Foreword

In the early years of my PhD research, we were excited to use a little more than 1,000 medical records for data analysis and knowledge discovery. The records were “hand-crafted” and “valuable-as-gold”. Those days, some of the advanced research methods needed an overnight run, but we had no issue in storing and accessing the data in the standard file system. This changed a lot in recent years. Today, machines and users generate data volumes of an unimagined scale. Typical examples are user-generated content in the (social) web, sensor data or operating data from machines. This development is massively supported by the increasing digitalization of our environment. Thereby, data is produced not only by digitalized processes but also by capturing many of our other daily private- and work-life situations. These different kinds of data sources determine that big data typically comes rich in variety and sometimes high velocity. In contrast to machine-generated data, the user-generated data typically lacks in quality and transparency. For industrial use cases, the collected data represents the “new oil” of our times since many value-added uses are promised from the utilization of the data. However, new challenges need to be tackled to reveal the data’s business potential.

Big data methods are already installed in end-user environments: Search engines and online shops, for example, use the collected mass data of users to predict the interestingness of links or shopping goods. In contrast, the digitalization of the industrial enterprise and the application of big data technologies are just in their beginnings. Current state-of-the-art machinery produces data streams continuously for monitoring systems. The collected data streams can be used to predict maintenance opportunities and to diagnose possible trouble spots. The effective use of this data demands for novel methods that are capable to handle large volumes of data possibly streamed, in different structures, and in different qualities.

The editors of the book, Martin Atzmueller, Samia Oussena, and Thomas Roth-Berghofer, uncover the relevant topics in big data for the engineering, analytics and the management of big data with a special focus on the enterprise. The book offers a selection of current research trends and methods.

Two chapters address the task of preprocessing big data: Stange and Funk discuss the question of the appropriate sample size in the chapter “How big does Big Data need to be” and Schmidt et al. present use cases and experiences with preprocessing industrial big data in their chapter “Data Preparation for Big Data Analytics”. A perspective on strategic and organizational issues in the context of Big Data is discussed by Roberts and Laurie in the chapter “Strategic management of data and challenges for organizations, strategy development and business value”.

The velocity of big data is determined by data streams. The related area of complex event processing is discussed by two contributions: A review of complex event processing methods is given by Mohamedali and Oussena in their chapter “Application of Complex Event Processing Techniques to Big Data Related
Foreword

to Healthcare. A Systematic Literature Review of Case Studies”. Wrench et al. discuss the relations of complex event processing and the closely related field of event stream processing in their chapter “Data Stream Mining of Event and Complex Event Streams”.

How to get the promised value from big data? Appropriate analytical methods help to uncover interesting patterns and correlations from the data. Kalaian et al. introduce analytics in their chapter “Descriptive and Predictive Analytical Methods for Big Data” by describing descriptive and predictive analytical methods. Chokkalingam and Vijayarani motivate research directions for analytics in their chapter “Research Challenges in Big Data Analytics”. The topics are motivated by practical examples. A dedicated analytics method is presented by Atzmuller et al. in their chapter “Big Data Analytics Using Local Exceptionality Detection”.

From the beginning of the research, the term “big data” was motivated by practical challenges and experiences. The book showcases a number of interesting use cases. The authors Sokolowski and Oussena review the applicability of big data methods in collaborative learning environments in their chapter “Using Big Data in collaborative learning”. The use of big data in an industrial application is reported by Klöpper et al. in “Data Modeling and Knowledge Discovery in Process Industries”. The application of big data methods on large technical text corpora is discussed by Furth and Baumeister in “Semantification of Large Corpora of Technical Documentation”. Interesting patterns can be extracted from large text corpora by summarization techniques, as shown by Meena and Gopalani in their chapter “Statistical Features for Extractive Automatic Text Summarization”. The analysis of big data collected from social networks is systematically introduced by Viol et al. in their chapter “A Framework to Analyze Enterprise Social Network Data”.

In summary, the book provides an interesting and current compilation of theoretical research questions, practical recommendations as well as industrial use cases in the emerging field of big data, and researchers and practitioners will benefit from the included contributions.

Joachim Baumeister
denkbare GmbH, Germany