DOI: 10.15199/13.2016.5.5

Big Data czyli analiza korelacji

(Big Data is a correlation analysis)

mgr KAROL NOWAKOWSKI, Bank Gospodarstwa Krajowego

prof. dr inż. WOJCIECH NOWAKOWSKI, Instytut Maszyn Matematycznych, Warszawa

Streszczenie

Big Data jest jednym z najważniejszych wyzwań współczesnej informatyki. Wobec zmasowanego napływu wielkich ilości informacji obecnych czasach pochodzących z różnych źródeł, konieczne jest wprowadzanie nowych technik analizy danych oraz rozwiązań technologicznych. Technologia Big Data dopiero się rodzi.

Słowa kluczowe: Big Data

Abstract

Big Data is a term frequently used in the literature, but still there is no consensus in implementations of such environments. In view of the massive influx of large amounts of information nowadays from various sources, it is necessary to introduce new data analysis techniques and technology. Then Big Data technology is only just being born.

Keywords: Big Data

Górnym ograniczeniem wielkości możliwego do zanalizowania zbioru stanowi wielkość pamięci masowej systemu komputerowego, na którym taką analizę będzie się prowadzić. Pierwszy nowożytny dysk twardy, którego powstanie zapoczątkowało *de facto* epokę danych cyfrowych, stworzono w IBM w roku 1956. Pojemność tego dysku wynosiła 3,75 MB. Dzisiejsze dyski twarde wykorzystywane w komputerach osobistych mają pojemności liczone w terabajtach (TB), a więc milionach megabajtów.

Nastąpił więc znaczny wzrost potencjalnej pojemności dysków twardych, jednak zapotrzebowanie na pojemności rośnie znacznie szybciej. Co prawda na największym dysku twardym komputera klasy PC o pojemności rzędu 2 TB możliwe jest zapisanie danych z ponad pięciu lat pomiaru np. światła w jednym pomieszczeniu. Ale przy analizie danych dla stu pokoi, możliwy czas zbierania danych skraca się do 19 dni.

Wielkie dane

Rozwój technologii powoduje gwałtowny wrost ilości produkowanych danych. Wielki Zderzacz Hadronów (LHC) zbudowany w ośrodku CERN w pobliżu Genewy generuje rocznie 30 petabajtów (PB) danych, czyli 30 miliardów megabajtów. Silniki samolotu Boeing 787 generują średnio pół terabajta danych na jeden lot (informacje te pozwalają zapobiegać kosztownym naprawom przez usuwanie drobnych odstępstw od normy zanim silnik zostanie unieruchomiony na ziemi, a tym bardziej w powietrzu). Podane przykłady odnoszą się wyłącznie do świata niewirtualnego, w którym gwałtowny przyrost danych dopiero nastąpi wraz z popularyzacją Internetu Rzeczy. W świecie wirtualnym przyrost danych jest jeszcze bardziej gwałtowny i trwa już ponad dziesięć lat.

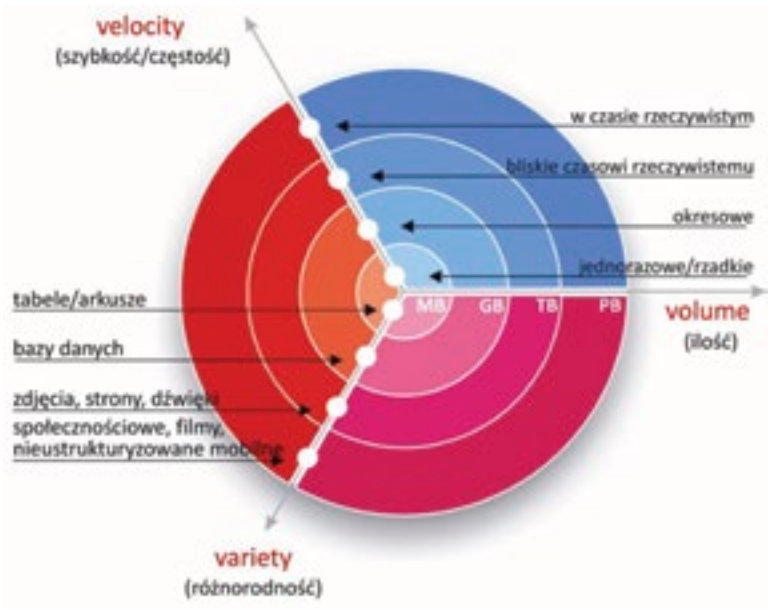
Pierwszym ograniczeniem, na które w XXI wieku napotyka cyfrowa analiza zjawisk jest więc zatem problem zbyt dużej ilości danych do analizy. Kolejnym problemem jest tempo przyrostu danych do analizy. VISA, centrum płatności z wykorzystaniem kart gromadzi informacje o ok. 100 milionach transakcji każdego dnia. Jeżeli rekord jednej transakcji będzie miał tylko 10 KB, to każdego dnia przybywać będzie 1 TB danych do analizy. Jeżeli przeprowadzenie wszelkich analiz obejmujących rozliczenia z punktami handlowymi, rozliczenia z bankami współpracującymi, uwzględnienie promocji i konkursów, uznanie bądź odrzucenia reklamacji czy wykonanie operacji *charge-back* zajmie więcej niż krótkie 24 godziny, to okaże się, że tempo analizy danych jest mniejsze niż tempo przyrastania danych!

W serwisie Twitter publikowanych jest 350 tys. ćwierknięć (czyli krótkich wypowiedzi użytkowników – *tweets*) na minutę. Głównym konkurentem Twittera (217 mln użytkowników) jest Facebook, który ma ponad 1,3 mld użytkowników! Przyrost danych w cyfrowym świecie przyspiesza w tempie uniemożliwiającym ich analizę z zachowaniem waloru aktualności jej wyników. Wskutek tego wyniki analizy takich zbiorów danych są przestarzałe już na długo przed ich publikacją.

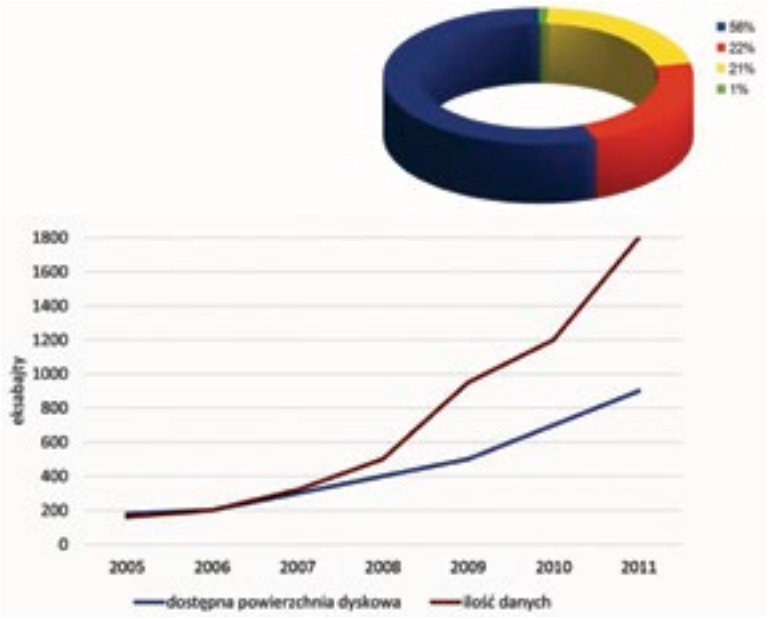
Big Data jako nowa skala w cyfrowej analizie zjawisk

Zbiory informacji, które są wykorzystywane do analizy muszą być strukturalizowane. W dzisiejszych realiach technicznych dane, które mają być analizowane muszą zostać opisane w wybranym metajęzyku i dopiero po dopasowaniu ich do odpowiednich wierszy (np. w językach SQL), celek (np. w programie Excel) czy tagów (np. w językach pochodzenia XML) mogą zostać poddane właściwej analizie. Niestety tylko 5% stworzonych dotychczas przez człowieka danych istniej w klasycznym zapisie alfanumerycznym, tzn. takim, który komputery mogą swobodnie odczytać (pozostałe dane to chociażby nagrania dźwięku albo obrazu, wymagające wcześniejszego przetwarzania).

Jedynie mały ułamek tych pięciu procent stanowią dane zapisane w sztywnie zdefiniowanych ramach wybranego



Rys. 1. *Big Data* – rozszerzanie w zakresie podstawowych 3 *v*. Źródło: datasciencecentral.com



Rys. 2. Tempo tworzenia danych i przyrostu powierzchni dyskowej. Zależność pomiędzy ilością tworzonych informacji a ilością tworzonego miejsca do jej przechowywania. Dane historyczne – do 2007 r. oraz prognostyczne – od 2008 r. Kolor niebieski – dostępna powierzchnia dyskowa, kolor czerwony – ilość danych. U góry po prawej struktura powierzchni dyskowej dostępnej w 2007 r: kolor niebieski: 56% – dyski twarde, czerwony: 22% – płyty CD, DVD, BD, żółty: 21% – napędy taśmowe, zielony: 1% – inne. Źródło: Gantz F. et all, 2008

formatu zapisu. I tylko ten ostatni mały wycinek istniejących danych może być analizowany z wykorzystaniem dzisiejszych metod komputerowej analizy. W języku angielskim to niezestrukturalizowanie określa się pojęciem *variety*, a proces strukturalizowania danych słowem *datafication* (tłumaczonym na język polski jako danetyzacja).

Volume, *velocity* oraz *variety* zostały już na początku XXI w. wskazane jako podstawowe problemy zarządzania danymi. Od tego czasu te 3 „v" [1] utożsamia się z wyjściową definicją *Big Data* jako takiego zbioru danych, który charakteryzuje się wielkimi rozmiarami, wielkim tempem przyrostu nowych danych oraz minimalnym stopniem entropii (*rys. 1*). Niezbędne jest dokładniejsze zdefiniowanie wskazanych trzech składowych. Problemem jest jednak wciąż brak jednoznacznej definicji *Big Data* – dziedzina ta jest zbyt młoda.

Wielkość (*volume*) graniczna zbioru, po przekroczeniu której zbiór będzie zaliczany do kategorii *Big Data* może być wyrażona w wartościach bezwzględnych lub względnych. W ankietach 1144 menedżerów oraz specjalistów IT o wskazanie uznawanej przez nich wielkości zbioru za wielkość należącą do kategorii *Big Data*. Ponad 50% respondentów wskazało rozmiar pomiędzy jednym jednym terabajtem, a jednym petabajtem, tylko 20% wskazywało na zbiory mniejsze od terabajta lub większe od petabajta (30% nie udzieliło odpowiedzi).

Trudność wskazania konkretnej wielkości uwidoczniona przez rozbieżność przytoczonych opinii po części wynika, jak się przypuszcza, z tempa wzrostu ilości danych. Stąd też odnotowujemy opinie, że z *Big Data* mamy do czynienia wtedy, kiedy odchodzimy od badania próby losowej z populacji, a przechodzimy do badania całej populacji (lub jej jak największej możliwej części). Czyli wtedy kiedy udaje się poddać analizie danych informacje o wszystkich badanych obiektach.

Prędkość (*velocity*) przyrostu danych charakterystyczna dla *Big Data* uniemożliwia prowadzenie ich analizy tradycyjnymi metodami obróbki cyfrowej ponieważ wyniki analizy kolejnych porcji danych pojawiałyby się wolniej niż napływałyby nowe informacje. Wal-Mart, amerykańska sieć supermarketów przetwarzała w 2010 r. więcej niż milion transakcji w każdej godzinie. Samo zgromadzenie takiej ilości danych w jednym centralnym zbiorze (pliku), a następnie wykonywanie na nim kwerend w czasie rzeczywistym sprawiałoby problemy techniczne. Tempo przyrostu danych jest już wyższe niż tempo przyrostu pojemności magazynowej pozwalającej na przechowywanie danych, co przedstawiono poniżej.

Trend ten nie powinien dziwić ponieważ już teraz istnieje olbrzymia ilość danych, których nie zapisujemy (np. nagrania przeprowadzonych rozmów telefonicznych, nagrania z kamer monitoringu, audycje radiowe i telewizyjne, itd.). Tworzona informacja musi jednak niemal zawsze zostać przeanalizowana, a analizowanie czegoś, czego nawet nie możemy zapisać wydaje się być zadaniem abstrakcyjnym.

Wykorzystywanie danych o wielkiej różnorodności (*variety*) jest możliwe dopiero po danetyzacji, czyli przetworzeniu do

formatu umożliwiającego analizę. Jeżeli dziennikarz posiada szafę z olbrzymią ilością materiałów drukowanych (zdjęć, rolek z filmami, wycinków z gazet, notatek odręcznych), to samo odszukanie czegokolwiek w takim zbiorze danych nastręcza wiele problemów. Jeśli dziennikarz ten przeprowadzi cyfryzację swojego zbioru, czyli zeskanuje i przeniesie wszystkie jego elementy do pamięci komputera a każdemu plikowi nada odpowiednią nazwę, to taki zbiór stanie się łatwiejszy do przeszukiwania.

Czy rzeczywiście taki zbiór danych stanie się łatwiejszy do przeszukiwania? Tak, chociaż w nieznacznym stopniu. Szukać w takim zbiorze możemy w dowolnym miejscu na świecie, kartki się nie wysypują. Ale dopiero danetyzacja będzie tym procesem, który przyniesie nam wartość dodaną. Za jej pomocą na zdjęciach i filmach przeprowadzimy proces rozpoznawania twarzy, gestów oraz miejsc; na filmach dodatkowo spróbujemy automatycznie rozpoznać ścieżki dźwiękowe, a wypowiadane w nich słowa przerobić na tekst literowy; wycinki z gazet poddamy procesowi rozpoznawania tekstu drukowanego, zaś notatki ręczne – procesowi rozpoznawania tekstu odręcznego oraz analizy pisma.

Po przeprowadzeniu danetyzacji możemy „odpytywać" zbiór na przykład nazwiskami i wtedy narzędzie do analizy *Big Data* poinformuje nas, które wycinki i notatki dotyczą danej osoby oraz gdzie można ją zobaczyć na zdjęciach i filmach. Różnorodność zbioru została w ten sposób opanowana dzięki wykorzystaniu narzędzi denetyzacji – czyli transformacji danych cyfrowych nieczytelnych dla komputera na takie, które pozostają czytelne, chociaż wciąż pozostają nieustrukturyzowane.

Dodatkowe wymiary zbiorów danych

Wraz z rozwojem *Big Data* zaczęto zwracać uwagę na dodatkowe wymiary zbiorów danych poza pierwotnym zestawem 3 *v* (*wielkością*, *prędkością przyrostu* oraz *różnorodnością*), czyli na *wiarygodność*, *nierównomierność*, *złożoność* oraz *wartość*.

Wiarygodność jest wymiarem zaproponowanym przez IBM, które zasugerowało, że analizowane dane nie muszą być poprawne, zgodne z prawdą ani też obiektywnie dobre. Z takim problemem spotkało się właśnie IBM (i... nie potrafiło mu sprostać) podczas tworzenia w latach 90-tych ubiegłego wieku systemu do tłumaczeń maszynowych IBM Candide, protoplasty Google Translate. Koncepcja takich tłumaczeń opiera się właśnie na danetyzacji jak największej ilości danych (słów, fraz, zdań, konstrukcji językowych) z języków, które mają być tłumaczone. IBM postanowił wykorzystać źródła najlepszej jakości – protokoły posiedzeń parlamentu Kanady prowadzone zawsze dwujęzycznie (po francusku i angielsku). Program tłumaczący korzystał z trzech milionów idealnie poprawnych gramatycznie i językowo zdań. Kiedy w 2006 r. uruchomiono system Google Translate, który korzystał z miliarda stron tłumaczeń często o bardzo niskiej wiarygodności i jakości (takich jak chociażby fora internetowe) system Candide okazał się mało skuteczny i po kilku latach IBM przestał go rozwijać. Bazujący na częściowo niepoprawnych danych Google Translate funkcjonuje do dzisiaj i nie ma konkurencji.

SAS Institute (światowy lider w zakresie analityki biznesowej oraz największy niezależny dostawca oprogramowania Business Intelligence) zaproponował uzupełnienie klasycznego modelu 3 *v* o *nierównomierność* oraz *złożoność*. Pod tym pierwszym pojęciem rozumieć należy nierównomierność (niestabilność) wszystkich pozostałych wymiarów *Big Data*, a więc przyjęcie założenia, że prędkość wzrostu danych w jednym okresie może być olbrzymia, a w innym pozostać na praktycznie zerowym poziomie i że różnorodność może rosnąć i maleć w dowolny sposób. Również, że wiarygodność danych w jednym momencie może być bardzo wysoka, a chwilę później (lub w innych warunkach) drastycznie niska. Przez złożoność, kolejny wymiar *Big Data* należy rozumieć, że żadne źródło danych, żaden format danych nie może zostać oddzielony i analizowany w oderwaniu od pozostałych, bo będzie to prowadzić do nieakceptowalnego odchylenia analizy. Patrząc chociażby tylko przez pryzmat wiarygodności SAS sugeruje, że błędne dane (czyli niską ich jakość) można naprawić przez dużą ilość danych, ale wyłącznie jeśli uwzględnimy wszystkie źródła i wszystkie formaty. Zatem nieuwzględnianie złożoności (i dogmatu kompletności danych) wzmacnia wady zbiorów *Big Data* i uniemożliwia ich skuteczną analizę, tzn. taką, która prowadzi do poprawnych wniosków.

Wartość to ostatni z powszechnie dyskutowanych wymiarów *Big Data*. Oracle wysunął pogląd, że zbiór wielkich ilości danych charakteryzuje się „zagęszczaniem wartości", że w momencie zbierania zawartość zbioru ma bardzo niską wartość – zarówno w stosunku do swojej wagi jak i w wartościach bezwzględnych. Dopiero po przeprowadzeniu analizy zbiór *Big Data* nabiera wartości, a więc może być wykorzystany do poprawienia efektywności, obniżenia kosztów czy zwiększenia strumienia dochodów. Znając wszystkie wymiary, tj. wielkość, prędkość, różnorodność, wiarygodność, nierównomierność, złożoność oraz wartość można by spróbować przedstawić pełniejsza definicję *Big Data*. Jednak nie. W przeprowadzonych przez SAP (*Systemanalyse und Programmentwicklung*) badaniach postawiono dyrektorom największych firm globalnych prośbę o przygotowanie definicji *Big Data*.

Rys. 3. Rodzaje definicji Big Data według dyrektorów badanych w kwietniu 2012 r.Kolor niebieski: 28% – gwałtowny przyrost informacji transakcyjnych, z uwzględnieniem danych od klientów i łańcucha dostaw, kolor czerwony: 24% – nowe technologie stworzone aby radzić sobie z wyzwaniami 3*v* w danych, kolor żółty: 19% – potrzeba przechowywania i archiwizowania danych dla zachowania zgodności z wymogami prawa i polityki bezpieczeństwa, kolor zielony: 18% – eksplozja nowych źródeł danych (media społecznościowe, urządzenia mobilne, maszyny), kolor brazowy: 11% – inna definicja. Źródło: Gandomi A., Haider M., 2014

Najistotniejszym wnioskiem z powyzszych badań jest brak jednomyślności nawet co do kierunku tworzenia definicji: 28% pytanych skupiło się na problemie jakim zjawiskiem jest *Big Data*, 24% – jakie technologie wiążą się z *Big Data*, dla 19% to przede wszystkim problem zakupu nowego sprzętu do magazynowania danych, dla 18% Big Data to zjawisko polegające na uzyskaniu większej ilości danych.

Sens pojęcia *Big Data* dobrze oddaje definicja sformułowana przez TechAmerica Foundation, która w wolnym tłumaczeniu brzmi „*Big Data jest pojęciem opisującym duże zbiory szybko rosnących, kompleksowych i zmiennych danych, których zdobywanie, magazynowanie, dystrybucja oraz analiza wymagają zaawansowanych technik i technologii*”. Wadą tej definicji jest pominięcie wszystkich dodatkowych wymiarów *big-data* i skupienie się wyłącznie na trzech podstawowych *v*.

Ciekawa definicja została sformułowana przez Chen P.C.L. oraz Zhang Ch.Y., w która mówi się, że z *Big Data* mamy do czynienia wtedy gdy zadziwiająca (oryg. *formidable*) jest perspektywa zdobycia danych, ich naprawy, analizy i wizualizacji z wykorzystaniem współczesnych technologii.

Ta młoda dziedzina analizy dużych zbiorów danych została już dostrzeżona w polskich publikacjach naukowych [2]: „*Big Data to określenie stosowane dla takich zbiorów danych, które jednocześnie charakteryzują się dużą objętością, różnorodnością, strumieniowym napływem w czasie rzeczywistym, zmiennością, złożonością, jak również wymagają zastosowania innowacyjnych technologii, narzędzi i metod informatycznych w celu wydobycia z nich nowej i użytecznej wiedzy”.*

Wpływ Big Data na proces analizy i analitycznego myślenia

W jaki konkretnie sposób *Big Data* może wpłynąć na proces analizy i sposób analitycznego myślenia specjalistów, a w szerszej perspektywie na gospodarkę i ekonomię. Proces analizy danych w przypadku zbiorów *Big Data* różni się znacząco od procesu zwykłej analizy statystycznej.

Podstawowe różnice to:

1. Odejście od analizy przyczynowo-skutkowej na korzyść badania samej korelacji,
2. Analizowanie całego zbioru danych zamiast analizowania precyzyjnej próby losowej.

Związek przyczynowo-skutkowy, zrozumienie istoty, reguły i zasady czy połączenie teorii z praktyką są podstawowymi narzędziami prowadzącymi do poznania dowolnego zjawiska.

Mechanik Boeing’a np. uczy się zasady działania sinika odrzutowego, jego budowy i zależności pomiędzy jego częściami. Ta wiedza zostaje uzupełniona o doświadczenie zdobyte przez lata praktyki, pracy na lotnisku przy serwisowaniu silników pod okiem starszych, bardziej doświadczonych kolegów.

Każdy samolot, który ląduje na lotnisku centralnym dla hubie przechodzi szybki przegląd, a ujawnione w wyniku tego przeglądu nieprawidłowości muszą zostać sklasyfikowane na wymagające naprawy natychmiastowej i na takie, którymi można się zająć w terminie późniejszym. Naprawa natychmiastowa wymaga uziemienia samolotu, odwołania lotu, znalezienia pasażerom innych połączeń i wypłaty odszkodowań. Gdy mechanik zabierze się za naprawę może się okazać, że nie ma odpowiednich części zamiennych, albo, że do tej operacji serwisowej nie ma wystarczającej wiedzy. Proces naprawy samolotu pasażerom może się więc wydłużać.

Analizy *Big Data* już dzisiaj zmieniają sposób obsługi silników Boeing’a. W trakcie pracy silnik generuje gigabajty danych, które są przesyłane w czasie rzeczywistym do centrali firmy. Tam komputery szukają odchyleń od normy, powtarzalnych sekwencji nienaturalnych odczytów czy sygnałów o awarii. W ten sposób wykrywa się nie tylko powstałą już awarię (ostrzegając pilota), ale wykrywa się stan, który nienaprawiony będzie prowadzić do awarii. Skoro wiemy, że w danym silniku wystąpi wkrótce awaria możemy już teraz wysłać niezbędne do naprawy części na lotnisko, na którym samolot wyląduje choćby jutro. Mechanik dostanie już teraz informację o tym jaką procedurę będzie musiał przeprowadzić i uzyska w ten sposób czas potrzebny na zapoznanie się z opisem odpowiedniej procedury i sposobem jej przeprowadzania. Zyskuje więc tym samym operator, mechanik, i pasażerowie.

Ale zysk ekonomiczny to tylko jeden z istotnych skutków posiadania wielkiej ilości danych i narzędzi do ich analizy. Ważniejszym skutkiem jest zmiana procesu kognitywnego – z opartego na związkach przyczynowo-skutkowych, na wykorzystywanie korelacji. Firma Boeing nie musi wiedzieć dlaczego wzrost ciśnienia o 0,4% w czujniku A7659.P połączony ze spadkiem temperatury oleju w czujniku T883.G w zakresie od -1,4°C do -2,7°C oznacza awarię czujnika 439, która może doprowadzić do uszkodzenia silnika. Firma Boeing wie, że tak po prostu jest, i że naprawa musi zostać przeprowadzona. Dla pasażerów również sposób podniesienia bezpieczeństwa podróży jest mniej istotny od samego faktu wzrostu poczucia bezpieczeństwa. Jak piszą Mayer-Schönberger i Cukier [3] „*możemy odkrywać schematy i korelacje, które umożliwiają nam nowe, bezcenne zrozumienie określonego zjawiska*” chociaż zrozumienie to nie jest oparte o naukowy związek przyczynowo-skutkowy, a o wielką bazę danych dotychczasowych lotów.

Drugą znaczącą zmianą w analizie rzeczywistości po przełączaniu się na badanie korelację jest w *Big Data* odejście od badania próby z populacji na badanie całej populacji. Statystycy, bazując na teoriach doboru próby zdefiniowali zasady tworzenia próby z populacji. Pionierem tego podejścia, który wielce zasłużył się dla tej koncepcji, był polski matematyk i statystyk Jerzy Spława-Neyman. To dzięki jego badaniom opublikowanym w 1934 r. wiemy, że dobra próba statystyczna to taka, która stworzona jest losowo, przy czym nie musi być ona bardzo duża pod warunkiem, że została stworzona zgodnie z regułami statystycznej losowości.

Powszechnie przyjmuje się, że prawidłowe badania statystyczne charakteryzują się błędem statystycznym nie większym niż trzy punkty procentowe. Taka dokładność wystarcza do wyciągania generalnych sądów o populacji, ale nie do odkrywania mniej licznych, choć nie mniej doniosłych związków.

W Ekwadorze, w prowincji Pichincha żyją osoby z karłowatością wynikającą z zaburzeń funkcjonowania hormonu IGF-1. W toku badań okazało się, że osoby te są odporne na raka. Ich populacja to 0,000004% ludzkości. Taka grupa ludzi nie zostałaby dostrzeżona w wynikach badania statystycznego przeprowadzonego na jakimkolwiek akceptowalnym poziomie istotności.

Próba losowa powstała jako konieczność, bowiem ilość danych wymagających analizy rosła na początku XX w dużym

tempie. Spis powszechny w USA prowadzony przez Census Bureau raz na 10 lat zabierał coraz więcej czasu – w 1880 r. zajął aż 8 lat i nie było wówczas nadziei na przyspieszenie tego procesu. Rozwój technologii wczesnych maszyn liczących poprawił sytuację, ale dopiero rezygnacja z badania populacji i ograniczenie się wyłącznie do próby losowej pozwoliło na urealnienie uzyskiwania wyników spisu. Obecnie, w sto lat później, możemy analizować wszystkie otrzymywane dane bez ograniczania się do prób statystycznych właśnie dzięki wykorzystaniu technologii *Big Data*. W rzeczywistości już to robimy. Różnice między badaniem statystycznym a *Big Data* przedstawiono poniżej.

Różnice pomiędzy badaniem statystycznym a badaniem *Big Data*

Badanie statystyczne	Badanie *Big Data*
korelacja	związek przyczynowo-skutkowy
próba losowa	całość populacji
dbałość o dokładność	dbałość o skalę
rozkład normalny	rozkład rzeczywisty
badacz sugeruje gdzie/czego szukać	algorytm szuka wszędzie

Rezygnacja z próby losowej na rzecz rejestracji wszystkich danych zmniejsza presję na niezawodność i dokładność pomiarów. To kolejna doniosła zmiana w prowadzonych analizach. Częściowo uzyskuje się to przez redundancję pomiarów. Jeżeli np. w fabryce zainstalowanych będzie dziesięć czujników temperatury przekazujących sygnały co dziesięć sekund, to wiarygodność i niezawodność tych czujników są kluczowe. Jeśli jednak zwiększymy liczbę czujników do tysiąca, a pomiary będą dokonywane co sekundę, to i tak uwzględniać będziemy jedynie odczyty uśrednione. Uśredniając wynik minimalizujemy błąd. Jeżeli jeden z tysiąca czujników zacznie przekazywać złe dane, błąd wyniesie 0,1%. Ale jeśli pozostałe czujniki w okolicy będą podawać prawidłowe odczyty, to ten jeden błędny zostanie zidentyfikowany przez algorytm programistyczny i przestanie być uwzględniany w dalszych pomiarach.

Ten sam stopień akceptacji błędów danych wykorzystuje Google Translate. Zamiast korzystać z idealnie poprawnych tłumaczeń posiedzeń parlamentu bierze pod uwagę wszystko, co może znaleźć w internecie. W ten sposób uczy się zarówno na danych poprawnych jak i na błędnych. Ale część danych błędnych zostaje w końcu „przykryta" danymi poprawnymi. Liczba błędów w końcowych tłumaczeniach jest ostatecznie mniejsza niż w danych wchodzących do algorytmu. Oczywiście można by ręcznie weryfikować poprawność dokumentu – jednak takie działanie byłoby ekonomicznie nieuzasadnione. Akceptujemy pewien błąd dlatego, że uzyskanie o kilka procent lepszego wyniku wymagałoby poniesienie kilkukrotnie większych nakładów.

Niestety nie zawsze możemy w takim stopniu uprościć uzyskiwanie danych. Jeżeli sprawdzamy stan naszego konta, to oczekujemy informacji podanej z dokładnością do jednego grosza. Wysokość środków na koncie można określić z bardzo dużą precyzją – analizując historię sald, wpłat i wypłat z konta. Nie możemy już tego zrobić, kiedy chcemy określić majątek jakiejś osoby. Zbyt wiele zmiennych nachodzi na siebie, całość informacji jest bardzo dynamiczna, a niektórych składników majątku nie da się jednoznacznie wycenić – można je jedynie oszacować jakąś wybraną metodą. Tak jak dla kierowcy nie istotne jest czy porusza się z prędkością 23.5 km/h czy 23.3 km/h, ale istotne pozostaje że porusza się zbyt wolno. Prędkość możemy przecież bardzo dokładnie i jednoznacznie określić – tyle, że nie jest to potrzebne.

Ostatnią zmianą w analizie, wynikająca z wykorzystywania na co dzień *Big Data*, to zmiana w nas samych. To zmiana, która jest trudniejsza do przewidzenia ponieważ stanowi wtórną reakcję społeczeństwa. Jak zachowają się użytkownicy internetu, którzy wpisując w wyszukiwarce nieprawidłowe frazy otrzymają prawidłowe wyniki? Jeśli np. wpiszą „odkużaecz workwe", otrzymają wyniki dla frazy „odkurzacze workowe". Czy takie naprawianie naszych błędów spowoduje, że zaczniemy się sami naprawiać i lepiej kontrolować czy też raczej, że komputerom na stopniowe przejmowanie inicjatywy w „myśleniu za nas". Jeśli przestaniemy się kontrolować, to może się okazać, że w nadchodzącej epoce „władzy danych" będziemy zdani na wszechobecnych tłumaczy, interpretatorów i analizatorów, którzy będą nam wyjaśniać dlaczego mamy robić to co akurat robimy.

Big Data niesie za sobą poza zmianami w samym sposobie myślenia czy prowadzenia analizy naukowej efekty ekonomiczne.

Ekonomiczne efekty wykorzystywania Big Data

W 2000 r. Luis von Ahn wymyślił sposób na oddzielenie ludzkiego ruchu w internecie od ruchu generowanego przez automatyczne roboty. Jego pomysłem była tzw. CAPTCHA (Completely Automated Public Turning Test to Tell Computers and Humans Apart), czyli wymuszenie wpisania krótkiego ciągu znaków przepisanych z obrazka. Zadanie banalne dla ludzkiego umysłu, nieosiągalne jednak dla robotów. Pomysł został szybko i szeroko zaakceptowany przez społeczność internetową. Po kilku latach autor zdał jednak sobie sprawę, ile złego wyrządził światowej ekonomii. Miliony użytkowników internetu z powodu jego pomysłu musiało codziennie przepisywać kod

Rys. 4. Zrzut ekranowy z mechanizmu reCAPTCHA Google (http://irevolution.net/2013/06/17/recaptcha-for-disaster-response)

z obrazka aby zalogować się do banku, pobrać poszukiwany plik, wypowiedzieć się na forum publicznym czy przesłać efekty swojej pracy. Zadanie było kompletnie bezproduktywne i nawiązywało raczej do Keynesowskiej koncepcji kopania i zasypywania dołów niż do poprawiającej się produktywności człowieka u progu XXI wieku.

Luis von Ahn znalazł jednak zastosowanie dla codziennie wykonywanej przez miliony internautów roboty. Zamiast przypadkowych ciągów znaków użytkownikom miały się pojawiać słowa, z którymi nie radziło sobie tradycyjne rozpoznawanie tekstu drukowanego. Jedno z prezentowanych słów było już rozpoznane i służyło jako test wiarygodności danego użytkownika. Drugie z prezentowanych słów nie było rozpoznane lub algorytm nie miał pewności poprawnego odczytania. Słowo nieznane było pokazywane wiele razy – tak długo, aż uzyskiwało kilkukrotne wskazanie jednej wersji „odczytanej". W ten sposób przykry bezproduktywny obowiązek został zamieniony w strumień przydatnych informacji. Mechanizm został wkrótce wykupiony przez Google i udostępniony jeszcze większej liczbie odbiorców w formie takiej jak niżej:

Firma Google zastosowała mechanizm do poprawiania danetyzacji skanowanych książek z domeny publicznej (projekt prowadzony jest pod nazwą Google Books) oraz archiwów New York Times. Po wielkim sukcesie mechanizm zastosowano również do rozpoznawania numerów budynków z usługi Google Street View, umożliwiając kierowanie korzystających z map użytkowników pod konkretny numer budynku dla każdej ulicy, która taką numerację posiada.

Korzyści z technologii Big Data

W [3] podjęto także próbę wyliczenia finansowej równowartości wykonywanej przez anonimowych internautów pracy. Po przyjęciu założeń, że rozpoznanie tekstu zajmuje średnio 10 sekund, kod jest przedstawiany do rozpoznania 200 mln razy, a minimalna stawka wynagrodzenia w USA wynosi 7,25 dolarów/godzinę wyprowadzono wniosek, że ten mechanizm oszczędza cztery miliony dolarów dziennie.

Schemat, w którym najpierw pojawiają się narzędzia lub dane, a dopiero później wymyśla się dla nich zastosowanie eko-

Rys. 5. Zasada otwartego dostępu do danych zaczyna być modą

nomiczne dość często powiela się w *Big Data*. W [4] zaprezentowano firmę Equifax, agencję oceniającą zdolność kredytową podatników w USA na podstawie ogólnodostępnych w internecie danych, ale z zastosowaniem odpowiednich modeli, które zostały wykryte dzięki analizie korelacji. Kamieniem milowym tego pomysłu był wynik badań, które w skrócie można by określić jako „swój ciągnie do swego". Okazało się na przykład, że osoby niewywiązujące się ze zobowiązań wobec banków chętniej na swoich znajomych wybierają osoby w podobnej sytuacji, niż takie które regularnie spłacają swoje długi i są godni zaufania. Wynik socjologicznej analizy powiązań międzyludzkich doprowadził do powstania firmy tworzącej oceny kredytowe.

Big Data przynosi nie tylko efekty czysto finansowe. Wczesne wykrywanie rozprzestrzeniania się wirusa ptasiej grypy (Google Flu Trends) oszczędziło trudne do oszacowania kwoty, które trzeba byłoby wydać na leczenie komplikacji pogrypowych, czy zbyt późno zdiagnozowanej choroby, ale także oszczędziło gospodarce negatywnych konsekwencji samej choroby i jej skutków.

Big Data jest też wykorzystywana do wykrywania nieprawidłowości, oszustw i ograniczania szarej strefy. W 2011 r. firma Xoom, organizująca międzynarodowe transfery pieniędzy dostrzegła dzięki analizie *Big Data* nieznaczny wzrost transakcji dokonanych kartami kredytowymi Discover w New Jersey. Dalsza analiza ujawniła grupę przestępczą wykorzystującą setki podrobionych kart.

Powyższe przykłady dotyczyły sytuacji, w której analizowane dane nie są publicznie dostępne, a monopolistyczna (w sensie posiadanych informacji) organizacja znajduje dla nich nowe zastosowanie. Nie jest tak jednak zawsze.

Serwis internetowy FlyOnTime.us połączył dwa ogólnodostępne źródła informacji – informacje o planowanych i faktycznych godzinach lotów ze wszystkich dużych lotnisk w USA (dostępne na stronie tychże lotnisk) oraz informacje o pogodzie na terenie lotniska (dostępne na wielu stronach internetowych). Na podstawie tej zebranej i połączonej bazy danych można było uzyskać odpowiedź na pytanie „*przy jakich warunkach pogodowych z jakim opóźnieniem należy się liczyć na danym lotnisku?*".

Po uzupełnieniu bazy o prognozy pogody dla lotniska serwis może więc prognozować opóźnienia. Taka informacja jest już wiele warta dla osób, które zastanawiają się ile zarezerwować czasu na przesiadkę na danym lotnisku, czy też jak rozplanować podróż łączoną z innymi środkami transportu, czy też po prostu – o ile wcześniej przylecieć żeby nie spóźnić się na zaplanowane spotkania. Bezwartościowe i nawet niearchiwizowane dane, łatwo dostępne w kilku miejscach w internecie po ich zarchiwizowaniu, złączeniu i odpowiednim obrobieniu matematyczno-wizualnym okazały się być strumieniem gotówki. Trudno znaleźć podobną cechę innych powszechnie znanych czynników produkcji, jak kapitał, ziemia, ludzie czy czas.

Dane – lub szerzej – wiedza są ekonomicznie zasobem specyficznym. Po pierwsze jest to dobro *nierywalizujące*, czyli takie którego wykorzystanie przez jedną osobę nie uniemożliwia wykorzystania ponownego przez inne osoby. Te same informacje o pogodzie mogą być wykorzystywane przez wiele innych firm do stworzenia najróżniejszych zastosowań w wielu branżach. Co bardziej istotne – wykorzystywanie danych nie powoduje ich zużycia i spadku ich wartości, ale przeciwnie.

Dane surowe zyskują wartość po ich obróbce (danetyzacji). Ich analiza przez jeden podmiot może ujawnić możliwość ich zastosowania w innych branżach co ponownie podniesie ich wartość. Dane archiwalne, które dzisiaj zdają się nie mieć żadnej wartości, mogą nagle ją uzyskać kiedy w przyszłości odkryta zostanie nowa korelacja i niezbędna będzie duża próba danych do jej potwierdzenia lub negacji.

To wszystko prowadzi do wniosku, że dane – nawet w formie analogowej, surowej – mogą okazać się kopalniami złota. Są zasobem, który mimo, że jest właściwie nieograniczony to ta nieograniczoność nie powoduje spadku jego wartości – a wręcz przeciwnie – jej wzrost z racji na wzrost liczby obszarów, na których może być zastosowany.

Rewolucja już wybuchła

Najbardziej rozwinięte kraje świata zaczynają dostrzegać olbrzymią ekonomiczną wartość danych. Unia Europejska chce udostępniać wiele wyjściowych zbiorów danych o gospodarce całego kontynentu (a nie tylko końcowe wyniki analizy z EuroStatu). Wielka Brytania otworzyła Open Data Institute, który stanowi wyłom w restrykcyjnym prawie dotyczącym danych – Crown Copyright. Stany Zjednoczone Ameryki Północnej za prezydentury Barracka Obamy poszły jeszcze dalej. W utworzonym serwisie internetowym data.gov wszystkie agencje federalne są zobowiązane udostępniać jak najwięcej informacji ile jest możliwe bez narażania interesów USA lub interesów obywateli USA. „*Gdyby powstały jakieś wątpliwości, przeważa zasada otwartości*" przeczytać możemy w memorandum wspomnianego prezydenta.

Jeden problem zwykle rodzi jednak następny. Skoro możemy pobrać darmowe dane z ogólnodostępnych źródeł, poświęcić na ich przekształcenie kilkadziesiąt godzin koszt pracy kilkunastu komputerów rzędu 20 tys., a następnie uzyskać giełdową kapitalizację na poziomie kilku milionów dolarów opartą twardo o wyniki naszej sprzedaży i długoletnie kontrakty na dostawę analiz, to jak odwzorować taką operację w będącej podstawą analiz gospodarczych rachunkowości? Już dziś czołowe firmy zajmujące się praktycznym wykorzystaniem *Big Data*. 75% wartości bilansowej mają w pozycji „wartości niematerialne i prawne". W klasycznym ujęciu rachunkowości bilansowej umyka trzy czwarte wartości firmy i jest zapisywana w kolumnie „inne".

Niewątpliwie obserwujemy budzenie się rewolucji „obrotu danymi", która już trwa i której nie można zatrzymać. Warto dokładniej przyjrzeć się obszarom zastosowania *Big Data* i długofalowym skutkom zastosowania tej technologii. A wielkie światowe firmy już wykorzystują analizy Big Data, zaś inni wielcy – publikują oprogramowanie narzędziowe do tych analiz [5].

Literatura

[1] Nowakowski K., Nowakowski W.: 2016, Big Data *ante portas*. Elektronika – konstrukcje, technologie, zastosowania, nr 2, str. 27–30, DOI: 10.15199/13.2016.2.5

[2] Tabakow M., Korczak J., Franczyk B.: 2014, Big data – definicje, wyzwania i technologie informatyczne. *Informatyka ekonomiczna*.

[3] Mayer-Schönberger V., Cukier K.: 2014, Big data - rewolucja, która zmieni nasze myślenie, pracę i życie. MT Biznes.

[4] Thurm S.: 2011, Next Frontier in Credit Scores: Predicting Personal Behaviour. *Wall Street Journal*, October.

[5] Nowakowski K., Nowakowski W.: 2016, Hadoop, narzędzie technologii Big Data i jego aplikacje. Elektronika – konstrukcje, technologie, zastosowania, nr 3, str. 33–36, DOI: 10.15199/13.2016.3.7