Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives

Zhi-Hua Zhou, Nitesh V. Chawla, Yaochu Jin, and Graham J. Williams

Abstract—"Big Data" as a term has been among the biggest trends of the last three years, leading to an upsurge of research, as well as industry and government applications. Data is deemed a powerful raw material that can impact multidisciplinary research endeavors as well as government and business performance. The goal of this discussion paper is to share the data analytics opinions and perspectives of the authors relating to the new opportunities and challenges brought forth by the big data movement. The authors bring together diverse perspectives, coming from different geographical locations with different core research expertise and different affiliations and work experiences. The aim of this paper is to evoke discussion rather than to provide a comprehensive survey of big data research.

Index Terms—Big data, data analytics, machine learning, data mining, global optimization, application

1 Introduction

BIG data is one of the "hottest" phrases being used today. Everyone is talking about big data, and it is believed that science, business, industry, government, society, etc. will undergo a thorough change with the influence of big data. Technically speaking, the process of handling big data encompasses collection, storage, transportation and exploitation. It is no doubt that the collection, storage and transportation stages are necessary precursors for the ultimate goal of exploitation through data analytics, which is the core of big data processing.

Turning to a data analytics perspective, we note that "big data" has come to be defined by the four V's — Volume, Velocity, Veracity, and Variety. It is assumed that either all or any one of them needs to be met for the classification of a problem as a Big Data problem. Volume indicates the size of the data, which might be too big to be handled by the current state of algorithms and/or systems. Velocity implies data are streaming at rates faster than that can be handled by traditional algorithms and systems. Sensors are rapidly reading and communicating streams of data. We are approaching the world of quantified self, which is presenting data that was not available hitherto. Veracity suggests that despite the data being available, the quality of data is still a major concern. That is, we cannot assume that with big data comes higher quality. In fact, with size comes quality issues, which needs to be either tackled at the data pre-processing stage or by the learning algorithm. Variety is the most compelling of all V's as it is presenting

- *Z.-H. Zhou is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (zhouzh@lamda.nju.edu.cn)*
- *N. V. Chawla is with the Department of Computer Science and Engineering and the Interdisciplinary Center for Network Science, University of Notre Dame, United States (nchawla@nd.edu)*
- *Y. Jin is with the Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom (yaochu.jin@surrey.ac.uk)*
- *G. J. Williams is with Togaware Pty Ltd and the Australian Tax Office, Camberra, Australia (Graham.Williams@togaware.com)*

data of different types and modalities for a given object in consideration.

Each of the V's is certainly not new. Machine learning and data mining researchers have been tackling these issues for decades. However, the emergence of Internet-based companies has challenged many of the traditional process-oriented companies—they now need to become knowledge-based companies driven by data rather than by process.

The goal of this article is to share the authors' opinions about big data from their data analytics perspectives. The four authors bring quite different perspectives with different research experiences and expertise, spanning computational intelligence, machine learning, data mining and science, and interdisciplinary research. Authors represent academia and industry across four different continents. This diversity brings together an interesting perspective and coverage on exploring data analytics in the context of today's big data.

It is worth emphasizing that this article does not intend to provide a comprehensive review about the state-of-the-art of big data research, nor to provide a future big data research agenda. The aim is to expose the authors' personal opinions and present their perspectives of the future based on their views. As such there will be necessarily limited evidential argument or literary support, given the rapidly changing landscape and significant lag of academic research reporting. Indeed, many important issues and relevant techniques are not specifically covered in this article, and are best left to survey papers.

While all authors have contributed to the overall paper, each author has focused on their particular specialities in the following discussions. Zhou covers machine learning, while Chawla brings a data mining and data science perspective. Jin provides a view from computational intelligence and meta-heuristic global optimization, and Williams draws upon a machine learning and data mining background applied as a practicing data scientist and consultant to industry internationally.

2 Machine Learning with Big Data

Machine learning is among the core techniques for data analytics. In this section we will first clarify three common but unfortunately misleading arguments about learning systems in the big data era. Then we will discuss some issues that demand attention.

2.1 Three Misconceptions

2.1.1 "Models are not important any more"

Many people today talk about the replacement of sophisticated models by big data, where we have massive amounts of data available. The argument goes that in the "small data era" models were important but in the "big data era" this might not be the case.

Such arguments are claimed to be based on empirical observations of the type illustrated in Fig. 1. With small data (e.g., data size of 10), the best model is about $x\%$ better than the worst model in the figure, whereas the performance improvement brought by big data (e.g., data size of 10^4) is $y\% \gg x\%$. Such observations can be traced over many years, as in [7], and predate the use of "big data." It is interesting to see that in the "big data era", many people take such a figure (or similar figures) to claim that having big data is enough to get better performance. Such a superficial observation, however, neglects the fact that even with big data (e.g., data size of 10^4 in the figure), there are still significant differences between the different models—models are still important.

Also, we often hear such arguments: As the figure shows, the simplest model with small data achieves the best performance with big data, and thus, one does not need to have a sophisticated model because the simple model is enough. Unfortunately, this argument is also incorrect.

First, there is no reason to conclude that the worst-performing model on the small data is

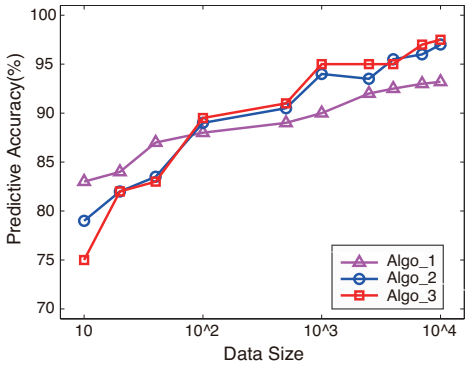

Fig. 1. Illustration: Model performance vs. Data size.

really the "simplest", and vice versa. Second, even if we assumed that the worst-performing model on the small data is really the simplest one, there is no support for the generalization of the argument that the simplest model will definitely achieve the best performance with big data in tasks other than the current empirical study.

If we take a look into [7] we can find that Algo_1 in Fig. 1 corresponds to a memory-based method, Algo_2 corresponds to Perceptron or Naïve Bayes, and Algo_3 corresponds to Winnow. It is hard to conclude that Winnow is simpler than the memory-based method; at least, the "simpler will be better" argument cannot explain why the performance of a Perceptron is better than that of the memory-based method on big data. A more reasonable explanation as to why the memory-based method is better on small data than Winnow but worse on big data may owe to its requirement of loading the data into memory. This is a memory issue and not whether a model is sophisticated or not.

The recent interest in deep learning [10], [26] provides strong evidence that on big data, sophisticated models are able to achieve much better performance than simple models. We want to emphasize that the deep learning techniques are not really new, and many ideas can be found from the 1990's [25], [32]. However, there were two serious problems that encumbered development at that time. First, the computational facilities available at that time could hardly handle models with thousands of parameters to tune. Current deep learning models involve millions or even billions of parameters. Second, the data scale at that time was relatively small, and thus models with high complexity were very likely to overfit. We can see that with the rapid increase of computational power, training sophisticated models becomes more and more feasible, whereas the big data size greatly reduces the overfitting risk of sophisticated models. From this sense, one can even conclude that in the big data era, sophisticated models become more favored since simple models are usually incapable of fully exploiting the data.

2.1.2 "Correlation is enough"

Some popular books on big data, including [37], claim that it is enough to discover "correlation" from big data. The importance of "causality" will be over taken by "correlation", with some advocating that we are entering an "era of correlation".

Trying to discover causality represents a great intention of searching for an in-depth understanding of the data. This is usually challenging in many real domains [44]. However, we have to emphasize that correlation is far from sufficient, and the role of causality can never be replaced by correlation. The reason lies in the fact that one invests in data analytics because one wants to get information helpful for making wiser decisions and/or taking suitable actions, whereas the abuse of correlation will be misleading or even disastrous.

We can easily find many examples, even in classic statistical textbooks, illustrating that correlation cannot replace the role of causality. For example, empirical data analysis on public security cases in a number of cities disclosed that the number of hospitals and the number

of car thefts are highly positively correlated. Indeed, car thefts increase almost linearly with the construction of new hospitals. With such a correlation identified, how would the mayor react to reduce car thefts? An "obvious" solution is to cease the construction of new hospitals. Unfortunately, this is an abuse of the correlation information. It will only decrease the opportunity of patients getting timely medical attention, whereas it is extremely unlikely to have anything to do with the incidence of car thefts. Instead, the increase of both the incidence of car thefts and the number of hospitals is actually affected by a latent variable, i.e., the residential population. If one believes that correlation is sufficient and never goes deeper into analysis of the data, one might serve as a mayor who plans to reduce car thefts by restricting the construction of hospitals.

Sometimes computational challenges may encumber the discovery of causality, and in such cases, discovering valid correlation, will be able to provide some helpful information. However, exaggerating the importance of "correlation" and taking the replacement of causality by correlation as a feature of the "big data era" can be detrimental, and lead to unnecessary, negative consequences.

2.1.3 "Previous methodologies do not work any more"

Another popular argument claims that previous research methodologies were designed for small data and they cannot work well on big data. This argument is often held by people who are highly enthusiastic for newly proposed techniques, thus they seek "totally new" paradigms.

We appreciate the search for new paradigms as this is one of the driving forces for innovative research. However, we highlight the importance of "past" methodologies.

Firstly, we should emphasize that researchers have always been trying to work with "big" data, such that what is regarded as big data today might not be regarded as big data in the future (e.g., in ten years). For example, in a famous article [31], the author expressed that "*For learning tasks with 10,000 training examples and more it becomes impossible ...*". The title of the paper "Making Large-Scale SVM Training Practical" implies that the goal of the article was "large-scale"—the experimental datasets in the paper mostly contained thousands of samples, and the biggest one contained 49,749 samples. This was deemed as "amazingly big data" at that time. Nowadays, few people will regard fifty thousand samples as big data.

Secondly, many past research methodologies still hold much value. We might consider [60], the proceedings of a KDD 1999 workshop. On the second page it is emphasized that "*implementation ... in high-performance parallel and distributed computing ... is becoming crucial for ensuring system scalability and interactivity as data continues to grow inexorably in size and complexity*". Indeed, most of the "current" facilitation for handling big data, such as high-performance computing, parallel and distributed computing, high efficiency storage, etc., has been used in data analytics for many years and will remain popular into the future.

2.2 Opportunities and Challenges

It is difficult to identify "totally new" issues brought about by big data. Nonetheless, there are always important aspects to which one hopes to see greater attention and efforts channeled.

First, although we have always been trying to handle (increasingly) big data, we have usually assumed that the core computation can be held in memory seamlessly. Whereas the current data size reaches to such a scale that the data becomes hard to store and even hard for multiple scans. However, many important learning objectives or performance measures are non-linear, non-smooth, non-convex and non-decomposable over samples. For example, AUC (Area Under the ROC Curve) [24],

and their optimizations, inherently require repeated scans of the entire dataset. Is it learnable by scanning the data only once, and if it needs to store something, the storage requirement is small and independent to data size? We call this "one-pass learning" and it is important because in many big data applications, the data is not only big but also accumulated over time, hence it is impossible to know the eventual size of the dataset. Fortunately, there are some recent efforts towards this direction, including [22]. On the other hand, although we have big data, are all the data crucial? The answer is very likely that they are not. Then, the question becomes can we identify valuable data subsets from the original big dataset?

Second, a benefit of big data to machine learning lies in the fact that with more and more samples available for learning, the risk of overfitting becomes smaller. We all understand that controlling overfitting is one of the central concerns in the design of machine learning algorithms as well as in the application of machine learning techniques in practice. The concern with overfitting led to a natural favor for simple models with less parameters to tune. However, the parameter tuning constraints may change with big data. We can now try to train a model with billions of parameters, because we have sufficiently big data, facilitated by powerful computational facilities that enable the training of such models. The great success of deep learning [10] during the past few years serves as a good showcase. However, most deep learning work strongly relies on engineering tricks that are difficult to be repeated and studied by others, apart from the authors themselves. It is important to study the mysteries behind deep learning; for example, why and when some ingredients of current deep learning techniques, e.g., pretraining and dropout, are helpful and how they can be more helpful? There have been some recent efforts in this direction [6], [23], [52]. Moreover, we might ask if it is possible to develop a parameter tuning guide to replace the current almost-exhaustive search?

Third, we need to note that big data usually contains too many "interests", and from such data we may be able to get "anything we want"; in other words, we can find supporting evidence for any argument we are in favor of. Thus, how do we judge/evaluate the "findings"? One important solution is to turn to statistical hypothesis testing. The use of statistical tests can help at least in two aspects: First, we need to verify that what we have done is really what we wanted to do. Second, we need to verify that what we have attained is not caused by small perturbations that exist in the data, particularly due to the non-thorough exploitation of the whole data. Although statistical tests have been studied for centuries and have been used in machine learning for decades, the design and deployment of adequate statistical tests is non-trivial, and in fact there have been misuses of statistical tests [17]. Moreover, statistical tests suitable for big data analysis, not only for the computational efficiency but also for the concern of using only part of the data, remain an interesting but under-explored area of research. Another way to check the validity of the analysis results is to derive interpretable models. Although many machine learning models are black-boxes, there have been studies on improving the comprehensibility of models such as rule extraction [62]. Visualization is another important approach, although it is often difficult with dimensions higher than three.

Moreover, big data usually exists in a distributed manner; that is, different parts of the data may be held by different owners, and no one holds the entire data. It is often the case that some sources are crucial for some analytics goal, whereas some other sources pose less importance. Given the fact that different data owners might warrant the analyzer with different access rights, can we leverage the sources without access to the whole data? What information must we have for this purpose? Even if the owners agree to provide some data,

it might be too challenging to transport the data due to its enormous size. Thus, can we exploit the data without transporting them? Moreover, data at different places may have different label quality, and may have significant label noise, perhaps due to crowdsourcing. Can we do learning with low quality and/or even contradictory label information? Furthermore, usually we assume that the data is identically and independently distributed; however, the fundamental *i.i.d.* assumption can hardly hold across different data sources. Can we learn effectively and efficiently beyond the *i.i.d.* assumption? There are a few preliminary studies on these important issues for big data, including [34], [38], [61].

In addition, given the same data, different users might have different demands. For example, for product recommendation, some users might demand that highly recommended items are good, and some users might demand that all the recommended items are good, while other users might demand all the good items have been returned. The computational, and storage loads of big data may be inhibitors to the construction of a model for each of the various demands separately. Can we build one model (a "general model" which can be adapted to other demands with cheap minor modifications) to satisfy the various demands? Some efforts have been reported recently in [35].

Another long-standing but unresolved issue is, in the "big data era", can we really avoid the violation of privacy concerns [2]? This is actually a long-standing problem that still remains open.

3 Data Mining/Science with Big Data

We posit again that big data is not a new concept. Rather, aspects of it have been studied and considered by a number of data mining researchers over the past decade and beyond. Mining massive data by scalable algorithms leveraging parallel and distributed architectures has been a focus topic of numerous workshops and conferences, including [1], [14], [43], [50], [60]. However, the embrace of the *Volume* aspect of data is coming to a realization now, largely through the rapid availability of datasets that exceed terabytes and now petabytes—whether through scientific simulations and experiments, business transactional data or digital footprints of individuals. Astronomy, for example, is a fantastic application of big data driven by the advances in the astronomical instruments. Each pixel captured by the new instruments can have a few thousand attributes and translate quickly to a peta-scale problem. This rapid growth in data is creating a new field called Astro-informatics, which is forging partnerships between computer scientists, statisticians and astronomers. This rapid growth of data from various domains, whether in business or science or humanities or engineering, is presenting novel challenges in scale and provenance of data, requiring a new rigor and interest among the data mining community to translate their algorithms and frameworks for data-driven discoveries.

A similar caveat also plays with the concept of *Veracity* of data. The issue of data quality or veracity has been considered by a number of researchers [39], including data complexity [9], missing values [19], noise [58], imbalance [13], and dataset shift [39]. The latter, dataset shift, is most profound in the case of big data as the unseen data may present a distribution that is not seen in the training data. This problem is tied with the problem of *Velocity*, which presents the challenge of developing streaming algorithms that are able to cope with shocks in the distributions of the data. Again, this is an established area of research in the data mining community in the form of learning from streaming data [3], [48]. A challenge has been that the methods developed by the data mining community have not necessarily been translated to industry. But times are changing, as seen by the resurgence of deep learning in

industry.

The issue with *Variety* is, undoubtedly, unique and interesting. A rapid influx of unstructured and multimodal data, such as social media, images, audio, video, in addition to the structured data, is providing novel opportunities for data mining researchers. We are seeing such data rapidly being collected into organizational data hubs, where the unstructured and structured cohabit and provide the source for all data mining. A fundamental question is related to integrating these varied streams or inputs of data into a singular feature vector presentation for the traditional learning algorithms.

The last decade has witnessed the boom of social media websites, such as Facebook, LinkedIn, and Twitter. Together they facilitate an increasingly wide range of human interactions that also provide the modicums of big data. The ubiquity of social networks manifests as complex relationships among individuals. It is generally believed that the research in this field will enhance our understandings of the topology of social networks and the patterns of human interactions [8], [18], [33], [36], [41], [54]. The relations among people affect not only social dynamics but also the broader dynamics of a variety of physical, biological, infrastructural and economic systems. While network theoretic techniques provide efficient means for analysis of data with complex underlying relationships, limitations in existing diffusion models are perhaps one of the main causes that restricts the extension of these methods to rather sophisticated application domains. However, these limitations are mainly due to the lack of capacity to adequately represent and process the imperfect data that are characteristic of such applications.

Our call to the community is to reconvene some of the traditional methods and identify their performance benchmarks on "big data". This is not about reinventing the wheel, but rather creating new paths and directions for groundbreaking research built on the foundations we have already developed.

3.1 From Data to Knowledge to Discovery to Action

Recent times have greatly increased our ability to gather massive amounts of data, presenting us with an opportunity to induce transformative changes in the way we analyze and understand data. These data exhibit a number of traits that have the potential to not only complement hypothesis-driven research but also to enable the discovery of new hypotheses or phenomena from the rich data, which could include spatial data, temporal data, observational data, diverse data sources, text data, unstructured data, etc.

Data of such extent and longitudinal character brings novel challenges for data-driven science for charting the path from data to knowledge to insight. This process of data-guided knowledge discovery will entail an integrated plan of descriptive analysis and predictive modeling for *useful insights* or hypotheses. These hypotheses are not just correlational but help explain an underlying phenomenon or help validate an observed phenomenon.

These discovered hypotheses or predictive analytics can help inform decisions, which include certain actions that can be appropriately weighed by the cost and impact of the action. The set of alternating hypotheses leads to scenarios that can be weighted situationally. Brynjolfsson et al. [11] studied 179 large companies and found that the companies that embraced data-driven decision making experienced a 5 to 6 percent higher level of productivity. The key difference was that these companies relied on data and analytics rather than solely on experience and intuition.

Healthcare is another area witnessing a significant application of big data. United Healthcare, for example, is expending effort on mining customer attitudes as gleaned from recorded voice files. The company is leveraging natural language processing along with text data to identify the customer sentiment and

satisfaction. It is a clear example of taking disparate big data, developing analytical models, and discovering quantifiable and actionable insights.

Big data presents unparalleled opportunities: to accelerate scientific discovery and innovation; to improve health and well-being; to create novel fields of study that hitherto might not have been possible; to enhance decision making by provisioning the power of data analytics; to understand dynamics of human behavior; and to affect commerce in a globally integrated economy.

3.2 Opportunities and Challenges

Big data is clearly presenting us with exciting opportunities and challenges in data mining research.

First, data-driven science and discovery should try to discover action-oriented insights that lead to charting new discoveries or impacts. Without understanding the nuances of one's data and the domain, one can fall into the chasm of simple and misleading correlation, sometimes leading to false discovery and insight. It is critical to fully understand and appreciate the domain that one is working in, and all observations and insights to be appropriately structured in that domain. It requires immersion of an individual in a domain to conduct feature engineering, data exploration, machine learning, and to inform system design and database design, and to conduct what-if analysis. This is not to say that a data scientist will be an expert in every aspect. Rather a data scientist may be an innovator in machine learning but well-versed in system design or databases or visualization or quick prototyping. But the data scientist cannot be divorced from the domain less they risk the peril of failing.

Second, algorithms are important, but before we jump on a journey of novel algorithms to tackle the four V's of big data, it is important for the community to consider the advances done hitherto, conduct a thorough empirical survey of them and then identify the potential bottlenecks, challenges and pitfalls of the existing state-of-the-art.

Third, any advances in scalable algorithms should be tied to the advances in architecture, systems, and new database constructs. We are witnessing a shift towards NoSQL databases, given the schema-free environment of the new data types and the prevalence of unstructured data. It is an opportunity for the algorithmic researchers to collaborate with systems/database researchers to integrate the machine learning or data mining algorithms as part of the pipeline to naturally exploit the lower constructs of data storage and computational fabric.

Fourth, a fundamental paradigm that is present in front of us is data-driven discovery. The data scientist must be the curious outsider who can ask questions of data, poke at the limitations placed on the available data and identify additional data that may enhance the performance of the algorithms at a given task. The hypothesis here is that there may be data external to the data captured by a given company, which may provide significant value. For example, consider the problem of predicting readmission for a patient on discharge. This problem of reducing readmission may find significant value by considering lifestyle data, which is outside of the patient's Electronic Medical Record (EMR).

We see these as some of the key opportunities and challenges that are specifically presented within the data mining with big data research context.

4 Global Optimization with Big Data

Another key area where big data offers opportunity and challenges is global optimization. Here we aim to optimize decision variables over specific objectives. Meta-heuristic global search methods such as evolutionary algorithms have been successfully applied to opti-

mize a wide range of complex, large-scale systems, ranging from engineering design to reconstruction of biological networks. Typically, optimization of such complex systems needs to handle a variety of challenges as identified here [12].

4.1 Global Optimization of Complex Systems

Complex systems often have a large number of decision variables and involve a large number of objectives, where the correlation between the decision variables may be highly nonlinear and the objectives are often conflicting. Optimization problems with a large number of decision variables, known as large-scale optimization problems, are very challenging. For example, the performance of most global search algorithms will seriously degrade as the number of decision variables increases, especially when there is a complex correlational relationship between the decision variables. Divide-and-conquer is a widely adopted strategy to deal with large-scale optimization where the key issue is to detect the correlational relationships between the decision variables so that correlated relationships are grouped into the same sub-population and independent relationships grouped into different sub-populations.

Over the past two decades, meta-heuristics have been shown to be efficient in solving multi-objective optimization problems, where the objectives are often conflicting with each other. The main reason is that for a population-based search method, different individuals can capture different trade-off relationships between the conflicting objectives, e.g., in complex structural design optimization [12]. As a result, it is possible to achieve a representative subset of the whole Pareto-optimal solution by performing one single run, in particular for bi- or tri-objective optimization problems. Multi-objective optimization meta-heuristics developed thus far can largely be divided into three categories, namely weighted aggregation based methods [28], Pareto-dominance based approaches [16] and performance indicator-based algorithms [5].

Unfortunately, none of these methods can work efficiently when the number of objectives becomes much higher than three. This is mainly because the number of total Pareto-optimal solutions becomes large and achieving a representative subset of them is no longer tractable. For the weighted aggregation approaches, it can become difficult to create a limited number of weight combinations to represent the Pareto-optimal solutions of a very high-dimension. For the Pareto-based approaches, most solutions in a population of a limited size are non-comparable. Thus, only few individuals dominate others and selection pressure for better solutions is lost. An additional difficulty is the increasingly large computational cost for performing the dominance relations when the number of objectives increases. Performance indicator-based approaches also suffer from high computational complexity, e.g., in calculating the hypervolume.

The second main challenge associated with optimization of complex systems is the computationally expensive processes of evaluating the quality of solutions. For most complex optimization problems, either time-consuming numerical simulations or expensive experiments need to be conducted for fitness evaluations. The prohibitively high computational or experimental costs make it intractable to apply global population-based search algorithms to such complex optimization problems. One approach that has been shown to be promising is the use of computationally efficient models, known as surrogates, to replace part of the expensive fitness evaluations [29]. However, constructing surrogates can become extremely challenging for large-scale problems with very limited data samples that are expensive to collect.

Complex optimization problems are often subject to large amounts of uncertainties, such as varying environmental conditions, system

degeneration, or changing customer demand [29]. Two basic ideas can be adopted to address the uncertainties in optimization. One is to find solutions that are relatively insensitive to small changes in decision variables or fitness functions, known as robust optimal solutions [29]. However, if the changes are large and continuous, meta-heuristics for tracking the moving optima will often be developed, which is known as dynamic optimization [42]. Different from the robustness approach to handling uncertainties, dynamic optimization aims to track the optimum whenever it changes. Theoretically this sounds perfect, but practically it is not desired for two reasons. First, tracking a moving optimum is computationally intensive, particularly if the fitness evaluations are expensive. Second, a change in the design or solution may be expensive and frequent changes are not allowed in many cases. To take these two factors into account, a new approach to cope with uncertainties, termed robustness over time, has been suggested [30]. The main idea is to reach a realistic trade-off between finding a robust optimal solution and tracking the moving optimum. That is, a design or solution will be changed only if the solution currently in use is no longer acceptable, and a new optimal solution that changes slowly over time, which is not necessarily the best solution in that time instant, will be sought.

4.2 Big Data in Optimization

Meta-heuristic global optimization of complex systems cannot be accomplished without data generated in numerical simulations and physical experiments. For example, design optimization of a racing car is extremely challenging since it involves many subsystems such as front wing, rear wing, chassis and tires. A huge number of decision variables are involved, which may seriously degrade the search performance of meta-heuristics. To alleviate this difficulty, data generated by aerodynamic engineers in their daily work will be very helpful to determine which subsystem, or even as a step further which part of the subsystem, is critical for enhancing the aerodynamic and drivability of a car. Analysis and mining of such data is, however, a challenging task, because the amount of data is huge, and the data might be stored in different forms and polluted with noise. In other words, these data are fully characterized by the four V's of big data. In addition, as fitness evaluations of racing car designs are highly time-consuming, surrogates are indispensable in optimization of racing vehicles.

Another example is the computational reconstruction of biological gene regulatory networks. Reconstruction of gene regulatory networks can be seen as a complex optimization problem, where a large number of parameters and connectivity of the network need to be determined. While meta-heuristic optimization algorithms have been shown to be very promising, the gene expression data for reconstruction is substantially big data in nature [51]. Data available from gene expression is increasing at an exponential rate [59]. The volume of data is ever increasing with developments in next generation sequence techniques such as high-throughput experiments. In addition, data from experimental biology, such as microarray data, is noisy, and gene expression experiments rarely have the same growth conditions and thus produce heterogeneous data sets. Data variety is also significantly increased through the use of deletion data, where a gene is deleted in order to determine its regulatory targets. Perturbation experiments are useful in reconstruction of gene regulatory networks, which, however, are another source of variety in biological data. Data collected from different labs for the same genes in the same biological network are often different.

It also becomes very important to develop optimization algorithms that are able to gain problem-specific knowledge during optimization. Acquisition of problem-specific knowledge can help capture the problem structure to perform more efficient search. For large-

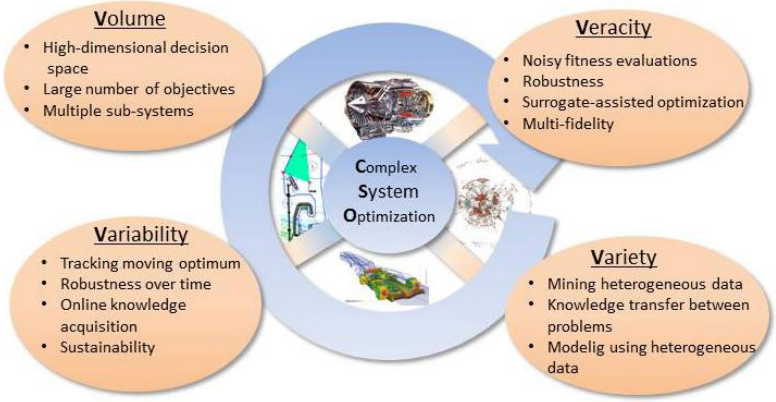

Fig. 2. Relationship between the challenges in complex engineering optimization and the nature of big data.

scale problems that have a large number of objectives, such knowledge can be used to guide the search through the most promising search space, and to specify preferences over the objectives so that the search will focus on the most important trade-offs. Unfortunately, sometimes only limited a-priori knowledge is available for the problem to be solved. It is therefore also interesting to discover knowledge from similar optimization problems or objectives that have been previously solved [20]. In this case, proper re-use of the knowledge can be very challenging. The relationship between the challenges in complex systems optimization and the nature of big data is illustrated in Fig. 2.

4.3 Opportunities and Challenges

As discussed above, big data is widely seen as essential for the success of the design optimization of complex systems. Much effort has been dedicated to the use of data to enhance the performance of meta-heuristic optimization algorithms for solving large-scale problems in the presence of large amounts of uncertainties. It is believed that the boom in big data research can create new opportunities as well as impose new challenges to data driven optimization. Answering the following questions can be central to converting the challenges posed by big data into opportunities.

First, how can we seamlessly integrate modern learning and optimization techniques? Many advanced learning techniques, such as semi-supervised learning [63], incremental learning [15], active learning [47] and deep learning [10] have been developed over the past decade. However, these techniques have rarely been taken advantage of within optimization with few exceptions, and they are critical in acquiring domain knowledge from a large amount of heterogeneous and noisy data. For optimization using meta-heuristics, such knowledge is decisive in setting up a flexible and compact problem representation, designing efficient search operators, constructing high-quality surrogates, and refining user preferences in multi-objective optimization.

Second, how can we formulate the optimization problem so that new techniques developed in big data research can be more efficiently leveraged? Traditional formulation of optimization problems consists of defining objective functions, decision variables and con-

straints. This works perfectly for small, well defined problems. Unfortunately, the formulation of complex problems is itself an iterative learning process. The new approaches to data analysis in big data can be of interest in simplifying the formation of complex optimization problems. For example, for surrogates to be used for rank prediction in population-based optimization, exact fitness prediction is less important than figuring out the relative order of the candidate designs. Might it be also possible to find meta-decision variables that might be more effective in guiding the search process than using the original decision variables?

Third, how do we visualize the high-dimensional decision space as well as the high-dimensional solution space to understand the achieved solutions and make a choice [27], [53]? How can techniques developed in big data analytics be used in optimization?

Overcoming these challenges in a big data framework will deliver significant advances to global optimization over the coming years.

5 Industry, Government and People with Big Data

We have presented above some of the technical challenges for research around the disciplines impacted and challenged by big data. In the end, we also must focus on the delivery of benefit and outcomes within industry, business and government. Over the decades, many of the technologies we have covered above, in machine learning, data mining, and global optimization, have found their way into a variety of large-scale applications. But what are the impacts we see today in industry and government of big data and how is this affecting and changing society and how might these changes affect our research across all these disciplines?

In this section we present a perspective on data analytics from the experiences in industry and government. The discussion is purposefully presented as a point of view of the future in practice rather than presenting a scientifically rigorous argument. We identify here areas where a focus from research might deliver impact to industry and government.

It is useful to reflect that over the past two decades we have witnessed an era where society has seen the mass collection of personal data by commercial interests and government. As *users*, we have been enticed by significant benefits to hand our data over to these organizations, and these organizations now house the big data that we have come to understand as the concept or the marketing term of the moment.

Google, Apple and Facebook, together with many other Internet companies that exist today, provide services ranging from the discovery of old friends to the ability to share our thoughts, personal details and daily activities publicly. With much of our email, our diaries and calendars, and our photos and thoughts and personal activities, now hosted by Google, for example, there is tremendous opportunity to identify and deal with a whole-of-client view on a massive scale. Combine that with our web logs, updates on our location and storage of our documents on Google Drive, and we start to understand the massive scope of the data collected about each of us, individually. These data can be used to better target the services advertised to us, using an impressive variety of algorithmic technologies to deliver new insights and knowledge.

Together, these crowdsourced data stores entice us to deliver our personal data to the data collectors in return for the sophisticated services they offer. The enticement is, of course, amazingly attractive, evidenced by the sheer number of users in each of the growing Internet ecosystems.

The customers that drive this data collection by these Internet companies are not the users of the services but are, for example, commercial advertisers and government services. The data is also made available to other organizations, purposefully and/or inappropriately.

As data have been *gradually* collected through centralized cloud services over time,

there is increasing need for broad discussion and understanding of the privacy, security, and societal issues that such collections of personal big data present. We have, as a society, reduced our focus on these issues and have come to understand that the centralized store of data is the only way in which we can deliver these desirable services.

However this concept of a centralized collection of personal and private data should be challenged. The centralized model primarily serves the interests of the organizations collecting the data—more so than the individuals. We are seeing a growing interest and opportunity to instead turn this centralized mode of data collection on its head, and to serve the interests of the individuals first, and those of the organizations second. The emergence of OwnClowd as a personally hosted replacement to Google Drive and DropBox is but one example of this trend.

The slowly growing rediscovery of the importance of privacy is leading to a shakeup of how we store data. We will see this leading to governments introducing new governance regimes for the Internet companies over time. We will also begin to see a migration of data not being stored centrally but being stored with the person about whom the data relates—the appropriate data owner. This *extreme data distribution* presents one of the greatest challenges to data scientists in the near future of big data.

We present three challenges presented by big data that relate to this emerging paradigm of extreme data distribution. The two initial challenges are the scale of our model building and the timeliness of our model building. The main focus is the challenge to society over the coming years where data are migrated from centralized to extremely distributed data stores. This is one of the more significant challenges that will be presented in the big data future, and one that has major impact on how we think about machine learning, data mining and global optimization.

5.1 Scaled Down Targeted Sub-Models

There has been considerable focus on the need to scale up our traditional machine learning algorithms to build models over the whole of the population available—and to build them quickly. Indeed, since the dawn of data mining [45] a key focus has been to scale algorithms. This is what we used to identify as the distinguishing characteristic between data mining (or knowledge discovery from databases) and the home disciplines of most of the algorithms we used: machine learning and statistics [56]. Our goal was to make use of all the data available, to avoid the need for sampling, and to ensure we capture knowledge from the whole population.

This goal remains with us today, as we continue to be obsessed with and able to collect data—masses of data—and thereby introduce new businesses and refine old ones. But of course we have for the past decades talked about big, large, huge, enormous, massive, humongous, data. The current fad is to refer to it as *big data* or *massive data* [40]. Irrespective, it is simply a lot of data.

In business and in government our datasets today consist of anything from 100 observations of 20,000 variables, to 20 million observations of 1,000 variables, to 1 billion observations of 10,000 variables. Such large datasets challenge any algorithm. While our research focus is generally on the algorithms, the challenges presented to the data collection, storage, management, manipulation, cleansing, and transformation are often generally much bigger (i.e., more time consuming) than presented by the actual model building.

Challenged with a massive dataset, what is the task, in practice, of the data scientist? In a future world, the learning algorithms will trawl through the massive datasets for us—they will slice and dice the data, and will identify anomalies, patterns, and behaviors. But how will this be delivered?

A productive approach will be to build on the successful early concept of ensemble model

building [55], but taken to a new massive scale. We are seeing the development of approaches that massively partition our datasets into many overlapping subsets, representing many different subspaces, often identifying behavioral archetypes. Within these subspaces we can more accurately understand and model the multiple behaviors exhibited by the entities of interest. The idea is not new [57] but has received scant attention over the years until now when we are realizing the need to slice and dice big data in sensible ways to uncover these multitudes of behavioral patterns. Today in industry and government this approach is delivering new models that are demonstrating surprisingly good results based on ensembles of thousands of smaller very different models.

The challenge here is how to identify the behavioral subspaces within which we build our models. Instead of building a single predictive model over the whole dataset we build a community of models that operate over the whole population as a single entity. This is done by understanding and dealing with the nuances and idiosyncrasies of the different sub-populations. It is within the much smaller sub-populations where the predictive models, for example, are built. The final model is then the ensemble of individual models applied to each new observation.

An example of this approach, deployed in practice, has analyzed over 2 million observations of 1000 variables to identify 20,000 such behavioral subspaces. The subspaces are created using a combination of cluster analysis and decision tree induction, and each subspace is described symbolically, each identifying and representing new concepts. The subspaces can overlap and individual observations can belong to multiple subspaces (i.e., a single observation may exhibit multiple behaviors). For each of the 20,000 subspaces, micro predictive models can be built, and by then combining these into an ensemble — using global optimization of multiple complex objective functions — we deliver an empirically effective global model, deployed, for example, to risk score the whole tax paying population as they interact with the taxation system.

We can (and no doubt will) continue to explore new computational paradigms like in-database analytics to massively scale machine learning and statistical algorithms to big data. But the big game will be in identifying and modeling the multiple facets of the variety of behaviors we all individually exhibit, and share with sizeable sub-populations, over which the modeling itself will be undertaken.

5.2 Right Time, Real Time, Online Analytics

The traditional data miner, in practice, has generally been involved in batch-oriented model building, using machine learning and statistical algorithms. From our massive store of historical data we use algorithms such as logistic regression, decision tree induction, random forests, neural networks, and support vector machines. Once we have built our model(s) we then proceed to deploy the models. In the business context we migrate our models into production. The models then run on new transactions as they arrive, perhaps scoring each transaction and then deciding on a treatment for that transaction—that is, based on the score, how should the transaction be processed by our systems?

The process is typical of how data mining is delivered in many large organizations today, including government, financial institutions, insurance companies, health providers, marketing, and so on. In the Australian Taxation Office, for example, every day a suite of data mining models risk score every transaction (tax return) received. The Australian Immigration Department [4], as another example, has developed a risk model that assesses all passengers when they check-in for their international flight to Australia (known as Advance Passenger Processing) using data mining models. Such examples abound through industry and government.

Today's agile context now requires more than this. The larger and older organizations, world wide, have tended to be much less agile in their ability to respond to the rapid changes delivered through the Internet and our data-rich world. Organizations no longer have the luxury of spending a few months building models for scoring transactions in batch mode. We need to be able to assess each transaction as the transaction happens, and to dynamically learn as the model interacts with the massive volumes of transactions that we are faced with as they occur. We need to build models in real-time to respond in real time and that learn and change their behavior in real-time.

Research in incremental learning is certainly not new. Incremental learning [15], [46], just-in-time or any-time learning [49], and data stream mining [21] have all addressed similar issues in different ways over the past decades. There is now increasing opportunity to capitalize on the approach. The question continues as to how we can maintain and improve our knowledge store over time, and work to forget old, possibly incorrect, knowledge?

The development of dynamic, agile learners working in real-time—that is, as they interact with the real world—remain quite a challenge and will remains a central challenge for our big data world.

5.3 Extreme Data Distribution—Privacy and Ownership

Having considered two intermediate challenges around big data, we now consider a game-changing challenge. The future holds for us the prospect of individuals regaining control of their data from the hands of the now common centralized massive stores. We expect to see this as an orderly evolution of our understanding of what is best for society as we progress and govern our civil society in this age of big data collection and surveillance and of its consequent serious risk to privacy.

Data ownership has become a challenging issue in our data rich world. Data collectors in the corporate and government spheres are learning to efficiently collect increasingly larger holdings of big data. However, with the help of civil libertarians, philosophers and whistle blowers, society is gradually realizing the need for better governance over the collection and use of data. Recent events like Wikileaks[1] and Edward Snowden[2] help to raise the level of discussion that is providing insight into the dangers of aggregating data centrally—with little regard about who owns the data.

We are well-aware of the dangers of single points of failure—relying on our data held centrally, as massive datasets, stored securely, and to be used by the data collectors only for the benefit of our society, and the individuals of that society. Yet, a single point of failure will mean that just one flaw or one breach can lead to devastating consequences on a massive scale. And with increasingly sophisticated methods of attack, it is increasingly happening. Even without sophistication, Snowden has demonstrated that simply having all the data stored in one location increases the risks significantly, to the detriment of governments, industry, and society as a whole.

After identifying risks. we often work towards strategies to mitigate those risks. An obvious strategy is to recoil from the inherently insecure centralization of the collection of data. Personal data need to move back to the individuals to whom the data belongs. We can increasingly collect and retain such data ourselves, under our control, as individuals, reducing the overall societal risk.

The services that are so attractive and that we have come to rely upon, the services provided by Google, Apple, Facebook, and many other Internet ecosystems, must still be delivered. The corporations must retain their ability to profit from the data, while the data itself is retained by the data owners. Under this future scenario, instead of centralizing all the computation, we can bring the computation to

1. http://en.wikipedia.org/wiki/Wikileaks
2. http://en.wikipedia.org/wiki/Edward_Snowden

intelligent agents running on our own personal devices and communicating with the service providers. Business models can still be profitable and users can regain their privacy.

Our personal data will be locked behind encryption technology, and the individuals will hold the key to unlock the data as they wish. The data will be served to our portable smart devices where it is decrypted and provides the services (mail, photos, music, instant messaging, shopping, search, web logs, etc.) we require. The data will be hosted in these personal encrypted clouds, running on mesh-network connected commodity smart devices. We see the beginnings of this change happening now with projects like the Freedom Box[3], OwnCloud[4], and the IndieWeb, and the widespread and massive adoption of powerful smartphones!

This view of distributed massive data brings with it the challenge to develop technology and algorithms that work over an *extreme data distribution*. How can we distribute the computation over this extreme level of data distribution and still build models that learn in a big data context?

The challenge of extreme distributed big data and learning is one that will quickly grow over the coming years. It will require quite a different approach to the development of machine learning, data mining and global optimization. Compare the approach to how society functions through an ensemble of humans. We are each personally a store of a massive amount of data and we share and learn from and use that data as we interact with the world and with other humans. So must the learning algorithms of the future.

5.4 Opportunities and Challenges

Identifying key challenges of big data leads one to question how the future of data collection might evolve over the coming years. As the pendulum reaches the limit of centralized massive data collection, in time and possibly sooner than we might expect, we will see the pendulum begin to swing back. It must swing back to a scenario where we restore data to the ownership and control of the individuals about whom the data relates. The data will be extremely distributed, with individual records distributed far and wide, just as the data owners are distributed far and wide across our planet.

With an extreme data distribution we will be challenged to provide the services we have come to expect with massive centralized data storage. Those challenges are surely not insurmountable, but will take considerable research, innovation, and software development to deliver.

The first challenge presented then becomes one of appropriately partitioning big data (eventually massive extreme distributed data), to identify behavioral groups within which we learn, and to even model and learn at the individual level. The second challenge is to refocus again on delivering learning algorithms that self-learn (or self-modify) in real-time, or at least at the right time, and to do this online. Finally, how do we deliver this in the context of extreme data distribution where the database records are now distributed far and wide and privacy protected, and how we might deliver learning agents that look after the interests of their "owner".

6 Wrap Up

From the data analytics perspectives we have presented here, there are many new opportunities and challenges brought by big data. Some of these are not necessarily new, but are issues that have not received the attention that they deserve. Here we recall some of the important/interesting issues:

- **Data size**: On one hand, we develop "one-pass learning" algorithms that require only one scan of the data with limited storage irrelevant to data size; on

3. http://en.wikipedia.org/wiki/FreedomBox
4. http://en.wikipedia.org/wiki/Owncloud

the other hand, we try to identify smaller partitions of the really valuable data from the original big data.

- **Data variety**: Data presents itself in varied forms for a given concept. It is presenting a new notion to learning systems and computational intelligence algorithm for classification, where the feature vector is multi-modal, with structured and unstructured text, and still the notion is to classify one concept from another. How do we create a feature vector, and then a learning algorithm with an appropriate objective function to learn from such varied data?
- **Data trust**: While data is rapidly and increasingly available, it is also important to consider the data source and if the data can be trusted. More data is not necessarily correct data, and more data is not necessarily valuable data. A keen filter for the data is a key.
- **Distributed existence**: Owners of different parts of the data might warrant different access rights. We must aim to leverage data sources without access to the whole data, and exploit them without transporting the data. We will need to pay attention to the fact that different sources may come with different label quality, there may be serious noise in the data due to crowdsourcing, and the *i.i.d.* assumption may not hold across the sources.
- **Extreme distribution**: Taking this idea even further, the unit-level data may be what we see as the level of data distribution, as we deal with issues of privacy and security. New approaches to modeling will be required to work with such distributed data.
- **Diverse demands**: People may have diverse demands whereas the high cost of big data processing may disable construction of a separate model for each demand. Can we build one model to satisfy the various demands? We also need to note that, with big data, it is possible to find supporting evidence to any argument we want; then, how to judge/evaluate our "findings"?
- **Sub-Models**: Diverse demands might also relate to diversity of the behaviors that we might be modeling within our application domains. Rather than one single model to cover it all, the model will consist of ensembles of smaller models that together deliver better understandings and predictions than the single, complex model.
- **Intuition importance**: Data is going to power novel discoveries and action-oriented business insights. It is important to still attach intuition, curiosity and domain knowledge without which one may become myopic and fall in the chasm of "correlation is enough". Computational intelligence should be tied with human intuition.
- **Rapid model**: As the world continues to "speed up", decisions need to be made more quickly because fraudsters can more quickly find new methods in an agile environment, model building must become more agile and real-time.
- **Big optimization**: Global optimization algorithms such as meta-heuristics have achieved great success in academic research, but have rarely been employed in industry. One major obstacle is the huge computational cost required for evaluating the quality of candidate designs of complex engineering systems. The emerging big data analytic technologies will remove the obstacle to a certain degree by reusing knowledge extracted from the huge amount of high-dimensional, heterogeneous and noisy data. Such knowledge can also be acquired with new visualization techniques. Big data driven optimization will also play a key role in reconstruction of large-scale biological systems.
- **Complex optimization**: Definition of decision variables, setup of the objectives and articulation of the constraints are three

main steps in formulating optimization problems before solving them. For optimization of complex systems, formulation of the optimization problem itself becomes a complex optimization problem. The big data approach might provide us new insights and methodologies for formulating optimization problems, thus leading to a more efficient solution.

In closing the discussion, we emphasize that the opportunities and challenges brought by big data are very broad and diverse, and it is clear that no single technique can meet all demands. In this sense, big data also brings a chance of "big combination" of techniques and of research.

ACKNOWLEDGMENTS

The authors want to thank the editor and anonymous reviewers for helpful comments and suggestions. This article presents a consolidation of some of the presentations and discussion of the panel "*Big Data: Real Challenges and Status*" hosted at IDEAL 2013; the very positive reflections from the audience for the panel motivated the writing of this article. The authors thank Hui Xiong, Kai Yu, Xin Yao and other participants for comments and discussion.

REFERENCES

[1] J. Abello, P. M. Pardalos, and M. G. Resende, *Handbook of Massive Data Sets*, ser. Massive Computing. Springer, 2002, vol. 4.

[2] C. C. Aggarwal and P. S. Yu, Eds., *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[3] C. C. Aggarwal, *Data streams: Models and algorithms*. Springer, 2007, vol. 31.

[4] Australian Department of Immigration, "Fact sheet 70 - managing the border," Internet, 2013. [Online]. Available: http://www.immi.gov.au/media/fact-sheets/70border.htm

[5] J. Bader and E. Zitzler, "HypE: An algorithm for fast hypervolume-based manyobjective optimization," *Evolutionary Computation*, vol. 19, no. 1, pp. 45–76, 2011.

[6] P. Baldi and P. J. Sadowski, "Understanding dropout," in *Advances in Neural Information Processing Systems 26*. Cambridge, MA: MIT Press, 2013, pp. 2814–2822.

[7] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001, pp. 26–33.

[8] A.-L. Barabási, *Linked: The New Science Of Networks*. Basic Books, 2002.

[9] M. Basu and T. K. Ho, *Data Complexity in Pattern Recognition*. London, UK: Springer, 2006.

[10] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[11] E. Brynjolfsson, L. Hitt, and H. Kim, "Strength in numbers: How does data-driven decision making affect firm performance?" *Available at SSRN 1819486*, 2011.

[12] T. Chai, Y. Jin, and S. Bernhard, "Evolutionary complex engineering optimization: Opportunities and challenges," *IEEE Computational Intelligence Magazine*, vol. 8, no. 3, pp. 12–15, 2013.

[13] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. US: Springer, 2005, pp. 853–867.

[14] N. V. Chawla *et al.*, *Learning on Extremes-Size and Imbalance-of Data*. Florida, US: University of South Florida, 2002.

[15] Q. Da, Y. Yu, and Z.-H. Zhou, "Learning with augmented class by exploiting unlabeled data," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec City, Canada, 2014.

[16] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[17] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[18] D. Easley and J. Kleinberg, *Networks, crowds, and markets*. Cambridge Univ Press, 2010.

[19] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, 2008.

[20] L. Feng, Y.-S. Ong, I. Tsang, and A.-H. Tan, "An evolutionary search paradigm that learns with past experiences," in *IEEE Congress on Evolutionary Computation*, Brisbane, QLD, Australia, 2012, pp. 1–8.

[21] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," *ACM SIGMOD Record*, vol. 34, no. 2, pp. 18–26, 2005.

[22] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, "One-pass AUC optimization," in *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, 2013, pp. 906–914.

[23] W. Gao and Z.-H. Zhou, "Dropout Rademacher complexity of deep neural networks," CORR abs/1402.3811, 2014.

[24] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic

curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

[25] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The "wake-sleep" algorithm for unsupervised neural networks," *Science*, vol. 268, no. 5214, pp. 1158–1161, 1995.

[26] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[27] H. Ishibuchi, M. Yamane, N. Akedo, and Y. Nojima, "Many-objective and many-variable test problems for visual examination of multiobjective search," in *IEEE Congress on Evolutionary Computation*, Cancun, Mexico, 2013, pp. 1491–1498.

[28] H. Ishibuchi and T. Murata, "Multi-objective genetic local search algorithm," in *Proceedings of IEEE International Conference on Evolutionary Computation*, Nagoya, Japan, 1996, pp. 119–124.

[29] Y. Jin and J. Branke, "Evolutionary optimization in uncertain environments - A survey," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 3, pp. 303–317, 2005.

[30] Y. Jin, K. Tang, X. Yu, B. Sendhoff, and X. Yao, "A framework for finding robust optimal solutions over time," *Memetic Computing*, vol. 5, no. 1, pp. 3–18, 2013.

[31] T. Joachims, "Making large-scale SVM training practical," in *Advances in Kernel Methods*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 169–184.

[32] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA: MIT Press, 1995.

[33] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 420–429.

[34] M. Li, W. Wang, and Z.-H. Zhou, "Exploiting remote learners in internet environment with agents," *Science China: Information Sciences*, vol. 53, no. 1, pp. 47–76, 2010.

[35] N. Li, I. W. Tsang, and Z.-H. Zhou, "Efficient optimization of performance measures by classifier adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1370–1382, 2013.

[36] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 243–252.

[37] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Transforms How we Work, Live, and Think*. Houghton Mifflin Harcourt, 2012. (Chinese translated version by Y. Sheng and T. Zhou, Zhejiang Renmin Press).

[38] M. Mohri and A. Rostamizadeh, "Stability bounds for stationary φ-mixing and β-mixing processes," *Journal of Machine Learning Research*, vol. 11, pp. 789–814, 2010.

[39] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[40] National Research Council, *Frontiers in Massive Data Analysis*. Washington D.C.: The National Academies Press, 2013.

[41] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[42] T. T. Nguyen, S. Yang, and J. Branke, "Evolutionary dynamic optimization: A survey of the state of the art," *Swarm and Evolutionary Computation*, vol. 6, pp. 1–24, 2012.

[43] B.-H. Park and H. Kargupta, "Distributed data mining: Algorithms, systems, and applications," in *The Handbook of Data Mining*, N. Ye, Ed. New Jersey, US: Lawrence Erlbaum Associates, 2002, pp. 341–358.

[44] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[45] G. Piatetsky-Shapiro and W. Frawley, Eds., *Proceedings of IJCAI89 Workshop on Knowledge Discovery in Databases*. Detroit, MI: AAAI, 1989.

[46] J. C. Schlimmer and D. Fisher, "A case study of incremental concept induction," in *Proceedings of the 5th National Conference on Artificial Intelligence*, Philadelphia, PA, 1986, pp. 496–501.

[47] B. Settles, "Active learning literature survey," Department of Computer Sciences, University of Wisconsin at Madison, Wisconsin, WI, Tech. Rep. 1648, 2009, http://pages.cs.wisc.edu/∼bsettles/pub/settles. activelearning.pdf.

[48] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. de Carvalho, and J. Gama, "Data stream clustering: A survey," *ACM Computing Surveys (CSUR)*, vol. 46, no. 1, p. 13, 2013.

[49] P. Smyth, "Recent advances in knowledge discovery and data mining (invited talk)," in *Proceedings of the 14th National Conference on Artificial Intelligence*, Providence, RI, USA, 1997.

[50] A. Srivastava, E.-H. Han, V. Kumar, and V. Singh, *Parallel Formulations of Decision-Tree Classification Algorithms*. US: Springer, 2002.

[51] S. A. Thomas and Y. Jin, "Reconstructing biological gene regulatory networks: Where optimization meets big data," *Evolutionary Intelligence*, vol. 7, no. 1, pp. 29–47, 2013.

[52] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Advances in Neural Information Processing Systems 26*. Cambridge, MA: MIT Press, 2013, pp. 351–359.

[53] D. Walker, R. Everson, and J. E. Fieldsend, "Visualising mutually non-dominating solution sets in many-objective optimisation," *IEEE Transactions on Evolutionary Computation*, vol. 17, no. 2, pp. 165–184, 2013.

[54] D. J. Watts, *Six degrees: The science of a connected age*. WW Norton, 2004.

[55] G. J. Williams, "Combining decision trees: Initial results from the MIL algorithm," in *Artificial Intelligence Developments and Applications*, J. S. Gero and R. B. Stanton, Eds. Elsevier Science Publishers B.V. (North-Holland), 1988, pp. 273–289.

[56] G. J. Williams and Z. Huang, "A case study in knowledge acquisition for insurance risk assessment using a KDD methodology," in *Proceedings of the Pacific Rim Knowledge Acquisition Workshop*, P. Compton, R. Mizoguchi, H. Motoda, and T. Menzies, Eds., Sydney, Australia, 1996, pp. 117–129.

[57] ——, "Mining the knowledge mine: The Hot Spots methodology for mining large, real world databases," in *Advanced Topics in Artificial Intelligence*, A. Sattar, Ed. Springer-Verlag, 1997, pp. 340–348.

[58] X. Wu and X. Zhu, "Mining with noise knowledge: error-aware data mining," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 4, pp. 917–932, 2008.

[59] M. Xiao, L. Zhang, B. He, J. Xie, and W. Zhang, "A parallel algorithm of constructing gene regulatory networks," in *Proceedings of the 3rd International Symposium on Optimization and Systems Biology*, D.-Z. Du and X.-S. Zhang, Eds., Zhangjiajie, China, 2009, pp. 184–188.

[60] M. J. Zaki and C.-T. Ho, Eds., *Large-scale parallel data mining*. Berlin, Germany: Springer, 2000.

[61] D. Zhou, J. C. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Cambridge, MA: MIT Press, 2012.

[62] Z.-H. Zhou, "Rule extraction: Using neural networks or for neural networks?" *Journal of Computer Science and Technology*, vol. 19, no. 2, pp. 249–253, 2004.

[63] X. Zhu, "Semi-supervised learning literature survey," Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, Tech. Rep. 1530, 2006, http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.