ARTYKUŁY POGLĄDOWE (REVIEW PAPERS)

Modele badawcze w ujęciem klasycznym a koncepcja analizy Big Data

(Research models in the classic approach and the Big Data analysis concept)

S Kasza [1,A,D], A Romaszewski [1,E], Z Kopański [1,D,F], W Uracz [2,B,E], F Furmanik [2,C], S Dyl [2,B], J Tabak [2,B]

1. Wydziału Nauk o Zdrowiu Collegium Medicum Uniwersytet Jagielloński
2. Collegium Masoviense – Wyższa Szkoła Nauk o Zdrowiu

Abstract - The authors pointed out that the collection of various data and information has always been an inherent element of human activity. The methods of collecting data that served administrative, tax and military purposes were presented in the past over the centuries. The last ten years have been a period of rapid technological development in which new problems arose, as well as the idea of solving them. New tools and methods of working with data allowed for a high quality jump in every field, including in the health and public health sectors. In the further part of the article, the authors presented differences in research models between the classic approach and the Big Data analysis concept. Elements were also discussed, which decided that modern times were called the information era, in which the acquisition of data radically changed its face. Along with the entry into the general use of digital devices and the dissemination of the Internet, it turned out that there is the possibility of extensive collection and analysis of personal data 24 hours a day. The article highlights the differences between the analysis of large data clusters in the traditional approach and in terms of modern techniques.

Key words - data and information gathering, traditional approach, big data analysis concept.

Streszczenie - Autorzy zwrócili uwagę, że gromadzenie różnorodnych danych i informacji zawsze było nieodłącznym elementem działalności człowieka. Przedstawiono w zarycie na przestrzeni wieków sposoby zbierania danych, które służyły celom administracyjnym, podatkowym i wojskowym. Ostatnie dziesięć lat jest okresem gwałtownego rozwoju technologicznego , w którym powstały nowe problemy, a także idee ich rozwiązywania. Nowe narzędzia oraz metody pracy z danymi pozwoliły na dokonanie dużego skoku jakościowego w każdej dziedzinie, w tym także w sektorze ochrony zdrowia i zdrowia publicznego. W dalszej części artykułu autorzy przedstawili różnice w modelach badawczych pomiędzy ujęciem klasycznym, a koncepcją analizy Big Data. Omówiono także elementy, które zadecydowały, że współczesne czasy nazwano erą informacyjną, w której pozyskiwanie danych diametralnie zmieniło swoje oblicze. Wraz z wejściem do użytku ogólnego urządzeń cyfrowych i rozpowszechnieniem Internetu okazało się, że istnieje możliwość szerokiego zbierania i analizowania danych osobowych przez całą dobę. W artykule podkreślono różnice pomiędzy analizą dużych skupisk danych w podejściu tradycyjnym i w ujęciu nowoczesnych technik.

Słowa kluczowe - gromadzenie danych i informacji, podejście tradycyjne, koncepcja analizy Big Data.

Wkład poszczególnych autorów w powstanie pracy— A-Koncepcja i projekt badania, B-Gromadzenie i/lub zestawianie danych, C-Analiza i interpretacja danych, D-Napisanie artykułu , E-Krytyczne zrecenzowanie artykułu, F-Ostateczne zatwierdzenie artykułu

Adres do korespondencji — Prof. dr Zbigniew Kopański, Wydziału Nauk o Zdrowiu Collegium Medicum Uniwersytet Jagielloński, Kraków, ul. Piotra Michałowskiego 12, PL-31-126 Kraków, e-mail: zkopanski@o2.pl

Zaakceptowano do druku: 29.08.2018.

WSTĘP

Gromadzenie różnorodnych danych i informacji zawsze było nieodłącznym elementem działalności człowieka. Zbieranie i analiza danych pozwalała na snucie wniosków podpartych dowodami. Choć początkowo nie były to zdigitalizowane cyfry na ekranie monitora, a raczej malunki na glinianych tabliczkach, pozwalały dowodzić prawdy o otaczającym nas świecie. Na przestrzeni wieków zbierane dane służyły głównie celą administracyjnym, podatkowym i wojskowym. Władcy narodów byli

świadomi siły wynikającej z dokładnie zbieranych i zapisywanych danych o każdym obywatelu, dlatego spisy powszechne wykorzystywano już w Babilonii czy Egipcie. Szczególny wzrost znaczenia wydobywania wiedzy z powierzonych informacji upowszechnił się w średniowieczu wraz z rozwojem gospodarki towarowopieniężnej i rozpowszechnieniem umiejętności czytania oraz pisania [1].

Pierwsze zastosowania masowej kolekcji danych w dziedzinie zdrowia publicznego miały miejsce w 1854 roku, kiedy fala cholery przetoczyła się przez Londyn. John Snow uważany za ojca współczesnej epidemiologii potrafił z zebranych danych na temat zarażonych domostw wysnuć wnioski, że lokalna studnia na Broad Street jest odpowiedzialna za występowanie nowych przypadków choroby. Przyczynił się tym samym do całkowitego zamknięcia skażonego obiektu oraz przywrócenia ładu epidemiologicznego w Londynie, bez znajomości przecinkowca cholery - bakterii powodującej epidemię. John Snow poświęcił swojej pracy kilka miesięcy zanim wpadł na rozwiązanie palącego problemu. Dzisiejsze systemy informatycznie i system pozycjonowania GPS mogłyby rozwiązać podobny problem w niespełna godzinę oszczędzając tym samym czasu i pracy, którą ojciec epidemiologii mógłby poświęcić na rozwiązywanie innych problemów. Nie miałby on jednak łatwego zadania, bowiem współcześnie liczba generowanych danych przekracza możliwości ich analizy. Zakres informacji, który jest zbyt obszerny i skomplikowany do przetwarzania tradycyjnymi modelami został określony mianem Big Data. Wielkość współczesnych wolumenów danych gromadzonych cyfrowo sprawia, że sposób wydobywania wiedzy wymaga nowego podejścia badawczego. Ostatnie dziesięć lat to dekada gwałtownego rozwoju technologicznego podczas, której narodziły się nowe problemy, a także idee ich rozwiązywania. Nowe narzędzia oraz metody pracy z danymi pozwalają na dokonanie dużego skoku jakościowego w każdej dziedzinie, także w sektorze ochrony zdrowia i zdrowia publicznego [2].

BIG DATA

Istnieją zauważalne różnice w modelach badawczych pomiędzy ujęciem klasycznym, a koncepcją analizy Big Data. Tradycyjne modele opierają się na stawianiu hipotez na podstawiepierwszych obserwacji. Akt ten zostaje dokonany jeszcze przed zebraniem i sklasyfikowaniem odpowiednich danych. Samo gromadzenie faktów odbywa się również w sposób często analogowy. Naukowcy zwykle stosują podejście tradycyjne w postaci „kartki i ołówka”, a co za tym idzie ankiet, kwestionariuszy i fizycznych pomiarów, w celu zebrania potrzebnych danych. Choć nieobce są im również nowoczesne technologie, które usprawniają cały proces to model ich przetwarzania pozostaje wciąż taki sam, a operacje na danych są często uproszczone. Postawione hipotezy za pomocą przetwarzania zgromadzonych informacji próbuje się potwierdzić lub obalić. Z kolei samo badanie może mieć charakter pełny (cała populacja) lub częściowy. Najczęściej z uwagi na trudność przeprowadzenia pełnego badania (np. ze względu na koszty czy czas), prowadzi się jego drugi typ. W procesie dochodzenia do wyników, pobiera się odpowiednią próbę losową z badanej populacji, z której z kolei po weryfikacji hipotezy, wysuwa się wnioski.. Próbki muszą mieć charakter reprezentacyjny, jednak nie zawsze jest to możliwe, a zdarza się, że naukowcy starają się naginać fakty dobierając stany potwierdzające lub zaprzeczające hipotezą jakie postawili. Zdarzają się też próby oszustwa i manipulacji wynikami badań. Istnieje także liczna grupa błędów metodologicznych, których autorzy badań muszą się wystrzegać, aby uzyskane wyniki byłby wiarygodne i możliwe do powtórzenia w innym punkcie badawczym, co jest podstawą współczesnej nauki [3].

ERA INFORMATYCZNA

Współcześnie wraz z nastaniem nowej ery określanej mianem informacyjnej, pozyskiwanie danych diametralnie zmieniło swoje oblicze. Wraz z wejściem do użytku ogólnego urządzeń cyfrowych i rozpowszechnieniem Internetu okazało się, że istnieje możliwość zbierania i analizowania danych osobowych przez całą dobę.[4] Ciągły napływ informacji zrodził nowe problemy, wobec których klasyczne procedury okazują się bezsilne. Dzisiejszy fenomen kwantyfikowania i zamiany wszelkich elementów rzeczywistości na dane cyfrowe, w celu ich późniejszej analizy nosi nazwę danetyzacji [5]. Termin został wykuty przez Viktora Mayer-Schonbergera i Kenneth Cukiera, dwóch naukowców zajmujących się tematem Big Data. W związku z rosnącą potrzebą analizy zapisanych na dyskach komputerów informacji zaczęto tworzyć nowe matematyczno-statystyczne metody i algorytmy, które umożliwiłyby wydobycie nieznanych wzorców ze skupisk faktów. Wysoce zaawansowane metody

wykraczają poza tradycyjne modele statystyczne. W celu ich aplikacji, wykorzystuje się systemy o dużej mocy obliczeniowej. Kluczową różnicą pomiędzy analizą dużych skupisk danych, a podejściem tradycyjnym, jest brak uprzednio postawionej hipotezy, która następnie podlega weryfikacji. „Nowe metody badawcze przekładają nacisk z poszukiwania przyczynowości na analizę korelacji" [3]. Bez znajomości podstaw i przyczyn zjawiska. Zastępują pytanie dlaczego? pytaniem co?. Obecnie przyczynowość spadła z piedestału głównego zainteresowania, ale z niej nie zrezygnowano (tabela 1).

Tabela 1. Różnice pomiędzy tradycyjnym, a nowym podejściem analizy danych Opracowane własne na podstawie [3-5]

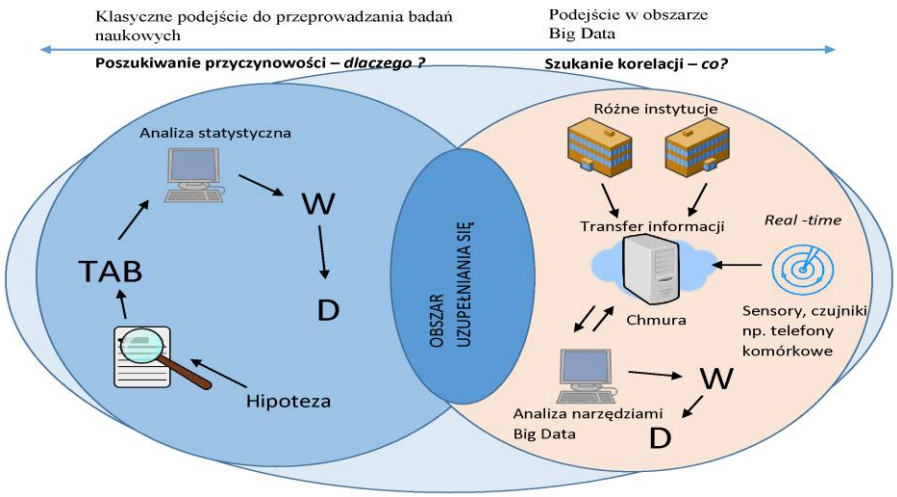
Sposób analizowania danych	Opis metody
Tradycyjny	Analogowa metoda gromadzenia danych; stawianie hipotezy przed zebraniem informacji; poszukiwanie przyczynowości i odpowiedzi na pytania dlaczego?; zależności liniowe
Kontekst Big Data	Masowe gromadzenie informacji z wielu źródeł, wykorzystanie nowoczesnych algorytmów, nacisk na analizę korelacji i odpowiedzi na pytanie co?; zależności nieliniowe, analiza real-time

Koncepcja Big Data pozwala na wielowątkową analizę danych pochodzących z różnych źródeł, której głównym przedmiotem zainteresowania jest człowiek. Metoda w głównej mierze opiera się na przewidywaniu ludzkich zachowań oraz środowiska wokół. Każdy z nas posiada skończoną liczbę reakcji, które można zapisać i scharakteryzować, dotyczy to także środowiska naturalnego. Określenie działań jest możliwe poprzez sygnały, które ludzie za sprawą różnych czynności wysyłają świadomie lub nieświadomie. Istnieją jednak ograniczenia z uwagi na dużą losowość i wypadkowość ludzkich poczynań. W literaturze pojęciowej wszystkie powstałe zakłócenia i przeciwności opisuje się mianem szumu (ang. noise). W chwili obecnej nie scharakteryzowano jeszcze granicy przewidywalności, ponieważ współczesne badania na temat określania zachowań widnieją jeszcze w początkowej fazie badawczej, ale istnieją już pierwsze przykłady ich zastosowań. Do tego zjawiska odnosi się na przykład tzw. „ciąża w markecie". W jednej z sieci amerykańskich supermarketów regularne zakupy robiła cała rodzina. W pewnym momencie ojciec rodu zaczął się zastanawiać, dlaczego do swojej skrzynki na listy od pewnego czasu zaczął otrzymywać reklamy produktów dla niemowlaków. Poirytowany tym faktem zarządzał wyjaśnień od sieci supermarketów, w której robili zakupy, a której ulotki otrzymywał ojciec. Wyjaśnienie było nie tylko zaskakujące dla samego ojca, ale i osób postronnych. Otóż córka, która robiła zakupy w tym hipermarkecie zaczęła kupować szampony i inne środki higieny osobistej, które posiadały delikatniejszy zapach niż dotychczas. System marketingowy sklepu zauważył tą różnice i połączył z faktem, że kobietom, które zachodzą w ciąże wyostrza się węch, z tego powodu wybierają delikatniejsze zapachy kosmetyków. W przedstawionej sytuacji system komputerowy szybciej wiedział o ciąży nastolatki, a niżeli sam ojciec, który nie został w tym czasie jeszcze poinformowany. Kwintesencją analizy dużych ilości danych jest porównywanie informacji zgormadzonych przez system z innymi osobami w grupie. Ogólnie rzecz ujmując im większa próba wykorzystywana do analizy tym większe szanse na odnalezienie określonych wzorców. Pojedyncze informacje behawioralne nie są w stanie zwizualizować konkretnych zależności pomiędzy zrachowaniami jednostek, a ich przyszłymi celami, czy zdarzeniami [6,7].

Dane osobowe zbierane przy pomocy współczesnych technologii mogą być wykorzystywane w każdej dziedzinie naszego życia. Niektóre odkrycia mogą być niezwykle zaskakujące i mogą prowadzić do nowych niespotykanych wcześniej wzorców lub podważać utarte przekonania. Zupełnym zaskoczeniem dla opinii publicznej były badania przeprowadzone prze duńskich naukowców nad korelacją pomiędzy występowaniem raka centralnego układu nerwowego. Okazało się, że dokonana analiza dużych danych wykazała, że taka zależność nie występuje. Dojście do otrzymanych rezultatów było możliwe głównie za sprawą łączenia istniejących danych oraz wykorzystania nowoczesnych analiz Big Data. Badana grupa obejmowała niemalże całą populację, a było to możliwe ze względu na skrupulatność Duńczyków w prowadzeniu rejestrów osób chorujących na raka, rejestrów poziomu wykształcenia i dochodów oraz danych udostępnionych o klientach przez operatorów komórkowych. Mimo skali badań, wszystkie dane zostały precyzyjnie uporządkowane, a wnioski płynące z wyników przyczyniły się do uzyskania bardzo cennej wiedzy

[5]. Zasadniczą cechą analizy wielkich baz danych jest też ich nie intuicyjność i pozorna nielogiczność obserwacji. Na poparcie tego twierdzenia należy przytoczyć badanie przeprowadzone przez Viktora Mayera Schoenbergera z Toronto na temat zakażeń u wcześniaków 24 godziny przed wystąpieniem objawów. Opracowane badanie przez profesora polegało na zbieraniu 1000 oznak życiowych na sekundę niemowlęcia przez superszybkie kamery. Paradoksalnie zapowiedzią zbliżającego się zakażenia nie było wymknięcie się parametrów życiowych spod kontroli, a ich znormalizowanie.

Niektóre grupy, a także specjaliści w zakresie gromadzenia danych postulują, zatem, aby każdy obywatel miał możliwość indywidualnego dysponowania danymi osobowymi. W ten sposób prócz własnego bezpieczeństwa mógłby osiągnąć również korzyści finansowe poprzez comiesięczną sprzedaż indywidualnych danych wielkim korporacją, które obecnie zbierają lub otrzymują je najczęściej zupełnie za darmo (rycina 1).



Rysunek 1. Różnica pomiędzy obszarem badań klasycznych, a Big Data [opracowanie własne]

Legenda:
Obrazek lupy – Badania terenowe i laboratoryjne, określenie próby
TAB – sporządzanie baz danych/tabel do analizy
W – Prezentacja wiedzy i wyników
D – działania, podejmowanie akcji, kierunków politycznych
Real – time- urządzenia nadające sygnał bez przerwy mogą być wykorzystane do analizy w czasie rzeczywistym z ang. real-time
Obszar uzupełniania się – to zakres, w którym metody są dla siebie kompatybilne. Przykład: zbieranie dużych wolumenów danych w badaniach laboratoryjnych metodami klasycznymi i ich pełna analiza bez określania próby, co przy badaniach molekularnych mogłoby nie dać pożądanego rezultatu

W przypadku zdarzeń medycznych przewidywanie przyszłości jest sprawą wielkiej wagi zdolnej ratować ludzkie życie [cyt. za 5,6].
W nurcie nowego być może przełomowego postępu w zakresie gromadzenia i badania danych, oprócz wielu głosów ekscytacji oraz zainteresowania, pojawiają się także opinie, które sugerują, że nowe techniki pozyskiwania informacji ograniczą naszą swobodę i zostaną wykorzystane głównie przez wielkie korporacje oraz służby wywiadowcze do kontrolowania naszych zachowań i poniekąd także myśli. Służy do tego min. personalizacja ogłoszeń i kreowanie dostępnych możliwości.

Nowy kierunek zmian w zakresie analizowania gromadzonych danych nie stwarza jednak konkurencji dla obecnych metod badawczych. Należy raczej spodziewać się, że dzięki nowym technikom obecne metody jeżeli nie komplementarne do obecnych to będą równolegle kreowały nowe zasoby wiedzy dla ludzkości.

Piśmiennictwo

1. Hołda A., MSR/MSSF w polskiej praktyce gospodarczej. Warszawa; Wyd. C.H. Beck, 2013.
2. Muin J, Ioannidis K, Ioannidis John P A. Big data meets public health. Science 2014; 346:105-110.

3. Mojrowski M. Big Data – Nowa era w analizie danych. [cytowany 20 września 2016]. Adres: http://businessit.pl/blog/big-data-nowa-era-w-analizie-danych
4. Papińska-Kacperek J. Usługi cyfrowe. Perspektywy wdrożenia i akceptacji cyfrowych usług administracji publicznej w Polsce. Łódź; Wyd. Uniwersytetu Łódzkiego, 2013.
5. Mayer-Schonberger V, Cukier K. Big Data. Rewolucja, która zmieni nasze myślenie. Warszawa; Wyd. MT Biznes sp. z.o.o., 2014.
6. Kieft M. Ile warte są nasze dane? [cytowany 20 września 2016]. Adres: https://www.google.com/search?source=hp&ei=ychVXOOHDoGTsA-GYjqog&q=Kieft++M.+Ile+warte+s%C4%85+nasze+dane%3F.+dokument%2C+Holandia%2C+2013&btnK=Szukaj+w+Google&oq=Kieft++M.+Ile+warte+s%C4%85+nasze+dane%3F.+dokument%2C+Holandia%2C+2013&gs_l=psy-ab.3...2366.2366..3993...0.0..0.250.250.2-1.....0....2j1..gws-wiz.....0.wK_y6vzYlPU
7. Duhigg C. How companies learn your secrets?. [cytowany 5 sierpnia 2016]. Adres: http://www.nytimes.com/2012/02/19/magazine/shoppinghabits.html?_r=2&hp=&pagewanted=all