Book review of 'Handbook of Big Data', edited by Peter Bühlmann, Petros Drineas, Michael Kane and Mark van der Laan

Richard J. Samworth
University of Cambridge

June 23, 2016

Fundamental research in methodology and theory for Big Data is currently progressing at such a rate that it can be extremely challenging for practitioners to keep up with the latest developments. It is crucial, therefore, that there is a regular process of synthesis, where the most important developments are made accessible to the widest audience possible. This edited book is conceived in this spirit, and it should be welcomed as a timely contribution.

Another feature of the modern data science landscape is that the breadth of the discipline has made research in this area increasingly collaborative; no single individual can be an expert in each of its constituent subfields. With this in mind, it is highly appropriate that the authors of the 24 chapters, like the editors, are drawn from a wide range of backgrounds. The chapters are loosely grouped under eight headings, beginning with 'General perspectives on Big Data'. Immediately the reader is drawn in by Richard Starmans' entertaining historical and philosophical reflections on the 'Unreasonable Effectiveness of Data'. Norman Matloff, a professor of computer science, continues the broad theme by emphasising that the need for Statistics is strengthened, not diminished, by the advent of Big Data. Subsequent headings include 'Data-centric, exploratory methods', 'Efficient algorithms', 'Graph approaches', 'Model fitting and regularization', 'Ensemble methods' and 'Causal inference'. There are many outstanding researchers among the contributors, and the overall quality of the articles is very high.

The editors state three very ambitious goals for their work: first, to identify modern, scalable approaches for Big Data; second, to identify areas with a pressing need for further

development; and third, to integrate current techniques across different disciplines, and facilitate greater inter-disciplinary communication and collaboration. To a large extent, I believe they have nevertheless succeeded in their goals. Some of the core ideas identified are divide-and-recombine (often called divide-and-conquer) methods, random projections, hashing, low-rank approximation, network structures, cross-validation, stochastic gradient descent, penalised estimation techniques, sample splitting and structural equation models. It is notoriously difficult to try to identify concrete grand challenges in this area, but several of the authors make efforts to outline open problems within their respective areas. Regarding the third aim, it seems to me that data scientists still have a key role in identifying the commonalities between the problems faced by practitioners in different fields. An excellent modern example of this is data scientists' realisation of the importance of the notion of sparsity for high-dimensional data.

The book contains a nice mix of philosophical musings, survey articles and cutting-edge research. It was designed as 'a useful resource for seasoned practitioners and enthusiastic neophytes alike'; some articles require more background than others, but each is self-contained, so enthusiastic neophytes are still left with plenty to get their teeth into. In summary, I am happy to recommend the book to those seeking to broaden their understanding of the underpinning methodologies for analysing Big Data.