PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 469 • 2017

Taksonomia 29
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Kamila Migdał-Najman, Krzysztof Najman
Uniwersytet Gdański
e-mails: kamila.migdal-najman@ug.edu.pl; krzysztof.najman@ug.edu.pl

BIG DATA = CLEAR + DIRTY + DARK DATA

BIG DATA = CLEAR + DIRTY + DARK DATA

DOI: 10.15611/pn.2017.469.13
JEL Classification: C38, C55, C81, C82

Streszczenie: Rozwój techniki teleinformacyjnej, Internetu i informatyki przy jednoczesnym spadku jednostkowych kosztów gromadzenia i przechowywania danych powoduje istotne ilościowe i jakościowe zmiany w podejściu zarówno do samych danych, jak i możliwości ich analizy. Ten coraz bardziej gęsty, ciągły i niestrukturyzowany strumień danych, nazywany Big Data, wywołuje współcześnie wiele emocji. Z jednej strony brak odpowiedniej ilości danych był zawsze wyzwaniem dla metod wnioskowania statystycznego i jednym z bodźców ich rozwoju. Jednak z drugiej strony, w dużych liczebnościach prób zawarte są liczne zagrożenia dla wiarygodności wnioskowania. W zbiorach takich, poza danymi o odpowiedniej jakości (Clear Data), znaczny udział mają dane nieprawdziwe, nieaktualne, zaszumione, często wielokrotnie zduplikowane, niekompletne lub błędne (Dirty Data), a także dane, o których jakości czy użyteczności nic nie wiadomo (Dark Data). Celem prezentowanych badań jest krytyczne przedstawienie struktury jakościowej zbioru Big Data.

Słowa kluczowe: Big Data, Clear Data, Dirty Data, Dark Data.

Summary: The development of technology data communications, the Internet and computer with the simultaneous decrease the unit costs of data collection and storage results in significant quantitative and qualitative changes in the approach to the same data, and the possibility of their analysis. The increasingly dense, continuous and unstructured data stream, called Big Data, evokes a lot of emotion today. On the one hand, the lack of adequate quantities of data has always been a challenge for the methods of statistical inference and one of the stimuli of their development. On the other hand, the large sets included threats to the reliability of the inference. In such collections, in addition to data of sufficient quality (Clear Data), the data which are inaccurate, outdated, noisy, often repeatedly duplicate, incomplete or erroneous (Dirty Data), as well as data about which quality or usability nothing is known (Dark Date) have a significante share. The aim of this study is to present the structure of the critical qualitative set of Big Data.

Keywords: Big Data, Clear Data, Dirty Data, Dark Data.

1. Wstęp

Aby skutecznie zarządzać państwem czy przedsiębiorstwem, odpowiednie służby zbierają i przetwarzają dane dotyczące ich funkcjonowania, a także otoczenia, w którym funkcjonują. Liczba i rodzaj zbieranych danych oraz zdolność ich magazynowania i przetwarzania bardzo się zmieniły w ciągu ostatnich 100 lat. Do końca lat 80. XX wieku[1] dane z badań masowych były zapisywane na kartach perforowanych, które z kolei pozwalały mechanicznym sumatorom na ich przetwarzanie. W latach 30. XX wieku urzędy w USA produkowały około 10 mln takich kart dziennie. Liczba ta wydaje się bardzo duża, a praca związana z ich przygotowaniem i analizą ogromna. Jednak jedna karta mieściła w sobie jedynie 70 do 80 bajtów danych, co daje około 670 MB danych dziennie. Gdyby traktować je jak jednorodny strumień danych, daje to około 8100 bajtów na sekundę dla całego USA. To mniej więcej tyle, ile mieści jedna współczesna płyta CD. W latach 50. pojawiły się bardziej dostępne dla biznesu maszyny cyfrowe, takie jak UNIVAC, i taśmy z niklowanej miedzi, które w połączeniu z odpowiednim urządzeniem zapisu i odczytu danych zdolne były do przechowywania znacznie większych ilości danych. Szybkość ich przetwarzania wzrosła gwałtownie do około 7200 bajtów na sekundę. W 1956 roku pojawił się na rynku pierwszy komputer (IBM 305 RAMAC) posiadający dysk twardy o pojemności 5 MB (na dysk twardy składało się pięćdziesiąt 24-calowych pojedynczych dysków), który wyceniono na 50 000 $. Jego szybkość znacznie przekraczała wszystkie znane wcześniej systemy. Średni czas dostępu do losowej ścieżki nie przekraczał 600 milisekund [http://www.pcworld.com]. Rozwój ten trwa nieprzerwanie do dnia dzisiejszego. Jego dynamika jest oszałamiająca. Przeciętny współczesny pendrive o pojemności 256 GB jest pojemniejszy od pierwszego dysku twardego ponad 50 000 razy. Jednocześnie cena za 1 MB jest mniejsza ponad 26 000 000 razy.

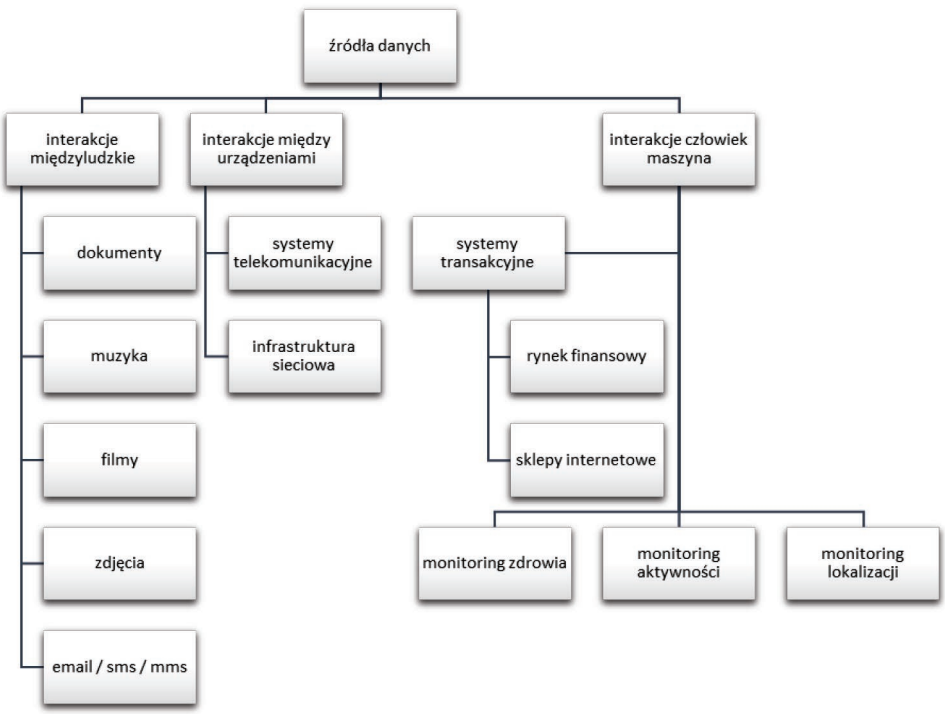
Opisany powyżej proces w XXI wieku doprowadził do powstania jakościowo nowej sytuacji. Ani pojemność nośników, ani ich cena, ani szybkość działania infrastruktury technicznej nie stanowią już istotnej bariery dla systemów gromadzenia i przechowywania danych. Możliwe stało się rejestrowanie praktycznie wszystkich danych powstających zarówno w sferze publicznej, jak i prywatnej. Nadeszła era Big Data.

2. Big Data

Obserwując uniwersum przestrzeni danych, można zauważyć, że niemal wszystkie rejestrowane obecnie dane pochodzą z trzech głównych źródeł. Pierwsze stanowią

[1] Karty perforowane były w użyciu do końca wieku XX. Powszechnie znany jest skandal, jaki wybuchł przy okazji wyborów prezydenckich w USA w 2000 roku. W stanie Floryda do liczenia głosów nadal stosowano tam karty perforowane, a ich niska jakość i wykryte błędy w sumowaniu głosów spowodowały konieczność ponownego ręcznego przeliczenia głosów.

dane będące efektem interakcji międzyludzkich (*human interaction data*). Tworzą je wszelkie formy komunikacji, takie jak wiadomości e-mail, SMS, wszelkie przesyłane dokumenty tekstowe, zdjęcia, filmy czy nagrania dźwiękowe. Drugim źródłem danych są interakcje między urządzeniami (*machine to machine data*), które stanowią infrastrukturę globalnej sieci komputerowej i wszelkie inne urządzenia za których pośrednictwem dane są rejestrowane czy przesyłane. Są to serwery, routery, przełączniki, urządzenia telekomunikacyjne, satelity, nadajniki, odbiorniki itp. Nawet gdyby żaden człowiek na świecie nie podejmował żadnej działalności, urządzenia te i tak generowałyby ogromną ilość danych. Trzecią kategorią są źródła pośrednie, łączące człowieka z urządzeniami, które dają mu dostęp do określonych usług (*human to machine data + transaction data*). Są to różnego rodzaju systemy transakcyjne, takie jak sklepy internetowe, usługi finansowe (np. transakcje giełdowe online, transakcje na rynku walutowym online) usługi mobilne (np. bankowość mobilna, zakup biletów komunikacji, na imprezy masowe), systemy monitorujące stan zdrowia, emocje, położenie, aktywność fizyczną, a także wszelkiego rodzaju interfejsy, za pomocą których komunikujemy się z innymi osobami lub urządzeniami. Strukturę tę w podstawowym ujęciu pokazano na rysunku 1.



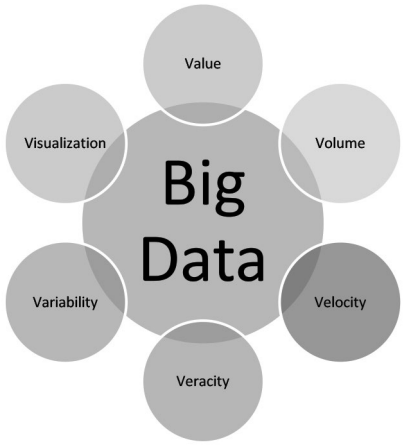
Rys. 1. Podstawowe źródła danych Big Data

Źródło: opracowanie własne.

Łączną ilość danych, generowaną przez wszystkie powyższe źródła, trudno oszacować. Pewnego pojęcia o skali zjawiska dostarczają badania firmy Go-Globe [http://www.go-globe.com/], które koncentrują się jednak tylko na aktywności obserwowanej w Internecie. Według tych szacunków w 2016 roku, w ciągu jednej minuty internauci przesyłali 150 mln wiadomości e-mail, 44 mln wiadomości tekstowych w komunikatorach internetowych, przesyłali ponad 2 mln zapytań do wyszukiwarki Google.com, 2 mln internautów przeglądały strony internetowe o adresach xxx.com, obejrzano 139 tys. godzin materiałów filmowych na YouTube.com, wysłuchano 39 tys. godzin muzyki na spotify.com. W tej samej minucie użytkownicy komputerów nagrali 2,6 mln płyt CD, kupili 4000 dysków pendrive, 2500 zbiorników z atramentem do drukarek, kupili 710 komputerów osobistych, 81 iPadów, 925 iPhone-'ów. W ciągu tej samej minuty powstało 38 ton odpadów elektronicznych. W ciągu sekundy globalna sieć powiększała się o około 30 GB danych. Nowojorska Giełda Papierów Wartościowych (The New York Stock Exchange) szacuje, że w ciągu jednej sesji rejestruje około 1 TB danych transakcyjnych. Z usług monitoringu zdrowia przez smartfony, specjalne opaski czy smartwatche korzysta na świecie więcej niż 500 mln osób, generując ogromną ilość danych.

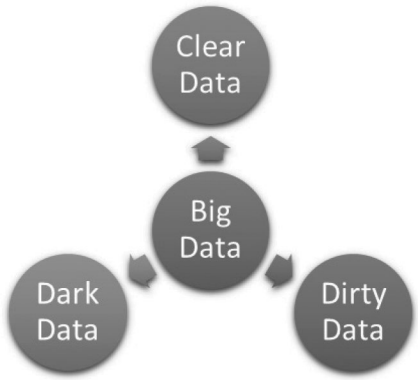
Ogromna szybkość (*velocity*) rejestrowania nowych danych jest jedną z cech tego zjawiska. Oczywistym skutkiem jest równie szybkie zwiększanie się ilości (*volume*) przechowywanych danych. Dane te nie mają jednak jednolitej struktury. Każde źródło, w zależności od dostawcy określonych usług, stosowanego sprzętu, uznawanych standardów czy obowiązującego prawa, generuje dane o własnej, niepowtarzalnej strukturze. Oznacza to, że globalny zbiór nie ma żadnej określonej struktury i jest bardzo zróżnicowany (*variety*). Sam strumień danych jest także źródłem zmienności (*variability*) w zasobach danych. Kolejne dane zmieniają wartość już zarejestrowanych danych, gdy użytkownik zmieni zdanie, treść wpisu na portalu społecznościowym czy kasując część wcześniej wprowadzonej treści. Część danych się dezaktualizuje i jest aktualizowana przez systemy automatyczne lub przez użytkowników. Oznacza to, że już zebrany zbiór danych może podlegać dalszym zmianom. Kolejną cechą współczesnych zbiorów danych jest niewielkie zaufanie (*veracity*), jakim darzą je decydenci, którzy na ich podstawie mają podejmować ważne decyzje. W zbiorach danych znajdują się dane niepełne, fałszywe, wielokrotnie powtórzone, błędne, a część danych do baz danych nie trafia wcale. Z badań prowadzonych przez McKinsey Global Institute, a także dostawców sprzętu i oprogramowania służącego do zbierania danych, takich jak SAS, Cisco, IBM [http://www.ibmbigdatahub.com/], wynika, że 1 na 3 decydentów nie ufa danym, na podstawie których podejmuje decyzje. Szacuje się, że koszt nieoptymalnych decyzji podjętych na podstawie istniejących baz danych kosztuje gospodarkę USA ponad 3 miliardy dolarów. Wielu badaczy uważa, że aby możliwe było przekształcenie zebranych danych w informacje, możliwa musi być ich wizualizacja (*visualization*). Możliwość wizualizacji oznacza bowiem, że jesteśmy w stanie analizować dane. Dopiero analiza pozwala poznać i zrozumieć strukturę danych. Bez tej wiedzy dane

są bezużyteczne. Formy graficzne pozwalają przy tym na maksymalne uproszczenie problemu i jego prezentację. Aby zbieranie danych nie pozostało tylko działaniem samym dla siebie, dane powinny dać się przekształcić w użyteczną informację. Tylko w ten sposób koszt poniesiony na zbieranie, przechowywanie i analizę danych może zwrócić się, tworząc nową wartość (*value*). To ta wartość w rzeczywistości stoi za stale zwiększającym się tempem zbierania danych. W wielu przypadkach jest to jedynie wartość potencjalna, oparta na nadziei, że gdy zbiór będzie odpowiednio duży, a nasza zdolność do jego zrozumienia i wykorzystania osiągnie użyteczny poziom, to wartość pojawi się niemal automatycznie.



Rys. 2. Główne elementy Big Data

Źródło: opracowanie własne.



Rys. 3. Struktura danych Big Data

Źródło: opracowanie własne.

Wszystkie elementy na rys. 2 tworzą nową jakość w technice zbierania, przechowywania i analizy danych, które nazywamy w skrócie Big Data.

Sam fakt istnienia zbioru Big Data to jeszcze zbyt mało, aby pojawiła się istotna wartość. W zbiorze zawarte są nie tylko dane, o których strukturze, pochodzeniu czy zawartości posiadamy jakąś wiedzę.

Znajdują się także dane powtórzone, niekompletne czy po prostu fałszywe. Poza nimi jest także istotna część, o której niewiele wiadomo, poza faktem ich istnienia. Te trzy części mogą zostać nazwane: Clear Data, Dirty Data i Dark Data. Tworzą one zbiór Big Data (por. rys. 3).

3. Clear Data

Zrozumienie zjawisk zachodzących w otaczającym nas świecie wymaga dostępu do danych o odpowiedniej jakości. Zdefiniowanie jakości danych nie jest jednak zadaniem łatwym. Według normy ISO 8402-1986 jakość to: „ogół cech i właściwości produktu lub usługi, który decyduje o zdolności zaspokojenia potrzeb zadeklarowanych lub domyślnych". Oznacza to, że w zależności o potrzeb użytkownika danych, ten sam zbiór potencjalnie może być uznany za wysokiej lub niskiej jakości. Aby możliwe było zaspokojenie potrzeb użytkownika danych, a więc aby zbiór danych mógł zostać nazwany zbiorem o wysokiej jakości, powinien zawierać dane[2]:

1. przydatne,
2. terminowe i punktualne,
3. dostępne i przejrzyste,
4. porównywalne,
5. spójne,
6. dokładne.

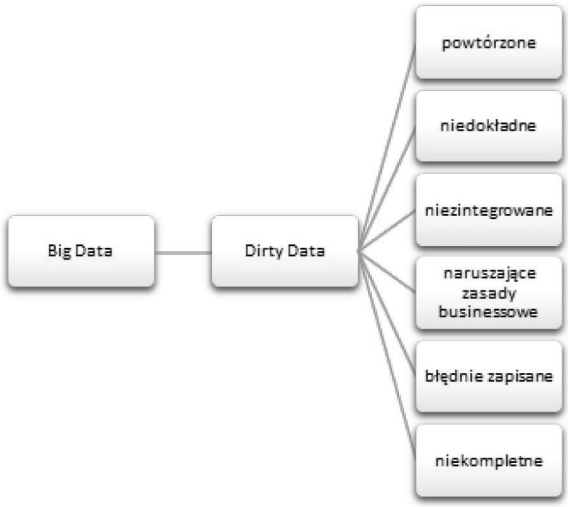
Zbiór danych, który charakteryzuje się powyższymi cechami, jest najbardziej pożądany przez analityka i decydenta. Jest to zbiór idealny, gotowy do analizy i wnioskowania, bez istotnego ryzyka popełnienia błędu wynikającego z samych danych. Dane tego rodzaju można nazwać Clear Data. W badaniach pierwotnych zbiór Clear Data jest abstrakcją, do której każdy badacz dąży. Jego uzyskanie jest możliwe dopiero po eliminacji Dirty i Dark Data.

4. Dirty Data

Zbiór Clear Data jest najbardziej pożądanym przez analityka typem danych. Stanowią one jednak niewielką część zbioru Big Data. Każdy badacz realizujący dowolne badanie empiryczne spotkał się z problemem błędów o charakterze losowym – wynikających z niedoskonałości mechanizmów probabilistycznych, i nielosowych

[2] Definicja jakości w statystyce na potrzeby grupy roboczej „Ocena jakości w statystyce", [Dokumenty Metodologiczne].

– wynikających z czynnika ludzkiego. Obciążenie próby, obciążenie w procesie estymacji, braki danych, pomyłki ankietera, pomyłki, niewiedza lub fałszowanie odpowiedzi przez respondenta, błędy na etapie wprowadzania, kodowania, analizy czy interpretacji danych są dobrze znane i powszechnie występujące.



Rys. 4. Źródła zanieczyszczeń Dirty Data

Źródło: opracowanie własne.

Media społecznościowe, z portalem Facebook na pierwszym miejscu, to największe źródła danych o internautach. Jest to także najczęściej wykorzystywane w praktyce źródło danych Big Data. Wielu badaczy skupia się na „tagach", „hashtagach", „lajkach", „komciach", „szerach", które charakteryzują strony i wypowiedzi użytkowników. Według analiz Networked Insights [http://info.networkedinsights.com/] ogromna część takich danych jest bezwartościowa, ponieważ wcale nie pochodzi od realnych użytkowników. Dane te (aż 53%) są generowane przez sztuczne boty (programy komputerowe podszywające się pod realnych użytkowników), osoby opłacane przez konkurencyjne firmy (23%) bądź przez nieaktywne konta (11%) lub są efektem działania spamerów czy celebrytów. Wszystkie te elementy zanieczyszczają zbiór danych, tworząc Dirty Data. Konsekwencje tego zanieczyszczenia są równie wielkie jak samo Big Data. Jeżeli nawet 90% danych [http://www.reachforce.com/blog] pochodzących z fanpage'ów w mediach społecznościowych to Dirty Data[3], a dane te stanowią podstawowe źródło danych o klientach, ich intencjach, preferencjach czy gustach, to jaką wartość mogą mieć tworzone ich profile?

[3] Może nawet w najgorszej odmianie – danych śmieciowych, których w żaden sposób nie można uzupełnić, poprawić czy uwiarygodnić.

Problem jest bardzo poważny. Już teraz od 50% do nawet 80% czasu, jaki badacze danych spędzają nad analizą Big Data, pochłania właśnie oczyszczanie Dirty Data.

Walka z Dirty Data to złożony proces obejmujący analizę problemu, oczyszczanie danych i zapobieganie powstawaniu zanieczyszczeń. Analiza problemu (*data profiling*) to statystyczny proces analizy danych pod kątem ich poprawności, kompletności, unikalności, spójności i racjonalności. Jest to proces, z którym statystyka radzi sobie względnie dobrze. Drugim elementem procesu jest oczyszczanie danych. Jest to proces uzupełniania, poprawiania i eliminacji niemożliwych do poprawienia danych ze zbioru. Proces ten jest trudny do zautomatyzowania, co w konsekwencji powoduje, że jest on organizacyjnie złożony, czasochłonny i kosztowny. Trzecim filarem jest zapobieganie powstawaniu błędów (*defect prevention*). Na podstawie poprzednich etapów identyfikuje się przyczyny, źródła, warunki i miejsca powstawania błędów. Planuje i wdraża się następnie mechanizmy zapobiegające ich powstawaniu.

5. Dark Data

O ile Dirty Data zawierały wiele błędów o różnym charakterze i źródle, możliwe były przynajmniej do częściowego wykorzystania dzięki procesowi ich oczyszczania. Zbiory Big Data zawierają jednak wiele danych, o których niewiele wiadomo. Często nie można zidentyfikować ich autora, miejsca, czasu ich powstania, nie wiadomo czego dotyczą, w jaki sposób są powiązane z innymi danymi. Zwykle nie mają określonej struktury, wewnętrznego porządku, mają surowy, nieprzetworzony charakter. Wiadomo że istnieją, jednak trudno powiedzieć, czego i w jaki sposób dotyczą. Te dane to Dark Data. Gartner w swoim słowniczku [http://www.gartner.com/it-glossary/], *Gartner IT Glossary*, definiuje Dark Data jako: „Zasoby informacyjne, gromadzone i przetwarzane[4] przez organizacje podczas ich codziennej aktywności biznesowej, które na ogół nie nadają się do wykorzystania w żadnym sensownym celu". Typowym przykładem są backupy danych archiwizowane przez przedsiębiorstwa. Zdecydowana większość z nich nigdy do niczego nie jest wykorzystywana. Backup danych „trzeba robić", ale ponieważ systemy komputerowe są obecnie w wysokim stopniu niezawodne, nie przydają się do niczego. Przedsiębiorstwo, wiedząc o tym, archiwizuje wszystkie dane, „jak leci", nie dbając o ich strukturę czy opis. Już w momencie ich powstania zakłada się, że nie będą użyte. Dark Data znacząco zwiększają wolumen Big Data, jednak nie tworzą żadnej wartości. Gromadzone są na wszelki wypadek.

6. Zakończenie

Zjawisko Big Data jest bardzo złożone i dynamiczne. W artykule poruszono jedynie problem struktury danych zbieranych w globalnych repozytoriach. Na zbiór Big

[4] Chodzi tu raczej o fakt, że są one wynikiem „jakiegoś" przetwarzania.

Data składają się nie tylko pożądane, łatwe do użycia Clear Data, ale także wymagające wiele zachodu Dirty Data i enigmatyczne Dark Data. Trudno jest jednoznacznie stwierdzić, jakie są ich proporcje, jednak wydaje się, że Clear Data to wyraźna mniejszość. Fakt ten jest wyraźnie widoczny w dysproporcji, jaka się wyraźnie uwidacznia między ilością zbieranych danych a ilością istotnych danych, które można przekształcić na wartościowe informacje. Możliwości techniczne i informatyczne zbierania i przechowywania danych znacznie wyprzedzają zdolność do ich analizy i wnioskowania na ich podstawie. Dysproporcja ta szybko rośnie. Wiceprezydent Google, Vinton Gray Cerf, przemawiając do zgromadzonych w San Jose członków American Association for the Advancement of Science [http://www.bbc.com/news/], mówił w 2015 roku m.in. o konieczności bieżącej pracy nad danymi, które już zgromadziliśmy. Miał na myśli przede wszystkim ciągłą weryfikację prawdziwości i aktualności danych. Przede wszystkim jednak ostrzegał przed nadchodzącymi „cyfrowymi, ciemnymi wiekami". Jako główny katalizator „Digital Dark Age" [http://www.slashgear.com/] wymienił **implozję** danych (Big Data), zdominowanych przez Dark Data. Zbieranie bezużytecznych, niemożliwych do analizy i wykorzystania danych, bez zachowania kontekstu i powiązania z innymi danymi, może w konsekwencji zniweczyć wszelkie plany, które leżą u podstaw rozwoju Big Data. Jest to jedno z ważniejszych współczesnych wyzwań także dla statystyki. Jeżeli mu nie sprostamy, oczekiwania i prognozy stóp zwrotu[5] z wdrażania rozwiązań Big Data po prostu się nie zrealizują. Oznaczałoby to wielkie marnotrawstwo sił i środków.

Literatura

Dokumenty Metodologiczne 4.2, Eurostat, Luksemburg, 2-3 października 2003.
http://info.networkedinsights.com/Dirty-Data-LP.html (4.11.2016).
http://www.bbc.com/news/science-environment-31450389 (4.11.2016).
http://www.gartner.com/it-glossary/(4.11.2016).
http://www.go-globe.com/ (4.11.2016).
http://www.pcworld.com.mx/Articulos/30148.htm (4.11.2016).
http://www.reachforce.com/blog/6-quick-dirty-data-stats/ (20.05.2017).
http://www.slashgear.com/a-digital-dark-age-is-coming-warns-father-of-the-internet-13368963/ (4.11.2016).
http://www.tcs.com/big-data-study/Pages/download-report.aspx (4.11.2016).
Networked Insights, 2012, *Big Data, Can Mean Big Insights*, Social Intelligence Report.
Networked Insights, 2015, *How Ditry is Big Data?*

[5] Wiele przedsiębiorstw spodziewa się znacznej, wynoszącej nawet ponad 60%, stopy zwrotu z wdrożeń technologii i analiz Big Data w swojej działalności [http://www.tcs.com/big-data-study/Pages/download-report.aspx].