społecznościowych, co równa się 27% globalnej penetracji, (https://www.slideshare.net/wearesocialsg/digital-in-2016).

kają informacji w przeglądarce. Dokładność tej metody sięga adresu IP, tj. konkretnego użytkownika.

Przedstawione powyżej przesłanki oznaczają, że rewolucja *Big Data* to nie tylko zmiana w postrzeganiu wyborów prezydenckich i innych, to również zmiana w postrzeganiu nauki. Dotychczas występująca narracja, dzieląca naukowców odwołujących się do metod ilościowych odpowiadających głównie a pytania (kto, ile), a jakościowych (dlaczego) będzie ulegała zmianie - niekoniecznie na lepsze. Przetwarzanie ogromnych ilości danych daje możliwość już nie tylko ustalaniu praw probabilistycznych, ale również indywidualnych cech w stosunku do jednej osoby oraz ustalanie portretu psychologicznego tłumaczącego zachowania badanej jednostki. Co znacząco wpływa na postrzeganie życia politycznego. Przykładem może służyć książka autorstwa Cathy O'Neil pt. *Weapons of Math Destruction: How Big Data Increases Inequality And Threatens Democracy,* w której analizuje czy algorytmy mogą uczynić świat bardziej sprawiedliwy, poprzez osądzanie każdego człowieka według tych samych zasad. Autorka wskazuje krytyczna postawę wobec takich rozwiązań. Twierdzi, że obecnie stosowane modele są nieprzezroczyste, nieuregulowane i sporne, ponieważ mogą wzmocnić proces dyskryminacji. Posługuje się przykładem biednego studenta, który nie może uzyskać kredytu na studia, ponieważ jego kod pocztowy pokazuje, że jest to zbyt ryzykowne dla banku. W następstwie zostaje odcięty od możliwości dalszej edukacji, która mogłaby go wyciągnąć z ubóstwa.

Kolejnym przykładem odnoszącym się do analizowanego zjawiska, jest książka autorstwa Hyunjoung Lee i Il Sohn pt. *Big Data w przemyśle. Jak wykorzystać analizę danych do optymalizacji kosztów procesów?* Obrazuje możliwości, jakie możliwe są już do zastosowania w części przedsiębiorstw czy firm produkcyjnych. Główna teza książki dotyczy, że zgodnie z badaniami firmy Gartner aż 80% procesów biznesowych w firmach będzie oparte na *Big Data* w 2020 r. co oznacza redukcje pracowników.

Monografii, raportów czy innych artykułów dotyczących zjawiska *Big Data* w zdrowiu (analiza rozpowszechniania się wirusa grypy, bądź zarządzenie szpitalem), edukacji (zmiana modelu nauczania oraz metodologii), innowacje (wydobycie ropy naftowej) czy bezpieczeństwie narodowym (bezpieczeństwo cybernetyczne) można znaleźć coraz więcej. Tendencja ta pokazuje coraz większe zainteresowanie analizowanym zjawiskiem. Pomijając korzyści jakie niesie za sobą, należy również rozważyć nowe, niebezpieczniejsze zagrożenia, poczynając od

końca prywatności jednostek, a kończąc na bezpieczeństwie narodowym.

Cytując M. Kosińskiego *„W ciągu kilkunastu lat w sieci setki milionów ludzi chcąc nie chcąc ujawniły i upubliczniły swoje przekonania, poglądy i pragnienia. Naukowiec, który dysponuje taką bazą danych, nie jest już zwykłym naukowcem. Ma do dyspozycji taką wiedzę, że w świecie nauki jest bogiem"* (Gostkiewicz, 2017).

Bibliografia

Chen M., Mao S., Liu Y. (2014), *Big data: a survey*, Mob. Netw. Appl. 19 (2)., http://pt.wkhealth.com/pt/re/lwwgateway/landingpage.htm;jsessionid=YwPL8R1QWLGTFgfnqJlvqMTZ2VHKVhxCQvcsyWvRSsvHRNBpgkbZ!-2019193196!181195628!8091!-1?sid=WKPTLP:landingpage&an=00134372-201409000-00002 (24.02.2017)

Czaputowicz, J. (2008), *Teorie stosunków międzynarodowych. Krytyka i systematyzacja*, Wydawnictwo Naukowe PWN, Warszawa

Dembiński B. (2013). *Późny Platon i Stara Akademia*, Derewiecki, Warszawa.

Douglas L. (2001), *3D Data Management: Controlling Data Volume, Velocity and Variety*, Meta Group, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (24.02.2017)

Ghemawat, S., Gobioff, H., Leung, S. T. (2003), *The Google file system. Proceedings of the nineteenth ACM Symposium on Operating Systems Principles*, ACM Press.

Gostkiewicz M. P*o zwycięstwie Trumpa i Brexicie Michal Kosinski zaczął odbierać maile z oskarżeniami: „To twoja wina". I nie bez powodu*, http://weekend.gazeta.pl/weekend/1,152121,21287773,po-zwyciestwie-trumpa-i-brexicie-michal-kosinski-zaczal odbierac.html?utm_source=facebook.com&utm_medium=SM&utm_campaign=FB_Gazeta (24.04.2017).

Grobler A. (2006), *Metodologia nauk,* Wydawnictwo Znak, Kraków.

Heller, M. (2011), *Filozofia Nauki*, Wprowadzenie, Petrus, Kraków.

Hermann M., Pentek T., Otto B. (2015), *Design Principles for Industrie 4.0 Scenarios: A Literature Review*, Technische Universität Dortmund

Hunter F. (2017), *Cambridge Analytica, the 'psychographic' data firm behind Donald Trump, eyes Australian move*, (23.01.2017), http://www.smh.com.au/federal-politics/political-news/cambridge-analytica-the-psychographic-data-firm-behind-donald-trump-eyes-australian-move-20161212-gt926e.html (24.02.2017).

Huurdeman A. (2003), T*he Worldwide History of Telecommunications,* John Wiley & Sons, New Jersey.

Kaznowski D. (2008), *Nowy marketing*, VFP Communications, Warszawa.
Konieczny, J. (2016), *Bezpieczeństwo zdrowia publicznego w zagrożeniach środowiskowych*, Wydawnictwo Naukowe WNPiD, Poznań.
Kosinski M., Stillwella D., Graepelb T. (2013), *Private traits and attributes are predictable from digital records of human behavior*, Proceedings of the National Academy of Sciences (PNAS).
Krysiak K. (2005), *Sieci komputerowe*, Helion
Kurzweil R. (2005), *The Singularity is Near*, Penguin Group.
Lapowsky I. (2016), *Here's How Facebook Actually Won Trump the Presidency*, (15.11.2017), https://www.wired.com/2016/11/facebook-won-trump-election-not-just-fake-news/?mbid=social_twitter (wejście: 24.02.2017)
Maciołek, C. (2017), *Zawód przyszłości. Smart data analytics potrzebny od zaraz,* https://www.hbrp.pl/b/zawod-przyszlosci-smart-data-analytics-potrzebny-od-zaraz/waSNriG1?utm_content=buffereabb3&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer (24.02.2017).
Mayer-Schonberger V., Cukier K. (2013), *Big Data: a revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt.
McPherson S. (2009), *Tim Berners-Lee: Inventor of the World Wide Web,*
„USA Today Lifeline Biographies".
Schwab K. (2017), *The Fourth Industrial Revolution*, Crown Publishing Group.
Sułek, M. (2010), *Prognozowanie i symulacje międzynarodowe*, Wydawnictwo Naukowe Scholar, Warszawa.
Vaughan-Williams N. (2005), *International Relations and the „Problem of History"*, Millennium: Journal of International Studies, 2005, vol. 34, nr 1.
White T. (2012), *Hadoop: The Definitive Guide*, „O'Reilly Media.
Youyou W., Kosinski M., Stillwell D. (2015), C*omputer-based personality judgments are more accurate than those made by humans*, Proceedings of the National Academy of Sciences (PNAS).
Youyou W., Schwartz A., Stillwell D., Kosinsk M. (2017), *Birds of a feather do flock together: behavior-based personality assessment method reveals personality similarity among couples and friends*, Psychological Science.

SUMMARY

The aim of this article is to describe and analyze the possibilities of using Big Data techniques (collection and analysis of data) in order

to win the presidential election by Donald Tramp. For this purpose, the author refers to the story of the development of the Internet as key to understanding the phenomenon of an unprecedented increase in the amount of information. As a result of the analyzed concept, changes are being made also in the field of methodology of sciences, especially concerning qualitative and quantitative approaches. Conclusions summarizing the article are theses behind the possibilities but also the dangers of data analysis techniques.

Nota o Autorze

Leonard Dajerling [leonard.dajerling@amu.edu.pl] – absolwent i doktorant Wydziału Nauk Politycznych i Dziennikarstwa UAM. Jego zainteresowania badawcze skupiają się wokół metodologii, logiki, historii nauki, filozofii nauki, *ubiquitous computing*, analizy i wizualizacji ustrukturyzowanych i nieustrukturyzowanych danych w Hadoop, machine learning. Zawodowo programista (język: Python) oraz analityk danych.