Bus Inf Syst Eng 59(5):311–313 (2017)
DOI 10.1007/s12599-017-0487-z

EDITORIAL

Responsible Data Science

Wil M. P. van der Aalst · Martin Bichler · Armin Heinzl

Published online: 26 June 2017
© Springer Fachmedien Wiesbaden GmbH 2017

1 Introduction

An increasing fraction of research reported in BISE (Business & Information Systems Engineering) is data-driven. This is not surprising since torrents of data are vigorously changing the way we do business, socialize, conduct research, and govern society (Hilbert and Lopez 2011; Manyika et al. 2011; White House 2016). Data are collected on everything, at every time, and in every place. The Internet of Things (IoT) is rapidly expanding, with our homes, cars, and cities becoming "smart" by using the collected data in novel ways. These developments are also changing the way scientific research is performed. Model-driven approaches are supplemented with data-driven approaches. For example, genomics and evidence-based medicine are revolutionizing the understanding and treatment of diseases. From an epistemological point of view, data-driven approaches follow the logic of the new experimentalism (Mayo 1996; Chalmers 2013) in which knowledge is derived from experimental observations, not theory. Information systems which exploit the combination of data availability and powerful data science techniques dramatically improve our lives by enabling new services and products, while improving their efficiency and quality. *However, there are also great concerns about the use of data* (van der Aalst 2016a, b). Increasingly, customers, patients, and other stakeholders are concerned about irresponsible data use. Automated data decisions may be unfair or non-transparent. Confidential data may be shared unintentionally or abused by third parties. Each step in the "data science pipeline" (from raw data to insights and knowledge) may create inaccuracies, e.g., if the data used to learn a model reflects existing social biases, the algorithm is likely to incorporate these biases. These concerns could lead to resistance against the large-scale use of data and make it impossible to reap the benefits of data science. Rather than to avoid the use of data altogether, we strongly believe that data science techniques, infrastructures and approaches need be made responsible by design. Over the last year the first author has been leading a Dutch initiative called *Responsible Data Science* (RDS), cf. http://www.responsibledatascience.org/. In the context of RDS, there are research projects and regular meetings to discuss new ways to make data science more responsible. We believe that the insights obtained from these discussions are also relevant for the BISE community. The data-driven nature of today's (business) information systems makes it essential to incorporate safeguards against irresponsible data use already in the requirements and design phases.

2 FACT: Fairness, Accuracy, Confidentiality, and Transparency

Responsible data science centers around four challenging questions (van der Aalst 2016a; Responsible Data Science Initiative 2016):

Prof. dr. ir. W. M. P. van der Aalst (✉)
Department of Mathematics and Computer Science (MF 7.103), Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: w.m.p.v.d.aalst@tue.nl

Prof. Dr. M. Bichler
Department of Informatics, Decision Sciences & Systems, Technical University of Munich (TUM), Boltzmannstr 3, 85748 Munich, Germany
e-mail: bichler@in.tum.de

Prof. Dr. A. Heinzl
Chair of General Management and Information Systems, University of Mannheim, 68161 Mannheim, Germany
e-mail: heinzl@uni-mannheim.de

www.manaraa.com

- Q1 fairness: data science without prejudice - how to avoid unfair conclusions even if they are true?
- Q2 accuracy: data science without guesswork - how to answer questions with a guaranteed level of accuracy?
- Q3 confidentiality: data science that ensures confidentiality - how to answer questions without revealing secrets?
- Q4 transparency: data science that provides transparency - how to clarify answers so that they become indisputable?

The terms fairness, accuracy, confidentiality, and transparency form the acronym FACT. This should not be confused with the well-known FAIR principles (Findable, Accessible, Interoperable, and Re-usable). Whereas FAIR looks at practical issues related to the sharing and distribution of data, FACT focuses more on the foundational scientific challenges.

Data science approaches learn from training data while maximizing an objective (e.g., the percentage of correctly classified instances). However, this does not imply that the outcome is fair. The training data may be biased or minorities may be underrepresented or individually discriminated. Even if sensitive attributes are omitted, members of certain groups may still be systematically rejected. Profiling may lead to further stigmatization of certain groups. Therefore, approaches are needed to detect unfair decisions (e.g., unintended discrimination) and to find ways to ensure *fairness*.

The abundance of data suggests that we should let the data "speak for themselves". Data science makes this possible, but at the same time analyses of data sets - large or small - often produce inaccurate results. In general, it is challenging to "let the data speak" in a reliable manner. If enough hypotheses are tested, one will eventually be true for the sample data used. If we have one response variable (e.g., "will someone conduct a terrorist attack") and many predictor variables ("eye color", "high school math grade", "first car brand", etc.), then it is likely that just by accident a combination of predictor variables explains the response variable for a given data set. Multiple testing problems are well-known in statistical inference, but often underestimated. Data science approaches should not just present results or make predictions, but also explicitly provide meta-information on the *accuracy* of the output.

Data science heavily relies on the sharing of data (Dwork 2011). If individuals do not trust the "data science pipeline" and worry about confidentiality, they will not share their data. The goal should not be to prevent data from being distributed and gathered, but to exploit data in a safe and controlled manner. *Confidentiality* questions need to be addressed both from a security perspective (polymorphic encryption and pseudonymization) and a legal/ethical perspective (e.g., perceptions and effects on the behavior of individuals). The focus should not be on circumventing the sharing of data, but on innovative approaches like confidentiality-preserving analysis techniques (e.g., techniques that work under a strict privacy budget).

Data science can only be effective if people trust the results and are able to correctly interpret the outcomes. Data science should not be viewed as a black box that magically transforms data into value. The journey from raw data to meaningful inferences involves multiple steps and actors, thus accountability and comprehensibility are essential for *transparency*.

Consider for example the recent attention and enthusiasm for deep learning. Breakthroughs make it possible to make better decisions; however, the neural networks used by the deep learning approach cannot be understood by humans. Hence, they serve as a black box that apparently makes good decisions, but cannot rationalize them. In several domains, this is unacceptable.

In most situations, *causal inference* is the goal of data analysis in business, but often enough correlation is confused with causality. Econometricians are well aware of this and have developed techniques for causal inference when a randomized controlled trial, the gold standard of causal inference, is not possible. Propensity score matching or inverse probability-weighed regression adjustment are just two approaches developed to combat the selection bias in observational data. While these techniques address the selection bias, their outcomes might still be far away from the results one would obtained with a randomized controlled trial as was recently illustrated by Gordon et al. (2016). This can lead to wrong interpretations of data and entirely spurious conclusions.

Simpson's paradox is another nice example to show how easy it is to give false advice even in the presence of "big" data. The paradox describes a phenomenon in which a trend appears in different groups of data but disappears or reverses when these groups are combined. It is frightening to see data scientists nowadays who seem not to be aware of the many pitfalls in the modeling of data. It takes years of training to acquire the skill set necessary to draw solid statistical inferences. Without this training, the likelihood of young and ambitious 'data scientists' making false claims is high.

3 Designing FACT-based Information Systems

Many consider (Big) data as the "new oil" which can be refined into new forms of "energy": insights, diagnostics, predictions, and automated decisions. However, the FACT

www.manaraa.com

challenges just described show that the careless transformation of "new oil" (data) into "new energy" (data science results) may negatively impact citizens, patients, customers, and employees. Systematic discrimination based on data, invasions of privacy, non-transparent life-changing decisions, and inaccurate conclusions can be viewed as new forms of "pollution". In van der Aalst (2016a) the term "green data science" was coined for cutting-edge solutions that enable individuals, organizations, and society to benefit from widespread data availability while ensuring Fairness, Accuracy, Confidentiality, and Transparency. Note that "green data science" does not refer to making data centers more energy efficient: It is about the possibly negative side effects that data may have on people's lives.

There might be an opportunity for Europe when it comes to green data science. Consider an "Internet Minute" (James 2016) with approximately:

- 1,000,000 Tinder swipes,
- 3,500,000 Google searches,
- 100,000 Siri answers,
- 850,000 Dropbox uploads,
- 900,000 Facebook logins,
- 450,000 Tweets sent,
- 7,000,000 Snaps received,
- etc.

All of the above activity is governed by software and hardware of US-based companies. In some countries this raises great concerns about competitiveness, privacy protection, etc. In the world of data, a few organizations seem to rule the world. Spotify (a Swedish) company is one of the rare exceptions (i.e., a successful non-US-based organization significantly contributing to today's internet traffic). However, the stricter laws in Europe can also create a competitive advantage. On 14 April 2016, the EU Parliament approved the general data protection regulation (GDPR) which aims to strengthen and unify data protection for individuals within the EU. This may provide a boost for new ways of using data without the "pollution" described before. However, this opportunity only exists if policy makers really want to invest. It is not sufficient to just bring in legislation, we also need technological breakthroughs.

4 Responsible Business and Information Systems Engineering

The BISE community should play an active role in making our next generation of information systems "green". Already during the design and requirements phases one should take into account questions related to fairness, accuracy, confidentiality, and transparency (FACT). Consider today's customer relationship management (CRM) systems, enterprise resource planning (ERP) systems, hospital information systems (HIS), learning management systems (LMS), etc. How can we make the next generation of these systems green? For example, should we add FACT elements to our modeling languages? How can FACT elements be embedded in our requirements? We hope to see future contributions to BISE addressing these questions!

References

Chalmers AF (2013) What is this thing called science? An assessment of the nature and status of science and its methods. McGraw Hill, New York

Dwork C (2011) A firm foundation for private data analysis. Commun ACM 54(1):86–95

Gordon B, Zettelmeyer F, Bhargava N, Chapsky D (2016) A comparison of approaches to advertising measurement: evidence from big field experiments at facebook. White paper, Kellog School of Management, Northwestern University, Evanston

Hilbert M, Lopez P (2011) The world's technological capacity to store, communicate, and compute information. Science 332(6025):60–65

James J (2016) Domo blog: data never sleeps 4.0. https://www.domo.com/blog/data-never-sleeps-4-0/. Accessed 11 June 2017

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, New York

Mayo DG (1996) Error and growth of experimental knowledge. University of Chicago Press, Chicago

Responsible Data Science Initiative (2016) Responsible data science. http://www.responsibledatascience.org/. Accessed 11 June 2017

van der Aalst W (2016a) Green data science: using big data in an "environmentally friendly" manner. In: Camp O, Cordeiro J (eds) Proceedings of the 18th international conference on enterprise information systems (ICEIS 2016), Science and Technology Publications, pp 9–21

van der Aalst W (2016b) Process mining: data science in action. Springer, Heidelberg

White House (2016) Artificial intelligence, automation, and the economy. (Report released by the Executive Office of the President in December 2016). https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.pdf. Accessed 11 June 2017

www.manaraa.com