# CONTENTS




## 3.18 Big Data and Analytics

### 3.18.1 Executive Summary

More data is collected, shared, and analyzed every day. The growing availability of data and demand for its insights hold great potential to improve many data-driven decisions, from the mundane to the strategic. But this growth also poses significant challenges, both technological and societal. To harness the deluge of data to beneficial use, we will need to address rapid changes in data acquisition, storage, and processing technologies; education, both of the analytics workforce and everyday users; and complex privacy issues.

### 3.18.2 Introduction

Imagine, if you will, the following scenario. You sit down at a restaurant for lunch, wondering what to order. You take a picture of the menu with your phone, starting a whirlwind of activity on your behalf. The software combs through the data it has collected during the day about you, such as your breakfast, exercise and calorie expenditure, blood pressure and blood sugar levels, etc. It combines this data with long-term information such as previous user reviews of the different dishes in this restaurant, your weight loss goals, food preference, and sensitiveness, and perhaps even your individual genetic properties. In milliseconds, it makes its top three suggestions, from which you choose your meal. In the meantime, data is collected anonymously in the aggregate about your choices and decisions, as well as other patrons'. The restaurant manager can use it to adjust the menu. Researchers can use it to better understand the relationship between nutrition, fitness, and health. Your friends can use it to obtain personalized food recommendations, and you can use it to track your progress toward your health goals.

This scenario, already more science than fiction, exemplifies how so-called "big data" can be used to seamlessly affect decisions from the prosaic (your choice of lunch) to the strategic (the FDA's nutrition guidelines). It represents but one of many opportunities envisioned for the large-scale analytics of diverse data. Big data is finally transitioning from the computer science and machine learning classrooms into numerous real-world scenarios in business, government and military, science, politics, medicine, climate, and personal analytics—a trend that we expect to grow rapidly through 2022.

### 3.18.3 State of the Art

Big data is exploding, with no signs of slowing down. The growth is manifest on two separate axes: more data is collected, and more data is shared. Growth along both axes is exponential, and

# CONTENTS



the combined growth results in a very rapid increase in total available data indeed. IDC estimates that the amount of data created and shared on the Internet will reach around 8 zettabytes by 2015.[2] Let's look at a few examples:

- Photos and videos taken and shared are growing at an exponential rate,[3] a result of three multiplicative trends. One, people are taking more photos, because more cameras are ubiquitously available through the proliferation of smart and feature phones. Two, these photos increasingly contain more data (pixels) through the rapid growth in sensor technology. And three, more people share their photos and videos on Facebook, YouTube, Twitter, Snapchat, and other fast-growth social networks.
- Crowd-sourced and individuals' data from day-to-day life is proliferating through a variety of mobile applications. Such information includes service reviews, traffic and geo-tagged check-ins, health and exercise metrics from wearable devices, etc. (see Section 3.2).
- More studies, more instrumentation, more simulations, and more sharing facilitated by the Internet (e.g., CERN's Grid) is translating into more scientific data.
- The increase in resolution of freely available elevation data has led to a lot more mapping.

2 IDC report "Extracting Value from Chaos," June 2011

3 See Kleiner Perkin's "Internet Trends 2013" report at http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013

### 3.18.4 Where We Think It Will Go

We see tremendous opportunities in big data. In the sciences, for example, the growth in experimental data and in simulations—the fourth paradigm of science—has already advanced our understanding of the universe and of life. The growth in ubiquity of mobile computing devices, as well as in the applications that collect data, means that much more data is available about a lot more people (and possibly *to* many more people). This data is often used in quotidian decisions such as picking a driving route. In business, much more data is collected on every aspect of operation, increasing efficiency, customer marketing, and pivoting to new markets.[4] In medicine, big data combined with rapid advances medical science can bring us to a point where all major health decisions are tailored to an individual's situation.[5]

If big data fulfills its promise, we think it will have tremendous impact in reducing uncertainty around large domains of decisions, both before they're made and, retrospectively, afterward, too.

### 3.18.5 Technological Challenges

The collection, organization, validation, interpretation, and management of large datasets present multiple technical and technological challenges. As the amount of data grows rapidly, additional

4 http://www.wired.com/insights/2013/07/putting-a-dollar-value-on-big-data-insights

5 http://spectrum.ieee.org/tech-talk/computing/software/predictive-analytics-and-deciding-who-should-receive-organ-transplants

CONTENTS



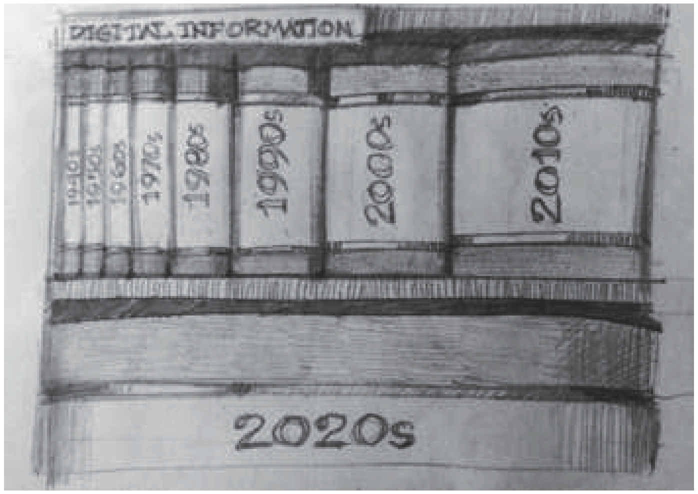
computational resources are required to process the data in a timely manner. This time and resource pressure is increased for ongoing analysis by a recurring deadline (such as daily business metrics) and even more so for interactive and exploratory data exploration.

Accordingly, computational resources dedicated to big data are growing explosively. The demand for data storage and processing can, however, grow faster than the underlying technologies. Take the recent growth of informatics-based scientific disciplines, for example. In genomics, advances in gene sequencer technology have brought down the cost and delay of sequencing to the point where many labs around the world can produce copious genetic data. Worldwide, we can now produce around 15 Pbytes of compressed genetic data per year, which is growing at a rate of 3x to 5x a year.[6] In high-energy physics, the Large Hadron Collider and other instruments at CERN alone produce a similar amount of data annually.[7]

Pervasive big data tools such as Hadoop are already prevalent in the analysis of these massive genomic databases. But despite the carefully designed scalability of the software tools, they are still limited by hardware constraints, such as power, acquisition, and operation costs; capacity growth in processors, hard disks, and networks; and increased complexity in management and cooling.

6 IEEE Spectrum 07-2013 "The DNA Data Deluge" http://spectrum.ieee.org/biomedical/devices/the-dna-data-deluge

7 http://home.web.cern.ch/about/computing




For example, CERN is planning for its Wigner datacenter to double its processing and storage capability in the next three years. Although impressive, this rate is a far cry from the growth rate of the data to be processed, creating an increasing gap between the amount of data to process and the hardware to process it. If cost weren't an issue in scaling the hardware, power still remains a stubborn constraint, limiting practical datacenter size to several MW. And even if that constraint were to be removed by advances in power efficiency, the speed of light effectively limits the scale of a datacenter to the tolerable limits of latency in data fetching before computation grinds to a halt.

### 3.18.6 Potential Technological Disruptions

To fully exploit the opportunity promised by big data, we must find ways to bridge the gap between data growth and processing capability.

On the hardware side, it is a simple matter to extrapolate current growth trends to predict increased storage density; continued processor

# CONTENTS



growth along Moore's law; and better power efficiency and networks. But the trends set in the past two decades, even if exponential in growth, are not disruptive enough to close this gap. It would take radical technological shifts to match resource growth with data growth, or we could experience a significant decline in the current rate of data growth and, with it, the predictive capabilities of its analysis.

Barring an unpredictable disruptive technology, a more feasible path to closing the gap is innovation in software. Big data software can be considered still in its infancy, with plenty of opportunities for growth. Areas of possible improvement include

- reducing the amount data to be processed: better compression; early detection of irrelevant data; and more effective sampling techniques;
- algorithmic effort/processing reduction: more efficient machine learning algorithms to produce predictions in less time, etc.;
- improving systems effort: better utilization and sharing of available hardware resources; and
- generating insightful analysis: something that produces much higher-level analytics and answers than is standard today.

Such innovations will not only reduce the requirements of labor and expertise from analysts but may in fact drive efficiency through more parsimonious representations of data.

### 3.18.7 Societal Challenges

Big data hardware and software are no panacea. In fact, they are currently useless without specialized human skills. These skills range from the selection, preparation, and cleaning of data; the exploration of different analyses on the data; the application of error checking and strong statistical reasoning to reduce bias and type I/II errors; and finally, the interpretation, visualization (for the higher-bandwidth visual sensory), and domain application of the results. Although we believe that many important parts of these processes can and will undergo increased automation, we do not foresee an elimination of the skilled human element. If anything, big data analytics falls in line with the workforce migration we observed in the past century from labor- to knowledge-intensive industries.

In a recent study, 88 percent of companies surveyed have already reported a talent shortage to successfully execute on big data initiatives.[8] Because big data is growing at such a rapid rate, along with the demand for data scientists and analysts, and because the skills required encompass a wide range of advanced computing, statistics, communication, and domain expertise, we may potentially face a critical shortfall in this workforce.[9] This challenge needs to be met by a correspondingly large challenge in workforce education, both in academia and industry.

Other aspects of the big data shift will require societal response. Perhaps the biggest one is the concern about eroding privacy and data leaks, with

8 http://thehiringsite.careerbuilder.com/2013/07/16/careerbuilder-big-data-study/

9 http://spectrum.ieee.org/podcast/at-work/tech-careers/is-data-science-your-next-career

CONTENTS



a potential for very significant personal, business, or military damage. The concentration of big data in the hands of governments also evokes concerns about the risk to democracy and civil rights. The challenge is then to find ways to collect, share, and benefit from big data technologies while still preserving the privacy, trust, and rights of the individuals whose data is collected.

### 3.18.8 Potential Disruptions

Academia is already mobilizing to develop new programs around big data and to train thousands of new data scientists and analysts.[10] Perhaps a more radical approach in workforce education is required to meet the rapid demand. Interestingly, one potential disruption in the training of the workforce in general, and big data in particular, also comes from a new scale-out field: MOOCs (refer to Section 3.4). Distributed online education, with its various levels of certification and cost, is already training many thousands of individuals in big data-related fields and showing a strong growth trend.[11]

One of the characteristics of MOOCs is that successful training no longer requires physical school attendance and is therefore independent of geography. This is just one aspect that may require employers too to radically adjust to the changing landscape of big data professionals, even if universities and MOOC train an adequate number of them. We may therefore experience a disruption from the traditional employment model of accredited employees sharing an office space. Instead, we may see the increasing demand for these professionals met by companies who successfully adapt to a distributed workforce of varying formal education.

Finally, we may find that the explosive success of big data may hinge on a significant disruption in the field of data security and privacy. There is certainly a technological challenge and opportunity here, to come up with provable standards of privacy and security. But there is also one that may require legal, normative, and educational changes to place acceptable limits on the use of big data.

### 3.18.9 References

G. Brumfiel, "High-Energy Physics: Down the Petabyte Highway," *Nature*, vol. 469, 19 Jan. 2011, pp. 282–83.

P. Webster, "Supercomputing the Climate: NASA's Big Data Mission," *CSC World*, Computer Sciences Corporation, 2012.

S. Shah, A. Horne, and J. Capella, "Good Data Won't Guarantee Good Decisions," *Harvard Business Review*, Apr. 2012.

"Data, Data Everywhere", *The Economist*, 25 Feb. 2010.

O.J. Reichman, M.B. Jones, and M.P. Schildhauer, "Challenges and Opportunities of Open Data in Ecology," *Science*, vol. 331, no. 6018, 2011, pp. 703–705; doi:10.1126/science.1197962.

S. Cherry, "Is Data Science Your Next Career?", *IEEE Spectrum*, 28 May 2013.

10 http://www.wired.com/insights/2013/07/the-growing-need-for-big-data-workers-meeting-the-challenge-with-training/

11 http://gigaom.com/2012/10/14/why-becoming-a-data-scientist-might-be-easier-than-you-think/