Data Sci. Eng. (2016) 1(1):1–3
DOI 10.1007/s41019-016-0005-1

EDITORIAL

Introduction to Data Science and Engineering

Elisa Bertino[1]

Published online: 25 February 2016

Recent technological advances in sensors, embedded devices, and smart mobile devices and in applications, such as social networks, are making possible to pervasively capture huge amounts of data from many different contexts in both the physical world and the cyber world. Technologies, such as cloud systems, high-performance computing, data analytics, make possible to not only to store huge amounts of data—often referred to as *big data*, but also to process data for many different purposes. Big data is poised to make possible applications that in the past were difficult if at all possible, like personalized healthcare, and enhance all aspects of human life and society, ranging from urban environments to energy and transportation infrastructures, and manufacturing. Data is increasingly playing a major role in addressing challenges such as food and water security and assuring the health of the various ecosystems. Data is today critical for scientific research. Many science and engineering areas are currently experiencing from a 100- to a 1000-fold increase in the volumes of generated data compared to only one decade ago. This data is produced by many sources including simulations, high-throughput scientific instruments, satellites, and telescopes. The availability of big data is revolutionizing how research is conducted and is leading to the emergence of a new paradigm of science based on data-intensive computing and analytics.

However, unlocking the power of data requires addressing several major challenges. Jagadish et al. [1] articulate a discussion of such challenges based on the notion of data management and analysis pipeline:

- *Data Acquisition and Recording*. It is critical to capture the context into which data has been generated in order to be able to filter out non-relevant data and compress data, to automatically generate metadata supporting a rich data description, and to track and record data provenance [2]. Capabilities for in-network data processing are also critical for effective and efficient data acquisition processes [3], especially when dealing with sensor-based data acquisition from rapidly changing environments.
- *Information Extraction and Cleaning*. Data may have to be transformed in order to extract information from it and to express this information in a form suitable for analysis. Data may also be of poor quality and/or uncertain. Data cleaning and data quality are thus critical.
- *Data Integration, Aggregation, and Representation*. Data may be very heterogeneous and have different metadata associated with it. Data integration, even in the more conventional cases, requires huge human efforts. Novel approaches that can improve the automation of data integration are critical as manual approaches will not scale to what is required for today and tomorrow large-scale multi-source data sets. Also different data aggregation and representation strategies may be needed for different data usages and data analysis tasks.
- *Query Processing and Analysis.* Suitable processing methods are needed able to deal with noisy, dynamic, heterogeneous, and untrustworthy data as well data characterized by complex relations. Supporting query processing and data analysis requires scalable data mining algorithms well integrated with query optimization techniques and powerful computing infrastructures.
- *Interpretation.* Data analysis results need to be interpreted by decision makers, and this may require users to be able to analyze the assumptions at each stage of

✉ Elisa Bertino
bertino@cs.purdue.edu

[1] Purdue University, West Lafayette, IN, USA

data processing and possibly re-trace the analysis steps. Rich data provenance is critical in this respect.

In addition, data privacy and security are critical requirements. Big data raises many privacy concerns as, for example, by combining multiple data sets one can easily re-identify privacy-sensitive data even this data has been anonymized. Data can also be used to create profiles of social groups; these profiles may be used for discriminating specific groups of individuals or even single individuals. In this respect, population privacy is as crucial. Data security raises challenging issues including scalable data security administration, management and integration of heterogeneous data security policies, and the security of data when hosted on clouds and managed on sensor networks.

Addressing the today and tomorrow's challenges of pervasive big data management and applications requires combining different disciplines, including computer science and engineering, statistics, and mathematics, as well human factors, cognitive psychology, and linguistics. Also as many data solutions ultimately depend on the specific applications of interest, multi-disciplinary approaches are required for the design and engineering of data-intensive applications.

The *Data Science and Engineering Journal (DSE)* has been created to provide a comprehensive international forum for original results in research, design, development, and assessment of technologies that timely address relevant challenges in data management and data-intensive applications. The journal will discuss problems and solutions at all levels of investigation across the data management and analysis pipeline as well as applications in a broad spectrum of domains. Open access to the journal articles, with no cost to the authors, will make sure that articles will be widely available to the research, industry, and practitioner communities.

The journal is off to an impressive start in this first issue. The papers, all invited, provide a broad perspective about the variety of research that can contribute to the development of effective and efficient data management technologies and data-intensive applications. Firmani, Mecella, Scannapieco, and Batini in "On the Meaningfulness of 'BigData' Quality" present an overview of the many dimensions of data quality and discuss such dimensions in the context of big data and for different types of data sources. This paper is an excellent starting point for everyone interested in the ever increasingly relevant topic of data quality. Sorias-Comas and Domingo-Ferrer in "Big Data Privacy: Challenges to Privacy Principles and Models," after a comprehensive overview of data privacy techniques, discuss the challenges that need to be addressed in order to apply and/or extend such techniques to address key challenges in big data privacy, namely composability, linkability, and low computational costs. This paper is an excellent reference for anyone interested in exploring new directions in data privacy. Hu, Xie, Lin, Wang, and Yu in "Clustering Embedded Approach for Efficient Network Inference" focus on the problem of analyzing information diffusion in social networks and present a framework based on clustering techniques. Extensive experimental results are reported in the paper assessing the efficiency of their approach on different data sources. Cui, Jiang, Huang, Xu, Gui, and Zhang in "POS: A High-Level System to Simplify RealTime Stream Application Development on Storm" focus on efficient and high-level techniques for processing streaming data. The goal of the POS system is to simplify the programming of streaming data applications. As new technological trends, such as the Internet-of-Things (IoT), are pushing the use of sensors and embedded devices, the availability of systems enhancing the development of data streaming applications is critical. The POS system makes an important step toward such direction. Future issues of *DSE* will include additional invited papers and special issues focusing on novel challenging research topics.

There are several individuals and groups who participated significantly in the proposal and creation of *DSE*. Professor Lizhu Zhou of Tsinghua University who was the Chair of the Database Technical Committee of China Computer Federation has been the main driving force of this project and will continue to do so in his role as managing editor and member of the advisory board of *DSE*. Professor Jianzhong Li from Harbin Institute of Technology, in his role as Editor in Chief, is actively working on shaping the contents of *DSE.* We are grateful to our distinguished colleagues that have accepted to serve as members of the advisory board: Michael J. Carey (University of California, Irvine, USA), Masaru Kitsuregawa (University of Tokyo, Japan), Rao Kotagiri (University of Melbourne, Australia), Beng-Chin Ooi (National University of Singapore, Singapore), Gerhard Weikum (Max Planck Institute for Informatics, Germany), Kyu-Young Whang (Korea Advanced Institute of Science and Technology, Korea), and Lizhu Zhou (Tsinghua University, China). *DSE* has also an impressive editorial boards including colleagues from different countries worldwide that we hope will help in making *DSE* an international worldwide forum for research in data science and engineering. We would like to thank Alfred Hofmann, Springer Vice-President Computer Science Publishing, and Celine Chang, Senior Editor of Springer Beijing Office, for their professional support in all the phases of the project. Last but not least, we would like to acknowledge Juirwen Argones—*DSE* editorial assistant.

As *DSE* is sponsored by the China Database Technical Committee, part of the China Computer Federation, we would like to extend our thanks to the China Database Technical Committee for the initiation and support of the project. We trust that *DSE* will become a key resource for researchers in China and abroad. Finally, we would like to acknowledge

the financial sponsorship by Nanjing Sinovatio Tech. Co. Ltd making possible to provide open access to *DSE* papers at no cost for the authors.

I hope you will enjoy this issue and many more issues to come.

References

1. Jagadish HV et al (2012) Challenges and Opportunities with Big Data. Available at http://cra.org/ccc/wpcontent/uploads/sites/2/2015/05/bigdatawhitepaper.pdf (downloaded on Jan. 301, 2016)
2. Sultana S, Bertino E (2015) A distributed system for the management of fine-grained provenance. J Database Manag 26(2):32–47
3. Bertino E, Nepal S, Ranjan R (2015) Building sensor-based big data cyberinfrastructures. IEEE Cloud Comput 2(5):64–69