WYDAWNICTWO UNIWERSYTETU ŁÓDZKIEGO

WSZYSTKO, CO TRZEBA WIEDZIEĆ!

- Snowden i WikiLeaks
- inteligentne miasta
- big data w biznesie

Dawn E. Holmes

BIG DATA

Tłumaczenie Robert Kowalczyk
Redakcja naukowa Piotr Fulmański

Original English language edition by
OXFORD UNIVERSITY PRESS

> KRÓTKIE WPROWADZENIE

BIG DATA

> *KRÓTKIE WPROWADZENIE*

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Dawn E. Holmes

BIG DATA

Tłumaczenie **Robert Kowalczyk**
Redakcja naukowa **Piotr Fulmański**

Original English
language edition by
OXFORD
UNIVERSITY PRESS

> *KRÓTKIE WPROWADZENIE*

Łódź 2021

Tytuł oryginału: *Big Data: A Very Short Introduction*

Rada Naukowa serii *Krótkie Wprowadzenie*
Jerzy Gajdka, Ewa Gajewska, Krystyna Kujawińska Courtney
Aneta Pawłowska, Piotr Stalmaszczyk

Redaktorzy inicjujący serii *Krótkie Wprowadzenie*
Urszula Dzieciątkowska, Agnieszka Kałowska

Tłumaczenie
Robert Kowalczyk

Redakcja naukowa
Piotr Fulmański

Opracowanie redakcyjne
Anna Surendra, Sebastian Surendra

Skład i łamanie
Munda – Maciej Torz

Projekt typograficzny serii
Tomasz Przybył

Projekt okładki
krzysztof de mianiuk

Zdjęcie wykorzystane na okładce: © Depositphotos.com/stillfx

Big Data: A Very Short Introduction was originally published in English in 2017.
This translation is published by arrangement with Oxford University Press.
Wydawnictwo Uniwersytetu Łódzkiego is solely responsible for this translation
from the original work and Oxford University Press shall have no liability for any
errors, omissions or inaccuracies or ambiguities in such translation or for
any losses caused by reliance thereon

Publikacja sfinansowana ze środków Wydawnictwa Uniwersytetu Łódzkiego

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.09310.19.0.M
Ark. wyd. 6,0; ark. druk. 9,75

Paperback ISBN Oxford University Press: 978-0-19-877957-5
ISBN 978-83-8220-061-4
e-ISBN 978-83-8220-062-1

Spis treści

Spis ilustracji

Przedmowa

Książki poświęcone dużym zbiorom danych[1] można podzielić na dwie kategorie: takie, które nie wyjaśniają kwestii, czym są duże zbiory danych, i takie, które wymagając usystematyzowanej wiedzy matematycznej, przeznaczone są tylko dla zaawansowanych studentów. Celem tej książki jest przedstawienie innego podejścia do kwestii, czym są duże zbiory danych i jak zmieniają świat; jaki wpływ mają na nasze codzienne życie, jak i na świat biznesu.

Kiedyś przez dane rozumiano kartki papieru, dokumenty, czasem zdjęcia, ale dzisiaj to coś znacznie więcej. Sieci społecznościowe generują duże ilości danych w formie obrazów, zdjęć i filmów. Zakupy przez Internet generują dane, kiedy podajemy nasz adres mailowy czy numer karty kredytowej. Jesteśmy w takim momencie historii, w którym gromadzenie i przechowywanie danych wzrasta w tempie niewyobrażalnym w stosunku do wcześniejszych dziesięcioleci i, jak zobaczymy dalej, nowe techniki analizy danych przekształcają je w użyteczne informacje. Podczas pisania tej książki odkryłam, że duże zbiory danych nie mogą być omawiane bez częstego odwołania się do tego, skąd pochodzą, co przechowują, a także bez ich analizy i użycia przez duże komercyjne firmy. Ponieważ w ośrodkach badawczych takich firm jak Google czy Amazon rozwijały się technologie związane z dużymi zbiorami danych, często będę się do nich odwoływała.

[1] Takie zbiory przyjęło się nazywać w literaturze anglojęzycznej terminem „big data", do czego nawiązuje tytuł niniejszej książki. W niektórych miejscach będziemy się tym terminem posługiwali, myśląc o „dużych zbiorach danych" [wszystkie przypisy pochodzą od tłumacza].

Pierwszy rozdział ogólnie zapoznaje czytelnika z różnorodnością danych, zanim będzie wyjaśnione, jak era cyfrowa doprowadziła do zmian w sposobie ich definiowania. Pojęcie dużych zbiorów danych wprowadzone jest nieformalnie przez ideę eksplozji danych, która obejmuje informatykę, statystykę i ich wzajemne powiązania. W rozdziałach od drugiego do czwartego wielokrotnie używam diagramów, które pozwalają wyjaśnić niektóre nowe metody wymagane w dużych zbiorach danych. Drugi rozdział poszukuje tego, co czyni duże zbiory danych wyjątkowymi, doprowadzając nas do lepszej definicji tego pojęcia. W rozdziale trzecim analizujemy kwestie związane z przechowywaniem i zarządzaniem dużymi zbiorami danych. Większości z nas znana jest konieczność robienia kopii zapasowych na osobistym komputerze. Ale jak tego dokonać w przypadku olbrzymiej ilości danych, które są obecnie generowane? Żeby odpowiedzieć na to pytanie, przyjrzymy się przechowywaniu danych i idei ich rozdzielenia pomiędzy grupy komputerów. Rozdział czwarty pokazuje, że duże zbiory danych są użyteczne tylko wtedy, gdy możemy wydobyć z nich istotne dla nas informacje. Zarys tego, jak dane przekształcane są w użyteczne informacje, podany jest z wykorzystaniem uproszczonych opisów kilku dobrze znanych technik.

Następnie przechodzimy do bardziej szczegółowych dyskusji na temat wykorzystania dużych zbiorów danych, rozpoczynając w rozdziale piątym od ich roli w medycynie. Rozdział szósty wyjaśnia praktyki biznesowe z analizą przypadków firm Amazon i Netflix, za każdym razem podkreślając różne cechy marketingu opartego na dużych zbiorach danych. W rozdziale siódmym przyglądamy się pewnym problemom związanym z bezpieczeństwem dużych zbiorów danych i ważności (konieczności) ich szyfrowania. Kradzież danych staje się dużym problemem i w tym miejscu przyjrzymy się niektórym znanym medialnie wydarzeniom, takim jak przypadek Snowdena i historia WikiLeaks. Na zakończenie rozdziału pokazano, w jaki sposób cyberprzestępczość stanowi problem wymagający brania pod uwagę w przypadku dużych zbiorów danych. W rozdziale ósmym rozważamy,

jak duże zbiory danych zmieniają społeczeństwo, w którym żyjemy, poprzez rozwój zaawansowanych technologicznie robotów i ich roli w środowisku pracy. Książka kończy się rozważaniami dotyczącymi inteligentnych domów i miast przyszłości.

W krótkim wprowadzeniu nie jest możliwe poruszenie wszystkich zagadnień, mam więc nadzieję, że czytelnik będzie pogłębiał swoją wiedzę w oparciu o polecane na zakończeniu w części *Lektura uzupełniająca* materiały.

Podziękowania

Kiedy wspomniałam Peterowi, że chciałam podziękować za jego wkład w powstanie tej książki, zasugerował, abym napisała: „Dziękuję Peterowi Harperowi, bez którego wkładu w sprawdzenie pisowni książki byłaby to zupełnie inna książka". Dodatkowo chciałabym podziękować mu za wiedzę w zakresie parzenia kawy oraz poczucie humoru! Wsparcie Petera jest nieocenione, zrobił dużo, dużo więcej i prawdą jest to, że bez jego nieustającej zachęty i konstruktywnego wkładu ta książka nie zostałaby napisana.

Dawn E. Holmes
kwiecień 2017 r.

Rozdział 1

Eksplozja danych

Czym są dane?

W 431 r. p.n.e. Sparta wypowiedziała wojnę Atenom. Tukidydes w swoim opisie wojny wyjaśnia, jak oblężone siły platejskie, lojalne w stosunku do Aten, planowały ucieczkę. Cel ten chciano osiągnąć, wspinając się i przechodząc przez mur otaczający Plateje, który został zbudowany przez siły peloponeskie kierowane przez Spartan. Aby to zrobić, potrzebowali wiedzieć, jak wysoki jest mur, po to, żeby skonstruować odpowiedniej wysokości drabiny. Większość muru peloponeskiego była pokryta chropowatym tynkiem z drobnych kamieni, ale znaleziono fragment, gdzie cegły były wyraźnie widoczne. W związku z tym dużej liczbie żołnierzy przydzielono zadanie liczenia warstw odsłoniętych cegieł w murze. Obliczenia były prowadzone w bezpiecznej, ale znacznej odległości od wroga, co wpływało na błędy rachunkowe, ale – jak wyjaśnia Tukidydes – biorąc pod uwagę, że wykonano wiele prób obliczeń, rezultat, który pojawiał się najczęściej, przyjęto za prawidłowy. Najczęściej pojawiający się wynik, który teraz nazwalibyśmy *dominantą*, został później użyty do obliczenia wysokości muru. Znając wymiary cegieł używanych w tym rejonie, platejanie byli w stanie skonstruować drabiny o wymaganej wysokości muru. To umożliwiło ucieczkę kilkuset ludziom, a ten epizod można uznać za najbardziej imponujący historyczny przykład pozyskiwania i analizy danych. Ale, jak zobaczymy dalej, pozyskiwanie, przechowywanie i analiza danych poprzedzała o stulecia nawet czasy Tukidydesa.

Na patykach, kamieniach i kościach odnalezione zostały nacięcia, które sięgają czasów górnego paleolitu. Choć nadal jest to przedmiotem dyskusji akademickiej, to nacięcia te uważane są za przykład danych reprezentujących liczby[2]. Być może najsłynniejszym tego przykładem jest kość z Ishango znaleziona w Demokratycznej Republice Konga w 1950 r., której wiek szacuje się na ok. 20 000 lat. Nacięcia te były różnie interpretowane, np. jako kalkulator czy kalendarz, choć są również opinie, że służyły do lepszego chwytania. Kość z Lebombo odkryta w latach 70. XX w. w Suazi jest jeszcze starsza i pochodzi z ok. 35 000 r. p.n.e. Z 29 nacięciami w poprzek ten fragment kości strzałkowej pawiana jest uderzająco podobny do kalendarzy umieszczanych na patykach przez Buszmenów w odległej Namibii, co sugeruje, że w rzeczywistości może to być metoda wykorzystywana do zapisu danych ważnych dla ich cywilizacji.

Podczas gdy interpretacja tych naciętych kości jest wciąż przedmiotem spekulacji, wiemy, że jednym z pierwszych dobrze udokumentowanych zastosowań danych jest spis ludności przeprowadzony przez Babilończyków w 3800 r. p.n.e. Ten spis powszechny systematycznie dokumentował liczbę ludności i towarów, takich jak mleko i miód, w celu zapewnienia informacji niezbędnych do obliczenia podatków. Pierwsi Egipcjanie również używali danych w postaci hieroglifów zapisanych na drewnie lub papirusie, w celu notowania dostaw towarów i śledzenia podatków. Wczesne przykłady używania danych w żadnym wypadku nie ograniczają się do Europy i Afryki. Inkowie i ich południowoamerykańscy poprzednicy, prowadząc statystyki do celów podatkowych i handlowych, używali zaawansowanego i złożonego systemu kolorowych sznurków wiązanych w supły, zwanych *quipu*, jako systemu obliczeń dziesiętnych. Te wiązane sznurki wykonane z jaskrawo barwionej bawełny lub wełny wielbłąda, pochodzą z trzeciego tysiąclecia przed naszą erą, i chociaż mniej niż tysiąc z nich przetrwało hiszpańską in-

[2] Zapisywane w postaci tzw. unarnego systemu liczbowego (ang. *telly marks*), czyli systemu, gdzie wartość liczbową literału otrzymujemy przez zsumowanie ilości wystąpień powtarzającego się symbolu.

wazję i późniejsze próby pozbycia się ich, należą do pierwszych znanych przykładów systemu do przechowywania dużych zbiorów danych. Obecnie opracowywane są algorytmy komputerowe w celu odkodowania pełnego znaczenia *quipu* i lepszego zrozumienia tego, w jaki sposób były wykorzystywane.

Pomimo że opisujemy te wczesne systemy liczbowe, używając słowa „dane", jest ono w zasadzie wyrazem liczby mnogiej pochodzenia łacińskiego, gdzie liczbą pojedynczą jest słowo „datum". „Datum" jest obecnie rzadko używanym słowem, a słowo „dane" (ang. *data*) jest używane zarówno w liczbie pojedynczej, jak i mnogiej[3]. *Słownik oksfordzki* przypisuje pierwsze znane użycie tego terminu XVII-wiecznemu angielskiemu duchownemu Henry'emu Hammondowi w kontrowersyjnym traktacie religijnym opublikowanym w 1648 r. Hammond użył w nim pojęcia „sterta danych" w znaczeniu teologicznym w nawiązaniu do niepodważalnych prawd religijnych. Ale chociaż ta publikacja wyróżnia się jako ta, która po raz pierwszy wprowadza użycie terminu „dane" w języku angielskim, nie posługuje się nim w nowoczesnym znaczeniu dla oznaczenia faktów i liczb dotyczących interesującej nas populacji. W dzisiejszym rozumieniu termin „dane" wywodzi się z rewolucji naukowej z XVIII w. reprezentowanego przez geniuszy, takich jak Priestley, Newton i Lavoisier. Po 1809 r. pojawiły się prace matematyków, takich jak Gauss i Laplace, którzy dali podwaliny pod współczesną metodologię statystyczną.

Na poziomie bardziej praktycznym dużą ilość danych zebrano w 1854 r. podczas wybuchu epidemii cholery na Broad Street w Londynie, co pozwoliło lekarzowi Johnowi Snowowi na zobrazowanie rozwoju epidemii. W ten sposób był w stanie poprzeć swoją hipotezę, że zanieczyszczona woda rozprzestrzeniała chorobę, co pozwoliło mu udowodnić, że to nie powietrze

[3] W języku angielskim często pisze się „data is", co na język polski należałoby przetłumaczyć „dane jest", a nie „dane są". Poprawnie powinno być „datum is" lub „data are". W tym przypadku jednak rzeczownik „data" używany jest jako tzw. rzeczownik zbiorowy (ang. *mass noun*) podobnie jak piasek czy deszcz, tzn. powiemy „dużo piasku", a nie „dużo piasków".

było przyczyną epidemii, jak wcześniej sądzono. Zbierając dane od lokalnych mieszkańców, ustalił, że wszyscy poszkodowani używają tej samej publicznej pompy wodnej. Następnie przekonał władze miejscowej parafii do jej unieruchomienia, przy czym cel ten osiągnięto poprzez usunięcie uchwytu pompy. Później Snow stworzył mapę pokazującą, że epidemia pojawiła się w skupiskach wokół pompy Broad Street. Kontynuował pracę w tej dziedzinie, zbierając i analizując dane, dzięki czemu obecnie jest znany jako pionier epidemiologii.

Kontynuując pracę Johna Snowa, epidemiolodzy i badacze społeczni coraz częściej uważają dane demograficzne za nieocenione źródło celów badawczych, a przeprowadzony obecnie w wielu krajach spis ludności pokazuje, że jest to cenne źródło informacji. Obecnie gromadzone są np. dane dotyczące urodzeń i zgonów, częstotliwości występowania różnych chorób i statystyki dotyczące dochodów i przestępstw, co nie było stosowane przed XIX w. Spis powszechny, który w większości krajów odbywa się co dziesięć lat, gromadzi coraz większe ilości danych, co doprowadza do sytuacji, w której ilość przetwarzanych danych przekracza możliwości ich rejestracji – ręcznej, prowadzonej za pomocą prostych maszyn liczących używanych wcześniej. Wyzwanie stojące przed przetwarzaniem tych stale rosnących ilości danych spisu powszechnego zostało w pewnym stopniu podjęte przez Hermana Holleritha podczas jego pracy w amerykańskim biurze do spraw spisu ludności.

Do momentu spisu powszechnego w Stanach Zjednoczonych w 1870 r. używano prostej maszyny liczącej, która w niewielkim stopniu ułatwiała pracę biura. Przełom nastąpił w czasie spisu powszechnego w 1890 r., kiedy użyto maszyny analityczno-liczącej[4] Hermana Holleritha do przechowywania i przetwarzania danych. Na przetworzenie danych ze spisu powszechnego w Stanach Zjednoczonych potrzebowano zwykle ok. ośmiu lat, natomiast użycie tego wynalazku skróciło czas do jednego roku. Maszyna Holleritha zrewolucjonizowała analizę spisu po-

[4] W oryginale: „punched cards tabulator", czyli maszyna analityczno-licząca wykorzystująca karty dziurkowane.

wszechnego w krajach na całym świecie, w tym w Niemczech, Rosji, Norwegii i na Kubie.

W końcu Hollerith sprzedał swoją maszynę firmie, która przekształciła się w IBM. W konsekwencji maszyna ta została udoskonalona i zaczęto sprzedawać ją na szeroką skalę. W 1969 r. American National Standards Institute (ANSI) ustandaryzował format karty kodów Holleritha (Hollerith Card Code), uznając jego wkład w powstanie karty perforowanej.

Dane w erze cyfrowej

Przed powszechnym użyciem komputerów dane ze spisu powszechnego, eksperymentów naukowych lub starannie zaprojektowane przykładowe ankiety i kwestionariusze były zapisywane na papierze – proces ten był czasochłonny i kosztowny. Zbieranie danych mogło nastąpić dopiero po tym, gdy naukowcy zdecydowali, na które pytania chcieli odpowiedzieć, przeprowadzając eksperymenty i ankiety, a uzyskane w ten sposób, wysoce ustrukturyzowane dane, zapisane na papierze w uporządkowanych wierszach i kolumnach, były następnie poddawane tradycyjnym metodom analizy statystycznej. W pierwszej połowie XX w. niektóre dane były przechowywane na komputerach, co częściowo ułatwiało tę wymagającą wielu nakładów pracę, a było możliwe dzięki powstaniu sieci WWW (w skrócie sieci Web) w 1989 r. oraz jej szybkiemu rozwojowi. W rezultacie coraz bardziej możliwe stało się generowanie, gromadzenie, przechowywanie i analizowanie danych w formie elektronicznej. Nieuchronnym skutkiem tego było pojawienie się problemów powodowanych przez bardzo dużą ilość danych udostępnianych przez Internet, które to problemy musiały zostać rozwiązane. Przyjrzymy się najpierw, jak możemy rozróżniać różne typy danych.

Dane, które uzyskujemy z sieci Web, można sklasyfikować jako: ustrukturyzowane (ang. *structured*), nieustrukturyzowane (ang. *unstructured*) lub częściowo ustrukturyzowane (ang. *semi-structured*).