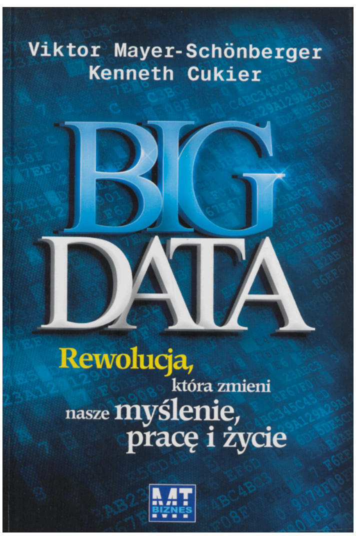
ALEKSANDRA MIKOŁAJSKA

Viktor Mayer-Schönberger, Kenneth Cukier, *BIG DATA: rewolucja, która zmieni nasze myślenie, pracę i życie*, przeł. M. Głatki, Warszawa: MT Biznes 2014, s. 280



Coraz częściej w literaturze spotykamy się z pojęciem *big data*. Analizując różnorodność technologii cyfrowych, zdajemy sobie sprawę, że internet dał nam możliwość komunikowania się m.in. ze światem. Co nam zatem da *big data*? Termin ten pojawił się w ostatnich latach i jest tak samo znany jak na przykład rozwiązania chmurowe. Według autorów omawianej książki *big data* „odnosi się do naszych nabytych umiejętności przetwarzania ogromnych ilości informacji, ich błyskawicznej analizy i wyciągania odkrywczych wniosków" (informacja na czwartej stronie okładki). Autorami publikacji są badacze internetu: pierwszy z nich jest profesorem Uniwersytetu w Oxfordzie, drugi dziennikarzem z „The Economist". Przedmiotem ich analizy są dane. Przyrost danych w dzisiejszym świecie jest ogromny. Jest to znaczące wyzwanie dla wielu instytucji, ponieważ są dla nich nowym obszarem działania, a przy takiej ilości powstaje problem z ich wykorzystaniem, przechowywaniem i zarządzaniem.

Książka składa się z dziesięciu rozdziałów. W pierwszym, zatytułowanym *Teraźniejszość* opisana została m.in. wyszukiwarka Google pod względem liczby archiwizowanych tam zapytań. Każdego dnia do wyszukiwarki kierowane są 3 miliardy zapytań. Dzięki odpowiedniemu oprogramowaniu z danych można wyczytać dużo więcej. Być może odpowiedzą one na pytanie, gdzie wybuchnie epidemia grypy H1N1, które linie mają najtańsze bilety lotnicze, w których supermarketach różne dobra i usługi są najtańsze. To zjawisko posługiwania się danymi i wyciągania z nich wniosków już istnieje.

Dane analizowane są od tysiącleci, czego dowodem są prowadzone od czasów biblijnych spisy powszechne. *Big data* jest wyzwaniem dla naszego życia i naszej interakcji ze światem. Autorzy książki stawiają tezę, że *big data* niesie ze sobą trzy zasadnicze zmiany.

Pierwsza z nich dotyczy analizy wszystkich danych, gromadzonych na różnych serwerach i przeznaczonych do badania. Kiedyś badania statystyczne opierały się wyłącznie na danych wybranych losowo. Obecnie są badane wszystkie dane. Druga zmiana dotyczy jakości tych danych. Są one zróżnicowane, nieuporządkowane. Tolerancja błędów jest do zaakceptowania, ważna jest ilość, a nie jakość. Ważniejsze jest poznanie jakiegoś kierunku rozwoju danego zjawiska, a nie jego szczegóły. Trzecia zmiana, bardzo istotna, prowadzi do odpowiedzi na pytanie „co się dzieje?", a nie „dlaczego?". Odkryte związki między danymi przemówią do nas i wskażą nam kierunek działania. Kiedyś ludzie mierzyli świat na podstawie prób losowych. Nie mieli narzędzi do tego, by dane przetworzyć. Obarczone one były dużymi błędami. Badania w podgrupach nie odzwierciedlały właściwej informacji o całości zjawiska. W obecnej erze cyfrowej problem jest rozwiązany. Przetwarzanie danych jest szybsze. W drugim rozdziale zatytułowanym *Więcej* jest o tym mowa. W ciągu sekundy można dokonać milionów obliczeń. Autorzy podają przykład zainstalowania w Chile w 2016 roku nowego teleskopu, który będzie gromadził określoną liczbę danych co pięć dni. Przedtem potrzebował na to kilka dziesięcioleci. Jeszcze niedawno badanie wycinka naszego DNA trwało dziesięć lat, od 2003 roku tylko trzy dni.

Co oznacza bezład w *big data*? Nad tym zastanawiają się autorzy w następnym rozdziale. Z pewnością przy tak ogromnej liczbie danych nie uniknie się chaosu. Liczba danych się zwiększa, wzrasta niedokładność. W *big data* nie chodzi jednak o precyzję, tylko o prawdopodobieństwo. Tak więc mając większą ilość danych i różne narzędzia do ich pomiarów, likwiduje się niedokładności. Należy pamiętać, że wydajność algorytmów poprawia się, im więcej jest danych. Przykładem tego jest tłumaczeniowa usługa Google, która dokonuje tłumaczeń między 60 językami, dzięki zastosowaniu olbrzymiej liczby algorytmów. W roku 1954 do przetłumaczenia pierwszego zdania przez IBM wykorzystano tylko sześć reguł gramatycznych. Nastąpiło to w okresie zimnej wojny. Przetłumaczone zdanie po rosyjsku brzmiało „Przekazujemy myśli za pośrednictwem mowy".

Stosując różne narzędzia, odpowiednie oprogramowanie, dokonuje się zmian w różnych dziedzinach życia. Najwięcej jednak dzieje się na rynku konsumpcyjnym. Walka o klienta trwa cały czas. Urzędowe statystyki są mało warte, ukazują się za późno. Każdy z nas chce za produkt zapłacić jak najmniej. Stworzony przez MIT projekt PriceStats dotyczy porównywania cen milionów produktów sprzedawanych na całym świecie. Jest ofertą dla banków i różnych podmiotów gospodarczych.

Dalej autorzy mówią o hierarchicznych systemach klasyfikacji, czego przykładem są m.in. katalogi biblioteczne. Przy małej liczbie danych klasyfikacja jest możliwa. Nie jest to jednak możliwe przy dużej ilości danych. Jako przykład autorzy podają serwis Flickr – do publikowania zdjęć w internecie. Do 2011 roku w serwisie tym było ponad 6 miliardów zdjęć. Ułożenie ich według jakiejś kategorii przedmiotowej byłoby niemożliwe. Klasyfikacja została zastąpiona tagowaniem. Każdy użytkownik, wysyłając swoje zdjęcie, tworzy i dołącza opis. Brak jest więc standardów. Nie ma ustalonej taksonomii. Na takich zasadach działają portale społecznościowe czy blogi. To uświadamia, że taksonomia jest przeżytkiem. Powinniśmy, według autorów, przyjąć ten rozgardiasz i zapomnieć o porządku rzeczy, do którego nasze umysły są przyzwyczajone od stuleci. Czy to oznacza, że będziemy bardziej doceniać niedbalstwo zamiast dokładności?

Big data narzuca nam zmiany, z którymi powinniśmy się pogodzić. Czy ten bezład bardziej zbliża nas do nowej rzeczywistości? Trzecią zmianą, o której piszą autorzy, jest poszukiwanie związków między danymi. Szukanie odpowiedzi na pytanie, jakie są przyczyny tego, co się dzieje, znajdziemy w rozdziale czwartym, zatytułowanym *Korelacja*. O jaką współzależność autorom chodzi? Czy raz zakupiona książka o kwiatach oznacza, że będziemy kupować tylko podobne pozycje? Autorzy tłumaczą, dlaczego grupa redaktorów, literatów pracujących dla Amazon została rozwiązana?

Otóż wyeliminowana została przez porównywanie związków między danymi. Zrobiły to komputery, które wykazały, że ich recenzje na stronie internetowej nie generowały większej sprzedaży. Zysk dla firm jest najważniejszy. Dlatego rekomendacje wykonane przez komputer wyparły recenzentów. Z tego przykładu wynika, że sprzedaż czegokolwiek powinna być od czegoś zależna. Otacza nas mnóstwo danych, więc współzależność jest wszędzie. Autorzy zadają pytanie, czy stawianie hipotez w przyszłości ma sens. W swoich twierdzeniach poszli tak daleko, iż uważają, że może nastąpić „koniec teorii naukowych". Później się z tego wycofują, bo przecież *big data* jest oparta na naukach ścisłych: chemia, fizyka czy matematyka to przede wszystkim teorie. Tak więc *big data* nie wyeliminuje teorii, ale spowoduje zmianę sposobu nadawania sensu naszemu światu. Potrzebne będą szybsze procesory i skuteczniejsze oprogramowanie.

Wiele aspektów naszej rzeczywistości przekładamy na dane. Dane prowadzą nas do nowego świata. Kolejny, piąty rozdział poświęcony jest *Danetyzacji*. Autorzy opisali w nim proces zapisywania danych, analizowania i reorganizowania. Obydwaj twierdzą, że dane powinny być wydobywane zewsząd, bo nigdy nie wiadomo, kiedy się przydadzą. Każdą informację mamy zmierzyć i zapisać. Ponownie autorzy wracają do przeszłości, przypominając nam, że liczenie i mierzenie są to najstarsze narzędzia cywilizacyjne i stanowią podstawę procesu danetyzacji.

Nie zdajemy sobie sprawy, kiedy nasze słowa stają się danymi. W 2004 roku firma Google ogłosiła, że każda strona wszystkich książek zostanie zeskanowana. Nie była to danetyzacja, ale cyfryzacja. Potem zrozumiano, że prawdziwa wartość kryje się w słowach. Potrzebna była nowa technologia, by tę ukrytą wartość wydobyć. Dokonało się to dzięki użyciu technologii OCR. Wszystkie znaki, wykresy, obrazy, litery zawarte w dokumentach są obecnie rozpoznawalne. Dzięki temu oprogramowaniu dowiadujemy się, kiedy słowa jako idee zostały użyte po raz pierwszy, jak ewoluowała myśl ludzka i jak się rozprzestrzeniała w różnych krajach. Danetyzacja sprawiła, że powstała nowa dziedzina naukowa zwana kulturomiką. Jest to komputerowa leksykologia, „nauka, która próbuje zrozumieć ludzkie zachowania i trendy kulturowe przy pomocy ilościowej i statystycznej analizy tekstów" (s. 116). To właśnie danetyzacja przyczyniła się również do wykrywania plagiatów.

Wielu wydawców popiera wersje elektroniczne, dla nich najważniejsza jest treść. To jest ich sposób na biznes. Nie dostrzegli jednak przez lata potrzeby, by treść zamienić w dane. Prawdziwy potencjał znajduje się w danych. Informacje, jakimi dzielimy się na portalach, takich jak Facebook czy Twitter, to nie tylko nic nieznacząca gadanina. Analizowanie zawartych tam treści odbywa się cały czas. Dane od tych portali kupują firmy, które stosują techniki tzw. nacechowania emocjonalnego, czyli gromadzą opinie klientów na temat wielu produktów. To może być wskazówką, np. czy film otrzyma prestiżową nagrodę lub też w jakim regionie świata wybuchnie epidemia.

Autorzy uważają, że dzięki *big data* będziemy patrzeć na świat składający się głównie z informacji, a nie przez pryzmat zjawisk społecznych i przyrodniczych.

Głównym użytkownikiem danetyzacji jest biznes, gdzie *big data* służy do tworzenia nowych wartości. Informacja zawsze była ważna. Pomagała w strategicznych transakcjach biznesowych. Teraz w epoce *big data* wszystkie dane mają wartość. Wiadomo jest już, że na sukcesy gospodarcze niektórych krajów wpłynęły właśnie analizy danych. O tym, jaka jest ich wartość i czy można ją wielokrotnie wykorzystać, dowiemy się w następnym rozdziale. Najcenniejszą wartością dla handlu jest dostęp do danych osobowych. Przekonujemy się o tym na co dzień, gdy jesteśmy proszeni o wypełnienie ankiet. Wtedy to handlowcy poznają nasze zainteresowania, zapoznają się z naszymi opiniami, zwyczajami czy kontaktami z innymi konsumentami. Tworzą swoje modele biznesowe nastawione na zysk. Dane będą przez nich wykorzystywane tak długo, dopóki będą rentowne.

Nikt się nie pozbywa starych stron internetowych ze starymi zapytaniami, bo nie wiadomo, kiedy mogą być ponownie wykorzystane przez firmy konsumenckie. Przechowywanie ich jest tanie. Innowacyjne firmy powinny o tym pamiętać.

Autorzy poruszają również w tym rozdziale problem łączenia danych w celu wydobycia innych wartości. Znane są przez użytkowników internetu tzw. *mushupy* – strony łączące w nowatorski sposób informacje z wielu źródeł. Między innymi w ten sposób badano związek wykorzystania telefonów komórkowych ze zwiększeniem ryzyka zachorowania na raka.

Dalej autorzy pokazują, jak ważne jest powtórne czy kolejne wykorzystanie danych. Prym wiedzie w tym wyszukiwarka Google, która zbiera dane z myślą o ich rozszerzeniu. Ich projekt Street View rejestrował nie tylko ulice i domy, ale zapisywane były dane GPS, sprawdzano dokładność map. Projekt obejmował również zapis nazw sieci WiFi prywatnych osób. Jedna podróż samochodem gromadziła zbiór danych, które firma wykorzystywała do innych celów.

W dalszej części rozdziału autorzy piszą o wartości danych resztkowych. Są to dane, które powstają jako uboczny produkt naszych działań w internecie. Jest to cyfrowy ślad naszych poszukiwań. Google tego nie lekceważy, wyznając zasadę „uczenia się z danych". Każde nasze działanie jest odbierane jako sygnał, który poddany analizie wraca do ich systemu. Powtórne wykorzystanie danych przybiera rozmaite formy. Na przykład przy tworzeniu poprawnej pisowni wyrazów.

Firma Google jest w tym najlepsza. Posiada najbardziej kompletny system sprawdzania pisowni wyrazów w różnych językach. Codziennie jest on udoskonalany dzięki naszym poszukiwaniom w wyszukiwarkach. Napisane błędnie słowa znajdują się w trzech miliardach zapytań kierowanych do wyszukiwarki. Tak więc to my wszyscy jesteśmy twórcami poprawnej pisowni, odpowiadając twierdząco na pytanie firmie Google „czy chodziło ci o". Odruchowo informujemy Google, jak pisze się dane słowo.

Google pokazuje nam, że stare zapytania nie są śmieciami. Są one agregowane i mogą być przydatne wydawcom, serwisom edukacyjnym czy też być źródłem przewagi konkurencyjnej.

Firmy Google i Amazon są pionierami *big data*, posiadają i udostępniają nam wiele danych. Należy jednak pamiętać, że pierwszymi instytucjami, które zaczęły gromadzić dane i mają ich najwięcej, są rządy państw. Instytucje publiczne, co jest rzeczą naturalną, zmuszają nas przecież do dostarczenia danych. Firmy prywatne muszą się o nie starać. Generalnie władze zawsze będą mieć więcej danych. W związku z tym powstają nowe inicjatywy Open Government Data (OGD), mające na celu udostępnienie danych obywatelom. Oczywiście tylko tych danych, które nie zagrażają bezpieczeństwu państwa i prywatności osób.

Przełomem w idei OGD na całym świecie był pierwszy dzień prezydentury Obamy, kiedy nakazał on agencjom federalnym udostępnić taką liczbę danych jak to możliwe. Efektem tej odważnej decyzji było powstanie strony data.gov. Wzorem USA w innych państwach też zaczęto wcielać ideę otwartego dostępu do danych rządowych. Dla zwykłego obywatela oznacza to, że może on od urzędników domagać się wglądu do danych, o które kiedyś nie miał odwagi nawet zapytać.

Przykładem wykorzystania otwartych danych rządowych są również strony, gdzie użytkownicy mogą dowiedzieć się, że np. niesprzyjająca pogoda może opóźnić loty samolotów. Jednostki, które nie gromadzą danych ani nie kontrolują ich przepływu, mają dostęp do nich i wykorzystują je, żeby stworzyć nową wartość.

Ostatnią kwestią poruszaną przez autorów w rozdziale o wartości danych jest ich wycena. Powstaje wiele giełd, które eksperymentują z wyceną danych. Różnica między wartością księgową a rynkową danych to wartości niematerialne i prawne. Wiele firm zrezygnowało z uwzględnienia ich w swoich bilansach. Wiele danych jest „uśpionych", nie do wyceny. Sprzedaje się licencje na nie. Brak jest konkretnego modelu wyceny. Sam fakt posiadania danych niewiele znaczy. Liczy się potencjał ukryty w danych i możliwość ich wielokrotnego użycia do nowych celów, czyli według autorów wartość opcyjna.

W następnym rozdziale – *Implikacje* – autorzy analizują sposób wykorzystania przez firmy *big data* do swoich działań. Opisują każdą kategorię posiadaczy danych, charakteryzując ich za pomocą trzech czynników: posiadanie danych, umiejętności i idee. Te firmy, które posiadają dane, nie zawsze wydobywają z nich wartość. Przykładem jest ponownie Twitter, który zdając sobie sprawę z posiadanego bogactwa danych, zaproponował innym firmom udzielenie licencji na wykorzystanie swoich danych. W drugiej grupie firm znajdują się firmy konsultingowe, analityczne, sprzedające nowe technologie. Charakteryzują się one kreatywnością, nie mają dostępu do danych, ale wiedzą, jak je wykorzystać. Natomiast trzecia grupa firm ma pozytywne nastawienie do *big data*. Założyciele tych firm i osoby tam pracujące mają wyjątkowe pomysły na wykorzystanie danych i wydobywają z nich nowe wartości. W związku z tym autorzy nadmieniają, że w ostatnich latach mamy do czynienia z nowym zawodem. Jest to badacz danych, ekspert od danych, który jest trochę statystykiem, programistą i infografikiem. To zawód z przyszłością. Będą się liczyły umiejętności wydobycia cennych informacji. Analizy eksperckie przeprowadzane w różnych branżach mogą stracić na znaczeniu. Specjalista od analizy danych nie musi być ekspertem np. w medycynie. Następstwa analizy danych mogą być różne. Autorzy podają wiele przykładów. Rozdział kończą rozważaniami na temat kwestii użyteczności danych i zmian struktury całych branż. Dotychczas, mówiąc o przewadze konkurencyjnej jakiejś firmy, mieliśmy na myśli wykorzystanie zasobów materialnych. Obecnie źródłem tej przewagi staje się *big data*. Wyścig małych i dużych firm już trwa.

Big data ma również swoje mroczne strony. To nowe zjawisko wpływa na naszą prywatność i poczucie wolności. Przeczytamy o tym w rozdziale *Zagrożenia*. Rozdział ten zaczyna się od omówienia działalności Stasi (służba bezpieczeństwa NRD), najbardziej rozbudowanego systemu inwigilacji obywateli, i porównania liczby akt tej służby z ilością gromadzonych o nas danych obecnie. Autorzy piszą o nieodpowiedzialnym przekazywaniu danych podmiotom, które wykorzystując swoją władzę, w przeszłości działały na szkodę obywateli. Niechlubnym tego przykładem jest użycie holenderskich akt cywilnych do aresztowań Żydów i ich eksterminacji. Kwestia odpowiedzialności za przekazywanie danych jest niezwykle ważna.

Dalej autorzy kontynuują wątek inwigilowania obywateli. Łatwo się domyślić, że jesteśmy pod ciągłą obserwacją, bo korzystamy z kart kredytowych, telefonów komórkowych, podajemy swój numer PESEL, jeśli sytuacja tego wymaga. W narzędzia, którymi się posługujemy, wbudowane jest oprogramowanie do przechwytywania danych. Cieszymy się, mając GPS w naszych smartfonach, ale tym samym policja zna naszą geolokalizację. Zachowanie naszej intymności życia prywatnego jest minimalne. Zgromadzone o nas dzisiaj informacje są większe niż kiedykolwiek wcześniej. W kolejnych częściach tego rozdziału autorzy omawiają ponownie portale społecznościowe, gdzie według nich „człowiek jest postrzegany jako suma relacji społecznych, interakcji online i ulubionych treści". Nie da się ukryć, że obecne sposoby inwigilowania przekazują ogrom informacji różnym służbom państwowym i przedsiębiorstwom. Nie muszą być one wykorzystane natychmiast, ale gromadzone wiadomości o danej osobie mogą być użyte później, np. przy popełnieniu przestępstwa. Kolejna część tego rozdziału poświęcona jest zagadnieniu kary. Czy *big* *data*, profilując osobę, odkrywając jej skłonności, wskaże np. policji potencjalnego przestępcę, mordercę, a policja aresztuje go za zamiar popełnienia morderstwa. Wykorzystanie prognoz w sądownictwie wydaje się nowym problemem epoki *big data*. Zjawisko to daje kuszące perspektywy i jest jednocześnie niehumanitarne. Każdemu z nas zależy na porządku publicznym i bezpieczeństwie, ale bronimy się przed wizją bycia społeczeństwem sportretowanym.

W ostatnich częściach tego rozdziału autorzy piszą o *dyktaturze danych i ciemnej stronie big data*. Z przeprowadzonych analiz wyłania się obraz gigantycznych liczb i ich niewłaściwego wykorzystania. Nie zawsze liczby odpowiadają za sukces w naszych działaniach. Oceny szkolne, życiorysy nie odzwierciedlają naszych umiejętności. Ważne są uczucia, nasze intuicje, czego najlepszym przykładem jest wprowadzenie na rynek przez Steve'a Jobsa iPoda, iPhone'a, iPada bez przeprowadzenia jakichkolwiek badań rynkowych.

Zjawisko *big data* wkroczyło w nasze życie prywatne tak bardzo, że niektóre modele prawne chroniące naszą prywatność są już mocno przestarzałe.

W przedostatnim rozdziale – *Kontrola* – autorzy rozważają kwestię szeroko pojętej kontroli *big data* i nowych regulacji prawnych.

Użytkownicy danych muszą być odpowiedzialni za swoje działania. Dotychczasowe prawo wobec nowych wyzwań nie daje instytucjonalnego wsparcia osobom poszkodowanym przez *big data* ani nie gwarantuje wolności działania osobom zajmującym się udostępnianiem danych. Na obecnym etapie brak jest definicji użycia kategorii danych w różnych okresach. Autorzy sugerują więc, żeby dla zbilansowania korzyści dla jednych i uniknięcia zagrożeń dla drugich wprowadzić czasową barierę wykorzystania danych, tak by np. reklamodawcy nie mogli zidentyfikować poszczególnych osób.

W dalszej części tego rozdziału autorzy dużo uwagi poświęcili nowym specjalistom – algorytmikom. Są to eksperci działający w takich dziedzinach jak matematyka, statystyka, informatyka, zatrudniani są w organizacjach, firmach, pilnują interesu tych firm i dbają o dobro osób, których dotyczą analizy. Będą oni również udzielać instytucjom rządowym wskazówek dotyczących wykorzystania *big data* w sektorze publicznym. Charakteryzować ich będzie bezstronność, dyskrecja, kompetencje i profesjonalizm. Niespełnienie tych standardów może zakończyć się sprawą sądową.

Pojawienie się audytu algorytmów, według autorów, to powstrzymanie wielu nieczystych działań *big data*. Rozdział ten autorzy kończą rozważaniami dotyczącymi wprowadzenia na początku XIX wieku przepisów antytrustowych, zapobiegających dominacji jednej firmy. W kontekście branży technologicznej i szeroko pojętej wymiany danych jest to bardzo ważne.

Omawiana książka kończy się rozdziałem *Przyszłość*. Konkluzje autorów są interesujące. Stwierdzają mianowicie, że ludzkość zmierzy się ze zmianami, które będą większe od epokowych odkryć z przeszłości, że wiedza kiedyś zdobyta służyła nam do zrozumienia przeszłości, natomiast dzięki danym będącym w posiadaniu firm będzie można przewidzieć przyszłość.

Może dzięki *big data* świat stanie się bardziej zrozumiały i znajdziemy rozwiązania globalnych problemów. Z pewnością *big data* zmienia nasze życie i sprawia, że nasze działania przynoszą więcej korzyści niż strat. Żyjemy w czasach, gdy możliwości techniczne wyprzedzają ludzką świadomość. Nie zwalnia nas to z prawa do kształtowania naszej przyszłości. Z narzędzia, jakim jest *big data*, należy mądrze i sprawnie korzystać, tak by chronić nasze człowieczeństwo – intuicję, zdrowy rozsądek, wolność wyboru. „Wielkość człowieka jest tym, czego nie ujawnią algorytmy i procesory" (s. 256).

Podsumowując, recenzowaną książkę można uznać za wyjątkową w ofercie publikacji przedstawiających nowe zjawiska. Cała praca poświęcona jest analizie danych. Dotyczy nowych technologii i społeczeństwa. Podział i kolejność poszczególnych rozdziałów są logiczne. Znajdziemy w niej wiele przekonujących przykładów zastosowania *big data* w działalności gospodarczej, działaniach rządowych i w życiu każdego z nas. Niektóre tezy autorów są kontrowersyjne i trudno się z nimi zgodzić. Książkę gorąco polecam.