International Journal of Data Science and Analytics (2018) 6:163–165
https://doi.org/10.1007/s41060-018-0153-7

EDITORIAL

Introduction to the special issue on Data Science in Europe

Peter Flach[1] · Myra Spiliopoulou[2] · Serge Allegrezza[3] · Matthias Böhmer[4] · Burkhard Hess[5] · Berthold Lausen[6]

Published online: 6 October 2018

It is our great pleasure to welcome the reader to this collection of papers devoted to current issues in Data Science viewed from a European perspective. This special issue of the *Journal of Data Science and Analytics* has its origin in the European Data Science Conference (EDSC), an invitation-only event organised by Professor Sabine Krolak-Schwerdt and her team in November 2016 in Luxembourg as the inaugural conference of the European Association for Data Science (EuADS). We thank the JDSA Editor-in-Chief, Professor Longbing Cao, for the opportunity to guest-edit this collection.

The conference programme of EDSC consisted of selected plenary talks, symposia, workshops and panel discussions on the following topics:

- The question of trust, transparency and provenance of data including where data come from and by which mechanisms trust in data might be achieved.
- Legal aspects of Data Science such as data protection, data privacy and data access, among others.
- The question how data scientists might navigate the complex chain from raw data to actionable outputs and how they might be supported by appropriate tools.
- The role of Data Science in medicine and health care with a specific focus on Open medical data and personalized Medicine.
- The question what makes Data Science different from other fields such as statistics and how to define the methodological substance of the field.

Prior to the conference, participants were invited to contribute position statements for presentation and discussion at the conference. They then were given the opportunity to submit revised and extended versions to this special issue. The submissions received went through rigorous peer review, resulting in twelve papers that are briefly described below. They can be broadly grouped into three main categories.

Perspectives on Data Science research and training

In 1962 John Tukey published an article entitled "The future of data analysis" in The Annals of Mathematical Statistics. In *What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences*, Christophe Ley and Stéphane Bordas take as their starting point the question what has really changed since then. Their paper can be seen as a dialogue between statistics and computational sciences, and their main conclusion is that 'Data Science enhances the traditional and more conservative world of Statistics with advanced algorithms' which is necessary to make sense of the large volumes of data we encounter today.

In his position paper *Data Science as a language: challenges for computer science*, Arno Siebes develops the viewpoint that Data Science is a way to conduct enquiries in 'datafied sciences' using computer science as a language, in much the same way as mathematics is the language of physics. He then goes on to consider the challenges this poses for computer science, and the future research directions these challenges entail.

Claus Weihs and Katja Ickstadt, in their paper *Data Science: The impact of statistics*, emphasise the importance of statistics in many steps of the Data Science pipeline, in particular data acquisition and enrichment, data exploration, data analysis and modelling, and validation and reporting. They argue that this importance is not always recognised and recommend statisticians to 'more offensively play their role in this modern and well accepted field of Data Science'.

✉ Peter Flach
Peter.Flach@bristol.ac.uk

[1] University of Bristol, Bristol, UK

[2] University of Magdeburg, Magdeburg, Germany

[3] Luxembourg Institute for Statistics and Economic Studies, Luxembourg, Luxembourg

[4] University of Luxembourg, Luxembourg, Luxembourg

[5] Max Planck Institute Luxembourg, Luxembourg, Luxembourg

[6] University of Essex, Colchester, UK

The next two papers consider the important issue of training people to become data scientists. Göran Kauermann and Thomas Seidl outline a possible curriculum based on the Data Science masters programme currently offered at their university in Munich, in their contribution *Data Science: a proposal for a curriculum*. The programme balances the statistics and computer science aspects of data analytics, while also covering ethical aspects as well as practical and communication skills.

Further insights from a running training programme are given by Gilbert Saporta in *Training data scientists: a few challenges*. He emphasises the necessity of continuous education and learning on the job to acquire and maintain the quite specific combination of skills that a data scientist needs: not just technical skills, but also a 'feeling for data' and business skills that cannot easily be taught in the classroom.

Data Science methodology, infrastructure and context

Data Science needs infrastructure and methodological innovations, while considering the relationship with neighbouring disciplines such as official statistics and law. An exemplar of a successful research infrastructure is described in *Data science at SoBigData: the European research infrastructure for social mining and big data analytics* by Valerio Grossi, Beatrice Rapisarda, Fosca Giannotti and Dino Pedreschi. They argue that managing and exploiting large data volumes require a complete architectural redesign with regard to data management, privacy and scalability and describe the approaches adopted in the SoBigData European project.

In *Declarative data analysis*, Hendrik Blockeel describes the process of data analysis as a pipeline in which questions arising in the domain of application are reformulated as statistical or data mining tasks, and the results obtained need to be translated into meaningful answers in the application domain. The goal of declarative data analysis is to automate these two translation steps as much as possible, so that users can analyse data in a goal-oriented manner without the need for detailed knowledge of the underlying techniques. The paper describes the state of the art and future directions towards this ambitious goal.

Building on his experience as Director-General (from 2008 to 2016) of Eurostat, a Directorate-General of the European Commission that provides official statistics for the EU, Walter Radermacher argues in *Official statistics in the era of big data: Opportunities and threats* the need for rethinking the role and quality of official statistics, widening its scope to go beyond mere production to include their usage and analysing scientifically how production and use interact.

Legal scholars Jan von Hein and Anna Bizer give their perspective in *Social media and the protection of privacy: Current gaps and future directions in European private international law*. They argue that privacy infringements on the internet are so far only partially governed by European Union law, with gaps that must be filled by laws of the member states. The article addresses these problems and develops a number of recommendations, taking into account the recent case law of the Court of Justice of the European Union.

Data Science in medicine and healthcare

Many contemporary societal challenges are related to public health, and while Data Science is poised to play an important role here there are also many challenges to overcome. In their contribution *Big data and precision medicine: Challenges and strategies with healthcare data*, Johann Kraus, Ludwig Lausser, Peter Kuhn, Franz Jobst, Michaela Bock, Carolin Halanke, Michael Hummel, Peter Heuschmann and Hans Kestler consider the promise of Data Science for precision medicine. They argue that major challenges arise not so much from algorithmic issues but rather from the need for integration of heterogeneous data sources, and propose a possible strategy by iterating processes of data harmonisation, semantic enrichment and analytics.

Continuing the data integration and quality theme, Ricardo Cruz-Correia, Duarte Ferreira, Gustavo Bacelar, Pedro Marques and Priscila Maranhão discuss in *Personalised medicine challenges: Quality of data* the challenges associated with managing digital health data. They particularly highlight the need for documenting the provenance of the data so that data scientists can fully understand how data were collected and what its context is.

Finally, in *Three controversies in health data science*, Niels Peek and Pedro Pereira Rodrigues discuss three issues that are heavily debated among health data scientists: whether data should be used only for the purpose for which they were collected; to which extent routine data sources and innovations in analytical methods alleviate the need to conduct randomised clinical trials; and questions of governance, privacy and trust when routine health data are made available for research. These three issues relate to core challenges for research with health data and define an essential research agenda for the health data science community.

In memoriam: Sabine Krolak-Schwerdt, 1958–2017

We dedicate this special issue to the memory of Professor Sabine Krolak-Schwerdt, the first president of the European Association for Data Science (EuADS), inspirator and organiser of the 2016 European Data Science Conference which led to this collection of papers. Professor Krolak-Schwerdt held a personal chair in Educational Measurement at the Faculty of Language and Literature, Humanities, Arts and Education at the University of Luxembourg, and positions in Germany prior to that. She sadly passed away in December 2017, shortly before her 59th birthday.

Sabine was at the forefront of several recent activities to build a European community of data scientists. In 2013 she was the inaugural conference chair of the European Conference on Data Analysis in Luxembourg, which has since established itself as a major conference bringing together methodological and applied data scientists from a wide range of disciplines, with subsequent ECDA conferences in Bremen (2014), Colchester (2015), Wroclaw (2017), Paderborn (2018) and Bayreuth (March 2019). Following on from ECDA 2013, Sabine worked tirelessly on establishing EuADS, formally located in Luxembourg, and in 2015 became its founding President. This special issue is just one of the many things that wouldn't have happened without Sabine's leadership.

In addition to being an outstanding and internationally recognised scholar fully dedicated to the scientific endeavour in all its forms, Sabine had a warm personality and a very pleasant and considerate manner which made it a privilege to work with her. She is sadly missed by us and many others, and we will strive to continue and develop her legacy embodied in EuADS and associated activities.