Letter from the Special Issue Editors

The prevalence of large volumes and varieties of accessible data is profoundly changing the way business, government and individuals approach decision making. Organizational big data investment strategies regarding what data to collect, clean, integrate, and analyze are typically driven by some notion of perceived value. However, the value of the data is inescapably tied to the underlying quality of the data. Although for big data, value and quality may be correlated, they are conceptually different. For example, a complete and accurate list of the books read on April 1, 2016 by the special editors of this issue may not have much value to anyone else. Whereas even partially complete and somewhat noisy GPS data from public transport vehicles may have a high perceived value for transport engineers and urban planners. In spite of significant advances in storage and compute capabilities, the time to value in big data projects often remains unacceptable due to the quality of the underlying data. Poor data quality is being termed as the dark side of big data, inhibiting the effective use of data to discover trusted insights and foresights.

Finding the nexus of use and quality is a multifaceted problem encompassing organizational and computational challenges. These challenges are often specific to the type of data (e.g. structured/relational, text, spatial, time series, social/graph, multimedia, RDF/web), the dimension of data quality (e.g. completeness, consistency, timeliness), and the preparatory processes (e.g. data acquisition, profiling, curation, integration) that precede the actual use of the data. Designing a practical strategy for tackling quality issues in big data requires data scientists to bring together these multiple aspects of data type, quality dimension and process within the context of their application setting.

In this special issue we have endeavoured to present recent research of some of the leading experts in the field of data quality with the aim of informing the design of such practical strategies. Out of the eight papers, four are on relational/structured data while the remaining four are on time series data, spatio-temporal data, micro-blog data and web data. The papers have targeted a number of data quality dimensions through a range of innovative approaches as outlined below.

The first two papers tackle data quality dimensions of **meta-data compliance** and **schema quality**. Sebastian Kruse, Thorsten Papenbrock, Hazar Harmouch, and Felix Naumann present data anamnesis as a means of meta-data discovery with an aim to assess the quality and utility of the underlying relational datasets. In the second paper, Henning Kohler, Sebastian Link and Xiaofang Zhou present a method for discovering meaningful certain keys, in the presence of **incomplete** and **inconsistent** data with an aim to tackle **redundancy** and maintain the integrity constraints of the underlying relational data.

The next two papers discuss data cleaning in the context of associated data transformation and curation activities. These works are instrumental in evaluating the effectiveness of data cleaning algorithms. A number of data quality dimensions are covered by these papers including **value**, **format and semantic consistency**, and **business rule compliance**. The paper by Ihab Ilyas proposes a decoupling between detecting data errors and the repairing of these errors within a continuous data cleaning life-cycle with humans in the loop. The paper by Patricia C. Arocena, Boris Glavic, Giansalvatore Mecca, Renee J. Miller, Paolo Papotti and Donatello Santoro, outlines the challenges and solutions for benchmarking data curation systems.

Spatio-temporal and time-series data are known to suffer from a variety of data quality problems, due to the correlated nature of the data. The paper by Juliana Freire, Aline Bessa, Fernando Chirigati, Huy Vo and Kai Zhao uses exploratory techniques and powerful visualizations to differentiate between error and feature in spatio-temporal data. Tamraparni Dasu, Rong Duan, and Divesh Srivastava in their paper on data quality for temporal streams use statistical distortion as a means of measuring data quality in a near-real time fashion. The papers address a range of data quality dimensions including **incompleteness**, **redundancy**, **inaccuracy** (e.g. GPS noise), **format consistency**, **dependency constraint violation**, **uniqueness** issues and handling **duplicates**.

The final two papers target the elusive data quality dimensions of **trustworthiness** and **understandability**. The paper by Wen Hua, Kai Zheng and Xiaofang Zhou focuses on improving the understandability of short-text as found in microblogs towards resolving entity ambiguity. The paper by Xin Luna Dong, Evgeniy Gabrilovich,

Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang presents a knowledge-based approach to estimating the trustworthiness of web sources.

Given the contextual nature of data quality research and the need to have reproducible results, we have compiled a list of the datasets used by the papers in the special issue. The list and links are available below.

We would like to thank all the authors for their insightful contributions and for also playing the dual role of reviewers of other papers, resulting in a synergized special issue. We hope you enjoy reading the papers as much as we enjoyed putting them together.

Link to datasets used in the papers

Data Anamnesis: Admitting Raw Data into an Organization
Datasets: MusicBrainz
Public Links: `https://musicbrainz.org`
Discovering Meaningful Certain Keys from Incomplete and Inconsistent Relations
Datasets: GO-termdb (Gene Ontology)
Public Links: `www.geneontology.org`
Datasets: IPI (International Protein Index)
Public Links: `www.ebi.ac.uk/IPI`
Datasets: LMRP (Local Medical Review Policy)
Public Links: `www.cms.gov/medicare-coverage-database`
Datasets: Naumann (benchmarks for FD mining)
Public Links: `https://hpi.de/naumann/projects/repeatability/data-profiling/fd.html`
Datasets: PFAM (protein families)
Public Links: `pfam.sanger.ac.uk`
Datasets: RFAM (RNA families)
Public Links: `rfam.sanger.ac.uk`
Datasets: UCI (Machine Learning Repository)
Public Links: `https://archive.ics.uci.edu/ml/datasets.html`
Benchmarking Data Curation Systems
Datasets: Real and synthetic dirty datasets from BART project
Public Links: `http://www.db.unibas.it/projects/bart`
Datasets: Real and synthetic integration scenarios from iBench project
Public Links: `http://dblab.cs.toronto.edu/project/iBench`
Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips
Datasets: TaxiVis
Public Links: `https://github.com/ViDA-NYU/TaxiVis`
Datasets: TLC Trip Record Data
Public Links: `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`
Quality-Aware Entity-Level Semantic Representations for Short Texts
Datasets: Wikipedia
Public Links: `https://dumps.wikimedia.org/enwiki`
Datasets: Probase
Public Links: `http://probase.msra.cn/dataset.aspx`

Shazia Sadiq, Divesh Srivastava
The University of Queensland (Sadiq), AT&T Labs-Research (Srivastava)