B. Mitschang et al. (Hrsg.): BTW 2017 – Workshopband,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 75

Big (and Small) Data in Science and Humanities

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

The importance of data has dramatically increased in almost all scientific disciplines over the last decade, e. g., in meteorology, genomics, complex physics simulations, biological and environmental research, and recently also in humanities. This development is due to great advances in data acquisition and data accessibility, such as improvements in remote sensing, powerful mobile devices, popularity of social networks and the ability to handle unstructured data (including texts). On the one hand, the availability of such data masses leads to a rethinking in scientific disciplines on how to extract useful information and on how to foster research. On the other hand, researchers feel lost in the data masses because appropriate data management, integration, analysis and visualization tools have not been available so far. However, this is starting to change with the recent development of big data technologies and with progress in natural language processing, semantic technologies and others that seem to be not only useful in business, but also offer great opportunities in science and humanities.

A typical scientific data analytics pipeline covers several crucial phases to allow for complex data processing, integration, analysis and visualization. First, relevant data need to be acquired and integrated from different, often heterogeneous data sources. This includes various forms of data such as measurements, textual documents, images, graph data or spatio-temporal data. Depending on the size and structure of data, suitable data storage solutions have to be selected, e. g., from the wide spectrum of NoSQL data stores. Before pushing data to the actual analytics phase, it is crucial to detect and eliminate data quality problems by applying a careful data cleaning and enrichment process. Data analysis typically comprises data aggregation and filtering as well as sophisticated analysis methods. By now, we have access to a huge set of big data frameworks (e. g., Apache Spark, Apache Flink, Apache Storm) that support batch and/or stream processing and already provide many algorithms for data mining, machine learning or graph analytics. However, generic systems do not always offer good solutions for specialized problems. Hence, individual domain-specific analysis methods might be adopted in the pipeline as well. Finally, experts need to interpret results and incorporate novel insights into their real-world applications. Since it is infeasible to review very large result data sets, a clear representation and visualization is indispensable. Moreover, interactive user interfaces are necessary, e. g., to give extensive user support or to incorporate user's feedback and know-how during the data analysis process.

[1] Universität Leipzig
[2] Friedrich-Schiller-Universität Jena
[3] Universität Stuttgart
[4] Philipps-Universität Marburg

The analysis of scientific data faces several challenges. One main challenge is the generation of *high quality* results. Big data analytics need to deal with large heterogeneous amounts of data impeding quality assurance during data processing and analysis. Usually, it is very challenging to assess the achieved quality since gold standard data sets and benchmarks are rare and often not fully representative for a considered real-world problem. Another important aspect is the *scalability* of the applied methods to very large and possibly growing or changing data sets. Data is often not static, but underlies *evolution* or covers inherent temporal aspects, e. g., when data is acquired periodically to enable temporal analysis. Further challenges arise when privacy requirements must be respected as it is often the case for personal data, e. g., in the medical domain. Another important challenge and aim is the realization of scientific data analysis workflows as flexible *end-to-end solutions*. In many domains, data analysis still faces the problem of interrupted, bit-by-bit data processing, although many manual steps could (easily) be automated. A manual execution of data analysis tasks is time-consuming and prone to human error. It is therefore crucial to build automated solutions that involve human interactions if and only if this is useful. It is further important to incorporate data *provenance* aspects and traceability of the adopted processes within scientific workflows to enable the *reproducibility* of analysis results. Finally, a key challenge is the *visualization* of intermediate and final results. Visual analytics can offer valuable new insights when humans reveal patterns and trends based on meaningful data visualization.

This workshop intends to bring together scientists from various disciplines with database researchers to discuss real-world problems in data science as well as recent big data technology. We selected seven contributions that address several challenges of data-driven analysis. The contributions cover scientific data from various domains, such as geographical, environmental and agricultural science, biodiversity, archeology, humanities and business. The proposed approaches address topics that are related to the analysis of temporal and spatial data, decision support, interactive data exploration, deep-learning, data quality, text analysis, scientific workflows as well as the scalability of the adopted methods.

Three papers present data analysis approaches to support users in their decisions within different domains. *Amara et al.* use a deep-learning-based approach to classify banana leaf diseases based on image data sets. The authors show the effectiveness of their approach under varying conditions, e. g., regarding illumination and resolution size. The proposed method mainly enables early disease recognition and decision support for farmers that need to identify banana plant diseases. *Elmamoozet et al.* follow a graph-based approach to model continuous mobility in museum exhibitions. The authors discuss challenges to cope with a huge amount of sensor data, graph dynamics, as well as heterogeneous user groups within the environment of a museum. The system aims to support curators in their guiding task as well as museum visitors. *Beilschmidt et al.* present an intuitive web-based user interface for interactive exploration in geosciences. The system deals with spatio-temporal data from biodiversity research. The approach addresses scientific users by supporting decisions based on visual analytics for effective data-driven science on large and heterogeneous data.

The paper of *Kaltenthaler et al.* presents a framework for supporting the scientific workflow for managing and analysing archeo-related species. In this domain, working online is usually

not possible since there is no reliable internet connection, e. g., in the field of an excavation. This makes synchronization of local and global data essential during the data gathering, sharing and analysis process. The provided tools are applied within German zooarchaeology projects.

Cornelia Kiefer analyzes and discusses important challenges of domain-specific text analysis. Domain-specific texts can show very different and varying characteristics compared to standard texts like newspaper articles. In her work, she hence reveals serious data quality problems when applying standard text analytics tools to domain-specific texts during all phases of the analytics workflow.

Two further papers address scalability and efficiency aspects in the context of data analysis and data generation. *Kemper et al.* present a scalable solution of a data generator in the context of graph-based business analytics. Data generation and simulation are necessary when no or little representative data is available within a domain. In order to scale for the generation of huge business process graphs, an existing system has been extended to allow for distributed execution based on Apache Flink and Gradoop. Within their experiments, the authors show the scalability of the system to large data sets. Furthermore, *Pascal Hirmer* presents a concept to optimize the efficiency of data-intensive Mashups. The optimized execution is currently based on Map Reduce. The approach includes policy annotations to decide which services need to be executed in particular steps of a data mashup pipeline. The accompanying Mashup tool supports domain experts in their explorative analysis tasks.

We are deeply grateful to everyone who contributed to this workshop, in particular, the authors, the reviewers, the BTW team, and all participants.

1 Workshop Organizers

Anika Groß (Universität Leipzig, DE)
Birgitta König-Ries (Friedrich-Schiller-Universität Jena, DE)
Peter Reimann (Universität Stuttgart, DE)
Bernhard Seeger (Philipps-Universität Marburg, DE)

2 Program Committee

Alsayed Algergawy (Uni Jena, DE)
Peter Baumann (Jacobs Universität, DE)
Matthias Bräger (CERN, CH)
Thomas Brinkhoff (FH Oldenburg, DE)
Michael Diepenbroeck (Alfred-Wegner-Institut, Bremerhaven, DE)
Jana Diesner (University of Illinois at Urbana-Champaign, US)
Christoph Freytag (Humboldt Universität Berlin, DE)
Michael Gertz (Uni Heidelberg, DE)
Anton Güntsch (Botanischer Garten und Botanisches Museum, Berlin-Dahlem, DE)

Thomas Heinis (Imperial College, London, UK)
Andreas Henrich (Universität Bamberg, DE)
Jens Kattge (Max-Planck-Institut für Biogeochemie, DE)
Alfons Kemper (TU München, DE)
Meike Klettke (Universität Rostock, DE)
Frank Leymann (Universität Stuttgart, DE)
Bertram Ludäscher (University of Illinois at Urbana-Champaign, US)
Alex Markowetz (Uni Bonn, DE)
Jens Nieschulze (Uni Göttingen, DE)
Eric Peukert (Universität Leipzig, DE)
Kai-Uwe Sattler (TU Ilmenau, DE)
Uta Störl (Hochschule Darmstadt, DE)
Andreas Thor (Hochschule für Telekommunikation Leipzig, DE)