2018

"Big Data or Big Brother?" That is the question now.

Johnson, Jeffrey; Denning, Peter; Delic, Kemal A.; Sousa-Rodrigues, David

ACM

Johnson, Jeffrey, et al. "Big data: big data or big brother? that is the question now." Ubiquity 2018.August (2018): 2.
http://hdl.handle.net/10945/60975

Ubiquity Symposium

Big Data

Big Data or Big Brother? That is the question now.

by Jeffrey Johnson, Peter Denning, Kemal A. Delic, and David Sousa-Rodrigues

Editor's Introduction

This ACM Ubiquity Symposium presented some of the current thinking about big data developments across four topical dimensions: social, technological, application, and educational. While 10 articles can hardly touch the expanse of the field, we have sought to cover the most important issues and provide useful insights for the curious reader. More than two dozen authors from academia and industry provided shared their points of view, their current focus of interest and their outlines of future research. Big digital data has changed and will change the world in many ways. It will bring some big benefits in the future, but combined with big AI and big IoT devices creates several big challenges. These must be carefully addressed and properly resolved for the future benefit of humanity.

Ubiquity Symposium

Big Data

Big Data or Big Brother? That is the question now.

by Jeffrey Johnson, Peter Denning, Kemal A. Delic, and David Sousa-Rodrigues

The "Big Data Symposium" began with the argument that the important thing about data is not that they are "big" in terms of volume, variety, velocity, but that they are digital. In principle, this means that huge amounts of data are available to anyone in the world over the internet. In practice much data are private and not available. Even private data are finding their way through the internet as many organizations find traditional transactional data have monetary value and are making them available on a commercial basis. The new generations of data that could be used in the social sciences represent a quantum shift from what was available to previous generations, but in practice those data are not freely available to researchers, when they are available at all. Digital data have enormous latent value and those who hold it are generally reluctant to make it available open source and free of charge.

The impact of digital data on society is very great and increasing. Social networks and big data determine what is noticed and acted upon. This is illustrated by the tweeting elected official speaking directly to the public, the emergence of targeted advertising, and "false news" that can influence the outcome of elections. Two billion people are signed up to Facebook [1], providing a free platform for rich interactions between families and other groups, while creating an amazing database of social connections and individual behaviors that is privately owned and controlled by the tech giant. But Facebook is not an anomaly, other social network sites also have large numbers of users and large privately owned databases of user behavior. The number of logged in users is estimated as follows: YouTube (1.5 billion), WeChat (889 million), Instagram (700 million), Twitter (328 million), and Snapchat (255 million) [2]. At the time of this writing, Google processes 3.5 billion searches per day, or 1.2 trillion searches per year[3]. These data give deep and new insights into human behavior. For example, Seth Stevens discovered while people tend to tell lies on Facebook, their Google searches reflect deep personal truths [4]. The same infrastructure that enables some people to maintain their social relationships is used by others to promote hate and intimidate. For example, the British House of Commons released a report in 2017 that focused on online hate crimes: "We took evidence

from Google (the parent company of YouTube), Twitter and Facebook on hate speech and extremism published on their platforms. … It was shockingly easy to find examples of material that was intended to stir up hatred against ethnic minorities on all three of the social media platforms that we examined—YouTube, Twitter and Facebook. … YouTube was awash with videos that promoted far-right racist tropes, such as anti-Semitic conspiracy theories" [5]. In another context, a 2013 Rand study [6] concluded, "Evidence from the primary research conducted confirmed that the Internet played a role in the radicalization process of the violent extremists and terrorists whose cases we studied ... our research supports the suggestion that the Internet may enhance opportunities to become radicalized, as a result of being available to many people, and enabling connection with like-minded individuals from across the world 24/7," and "secondly, our research supports the suggestion that the Internet may act as an 'echo chamber' for extremist beliefs; in other words, the Internet may provide a greater opportunity than offline interactions to confirm existing beliefs."

The information required for policy is massive in scale, fine-grained in resolution, and distributed over many data sources. Here policy includes city and transportation planning, provision of health and welfare, control of epidemics, financial regulation, and much else that keeps enables our societies to function. This requires a new understanding of digital data for policy applications.

Digital Data Technology

New devices, network, and storage technologies combined with new platform architectures have enabled and driven the huge growth in digital data. Already a new generation of technological innovation is under way to support new generations of digital data processing; HPC (high-performance computing) is a strategic resource for Europe's future as it allows researchers to study and understand complex phenomena while allowing policy makers to make better decisions and enabling industry to innovate in products and services. The European Commission believes societal, scientific, and economic needs are the drivers for the next generation of HPC, computing with exascale performance (10 to the power of 18 floating point operations per second). Modern scientific discovery requires very high computing power and the capability to deal with huge volumes of data. Industry and SMEs are increasingly relying on the power of supercomputers to invent innovative solutions, reduce costs, and decrease time to market for products and services; HPC is part of a global race. Many countries (USA, Japan, Russia, China, Brazil, and India) have announced ambitious plans for building the next

generation of HPC with exascale performance and deploying state-of-the-art supercomputers [7].

Data Science Applications

Apart from individuals using the internet for antisocial purposes, criminal organizations are exploiting weaknesses in the new and poorly understood consequences of digital data. For example, machine learning can be subverted, while corporate security is a big data problem.

More positively, large heterogeneous digital data are essential in public administration and commerce. This brings new requirements and technical challenges: "One of the key challenges in building systems to support policy informatics is information integration. Synthetic information environments (SIEs) present a methodological and technological solution that goes beyond the traditional approaches of systems theory, agent-based simulation, and model federation. An SIE is a multi-theory, multi-actor, multi-perspective system that supports continual data uptake, state assessment, decision analysis, and action assignment based on large-scale high-performance computing infrastructures. An SIE allows rapid course-of-action analysis to bound variances in outcomes of policy interventions, which in turn allows the short time-scale planning required in response to emergencies such as epidemic outbreaks."

Data Science and Education

Although digital data has huge latent value, extracting that value is becoming increasingly difficult. Data science is required to transform big data into useful, valuable information. Data science is driven by practical problems and involves finding relevant data, data preparation, data analysis, and data visualisation. The first of these includes many data identification processes, including manual and automated search, but also approaching data owners to identify possible data sources and negotiating the costs and conditions for its use. The heterogeneous data produced by the search process usually means the various data sets are incomplete and inconsistent, sometimes requiring considerable data cleaning work in the data preparation stage. Following this, the data can be presented to a wide variety of analytic programs in the data analysis phase. Analysis can involve establishing statistical distributions and relationships, building models to investigate systems, including mathematical and agent based modeling, and increasing the use of deep-learning techniques. The applications-driven

nature of data science means that visualization is extremely important for understanding the output of the applications stage, and communicating the results to clients and stakeholders. In other words, data engineering is as important as data science.

Data science covers a wide variety of hard and soft knowledge and skills, and data-science education has become a very dynamic area in the last few years with data scientist commanding more than twice average incomes. The hard skills include programming in a variety of high- and low-level experience with platforms and software, including Hadoops and the Amazon and Apache families of platform software, mathematical and statistical modeling, machine learning, and so on. Essential soft skills include team working, independence and self-direction, enthusiasm, and the ability to work in fast-moving environments, and so on.

Given the enormous potential value to be abstracted from data using new state-of-the-art methods, there is great demand for data science education. For example, in the U.K. there are more than 100 master's courses lasting one year with fees in the order of 8,000-15,000 pounds. Also the are many commercial boot camps giving intense data science education over a period of a few months for fees up to 15,000 pounds for residential courses. Worldwide there are many MOOCs (massive open online courses) in many areas relevant to data science, where these may be free or subject to fees in the order of 50 to a few hundred dollars.

Many organizations do not have data science departments and are unaware of the added value that data science can bring. Apart from highly trained data scientists there is need for the "data science bridge person," someone from the business side of an organization trained to know enough about data science to be able to hold productive conversations with specialists, either within or without their organizations. Currently there is no education for this requirement. The Da.Re. project is creating a blended course of 80 hours online education followed by 70 hours face-to-face education learning how to apply their technical knowledge to practical case-study problems [8].

The Limitations of Big Data, AI, and Deep Learning

After nearly a decade, big data continues to attract great interest, as its added value becomes better understood. However, the harvest of low hanging fruit of big data is over and new challenges lie ahead. Although very impressive, what can be achieved with cutting edge technologies such as deep learning has been achieved and future progress in this area will be more incremental. Artificial intelligence (AI) caused great interest in the 1970s and 1980s, but

fell short of the claimed benefits. Despite great achievements such as Big Blue beating the world champion chess player in 1996 the gap between machine intelligence and human intelligence remains immense. Deep learning through neural networks has also resulted in astonishing achievements including the advertisements spookily targeted at the user as part of the implicit Faustian pact they make with the suppliers of "free" web services. However, the difference between deep learning and human learning and analysis is also immense. The reason is that the underlying computational models are not rich enough to represent the complexities of the systems being investigated. Put simply, we have no reliable science of multilevel social and economic systems and it is unlikely that such a science will emerge from the current generation of AI supported neural networks operating on massive digital data sets.

Data Privacy

Abuse of personal data is a problem that has been tackled head-on by the European Commission. On May 25, 2018 a new directive was implemented across Europe known as the General Data Protection Regulation (GDPR) The benefits for citizens include: protection of and tools for gaining control of one's personal data, strengthening citizens' rights and building trust, a "right to be forgotten" easier access to one's data, the right to know when one's data has been hacked, data protection by design and by default, and stronger enforcement of the rules [9].

Conclusion

We shall conclude with a few departing, warning thoughts:

- The major proportion of researchers are pushing positive, beneficial, and optimistic outcomes of big data research, while downplaying possible misuses and manipulations. In particular, big commercial companies are amassing troves of private data claiming no interest in personal details, while in reality selling, exchanging, or misusing such data. One can easily imagine what would happen if medical, financial, and behavioral data fused for the targeted individual fell into hands of bankers, insurers, politicians, or criminals. Mayhem would follow, no doubt.

- Significant marketing money is spent on big data publicity claiming some major breakthrough may happen once we deploy advancing AI algorithms on big data troves. However, recent research disclosed only one project in six is put in production and used. Therefore, one should carefully consider all aspects before launching big data project—the failure rate is pretty high.

- Thinking about big data fed from the forthcoming big IoT infrastructures with huge streams of real-world data, one should especially consider avoidance of the big brother infrastructure enabling surveillance at unprecedented levels, depth, and size. Having thought about benefits, we should always consider possible creation of the monster that could be used against our will or original intention. Will regulatory initiatives such GDPR be strong enough to counter such challenges?

- Data science is essential to find and extract the enormous latent value of digital data. Data science as it exists today is limited by the range of techniques and software currently in use. The future lies in new science and engineering that can go far beyond the current limitations of AI, deep learning, and human-computer interaction. There are reasons to be optimistic.

Big data will bring some big benefits in the future, but combining it with big AI (advanced, efficient algorithms) and big IoT (omnipresent, always connected) devices brings several big challenges. These must be carefully addressed and properly resolved for the future benefit of humanity. The future of big data is very human, and if we are wise it can help us face the challenges ahead and improve global social and economic well being.

About the Authors

Jeffrey Johnson is Professor of Complexity Science and Design at the Open University in the UK, Deputy President of the UNESCO UniTwin Complex Systems Digital Campus (http://www.cs-dc.org), Past-President of the Complex Systems Society (https://cssociety.org), and a partner in the European Erasmus Da.Re. Project (Data Science Pathways to Re-imagine Education, http://dare-project.eu). His research interest is in the dynamics of complex multilevel social and environmental systems, and in systems thinking and computational complex systems science in policy and management. He has published many scientific books and papers, and supervised a

large number of doctoral students. He has many years experience of online education including creating MOOCs on global systems science and policy, systems thinking and complexity, and data science. He is CEO of Vision Scientific Ltd. and had led many scientific and commerical research and development projects. He is an associate editor of ACM *Ubiquity*.

Peter J. Denning is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at the Naval Postgraduate School in Monterey, California, is Editor of ACM *Ubiquity*, and is a past president of ACM. The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

Kemal A. Delic is a senior technologist and practicing enterprise architect with Hewlett-Packard Co. He is Visiting Senior Fellow with The Open University, UK. He is also adjunct professor at IAE business school at Grenoble University. He serves as associate editor to ACM *Ubiquity* magazine. He acted as an advisor and consultant to European Commission on FET programs. He holds two U.S. and one E.U. patent. Delic lives in Grenoble, France. He holds a Dipl. El. Ing. degree from the University of Sarajevo. During the last 30-plus years he has worked mainly in the area of very large-scale systems—either on industrial, commercial products or academic research. His principal interest is in architecture, design, and engineering of large-scale systems. From 1989 to 1990, and again from 1994 to 1996, he was a visiting professor and researcher with CNR – Italian National Research Council – IEI institute working on error detection through signature analysis (PDCS Project), Bayesian Belief Network models in safety arguments (EU SHIP Project), and marketing with mobile intelligent agents (EU MIA Project). He has published 100-plus papers, articles, and essays in journals, magazines, and conferences. He has given talks, delivered lectures, and organized workshops and conferences. In the 1980s he published original book on Pattern Recognition Principles and more recently contributed three book chapters on “Enterprise Knowledge Clouds” (2010, 2011). His recent research interest is in the science and practice of hybrid complex systems. Big science, big data and the digital economy are his most recent interests.

David Sousa-Rodrigues is a professor of Computer Science, Complexity and Design at the European University (https://www.europeia.pt/), in Lisbon, and at Polytechnic Arts and Design School (http://www.esad.ipleiria.pt/) in Caldas da Rainha, Portugal. He is a member of the Centre for Complexity and Design research group of the Open University in the U.K., and was member of Complex Systems Society permanent council. His research interests are distributed multilevel complex systems and bio-inspired computational modeling. Specific topics of interest

include social networks, evolutive computation, machine learning, data analytics, community detection, cellular automata, emergence, and stigmergy.

DOI: 10.1145/3158352

References

[1] Nowak, M., and Spiller, G. Two billion people coming together on Facebook. Facebook, June 27, 2017; https://newsroom.fb.com/news/2017/06/two-billion-people-coming-together-on-facebook

[2] Constine, J. Facebook now has 2 billion monthly users… and responsibility. *TechCrunch*, June 27, 2017; https://techcrunch.com/2017/06/27/facebook-2-billion-users

[3] Google Search Statistics. InternetLiveStats; http://www.internetlivestats.com/google-search-statistics (Accessed September 7, 2017).

[4] Stevens-Davidowitz, S. *Everybody Lies, and What the Internet Can Tell Us About Who We Really Are*. Harper Collins, New York, 2017.

[5] Home Affairs Committee. House of Commons. Hate Crime: Abuse, Hate and Extremism Online. Fourteenth Report of Session 2016–17. HC 609. May 1, 2017; https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/609.pdf

[6] von Behr, I., Reding, A., Edwards, C., and Gribbon, L. Radicalisation in the Digital Era - The use of the internet in 15 cases of terrorism and extremism. Rand Europe, 2013; https://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR453/RAND_RR453.pdf

[7] European Commission. High-Performance Computing (HPC). 2016; https://ec.europa.eu/programmes/horizon2020/en/h2020-section/high-performance-computing-hpc

[8] Cristali, C. et al. New Big Data Initiatives: Towards a Data-driven Mind-Set. Intellectual Output 1. Data Science Pathways to Re-Imagine Education (Da.Re.) Project; www.dare-project.eu

[9] European Commission. Questions and Answers - Data Protection Reform. December 21, 2015; http://europa.eu/rapid/press-release_MEMO-15-6385_en.htm