A Review of *Handbook of Statistics 33: Big Data Analytics*

Steven Noel

Founded in the early 1980s by P. R. Krishnaiah, the *Handbook of Statistics* is a series of statistics reference books, with each volume devoted to a topical focus area. Given that statistics has branched into so many subjects, and continues to branch at an accelerated rate, this series of handbooks provides a venue for comprehensive treatment of recent developments in topical areas. The series is written for practicing statisticians and other researchers who need to employ such statistical methodology. As such, it has special orientation towards application rather than pure theory, and assumes the reader has a good grasp of statistics fundamentals.

The installments of the *Handbook of Statistics* series are edited volumes, with chapters being contributed by prominent workers in the volume's focus area. The series editor is the eminent statistician C. R. Rao. Along with professor Rao, *Volume 33: Big Data Analytics* is edited by Venu Govindaraju (State University of New York) and Vijay V. Raghavan (University of Louisiana at Lafayette). This volume focuses on Big Data challenges in a variety of domains, and how innovative statistics-based approaches can accelerate discovery, help glean insights, and support more intelligent decisions. This volume has 14 chapters, with 38 contributing authors. The volume's overall content is divided into Modeling and Analytics (7 chapters) and Applications and Infrastructure (7 chapters).

The first 3 chapters are written by researchers from the State University of New York (SUNY) at Buffalo. In the first chapter, Govindaraju et al propose methods for accelerated discovery and meta-learning (learning how to best learn) from large scientific document corpuses, through probabilistic topic models, clustering, information retrieval, and visualization. In Chapter 2, Biondini gives an in-depth introduction to variance reduction techniques for simulating rare events though Monte Carlo methods, in which very small probabilities need to be accurately estimated. Chapter 3 (by Pokhriyal et al) is a large-scale experimental study that explores whether a person's cognitive state can be used as a biometric trait for authentication in pervasive computing.

In the fourth chapter, Kwac and Rajagopal of Stanford cast the targeting of business customers as a demand-reduction problem from electrical power management, which they solve by applying data mining techniques for big data. In Chapter 5, Chan and Treleaven (University College London) improve prediction performance of large-scale retail recommender systems through an approach for continuous model selection that adapts to rapidly changing business dynamics. Chapter 6 by Shoaran et al (University of Tabriz and University of Victoria) describes "zero-knowledge privacy" for the protection of personal information when releasing summarizations of graph and social network databases. In Chapter 7, Djuric et al (Yahoo Labs and Temple University) describe a new linear solver for support vector machines, for efficient training of confidence-weighted classifiers on big data platforms.

Part B (Applications and Infrastructure) begins with Chapter 8 by Pyne et al of the C. R. Rao Advanced Institute of Mathematics, Statistics and Computer Science (at University of Hyderabad) and Virginia Tech. This chapter describes diffusion in network models for understanding the complex dynamics of epidemics in large-scale populations. Chapter 9 by Gudivada et al (East Carolina University, Marshall University, and University of Louisiana at Lafayette) examines how Big Data-enabled research supports a variety of tasks in natural language processing, e.g., statistical machine learning. In Chapter 10, Chandola et al (SUNY at Buffalo, North Carolina State University, and Northeastern University) explore big data requirements for four popular algorithms for spatial and spatiotemporal data mining – spatial autoregressive models, Markov random field classifiers, Gaussian process learning models, and mixture models.

Chapter 11 by Galas (University College London) describes a big data environment and multi-agent simulation framework for research in finance and economics, e.g., for modeling trading agents and market exchanges. Chapter 12 by Moise and Shestakov (Cray and Bright Computing) investigates a variety of mechanisms for improving Hadoop performance over terabytes of data, using image similarity search as a case study. In Chapter 13, Lehmann et al (Universitat Pompeu Fabra and Yahoo Labs) study inter-site user engagement in networks of web sites operated by large online service providers. Finally, Chapter 14 by Castellana et al (Pacific Northwest National Laboratory, Context Relevant, and NVIDIA Research) describes a novel software stack that implements graph databases on top of commodity, high-performance clusters, e.g., converting graph queries into parallel graph-matching algorithms.

Overall, this volume addresses a variety of challenges regarding Big Data, including capture, storage, search, analysis, sharing, visualization, and privacy. Data volume, velocity, and variety are increasing at a tremendous rate, and traditional database and software techniques are no longer adequate. We need to be able to derive insights and create new knowledge from massive quantities of data from diverse sources, e.g., ubiquitous sensors, smart devices, social networks, and the internet of things. This volume describes a number of emergent advances in statistics, machine learning, data mining, and distributed systems to analyze and reason with this data. This in turn has the potential to greatly improve enterprise operations and enable faster, more intelligent decision making.