Workshop Big (and Small) Data in Science and Humanities (BigDS 2019)

Friederike Klan[1] Birgitta König-Ries[2] Peter Reimann[3] Bernhard Seeger[4] Anika Groß[5]

Over the last decade, we have witnessed a still ongoing digital transformation of science, society and economy. Advances in data acquisition and the expansion of the internet to an ubiquitous medium led to the era of Big Data, which is characterized by the availability of a huge and ever increasing volume of complex, interlinked and heterogeneous data. Remote and ground-based sensors in earth observation for example produce petabytes of data with increasing spectral, temporal and spatial resolution. Social media users generate content at a high rate. Information and knowledge encoded in those data have an enormous value potential, that if revealed, could help to better understand the mechanisms underlying complex systems such as the human society or our earth, to generate innovations and to make well-founded decisions.

Thus, the importance of data has dramatically increased not just in economy but also in almost all scientific disciplines, e.g. in meteorology, genomics, complex physics simulations, biological and environmental research, and recently also in humanities. The unprecedented availability of data stimulates a rethinking in scientific disciplines on how to extract useful information and on how to foster research. At the same time researchers face severe challenges in leveraging data, since appropriate data management, integration, analysis and visualization tools have not been available so far. Recent advances in the development of big data technologies and the progress in machine learning, semantic technologies and other areas seem to be not only useful in business, but also offer great opportunities in science and humanities. Scientific workflows need to be realized as flexible end-to-end analytic solutions to allow for complex data processing, integration, analysis and visualization of Big Data in various application domains.

The need to discuss real-world problems in data science as well as recent advances in big data technology with database researchers and scientists from various disciplines led to the first and second edition of the workshop on Big (and Small) Data in Science and Humanities

[1] DLR Institut für Datenwissenschaften
[2] Friedrich-Schiller-Universität Jena
[3] Universität Stuttgart
[4] Philipps-Universität Marburg
[5] Hochschule Anhalt

doi:10.18420/btw2019-ws-10

(BigDS) at BTW 2015[6] and 2017[7]. This years third edition of the BigDS workshop[8] co-located with the 18th symposium of "Database systems for Business, Technology and Web" accommodates the still growing interest in methods to efficiently and effectively manage and analyze Big Data. With workshop contributions from various disciplines we hope to promote the dialog between domain experts and data scientists.

The workshop program included a keynote talk on digital humanities by *Andreas Henrich*, where he discussed current approaches, challenges and applications in the context of data integration, data federation and data analysis for humanities. We further selected six contributions that address different challenges in the context of data-driven analytics. The papers contribute to the management and analysis of data from various domains, such as mobile data, automobile data, textual data like legal texts and bibliographic data as well as ecological data. The proposed approaches are related to the analysis and use of complex graphs and ontologies, item set mining and entity extraction as well as evaluation and quality criteria.

Two papers focus on methods and models in the context of data analytics. *Rost et al.* present an extension of the graph data management tool Gradoop to support temporal graph analytics. They added time properties to vertices, edges and graphs and used them within graph operators, e.g. to analyze temporal citation patterns as presented in a bibliographic usage scenario. *Spieß and Reimann* analyzed the regulation and control of vehicle components in automotive series production. They developed an adapted item set mining approach in order to successfully perform association analysis for the domain-specific problem of automatically identifying vehicles with high risk of failure.

Three papers deal with the analysis and extraction of information from textual data. *Cornelia Kiefer* presents and discusses quality indicators for textual data. Beside the quality of texts themselves, her aim is to predict the quality of text analysis results and to decide whether default text mining modules are likely to deal with the textual data or not. In her evaluation she investigates texts, e.g. from production, news and tweets using the proposed quality indicators. The goal of *Wehnert et al.* is to provide a decision support system for legal regulations, e.g. to inform companies about relevant regulatory changes that need to be considered. In this work, they use linked laws from their ontology of legal textbooks and developed a context selection mechanism to help users navigating in their legal knowledge base, e.g. to find all applications of a law. *Udovenko et al.* present a hybrid approach to extract entities from scientific publications in the ecological domain. They propose a framework including the use of domain-related ontologies for entity annotation, and run an initial evaluation for entity extraction from publications on biodiversity.

Finally, *Steinberg et al.* present a comparative evaluation for different software solutions that support the form-based collection of mobile data. Nowadays, mobile devices heavily

[6] http://www.btw-2015.de/?dms

[7] http://btw2017.informatik.uni-stuttgart.de/?pageId=BigDS

[8] https://btw.informatik.uni-rostock.de/index.php/de/call-for-workshops/bigds

support the data collection process, and users often build on existing infrastructure and software to collect and submit data. The paper reports on experiences with respect to the whole data collection workflow and compares eight existing tools in terms of their features and characteristics.

All contributions of this year's BigDS workshop give new domain-relevant insights and promote the use of generic as well as domain-specific methods for scientific data management and analytics. We would like to thank everyone who contributed to the workshop, in particular, the authors, the keynote speaker Andreas Henrich, the BigDS program committee, the BTW team, as well as all participants.

Workshop Organizers

Anika Groß (Hochschule Anhalt, DE)
Friederike Klan (DLR-Institut für Datenwissenschaften, DE)
Birgitta König-Ries (Friedrich-Schiller-Universität Jena, DE)
Peter Reimann (Universität Stuttgart, DE)
Bernhard Seeger (Philipps-Universität Marburg, DE)

Program Committee

Alsayed Algergawy (Friedrich-Schiller-Universität Jena, DE)
Peter Baumann (Universität Bremen, DE)
Matthias Bräger (CERN, CH)
Thomas Brinkhoff (Jade Hochschule, DE)
Jana Diesner (University of Illinois at Urbana-Champaign, US)
Johann-Christoph Freytag (Humboldt-Universität zu Berlin, DE)
Michael Gertz (Universität Heidelberg, DE)
Thomas Heinis (Imperial College London, UK)
Andreas Henrich (Otto-Friedrich-Universität Bamberg, DE)
Alfons Kemper (Technische Universität München, DE)
Jens Nieschulze (Georg-August-Universität Göttingen, DE)
Eric Peukert (Universität Leipzig, DE)
Norbert Ritter (Universität Hamburg, DE)
Kai-Uwe Sattler (Technische Universität Ilmenau, DE)
Holger Schwarz (Universität Stuttgart, DE)
Uta Störl (Hochschule Darmstadt, DE)
Andreas Thor (Hochschule für Telekommunikation Leipzig, DE)