Tworzenie i wykorzystywanie wielkich baz danych

Czy bać się baz?

DOMINIK BATORSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
Uniwersytet Warszawski
db@uw.edu.pl
Dr Dominik Batorski jest badaczem zjawisk i procesów społecznych związanych z Internetem.

Dzięki nowym technologiom możliwe staje się gromadzenie i przetwarzanie olbrzymich ilości danych. Stwarza to ciekawe możliwości, ale i wyzwania. Może także doprowadzić do zmiany roli i charakteru samej nauki

Zbieranie i wykorzystywanie dużych danych dotyczy praktycznie każdej dziedziny nauki, nawet humanistyki. Coraz większe zbiory informacji na temat osób, rzeczy, zdarzeń, a także relacji między nimi produkowane są zarówno przez ludzi, jak i urządzenia. Gromadzone są one nie tylko w nauce, ale także przez firmy i instytucje. Rosną także możliwości ich przetwarzania.

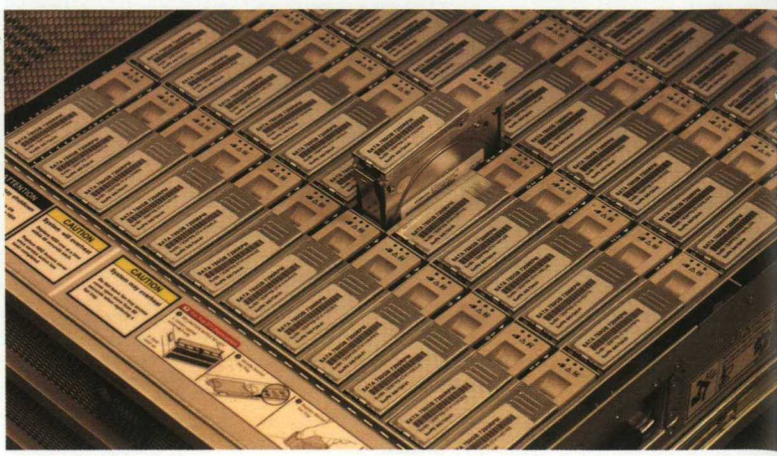
Coraz szybszy wzrost

Wszystko to jest możliwe dzięki rozwojowi nowych technologii. To dzięki nim możliwe staje się wykorzystanie danych, które wcześniej nie były zbierane lub ich przetwarzanie było praktycznie niemożliwe, w szczególności tych interesujących dla nauk społecznych.

Każde zachowanie w Internecie przy użyciu telefonów komórkowych czy innych urządzeń pozostawia ślad elektroniczny. W logo serwisów internetowych, historii wyszukiwania, serwisach społecznościowych czy zawartości stron internetowych zbierane są dane o zainteresowaniach, używanych usługach, komunikacji i relacjach społecznych. Gromadzone są dokumenty, informacje o e-handlu, transakcjach bankowych i giełdowych oraz inne dane finansowe. Sieci handlowe analizują co, kiedy i w jakich konfiguracjach jest kupowane. Każda transakcja kartą płatniczą lub kredytową jest rejestrowana w systemach bankowych. A informacje o korzystaniu z usług, połączeniach, a także przemieszczaniu się pomiędzy stacjami bazowymi telefonii komórkowej są gromadzone przez operatorów.

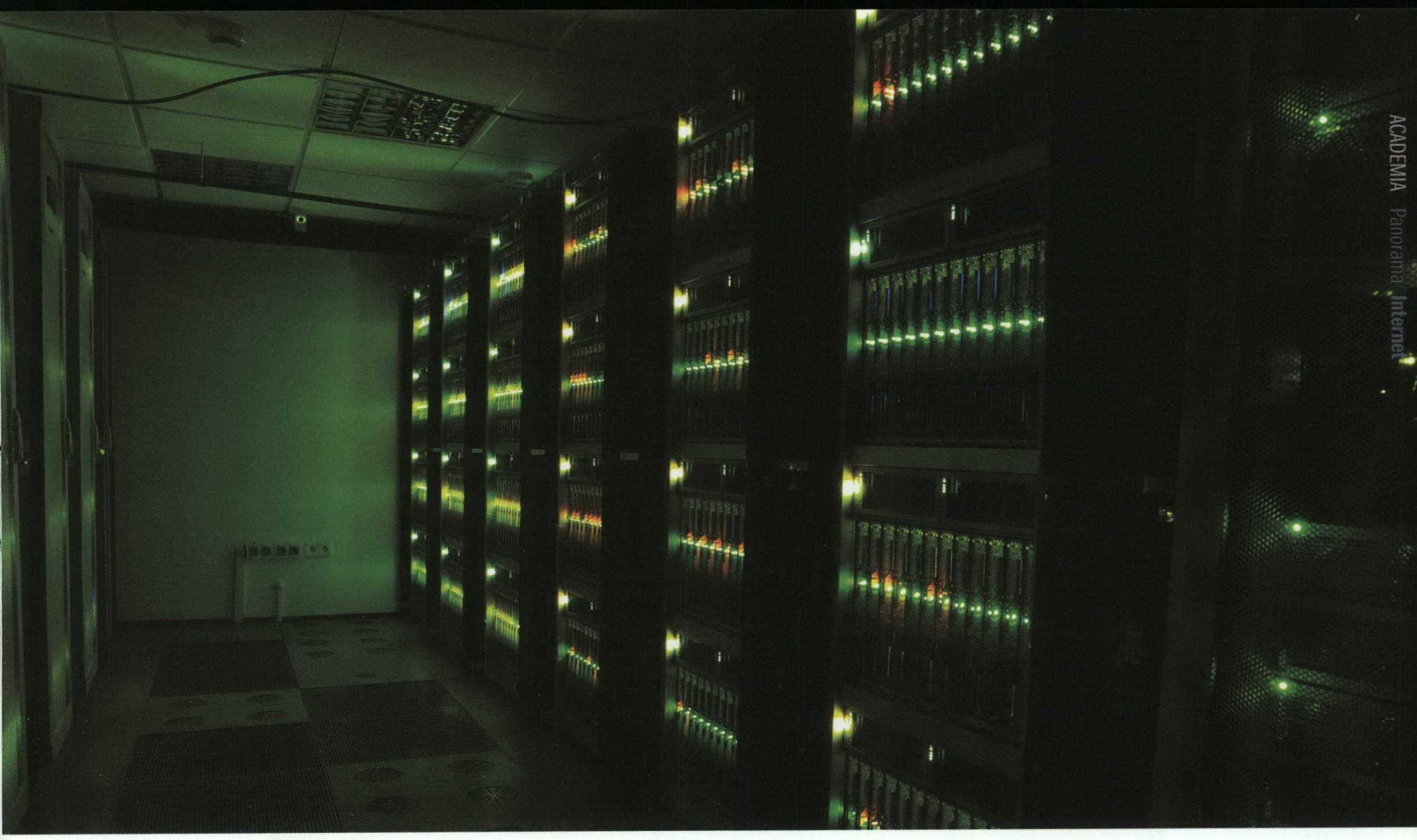
Jako źródła coraz częściej wykorzystywane są różnego rodzaju chipy, czujniki, sensory i kamery, a także satelity. Umożliwiają one gromadzenie danych atmosferycznych, astronomicznych, medycznych, a także genetycznych i biologicznych. Wszystko to przekłada się na ogromne ilości danych, które są zbierane przez instytucje naukowe i firmy. Rośnie także znaczenie ich przetwarzania.

Ilość danych gromadzonych na świecie zwiększa się wykładniczo (McKinsey Global Institute 2011). Według szacunków w 2007 roku przekroczyła ona globalną ilość przechowywanych już w roku poprzednim. Całkowita ilość danych dla 2009 roku to 800 exabajtów (exa=10^{18}). Już w połowie 2008 roku liczba unikatowych adresów stron internetowych indeksowanych przez Google przekroczyła bilion, a liczba zapytań wpisywanych do wyszukiwarki dziennie wynosiła około 2 miliardów. Sloan Digital Sky Survey (SDSS), począwszy od 2000 roku, zbiera około 200 GB danych dziennie, gromadząc dotychczas prawie 150 terabajtów informacji. Wielki zderzacz hadronów (LHC) w samym 2010 roku dostarczył 13 petabajtów danych (10^{15}). Facebook przetwarza codziennie około 500 terabajtów danych, użytkownicy wy-



Bartosz Niezgódka

Macierz dyskowa – urządzenie zawierające do kilkuset dysków do przechowywania dużych danych



Bartosz Niezgódka

Serwery w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego

mieniają ponad 2,5 miliarda treści i wgrywają około 300 milionów zdjęć. Przykłady wielkich danych można by mnożyć.

Ilość gromadzonych danych będzie się nadal zwiększała. I to coraz szybciej. Wraz z rozwojem tzw. sieci rzeczy i upowszechnieniem wykorzystania różnego rodzaju sensorów możliwe stanie się zbieranie dokładnych danych dotyczących stanu oraz zachowań ludzi, urządzeń i innych obiektów fizycznych. Od monitorowania stanu zdrowia osób po analizy sytuacji pogodowej.

Tak ogromne ilości danych wymagają też odpowiedniego do nich podejścia. Coraz częściej określa się je zbiorczym pojęciem „dużych danych" (ang. Big Data). Termin ten jest do pewnego stopnia elementem marketingu twórców rozwiązań służących do zbierania, przechowywania i analizowania danych. Z drugiej strony służy też podkreśleniu bezprecedensowej objętości gromadzonych informacji, prędkości ich przyrostu, a także różnorodności. Wiele danych jest zbieranych i przetwarzanych w czasie rzeczywistym. Często są to też dane nieustrukturyzowane i bardzo różnego typu – nie tylko liczbowe, ale też tekstowe, obrazy, wideo, audio, dane geolokalizacyjne itp.

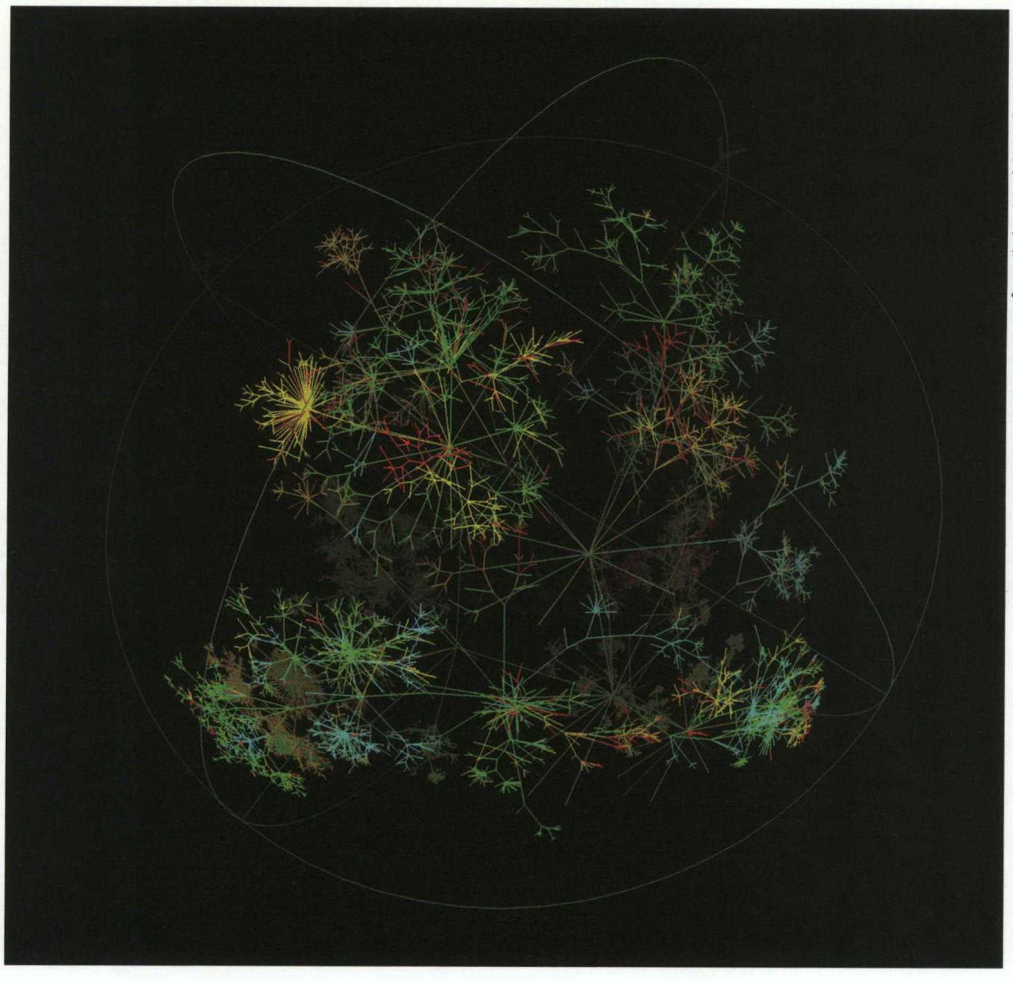
Nowe wyzwania

Równolegle do wzrostu ilości danych rosną też możliwości ich przechowywania i analizowania. Coraz większa część informacji jest bowiem zdigitalizowana. W 2000 roku 25% informacji na świecie przechowywano w formie cyfrowej, a w 2007 roku już 94%. Bardzo szybko rośnie też moc obliczeniowa komputerów, podwajając się zgodnie z prawem Moore'a co mniej więcej 18 miesięcy.

Rozwijane są także nowe techniki przetwarzania dużych danych. Sam fakt ich posiadania niewiele znaczy, dopóki nie odkryje się wiedzy w nich ukrytej. W przypadku ogromnych baz danych analiza nie byłaby możliwa z wykorzystaniem tradycyjnych metod i programów statystycznych. Dopiero niesłychany rozwój mocy obliczeniowej komputerów oraz – dynamiczny od 30 lat – sztucznej inteligencji i dziedzin pokrewnych takich jak uczenie maszynowe (ang. *machine learning*) czy *data mining*, a w ostatnich latach techniki MapReduce umożliwił inteligentną i automatyczną eksplorację dużych wolumenów danych. Dzięki technologiom takim jak Apache Hadoop możliwe stało się przetwarzanie rozproszonych danych.

Ilościowy wzrost objętości informacji wymaga często jakościowo innego podejścia do ich wykorzystania i analizy. Wynika to także z tego, że choć większość gromadzonych informacji występuje w postaci danych liczbowych lub tekstowych, to ciągle rośnie udział danych do tej pory niestandardowych – przede wszystkim multimedialnych. Stąd też coraz większe zapotrzebowanie i szybszy rozwój technik automatycznego przetwarzania danych o różnej strukturze, tekstu, obrazu i dźwięku.

Przetwarzanie ogromnych ilości danych wymaga coraz większych kompetencji anali-



Cooperative Association fo Internet Data Analysis (CAIDA)

Wizualizacja struktury połączeń w Internecie

tycznych, umiejętności programowania połączonej ze znajomością technik statystycznych, a także pracy z bazami – zdolności odpowiadania na trudne pytania przy użyciu danych i odpowiednich metod ich przetwarzania oraz jasnego komunikowania, w tym wizualizacji wyników. Dlatego też coraz częściej mówi się o osobnej dyscyplinie, tzw. Data Science. Rośnie też liczba uczelni oferujących studia w tym zakresie, ale jeszcze szybciej rośnie liczba ofert pracy dla data scientist. Zapewnienie podaży osób o pożądanych kompetencjach, będącej w stanie zaspokoić popyt ze strony firm i instytucji, jest wyzwaniem, na które większość uczelni wyższych w Polsce nie jest przygotowana.

Rozwiązania bez wyjaśniania

Szanse gromadzenia zupełnie nowych danych, jak i przyrost ich ilości przyczyniają się do ogromnego wzrostu możliwości rozwiązywania problemów. Ich znaczenie zaczyna być widoczne w prawie każdej dyscyplinie naukowej, a także poza nauką. W biznesie dane i umiejętności ich wykorzystania decydują o uzyskaniu przewagi konkurencyjnej. W coraz większej liczbie dziedzin zaczyna dominować podejście typu *data driven*, według którego działania powinny być podejmowane w oparciu o dane, a nie wyłącznie na podstawie intuicji czy doświadczenia. W sferze publicznej coraz popularniejsze staje się również *evidence based policy*. Tworzonych jest też wiele nowych zautomatyzowanych usług działających na podstawie danych, jak choćby rozwiązania pozwalające na tworzenie inteligentnych budynków i miast (tzw. *smart city*).

Jednak potencjalnie znacznie istotniejsze wydaje się to, że wykorzystanie dużych danych daje także możliwość zupełnie innego sposobu rozwiązywania problemów. Doskonałym przykładem jest tu udostępniana przez firmę Google usługa tłumaczenia tekstów między różnymi językami, działająca wyłącznie dzięki prostym regułom statystycznym i ogromnym zbiorom tekstów, w tym takich, o których wiadomo, że jest

to ta sama treść w różnych językach. Podobnie z narzędziami automatycznego poprawiania błędów pisowni, które nie wymagają znajomości języka, wykorzystując jedynie dane o błędach popełnianych chociażby przy wpisywaniu haseł w wyszukiwarkę i ich poprawianiu.

Jak zwraca uwagę David Weinberger w wydanej niedawno książce „Too Big to Know" (2012), dostępność dużych danych przyczyniać się może do zmiany charakteru nauki i roli teorii. Kiedyś zbieranie danych było znacznie trudniejsze, dlatego kluczowe znaczenie miało tworzenie teorii, które pozwalały opisywać prawa przyrody i obserwowalne zależności, a dzięki temu ułatwiały przewidywanie faktów. Z drugiej strony podejście takie powodowało, że słabo radzono sobie z analizowaniem bardzo złożonych zjawisk. Obecnie zbieranie danych jest znacznie prostsze. Jednocześnie analizowanie złożonych układów często nie pozwala na określenie ogólnych zależności. Łatwiej jest więc opisać układ za pomocą danych, niż wyjaśnić jego funkcjonowanie. Budowaniu przewidywań służą zaś symulacje komputerowe i modelowanie zachowań układu. Pojawia się tym samym możliwość znajdowania rozwiązań bez wyjaśniania samych zjawisk.

Dostępność danych przyczynia się do zmiany sposobu uprawiania nauki i wytwarzania wiedzy, jednak pojawiające się w konsekwencji takich obserwacji głosy o możliwym „końcu teorii" wydają się przesadzone. Niewątpliwie zamiast modelu, w którym najpierw stawiane były hipotezy, a następnie gromadzone dane pozwalające na ich falsyfikację, coraz częściej spotykamy się z sytuacją, w której najpierw są dane, a dopiero później następuje budowanie teorii.

Z tą zmianą związana jest też największa krytyka wykorzystania dużych danych (por. np. Boyd i Crawford 2011). Dobrych przykładów dostarczają nauki społeczne - gdzie mimo niewątpliwych szans, jakie stwarza pozyskiwanie danych o zachowaniu użytkowników w środowiskach cyfrowych, trzeba jednocześnie pamiętać, że dane te są zwykle jedynie fragmentaryczne. Wielu zachowań nie sposób przy ich użyciu wyjaśnić, a jednocześnie pojawiają się problemy z reprezentatywnością, ponieważ dostępność danych dla różnych podgrup w ramach interesującej populacji jest różna. Dane mogą być też stronnicze lub skrzywione ze względu na kontekst, w którym powstają. Ograniczenia te powinny być uwzględniane przy prowadzeniu analiz i interpretacji rezultatów.

Warto też zauważyć, że w przypadku wielu typów danych dostęp do nich mają głównie firmy i instytucje, które je gromadzą. W konsekwencji w niektórych dziedzinach ciężar wytwarzania wiedzy przesuwa się do biznesu. Stworzy to nowe wyzwania dla uczelni wyższych, w większym stopniu wymuszając współpracę między nauką a biznesem.

Istotnym źródłem danych dla prac naukowych mogą być również dane publiczne. W coraz większej liczbie krajów (m.in. Stanach Zjednoczonych i Wielkiej Brytanii) dane gromadzone przez instytucje publiczne są udostępniane. W ten sposób stwarza się możliwość ich wtórnego wykorzystania w nauce, biznesie czy przez organizacje pozarządowe (ang. *open government data*). Potrzeba lepszego udostępnienia danych publicznych zaczyna być dostrzegana również w Polsce.

Czas zmian

Efektem ubocznym wzrostu znaczenia danych jest to, że w niektórych sytuacjach sam fakt ich zbierania i przetwarzania może prowadzić do zmiany logiki funkcjonowania całego systemu. Doskonałym przykładem może być wprowadzenie zestandaryzowanych pomiarów w szkolnictwie (testy gimnazjalne, nowa matura), które co prawda umożliwiło szersze porównywanie, ale jednocześnie zmieniło to, na co kładziony jest nacisk w edukacji. Pomiar i gromadzenie informacji może mieć charakter dyscyplinujący i być elementem kontroli, ale zwiększona kontrola niekoniecznie musi poprawiać jakość efektów funkcjonowania systemu.

Innym przykładem zmiany działania systemu w efekcie pomiaru może być obecna reforma nauki, która poprzez działania na rzecz kwantyfikacji dorobku naukowego będzie mieć niewątpliwe efekty dla funkcjonowania osób i instytucji działających w obszarze nauki. ■

Chcesz wiedzieć więcej?

Boyd D. i Crawford K. (2011). *Six Provocations for Big Data*. Zaprezentowane na konferencji: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, dostępne przez SSRN-id1926431.

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Weinberger D. (2012). *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room*. Basic Books.