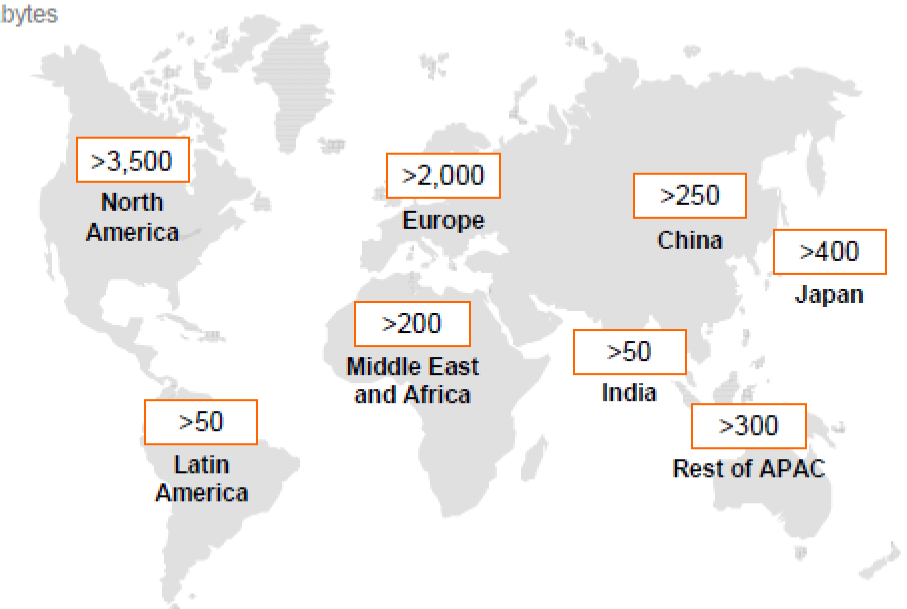
Big Data: Introduction

The catchword *Big Data* refers to (the more recent phenomenon of) massive, constantly growing and often unstructured datasets from diverse sources, stored in remote server farms, and mined for meaning. Many internet-based companies hold petabytes of data – Yahoo alone has over 25 petabytes[1]. In 2009, more than one petabyte (10^{15} bytes) of personal location-based data was generated across the globe[2]. In 2010, every continent on the planet produced in excess of 50 petabytes of data.

Amount of new data stored varies across geography

New data stored[1] by geography, 2010

Petabytes



1 New data stored defined as the amount of available storage used in a given year; see appendix for more on the definition and assumptions.

SOURCE: IDC storage reports; McKinsey Global Institute analysis

By some accounts[3], 90% of the recorded data in the world today has been created in the last two years - and the rate is accelerating. Big data is big, but also diverse. Big data sources include environmental sensor networks, traffic monitoring systems, mobile phones, satellite imagery, video surveillance, posts to social media sites, transaction records of online purchases, electronic health records, and satellite imagery – just to list a few important ones.

At the root of Big Data are, most often, repeated observations/measurements over time and space. For example, a retailer might have thousands of stores and tens of thousands of products and millions of customers but logs billions upon billions of individual transactions a year. Health care is even more extreme: FMRI neuroimaging studies can generate hundreds of gigabytes of data in a single experiment.

[1] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, The Hadoop Distributed File System, IEEE2010.

[2] McKinsey Global Institute, The next frontier for innovation, competition, and productivity, 20112011, p. 87.

[3] http://www-01.ibm.com/software/data/bigdata/

Many scientific domains have a long history of generating vast amounts of data over time and space. Astronomy[4] and environmental monitoring are good examples. Astronomy in particular is data-rich due to systematic observations of the sky over a range of wavelengths. These data are then often paired with numerical simulations producing comparable volumes of information. Indeed, many of the technical challenges facing Big Data today have precedents in Astronomy (database design and federation, data mining and advanced visualization, for example).

Big Data is big along at least two dimensions: Variety and velocity/volume. Big Data encompasses many kinds of data in a plethora of formats including text, audio, video, click streams, log files and more. Big data needs lots of storage space and bandwidth. Volume and velocity have increased so much that new storage and access technologies were created to deal with them. Indeed, the prevalence of Big Data is a direct consequence of these new storage and data access systems.

The Hadoop Distributed File System, for example, is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. Importantly, Hadoop distributes data and the processing of this data across thousands of host computers, performing computations in parallel. In a large Hadoop cluster, thousands of servers (aka server farm) host directly attached storage and execute computational tasks. The advantage of this distributed approach is that it can accommodate large amounts of data and it can scale to continuously accommodate increasing data volumes while remaining economical to service at every step.

Big Data poses challenges not only to access, storage and transmission, but also to analysis. Finding pertinent information in petabytes of data has challenged researchers from many disciplines to consider new approaches to analysis and inference. While many contributions have been effective in creating the current culture of large data analytics, some algorithms have had more impact than others. In the domain of online search, the well-known PageRank algorithm originally devised by Google founders in 1998 relies on hyperlinks. PageRank produces a static ranking of Web pages based on the level of connectedness to other pages. The algorithm relies on the mass behavior of web users by deriving from a vast link structure an indicator of an individual page's pertinence to a given search term. Other less known but equally important data mining algorithms include K-means, Support Vector Machines, Apriori, Expectation Maximization, AdaBoost, K- Nearest Neighbors, and Naive Bayes[5].

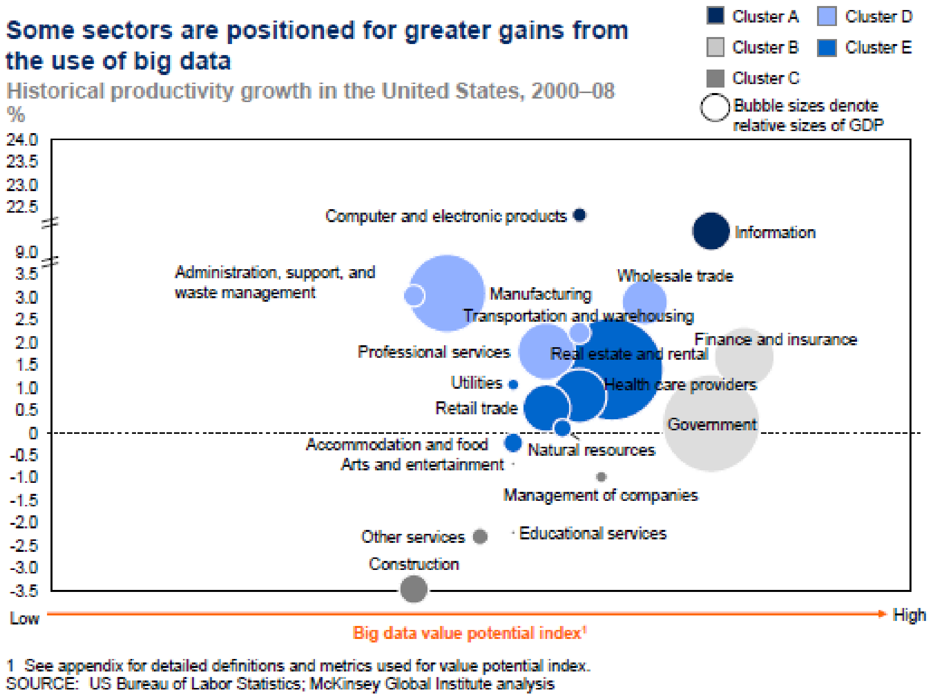
Data mining efforts have also led to new approaches to making data 'intuitive'. On the user level, features such as the 'like' button allow people to share their approval of sites (while feeding the data mining industry). New data visualization approaches attempt to make vast data sets intuitive by allowing people to visually parse large amounts of complex information. Attempts include new ways of dealing with the trade-offs between time multiplexing vs. space multiplexing techniques as well as an emphasis on context-sensitive use of data in two and three dimensions. Biosciences are also heavily invested

[4] Robert J. Brunner, S.George Djorgovski, Thomas A. Prince, Alex S. Szalay, Massive datasets in astronomy. arXiv.org, Jun 2001

[5] Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems, vol. 14, no. 1, 2008, p. 1-37

in new visualization approaches to deal with the diversity and interconnectedness in genomic, molecular and chemical data.

Without a doubt much of the current interest in Big Data stems from the numerous commercial opportunities it is creating. The marketing research firm McKinsey estimates that big data is worth approximately $300 billion annually to the growing E-health industry in the United States alone.



Some sectors are positioned for greater gains from the use of big data

Historical productivity growth in the United States, 2000–08

1 See appendix for detailed definitions and metrics used for value potential index.
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Contrary to the enthusiasm in the business and marketing industries, privacy advocates and activists are alarmed by Big Data. A particularly thorny set of issues are the under-controlled activities of commercial data brokers who collect and sell personal information to private companies and governments, and generate new and under-specified risks to due process and privacy in the process.

Government misuse of Big Data is also of concern, with examples of police misuse of databases, and violations involving even local police officers abounding in the media. In general the concern is that the uncontrolled collection of information on individuals will upset the balance between government and individuals, generating a more oppressive[6] power structure. And companies are using increasingly invasive marketing techniques to maintain control over data flows, sometimes against the will of their customers. OnStar, for example, changed its service contact such that vehicles of owners who no longer subscribe could still be monitored via the system's still-active two-way cellular link[7]. It is one thing to trade privacy for security in full knowledge of the exchange, and another thing be stuck in privacy compromising technology against one's will – and after the expiration of the original contract. It is the perceived value of the data (in this case of interest to government and insurance agencies) that is now leading to these contractual

[6] Hoofnagel, C.J., Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect, Process, and Package Your Data for Law Enforcement, 29 N.C.J. Int'l L. & Com. Reg. 595 (Summer 2004), p. 1-31

[7] http://wheels.blogs.nytimes.com/2011/09/22/changes-to-onstars-privacy-terms-rile-some-users/?hpw

‘innovations’. Additionally, hackers have shown on more than one occasion that such ‘service data’ as OnStar collects, even when anonymized, can be cross-referenced with other databases to identify individual users. And some cyber security researchers have recently shown how a car’s wireless connections can be exploited, to break into a vehicle[8].

In academia, the consequences of Big Data for privacy design are currently being hotly debated. What is becoming clear is that privacy is as much about direct data as it is about the context in which any kind of data is collected, analyzed and stored. This means that methodologies for Big Data need to be thought and implemented contextually, with the people who might be negatively affected by the ‘results’ in mind. That is why the privacy question is by necessity also one of ethics[9]. Privacy, then, is not primarily an issue of control of raw data, and it is not a property of data, nor is there an algorithm for privacy. Rather, privacy is a collective understanding of a social situation’s boundaries with the implicit and accountable acknowledgement of how to operate within said boundaries. Because these boundaries are fluid and subject to negotiation, all actors in the Big Data sector should make their assumptions and operating principles openly known.

Much of the added value from mining the data comes from the generation of information from seemingly insignificant data. Even without collecting personal information, detailed patterns of online behavior can be established simply by recording behavior-based activities (such as ‘click events’) from a given IP address. This constitutes a new kind of indirect observation, supplied involuntarily and unknowingly by most users, that nonetheless impinges directly on a (clicking) person’s privacy.

Because of the inferential nature of this knowledge, however, it falls outside of currently existing legal frameworks. Furthermore, this indirect data from inference is accompanied by a growing amount of ‘remixed data’, data from different sources which is then remixed by different data brokers for various purposes. While the practice of data remix is wide-spread, the consequences can remain opaque. In remixing operations, obvious mistakes are less problematic than ‘creative’ solutions that combine seemingly compatible data sets from completely different contexts, for example.

Sociologists in particular have voiced additional unease about the impact of Big Data. One concern is the lack of a formal framework and theory of Big Data. All insights from Big Data rely on ad hoc empirical approaches and inference methods that assume that the data in Big Data is a ‘true’ and representative sample of the world. It should be obvious that there is no ‘pure’ data in Big Data and not all data is created equal. Furthermore, some parties (adolescent well educated males on technology blogs, for example) are clearly overrepresented. Many inferences from Big Data inquiries do not take this adequately into account.

[8] http://www.isecpartners.com/storage/docs/presentations/iSEC_BH2011_War_Texting.pdf

[9] Boyd, D., "Privacy and Publicity in the Context of Big Data." WWW. Raleigh, North Carolina, April 29 2010.

Further Reading:

- Agrawal, D., Das, S., Abbadi, A. E., Big Data and Cloud Computing: Current State and Future Opportunities, Proceedings of the 14th International Conference on Extending Database Technology, Mar 21-24, 2011, Uppsala, Sweden, p. 530-533
- Bollier, D., The Promise and Peril of Big Data, The Aspen Institute, Communications and Society Program, Washington, DC, 2010
- Boyd, D., "Privacy and Publicity in the Context of Big Data." WWW. Raleigh, North Carolina, April 29 2010.
- Hoofnagel, C.J., Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect, Process, and Package Your Data for Law Enforcement, 29 N.C.J. Int'l L. & Com. Reg. 595 (Summer 2004), p. 1-31
- ISec Partners, WarTexting: Weaponizing Machine 2 Machine, 2011 http://www.isecpartners.com/storage/docs/presentations/iSEC_BH2011_War_Texting.pdf
- Lee, C.P., Dourish, P., Mark, G., The Human Infrastructure of Cyberinfrastructure, CSCW'06, November 4–8, 2006, Banff, Alberta, Canada, p. 483-492
- McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity, 2011
- Wynholds, L., Fearon, D., Borgman, C.L., Awash in Stardust: Data Practices in Astronomy, Proceedings of the 2011 iConference, Feb 8-11, 2011, p. 802-803
- Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems, vol. 14, no. 1, 2008, p. 1-37