

## ARTYKUŁY POGLĄDOWE (REVIEW PAPERS)

# Problemy analityczne i ograniczenia systemów informatycznych

(Analytical problems and limitations of information systems)

S Kasza<sup>1,A,D</sup>, A Romaszewski<sup>1,E</sup>, Z Kopański<sup>1,D,F</sup>, W Uracz<sup>2,B,E</sup>, F Furmanik<sup>2,C</sup>, S Dyl<sup>2,B</sup>, J Tabak<sup>2,B</sup>

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

**Abstract** - The authors emphasized that the main limitation in the processing of large databases is the ubiquitous "noise", which affects the quality of data and constitutes a serious obstacle on the way to learn reliable correlations and the selection of appropriate parameters. Today, there is a huge concentration of data flooding the source of raw data. The most spectacular example is the Internet, which multiplies its volume with each passing year. The errors formed with the analysis of clusters called jargon referring to large databases, the so-called Big Errors show the first problems in the application of new technology in the health and public health segments. Another serious limitation in making the right decisions based on Big Data is the problem of data unification. The format of file storage varies between organizations, which makes it very difficult and lengthy or completely impossible to analyze available data. Unification of data is important due to many reasons, but the most important seems to be in case of outbreak and coordinated action throughout the country. Another problem is the methodology of analyzing large databases in health care. All these issues are discussed in more detail in the above article.

**Key words** - big databases, limitations in processing.

**Streszczenie** - Autorzy podkreślili, że głównym ograniczeniem w przetwarzaniu wielkich baz danych jest wszechobecny „szum”, który wpływa na jakość danych i stanowi poważną przeszkodę na drodze do poznania wiarygodnych korelacji oraz doborze odpowiednich parametrów. Współcześnie istnieje ogromne skupisko danych zalewających źródła surowych danych. Najbardziej spektakularnym przykładem jest Internet, który z każdym rokiem z wielokrotną swoją objętość. Tworzące się przy analizie skupisk danych błędy nazwane żargonem odnoszącym się do wielkich baz danych tzw. „Big Errors”, wykazują pierwsze problemy w stosowaniu nowej technologii w segmencie zdrowia oraz zdrowia publicznego. Innym poważnym ograniczeniem w podejmowaniu właściwych decyzji bazujących na Big Data jest problem unifikacji danych. Format przechowywania plików różni się między organizacjami, co bardzo utrudnia i wydłuża bądź całkowicie uniemożliwia analizę dostępnych danych. Ujednolicenie danych jest istotne ze względu na wiele powodów, ale najbardziej ważne wydaje się w przypadku wybuchu epidemii i skoordynowanego działania na terenie całego kraju. Kolejnym problemem jest metodologia analizy wielkich baz danych w ochronie zdrowia. Wszystkie te zagadnienia zostają szerzej omówione w powyższych artykule.

**Słowa kluczowe** - wielkie bazy danych, ograniczenia w przetwarzaniu.

**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.

## ROLA „SZUMU”

Głównym ograniczeniem w przetwarzaniu wielkich baz danych jest wszechobecny „szum”, który wpływa na jakość danych i stanowi poważną

przeszkodę na drodze do poznania wiarygodnych korelacji oraz doborze odpowiednich parametrów. Pod pojęciem hasła kryje się ogromne skupisko danych zalewających źródła surowych danych. Najbardziej spektakularnym przykładem jest

Internet, który z każdym rokiem zwielokrotnia swoją objętość. Tworzące się przy analizie skupisk danych błędy nazwane żargonem odnoszącym się do wielkich baz danych tzw. „Big Errors”, wykazują pierwsze problemy w stosowaniu nowej technologii w segmencie zdrowia oraz zdrowia publicznego. [1-3]

W kontekście pomiarów populacyjnych w roku 2013 zachorowalność na grypę w Stanach Zjednoczonych mierzona przez aplikacje badające internet, w szczególności Google Flu, drastycznie, bo ponad 2 krotnie przewyższyła szacunki, w porównaniu do zachorowalności mierzonej standardowymi metodami. Błędów w pomiarze dopatrywano się w podatności algorytmu na medialną panikę, model biznesowy Google oraz sposób wyszukiwania, a także grupy wpływów, które wykorzystują internet na własny użytek. Niestety wraz z rozwojem tej technologii analizy i pomiaru liczba potencjalnych fałszywych alarmów może systematycznie rosnąć, nie tylko w wymiarze zdrowia publicznego. Fałszywe korelacje oraz błędne przekonania w badaniach obserwacyjnych ekologicznych mogą ulec zwielokrotnieniu, a ilość korelacje typu „*kolonie pszczoł miodnych w sposób odwrotnie proporcjonalny korelują z młodocianymi aresztowanymi za marihuanę*”, mogą się mnożyć. Cała siła w Big Data polega na odnajdywaniu współzależności, nie w wykazywaniu czy badane korelacje mają jakikolwiek sens. [4] Znalezienie sygnału odpowiadającego hipotezie to dopiero pierwszy krok w kierunku do finalizacji badań pod przewodnictwem nowych technologii. [5-7]

## UNIFIKACJA DANYCH

Innym poważnym ograniczeniem w podejmowaniu właściwych decyzji bazujących na Big Data jest problem unifikacji danych. Format przechowywania plików różni się od organizacji do organizacji, co bardzo utrudnia i wydłuża bądź całkowicie uniemożliwia analizę dostępnych danych. Ujednolicenie danych jest istotne ze względu na wiele powodów, ale najbardziej ważne wydaje się w przypadku wybuchu epidemii i skoordynowanego działania na terenie całego kraju. Na tle epidemii warto wspomnieć też o problemie współdzielenia się danymi pomiędzy różnymi organizacjami, które niejednokrotnie nie są skore do współpracy z innymi podmiotami. Obawiają się

naruszenia obowiązujących praw oraz konkurencji, która może wykorzystać uzyskanie informacji. Pogarsza to bezpieczeństwo systemu zdrowotnego, a także wykrywalność oszustw.

Po mimo zbliżonych algorytmów i modeli wykorzystywanych w przetwarzaniu wielkich baz danych oraz baz standardowych. Interfejsy użytkownika pomiędzy tymi narzędziami znacząco się różnią. Tradycyjne programy do analiz zdrowotnych przeobraziły się w proste i transparentne narzędzia. Z drugiej strony programy do przeglądu Big Data są w tym momencie bardzo złożone, wymagające programistycznie oraz wymagają wiedzy i umiejętności z kilku dyscyplin. Powodem takiego stanu rzeczy jest pochodzenie pierwotnych aplikacji oraz natura wprowadzanych informacji, która jest chaotyczna, wieloźródłowa i nieczysta, w przeciwieństwie do tradycyjnych narzędzi business intelligence (BI). Narzędzia powstałe do analizy wielkich baz danych zostały zapoczątkowane przez programistów z sektora otwarto źródłowego oprogramowania. Brak przyjazności użytkownikowi, realnego wsparcia oraz narzucana przez wydawców droga rozwoju oprogramowania nie sprzyja wykorzystaniu tej technologii w ochronie zdrowia. Dlatego wielu naukowców postuluje, aby wzmocnić relacje pomiędzy programistami odpowiedzialnymi za programowanie nowych aplikacji, a specjalistami z zakresu zdrowia i zdrowia publicznego. Za odpowiednie rozwiązanie uważa się uproszczenie narzędzi i nadanie im graficznego interfejsu, stworzenie modułów do samodzielnej nauki i podręczników dla osób, które nie miały styczności z językiem programowania, a także utworzenie bogatych programów dla całego zakresu specjalistów w systemie zdrowotnym. Przeciętny pracownik opieki zdrowotnej nie jest naukowcem związanym z informatyką. Za przykład dobrego modelu podaje się The Ebola Response od Centers for Disease Control and Prevention (CDC). Big Data powinno pozwolić naukowcom, jak i również innym specjalistom odsłonić nową wiedzę, która będzie wykorzystywana w praktyce. [8-10]

Ostatecznym problemem, który ogranicza wykorzystanie Big Data na tle systemowym jest brak krajowych hurtowni danych, gdzie wszystkie organizacje z sektora mogłyby wysyłać informacje. Niestety podobny system jest bardzo czasochłonny i kosztowny w realizacji. [11]

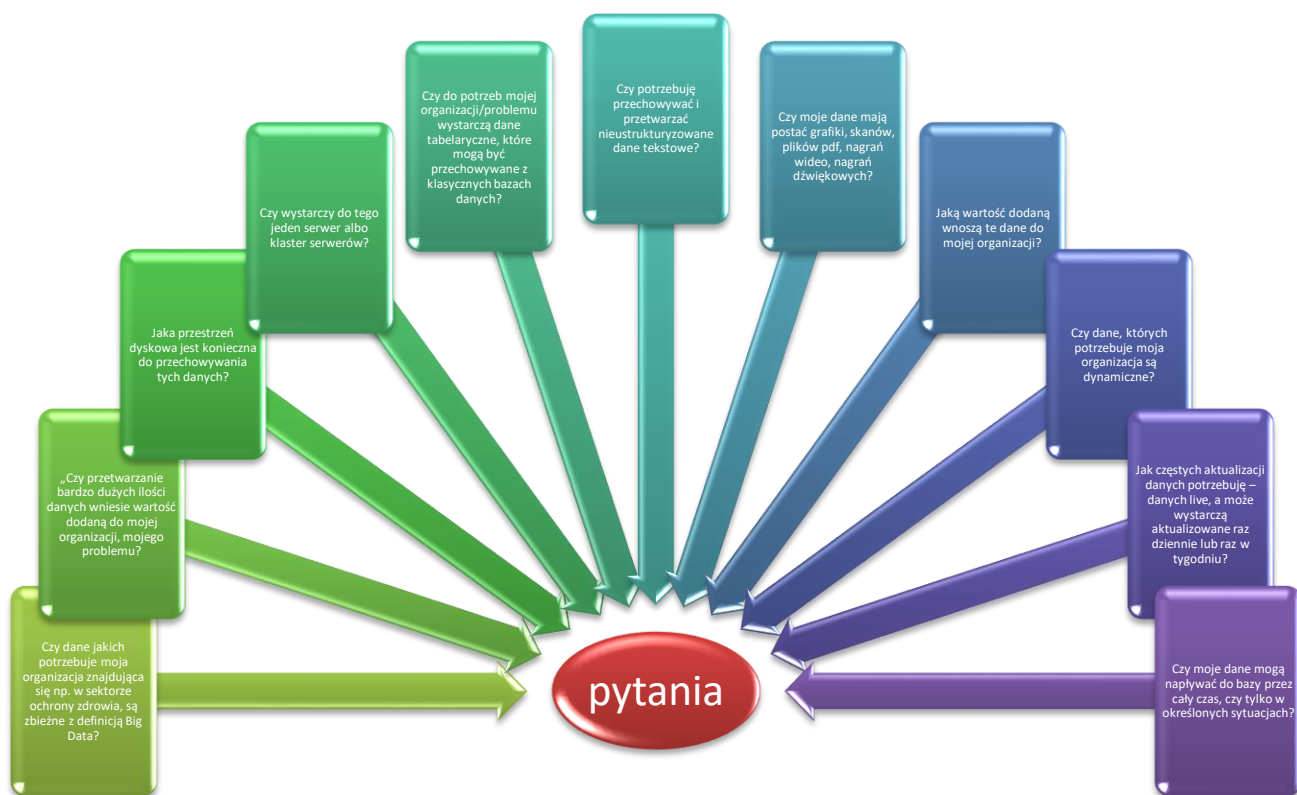
## METODOLOGIA ANALIZY WIELKICH BAZ DANYCH W OCHRONIE ZDROWIA

Analiza Big Data coraz śmielej wkracza na pole związane z ochroną zdrowia. Obszary badawcze zarezerwowane do tej pory dla standardów statystycznych i badań epidemiologicznych oraz klinicznych, zostają skonfrontowane bądź uzupełnione o nowe odkrycia w dziedzinie wielkich baz danych. [11,12]

Zanim zostanie przeprowadzone wnioskowanie oparte na wielkich zbiorach danych w przedsiębiorstwie, organizacji lub badaniu, w sektorze opieki zdrowotnej, należy udzielić odpowiedzi na pytania, które w sposób zasadniczy odpowiedzą na sensowność wdrożenia systemu i zastosowania tego typu analizy informacji.

Opłacalność inwestycji zależy od specyfiki biznesu lub zadania do rozwiązania. Lista 11 pomocniczych pytań została przedstawiona na rycinie 1.

W piśmiennictwie można też wyróżnić metodologię, która bezpośrednio odnosi się do sektora ochrony zdrowia, a jej zastosowanie ułatwia pracę nad problemem wymagającym użycia Big Data. Procedura składa się z 4 głównych kroków, który każdy z nich jest podzielony na kolejne etapy. W punkcie pierwszym interdyscyplinarny zespół określa koncepcje nowego projektu doszukując się potrzeb dla, których to właśnie ta metoda będzie odpowiednia. Wymienia znaczenie projektu biorąc pod uwagę, kompromis pomiędzy skalą, kosztami i innymi ważnymi zmiennymi. Krok drugi przedstawia założenia rozwojowe inicjatywy.



Rycina 1. Pytania dotyczące sensowności wdrożenia systemu i zastosowania tego typu analizy informacji [11,12]

Na tym etapie rozbudowuje się pierwotną koncepcję o dodatkowe odpowiedzi na pytania: Jaki problem jest rozważany?, Dlaczego jest ważny i interesujący? Dlaczego podejście związane z Big Data? (ponieważ praca, koszty i nakłady, które trzeba ponieść są dużo wyższe niż przy stosowaniu tradycyjnych metod).

Grupa projektowa powinna podać również informacje na temat tła problemu oraz wstępne analizy przeprowadzone na tym polu. Etap 3 szczegółowa metodologia działania. Przedstawione zadania to jedynie propozycje i nie są ścisłym przewodnikiem analizy Big Data, a jedynie sugestiami, które warto wziąć pod uwagę. Istotnym znaczeniem dla całego projektu będzie wybór narzędzi oraz metod algorytmicznych. W ostatnim punkcie odkrycia zostają ewaluowane oraz zweryfikowane, a także przedstawione interesariuszom, którzy wykorzystają otrzymane dane w celu podjęcia odpowiednich decyzji. Na każdym przedstawionym kroku powinien istnieć też system raportujący informacje zwrotne do grupy projektowej, aby uniknąć pomyłek. [2,8,11,12]

Tabela 1. Zarys analizy Big Data w ochronie zdrowia - metodologia. Opracowanie własne na podstawie [2,8,11]

|               |                                                                                                      |
|---------------|------------------------------------------------------------------------------------------------------|
| <b>Krok 1</b> | Wyrażenie koncepcji                                                                                  |
|               | Ustalenie potrzeb związanych z projektem wymagającym analizy dużych danych bazując na koncepcji „4V” |
| <b>Krok 2</b> | Założenie                                                                                            |
|               | Jaki problem jest rozważany?                                                                         |
|               | Dlaczego jest ważny i interesujący?                                                                  |
|               | Dlaczego podejście związane z Big Data?                                                              |
| <b>Krok 3</b> | Materiały przygotowawcze                                                                             |
|               | Metodologia                                                                                          |
|               | Propozycje działania                                                                                 |
|               | Selekcja zmiennych                                                                                   |
|               | Zbieranie danych                                                                                     |
|               | ETL i transformacja danych                                                                           |
|               | Wybór narzędzi/platformy                                                                             |
|               | Model koncepcyjny                                                                                    |
|               | Techniki analityczne                                                                                 |
|               | Powiązania, klastry, klasyfikacje itd                                                                |
| <b>Krok 4</b> | Rezultaty i wgląd                                                                                    |
|               | Wdrożenie                                                                                            |
|               | Ewaluacja i weryfikacja                                                                              |
|               | Testowanie                                                                                           |

## PIŚMIENNICTWO

1. IBM. Data Driven Healthcare Organizations Use Big Data Analytics for Big Gains. 2013. [cytowany 5 sierpnia 2018]. Adres: <https://www.ibmbigdatahub.com/whitepaper/data-driven-healthcare-organizations-use-big-data-analytics-big-gains>
2. Muin J, Ioannidis K, Ioannidis John P A. Big data meets public health. Science 2014; 346:105-110.
3. Chun-Hai Fung I, Zion Tsz Ho Tse, King-Wa Fu. Converting Big Data into public health. Science 2015; 5:347-353.
4. Schmarzo B. Big Data. Understandinga How Data Powers Big Business. Indianapolis; John Wiley & Sons Inc., 2013.
5. Głowacka E. Ekologia informacji – sposób na choroby informacyjne? [cytowany 5 sierpnia 2018]. Adres: [http://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CC4QFjAA&url=http%3A%2F%2Fkonferencja.biblio.cm.umk.pl%2Ffileadmin%2Fpelne\\_teksty%2Fnowy\\_ekologia\\_inf..doc&ei=Cb05UpXqKceB4gTR\\_YCADA&usq=AFQjCNHfkaNyNf7kuPUaNUoSny\\_m-mWBGg&sig2=Rfk7a8OKXoadjmwA3yl8cw&bvm=bv.52288139,d.bGE](http://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CC4QFjAA&url=http%3A%2F%2Fkonferencja.biblio.cm.umk.pl%2Ffileadmin%2Fpelne_teksty%2Fnowy_ekologia_inf..doc&ei=Cb05UpXqKceB4gTR_YCADA&usq=AFQjCNHfkaNyNf7kuPUaNUoSny_m-mWBGg&sig2=Rfk7a8OKXoadjmwA3yl8cw&bvm=bv.52288139,d.bGE) [odczyt: 31.08.2013].
6. Tadeusiewicz R. Społeczność internetu. Warszawa; Akademicka Oficyna Wydawnicza Exit, 2002.
7. Tadeusiewicz R. W dymie i we mgle...[cytowany 25 sierpnia 2018]. Adres: <http://www.solidarnosc.org.pl/~ksn/Docs/rystad.pdf>
8. Marconi K, Dobra M, Thompson C. The use of Big Data in Healthcare. In: Liebowitz J. Big Data and Business Analytics. Boca Raton; CRC Press, 2013: 229–248.
9. Fan W, Bifet A. Mining Big Data: Current Status, and Forecast to the Future. New York; Wyd. ACM SIGKDD Explorations Newsletter 2012.
10. Diebold F. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive. Pennsylvania; Penn Institute for Economic Research, Department of Economics. University of Pennsylvania, 2012.
11. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inform Sci Syst 2014; 2: 3-10.
12. Redakcja. Co to jest Big Data? [cytowany 16 sierpnia 2016] Adres: <http://bigdataweb.pl/co-to-jest-big-data/>