New technologies for generating giant databases with applications in business and science

The Burden of Big Data

DOMINIK BATORSKI
University of Warsaw
Interdisciplinary Centre for Mathematical and Computer Modelling, University of Warsaw
db@uw.edu.pl
Dr. Dominik Batorski studies social processes in the Internet age.

New technologies are making it possible to gather and process huge amounts of data. This opens up fascinating possibilities as well as new challenges; it could also lead to a shift in the role and nature of science itself

Gathering and processing high volumes of data is relevant in all branches of science, including the humanities. Increasingly large collections of information on people, objects, and events, and the relationships between them are being created by people and by machines, both for research purposes and for companies and institutions. Our ability to process this data is also improving by leaps and bounds.

Ever faster growth

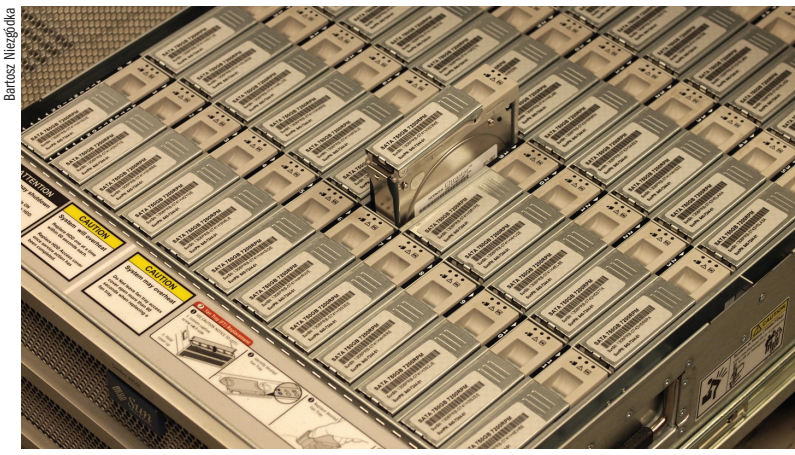
This is all made possible by the development of state-of-the-art technologies used to process data which previously went uncollected, or whose processing was previously impractical. This trend has affected the social sciences in particular.

Each activity performed online, using a mobile phone or another electronic device leaves behind a digital trace. Logs stored by web pages, search histories, social networks, and website content are used to gather data on browsing habits, services accessed, types of communication and social interactions. Data gathered includes documents, information on online purchases, banking and stock market transactions, and other financial data. Commercial networks analyze purchasing patterns: what is being bought, where, and under what circumstances. Each transaction with a credit or debit card is recorded by banking systems, while information on services, connections, and movement between base stations for mobile phone networks is accumulated by operators.

Data is also increasingly being generated by various types of chips, sensors, and cameras, as well as satellite networks. Some of the types of data being collected, for instance, are atmospheric, astronomical, medical, genetic, and biological data. This means that as vast amounts of data are being compiled by research institutions and companies, the significance of data processing increases.

The amount of data gathered globally has grown exponentially (McKinsey Global Institute 2011). According to estimates, the amount of new data created in 2007 exceeded the overall volume of data that had been generated and stored in all previous years. The total volume of data generated in 2009 was around 800 exabytes (1 exabyte = 1x1018 bytes). By mid-2008, the number of unique web addresses indexed by Google exceeded a trillion, while the number of queries entered into the search engine was around 2 billion every day. Since 2000, the Sloan Digital Sky Survey (SDSS) has been gathering around 200GB of data daily, accumulating close to 150 terabytes of data so far. The Large Hadron Collider (LHC) generated 13 petabytes (1015) of data in 2010 alone. Facebook processes around 500 terabytes of data every day, with the users exchanging over

Bartosz Niezgódka



A disk matrix – a devise containing up to several hundred physical disks for storing big data

Bartosz Niezgódka

Servers at Warsaw University's Interdisciplinary Centre for Mathematical and Computational Modelling (ICM)

2.5 billion posts and uploading around 300 million photos daily.

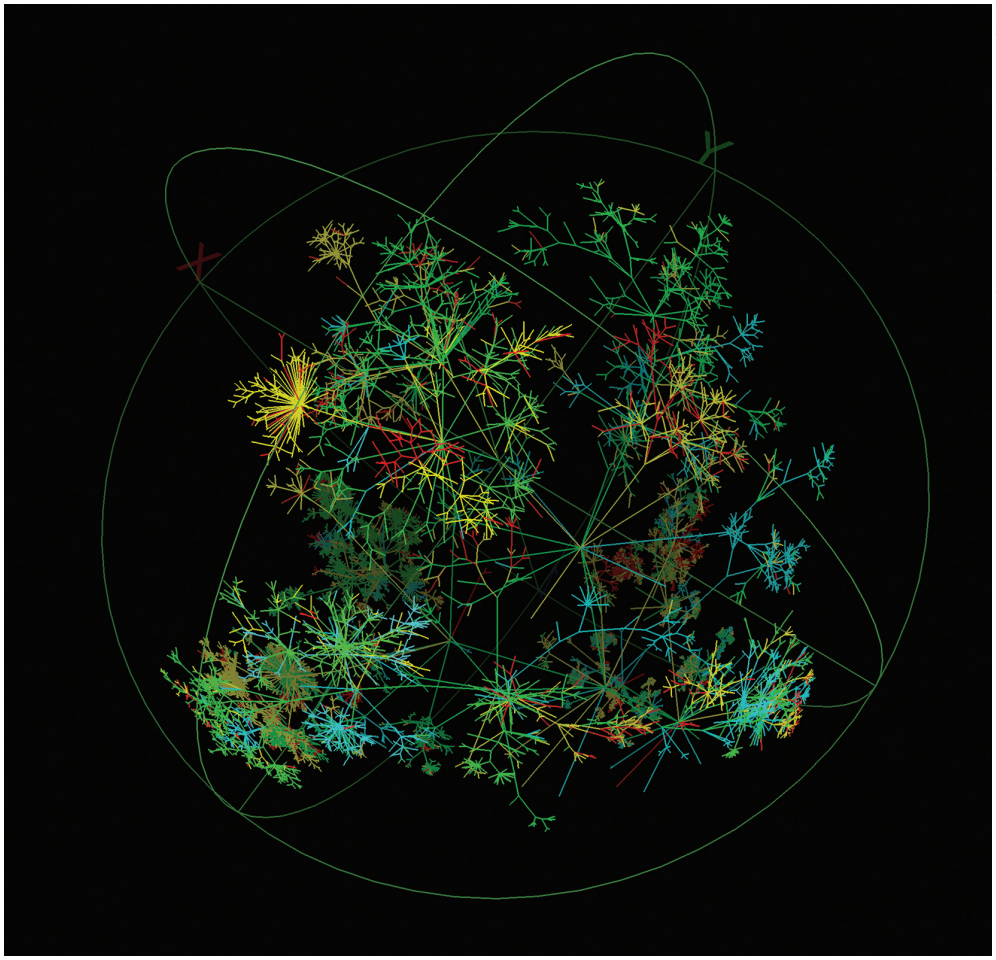
There are countless other examples of such impressively large figures, yet the volume of user-generated data will continue to grow at an ever-increasing rate. Alongside the development of the "Internet of Things" and the growing use of various sensors and devices connected to the Internet, it will be increasingly possible to gather data on the behavior and status of people, equipment and other physical objects, from monitoring health of individuals to weather analysis.

Such high volumes of data, frequently described as "Big Data," require a particular approach. To some extent, this term is a marketing element used by experts working on collecting, storing, and analyzing data. It also serves to stress the unprecedented volume of information, its rate of growth, and its sheer diversity. Much of the data is collected and processed in real time. Frequently it is unstructured data of a range of types; not just numerical, but also text, image, video, audio, geolocation data, and so on.

New challenges

As the amount of data generated increases, so do the methods of its storage and analysis, since the majority of information is now digitized. According to McKinsey, 25% of data available globally in 2000 was stored digitally; by 2007, this reached 94%. Processing power is also increasing rapidly, doubling roughly every 18 months in keeping with Moore's law.

New technologies for processing Big Data are also currently being developed in order to discover and make use of the information it contains. For huge databases, traditional methods and statistical programs are insufficient. Intelligent and automatic exploitation of such large volumes of data have been rendered possible by the unprecedented development of computing power over recent decades, by rapid progress in artificial intelligence and associated disciplines such as machine learning and data mining, and more recently the MapReduce model. Software tools and solutions such as Apache Hadoop make it possible to process distributed data.



Cooperative Association for Internet Data Analysis (CAIDA)

A visualization of the structure of global Internet connections

A quantitative increase in data frequently necessitates a qualitatively different approach to its processing and analysis. Although most information exists as either numerical or text data, the amount of previously non-standard data – mainly multimedia formats – has been increasing. This means growing demand and faster development in technologies for automatic processing of data with a variety of structures, containing text, images and sounds.

Processing vast amounts of data demands improved analytical capabilities, programming abilities combined with an understanding of statistical techniques, database skills including using data to answer difficult queries, and appropriate methods for data processing and for the clear communication of results, including visualization. As a result, data science is frequently discussed as a separate discipline. There are growing numbers of universities offering courses in this field, and the numbers of jobs available to data scientists are increasing at an even faster rate. Being able to provide sufficient numbers of people with the skills required to meet the demand from companies and institutions is a challenge which most of Poland's universities are not currently able to meet.

Solutions without explanation

The opportunities for gathering brand new data and the rapid growth in data volume greatly augment our problem-solving abilities. The significance of this data is becoming increasingly clear in most scientific disciplines, as well as further afield. In business, data and our ability to process it can be decisive in maintaining a competitive advantage. Growing numbers of disciplines are becoming dominated by the data-driven approach, according to which actions should be taken on the basis of data rather than purely on intuition or experience. In the public sphere, evidence-based policy is also gaining in popularity. Additionally, numerous new automated services based on the use of data are being created, such as solutions for creating smart buildings and cities.

However, even more significantly, using large volumes of data also makes it possible to devise completely new ways of solving problems. An excellent example is the Google Translate service, operating purely on the basis of simple statistical rules and huge collections of texts, including those with the same content in different languages. Automatic spellcheck tools operate on a similar basis; they do not require any knowledge of language as such, and simply use information about errors made during tasks such as entering search terms into browsers, and their corrections.

David Weinberger notes in his recently published book Too Big to Know (2012) that access to big data may even cause a shift in the character of science and the role of theory. In the past, data collection used to be far more difficult, therefore it was essential to formulate theories describing the laws of nature and observable relationships in order to support forecasting. However, this approach meant that scientists were not particularly adept at analyzing very complex phenomena. Today, data collection is far simpler, while the analysis of complex systems frequently does not make it possible to suggest general dependencies. As such, it is easier to use data simply to describe a system than to explain how it functions. Computer simulations and models of system behavior can be used in forecasting, making it possible to find solutions without actually explaining the underlying phenomena.

The availability of data has an impact on the way science is practiced and knowledge is generated, although claims about a potential "end of all theory" seem to be exaggerated. It is true that the model in which a hypothesis is first posed, then data is gathered in order to verify it, is increasingly being replaced by a situation where the data is gathered first and a theory is built up around it later.

This is also related to the greatest criticism against big data as a source of information (cf. Boyd and Crawford 2011). Social sciences are a good example: in spite of the indisputable opportunities provided by data gathered on the behavior of users of digital media, it should be remembered that this data tends to be fragmented. It cannot be used to explain certain types of behavior, and it raises issues of representativeness, since the availability of data varies for different subgroups of the population studied. Data can also be biased or skewed depending on the context in which it is generated. Such limitations must to be factored into the analysis and interpretation of results.

It is also worth noting that many types of data are often only available to the companies and institutions that collect it. As a result, in certain fields the weight of creating knowledge is shifting towards business. This creates new challenges for universities, encouraging cooperation between research and business.

Data in the public domain is also an important source of information for academic papers. In increasing numbers of countries (such as the US and the UK), data gathered by public institutions is made available to business, research, and non-governmental organizations. The process is known as open government data. This growing need to make public data more widely available is becoming increasingly observable in Poland.

Time for change

One side-effect of this rapid increase in the amount of data is that in certain situations, the mere fact of its collection and processing can alter the logic by which the entire system functions. One example is the introduction of standardized tests in education at high school level, making it possible to conduct broader comparisons; however, this has also brought about a shift in key focus points in education. Gathering and managing information can introduce an element of discipline and control, although increased levels of control may not in themselves improve the quality of the functioning of the system.

Another example of how measurement-taking can alter a system can be found in the ongoing reforms of Poland's research sector, which, through its drive to quantify scientific achievements, will significantly affect how science is done in Poland. ■

Further reading:

Boyd, Danah i Kate Crawford. (2011). Six Provocations for Big Data. Presentation at the conference "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society," accessible via SSRN-id1926431

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

Weinberger, David. (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.