Preface

The increased capacity of contemporary computers allows the gathering, storage and analysis of large amounts of data which only a few years ago would have been impossible. These new data are providing large quantities of information, and enabling its interconnection using new computing methods and databases. There are many issues arising from the emergence of big data, from computational capacity to data manipulation techniques, all of which present challenging opportunities. Researchers and industries working in various different fields are dedicating efforts to resolve these issues. At the same time, scholars are excited by the scientific possibilities offered by big data, and especially the opportunities to investigate major societal problems related to health, privacy, economics, business dynamics and many more. These large amounts of data present various challenges, one of the most intriguing of which deals with knowledge discovery and large-scale data mining. Although these vast amounts of digital data are extremely informative, and their enormous possibilities have been highlighted on several occasions, issues related to optimization remain to be addressed. For example, formulation of optimization problems of unprecedented sizes (millions or billions of variables) is inevitable.

The main objective of this book is to provide the necessary background to work with big data by introducing some novel optimization algorithms and codes capable of working in the big data setting as well as introducing some applications in big data optimization for interested academics and practitioners, and to benefit society, industry, academia and government.

To facilitate this goal, chapter "Big Data: Who, What and Where? Social, Cognitive and Journals Map of Big Data Publications with Focus on Optimization" provides a literature review and summary of the current research in big data and large-scale optimization. In this chapter, **Emrouznejad and Marra** investigate research areas that are the most influenced by big data availability, and on which aspects of large data handling different scientific communities are working. They employ scientometric mapping techniques to identify who works on what in the area of big data and large-scale optimization problems. This chapter highlights a major effort involved in handling big data optimization and large-scale data mining

which has led to several algorithms that have proven to be more efficient, faster and more accurate than earlier solutions.

This is followed by a comprehensive discussion on setting up a big data project in chapter "Setting Up a Big Data Project: Challenges, Opportunities, Technologies and Optimization" as discussed by **Zicari, Rosselli, Ivanov, Korfiatis, Tolle, Niemann and Reichenbach**. The chapter explains the general value of big data analytics for the enterprise and how value can be derived by analysing big data. Then it introduces the characteristics of big data projects and how such projects can be set up, optimized and managed. To be able to choose the optimal big data tools for given requirements, the relevant technologies for handling big data, such as NoSQL and NewSQL systems, in-memory databases, analytical platforms and Hadoop-based solutions, are also outlined in this chapter.

In chapter "Optimizing Intelligent Reduction Techniques for Big Data", **Pop, Negru, Ciolofan, Mocanu, and Cristea** analyse existing techniques for data reduction, at scale to facilitate big data processing optimization. The chapter covers various areas in big data including: data manipulation, analytics and big data reduction techniques considering descriptive analytics, predictive analytics and prescriptive analytics. Cyber-Water cast study is also presented by referring to: optimization process, monitoring, analysis and control of natural resources, especially water resources to preserve the water quality.

Li, Guo and Chen in the chapter "Performance Tools for Big Data Optimization" focus on performance tools for big data optimization. The chapter explains that many big data optimizations have critical performance requirements (e.g., real-time big data analytics), as indicated by the velocity dimension of 4Vs of big data. To accelerate the big data optimization, users typically rely on detailed performance analysis to identify potential performance bottlenecks. To alleviate the challenges of performance analysis, various performance tools have been proposed to understand the runtime behaviours of big data optimization for performance tuning.

Further to this, **Valkonen**, in chapter "Optimising Big Images", presents a very good application of big data optimization that is used for analysing big images. Real-life photographs and other images, such as those from medical imaging modalities, consist of tens of million data points. Mathematically based models for their improvement—due to noise, camera shake, physical and technical limitations, etc.—are moreover often highly non-smooth and increasingly often non-convex. This creates significant optimization challenges for application of the models in quasi-real-time software packages, as opposed to more ad hoc approaches whose reliability is not as easily proven as that of mathematically based variational models. After introducing a general framework for mathematical image processing, this chapter presents the current state-of-the-art in optimization methods for solving such problems, and discuss future possibilities and challenges.

As another novel application **Rajabi and Beheshti**, in chapter "Interlinking Big Data to Web of Data", explain interlinking big data to web of data. The big data problem can be seen as a massive number of data islands, ranging from personal, shared, social to business data. The data in these islands are becoming large-scale, never ending and ever changing, arriving in batches at irregular time intervals. In

this context, it is important to investigate how the linked data approach can enable big data optimization. In particular, the linked data approach has recently facilitated accessibility, sharing and enrichment of data on the web. This chapter discusses the advantages of applying the linked data approach, toward optimization of big data in the linked open data (LOD) cloud by: (i) describing the impact of linking big data to LOD cloud; (ii) representing various interlinking tools for linking big data; and (iii) providing a practical case study: linking a big data repository to DBpedia.

Topology of big data is the subject of chapter "Topology, Big Data and Optimization" as discussed by **Vejdemo-Johansson and Skraba**. The idea of using geometry in learning and inference has a long history going back to canonical ideas such as Fisher information, discriminant analysis and principal component analysis. The related area of topological data analysis (TDA) has been developing in the past decade, which aims to extract robust topological features from data and use these summaries for modelling the data. A topological summary generates a coordinate-free, deformation invariant and a highly compressed description of the geometry of an arbitrary data set. This chapter explains how the topological techniques are well suited to extend our understanding of big data.

In chapter "Applications of Big Data Analytics Tools for Data Management", **Jamshidi, Tannahill, Ezell, Yetis and Kaplan** present some applications of big data analytics tools for data management. Our interconnected world of today and the advent of cyber-physical or system of systems (SoS) are a key source of data accumulation—be it numerical, image, text or texture, etc. SoS is basically defined as an integration of independently operating, non-homogeneous systems for a certain duration to achieve a higher goal than the sum of the parts. Recent efforts have developed a promising approach, called "data analytics", which uses statistical and computational intelligence (CI) tools such as principal component analysis (PCA), clustering, fuzzy logic, neuro-computing, evolutionary computation, Bayesian networks, data mining, pattern recognition, etc., to reduce the size of "big data" to a manageable size. This chapter illustrates several case studies and attempts to construct a bridge between SoS and data analytics to develop reliable models for such systems.

Optimizing access policies for big data repositories is the subject discussed by **Contreras** in chapter "Optimizing Access Policies for Big Data Repositories: Latency Variables and the Genome Commons". The design of access policies for large aggregations of scientific data has become increasingly important in today's data-rich research environment. Planners routinely consider and weigh different policy variables when deciding how and when to release data to the public. This chapter proposes a methodology in which the timing of data release can be used to balance policy variables and thereby optimize data release policies. The global aggregation of publicly-available genomic data, or the "genome commons" is used as an illustration of this methodology.

Achieving the full transformative potential of big data in this increasingly digital and interconnected world requires both new data analysis algorithms and a new class of systems to handle the dramatic data growth, the demand to integrate structured and unstructured data analytics, and the increasing computing needs of

massive-scale analytics. **Li**, in chapter "Big Data Optimization via Next Generation Data Center Architecture", elaborates big data optimization via next-generation data centre architecture. This chapter discusses the hardware and software features of High Throughput Computing Data Centre architecture (HTC-DC) for big data optimization with a case study at Huawei.

In the same area, big data optimization techniques can enable designers and engineers to realize large-scale monitoring systems in real life, by allowing these systems to comply with real-world constrains in the area of performance, reliability and reliability. In chapter "Big Data Optimization Within Real World Monitoring Constraints", **Helmholt and der Waaij** give details of big data optimization using several examples of real-world monitoring systems.

Handling big data poses a huge challenge in the computer science community. Some of the most appealing research domains such as machine learning, computational biology and social networks are now overwhelmed with large-scale databases that need computationally demanding manipulation. Smart sampling and optimal dimensionality reduction of big data using compressed sensing is the main subject in chapter "Smart Sampling and Optimal Dimensionality Reduction of Big Data Using Compressed Sensing" as elaborated by **Maronidis, Chatzilari, Nikolopoulos and Kompatsiaris**. This chapter proposes several techniques for optimizing big data processing including computational efficient implementations like parallel and distributed architectures. Although Compressed Sensing (CS) is renowned for its capability of providing succinct representations of the data, this chapter investigates its potential as a dimensionality reduction technique in the domain of image annotation.

Another novel application of big data optimization in brain disorder rehabilitation is presented by **Brezany, Štěpánková, Janatoá, Uller and Lenart** in chapter "Optimized Management of BIG Data Produced in Brain Disorder Rehabilitation". This chapter introduces the concept of scientific dataspace that involves and stores numerous and often complex types of data, e.g. primary data captured from the application, data derived by curation and analytics processes, background data including ontology and workflow specifications, semantic relationships between dataspace items based on ontologies, and available published data. The main contribution in this chapter is applying big data and cloud technologies to ensure efficient exploitation of this dataspace, namely novel software architectures, algorithms and methodology for its optimized management and utilization.

This is followed by another application of big data optimization in maritime logistics presented by **Berit Dangaard Brouer, Christian Vad Karsten and David Pisinge** in chapter "Big data Optimization in Maritime Logistics". Large-scale maritime problems are found particularly within liner shipping due to the vast size of the network that global carriers operate. This chapter introduces a selection of large-scale planning problems within the liner shipping industry. It is also shown how large-scale optimization methods can utilize special problem structures such as separable/independent sub-problems and give examples of advanced heuristics using divide-and-conquer paradigms, decomposition and mathematical programming within a large-scale search framework.

On more complex use of big data optimization, chapter "Big Network Analytics Based on Nonconvex Optimization" focuses on the use of network analytics which can contribute to networked big data processing. Many network issues can be modelled as non-convex optimization problems and consequently they can be addressed by optimization techniques. **Gong, Cai, Ma and Jiao**, in this chapter, discuss the big network analytics based on non-convex optimization. In the pipeline of non-convex optimization techniques, evolutionary computation gives an outlet to handle these problems efficiently. Since network community discovery is a critical research agenda of network analytics, this chapter focuses on the evolutionary computation-based non-convex optimization for network community discovery. Several experimental studies are shown to demonstrate the effectiveness of optimization-based approach for big network community analytics.

Large-scale and big data optimization based on Hadoop is the subject of chapter "Large-Scale and Big Optimization Based on Hadoop" presented by **Cao and Sun**. As explained in this chapter, integer linear programming (ILP) is among the most popular optimization techniques found in practical applications, however, it often faces computational issues in modelling real-world problems. Computation can easily outgrow the computing power of standalone computers as the size of problem increases. The modern distributed computing releases the computing power constraints by providing scalable computing resources to match application needs, which boosts large-scale optimization. This chapter presents a paradigm that leverages Hadoop, an open-source distributed computing framework, to solve a large-scale ILP problem that is abstracted from real-world air traffic flow management. The ILP involves millions of decision variables, which is intractable even with the existing state-of-the-art optimization software package.

Further theoretical development and computational approaches in large-scale unconstrained optimization is presented by **Babaie-Kafaki** in chapter "Computational Approaches in Large–Scale Unconstrained Optimization". As a topic of great significance in nonlinear analysis and mathematical programming, unconstrained optimization is widely and increasingly used in engineering, economics, management, industry and other areas. In many big data applications, solving an unconstrained optimization problem with thousands or millions of variables is indispensable. In such situations, methods with the important feature of low memory requirement are helpful tools. This chapter explores two families of methods for solving large-scale unconstrained optimization problems: conjugate gradient methods and limited-memory quasi-Newton methods, both of them are structured based on the line search.

This is followed by explaining numerical methods for large-scale non-smooth optimization (NSO) as discussed by **Karmitsa** in chapter "Numerical Methods for Large-Scale Nonsmooth Optimization". NSO refers to the general problem of minimizing (or maximizing) functions that are typically not differentiable at their minimizers (maximizers). NSO problems are in general difficult to solve even when the size of the problem is small and the problem is convex. This chapter recalls two numerical methods, the limited memory bundle algorithm (LMBM) and the

diagonal bundle method (D-BUNDLE), for solving large-scale non-convex NSO problems.

Chapter "Metaheuristics for Continuous Optimization of High-Dimensional Problems: State of the Art and Perspectives" presents a state-of-the-art discussion of metaheuristics for continuous optimization of high-dimensional problems. In this chapter, **Trunfio** shows that the age of big data brings new opportunities in many relevant fields, as well as new research challenges. Among the latter, there is the need for more effective and efficient optimization techniques, able to address problems with hundreds, thousands and even millions of continuous variables. In order to provide a picture of the state of the art in the field of high-dimensional continuous optimization, this chapter describes the most successful algorithms presented in the recent literature, also outlining relevant trends and identifying possible future research directions.

Finally, **Sagratella** discusses convergent parallel algorithms for big data optimization problems in chapter "Convergent Parallel Algorithms for Big Data Optimization Problems". When dealing with big data problems it is crucial to design methods able to decompose the original problem into smaller and more manageable pieces. Parallel methods lead to a solution by concurrently working on different pieces that are distributed among available agents, so as to exploit the computational power of multi-core processors and therefore efficiently solve the problem. Beyond gradient-type methods, which can of course be easily parallelized but suffer from practical drawbacks, recently a convergent decomposition framework for the parallel optimization of (possibly non-convex) big data problems was proposed. Such framework is very flexible and includes both fully parallel and fully sequential schemes, as well as virtually all possibilities in between. This chapter illustrates the versatility of this parallel decomposition framework by specializing it to different well-studied big data optimization problems such as LASSO, logistic regression and support vector machines training.

January 2016   Ali Emrouznejad