KI - Künstliche Intelligenz (2018) 32:3–8
https://doi.org/10.1007/s13218-017-0523-7

EDITORIAL

From Big Data to Big Artificial Intelligence?

Algorithmic Challenges and Opportunities of Big Data

Kristian Kersting[1] · **Ulrich Meyer**[2]

Abstract
Big Data is no fad. The world is growing at an exponential rate, and so is the size of data collected across the globe. The data is becoming more meaningful and contextually relevant, breaks new ground for machine learning and artificial intelligence (AI), and even moves them from research labs to production. That is, the problem has shifted from collecting massive amounts of data to understanding it, i.e., turning data into knowledge, conclusions, and actions. This Big AI, however, often faces poor scale-up behaviour from algorithms that have been designed based on models of computation that are no longer realistic for Big Data. This special issue constitutes an attempt to highlight the algorithmic challenges and opportunities but also the social and ethical issues of Big Data. Of specific interest and focus have been computation- and resource-efficient algorithms when searching through data to find and mine relevant or pertinent information.

1 About the Special Issue

The amount of newly generated data per year is huge and keeps on growing tremendously[1]: from about 150 Exabytes in 2005 (worldwide) to approximately 1200 Exabytes in 2010. Nowadays, we create 2.5 quintillion bytes of data every day[2]. Twitter users generate over 500 million tweets every day[3], and a similar amount of images is uploaded to Facebook[4]. In 2016, the Facebook graph, which reflects the friendship relation between Facebook users, features more than a billion nodes and over hundreds of billions friendship edge[5]. And, the size of the indexed World Wide Web (estimated via the size of Googles index) is over 45 billion web pages[6], and Google alone performs several billion searches on it every day.

A similar data explosion can be observed in the scientific world, for example in genetics, biology, or particle physics, as well as ever increasing digitized text collections. For instance, particles collide in the large hadron collide (LHC) detectors approximately 1 billion times per second, generating about one petabyte of collision data per second. Even though only the most "interesting" events can be stored and processed, CERN data center has already accumulated over 200 petabytes of filtered data[7].

Together with advances in machine learning and AI, this data has the potential to lead to many new breakthroughs. For example, high-throughput genomic and proteomic experiments can be used to enable personalized medicine. Large data sets of search queries can be used to improve

✉ Kristian Kersting
kersting@cs.tu-darmstadt.de

Ulrich Meyer
umeyer@ae.cs.uni-frankfurt.de

[1] Department for Computer Science and Centre for Cognitive Science, TU Darmstadt, Germany

[2] Institute for Computer Science, Goethe-Universitt Frankfurt am Main, Frankfurt, Germany

[1] The Economist, issue of Feb 25th, 2010

[2] IBM's "10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations", https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN, accessed Dec 3, 2017

[3] http://www.internetlivestats.com/twitter-statistics/, accessed Dec 3, 2017

[4] http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9, accessed Dec 3, 2017

[5] https://research.fb.com/three-and-a-half-degrees-of-separation/, accessed Dec 3, 2017

[6] http://www.worldwidewebsize.com/, accessed Dec 3, 2017

[7] https://home.cern/about/updates/2017/07/cern-data-centre-passes-200-petabyte-milestone, accessed Dec 3, 2017

information retrieval. Historical climate data can be used to understand global warming and to better predict weather. Large amounts of sensor network readings and hyperspectral images of plants can be used to identifying drought conditions and to gain insights into when and how stress impact plant growth and development and in turn how to counterattack the "world hunger" problem.

Unfortunately, we often face poor scale-up behaviour from algorithms that have been designed based on models of computation that are no longer realistic for Big Data. This implies challenges like algorithmic exploitation of parallelism (multicores, GPUs, parallel and distributed systems, etc.), handling external and outsourced memory as well as memory-hierarchies (clouds, distributed storage systems, hard-disks, flash-memory, etc.), dealing with large scale dynamic data updates and streams, compressing and processing compressed data, approximation and online processing respectively mining under resource constraints, increasing the robustness of computations (e.g., concerning data faults, inaccuracies, or attacks) or reducing the consumption of energy by algorithmic measures and learning. Only then Big Data will truly open up unprecedented opportunities for both scientific discoveries and commercial exploitation across many fields and sectors. That is, only then we can achieve Big AI.

This special issue of the *German Journal of Artificial Intelligence* (KI) constitutes an attempt to highlight the recent progress made towards meeting these algorithmic challenges but also touches upon the social and ethical issues raised by doing so. Specifically, the special issues draws upon an open call for contributions and the progress made within two national research initiatives established by the German Science Foundation (DFG) in recent years: the Priority Programme SPP 1736 "Algorithms for Big Data" and the Collaborative Research Center CRC 876 "Providing Information by Resource-Constrained Analysis".

The priority programme (in German: Schwerpunktprogramm) SPP1736 aims at meeting the challenges mentioned above by bringing together expertise from different areas. On the one hand recent hardware developments and technological challenges need to be appropriately captured in better computational models. On the other hand, both common and problem specific algorithmic challenges due to Big Data are to be identified and clustered. Considering both sides, a basic toolbox of improved algorithms and data structures for Big Data is to be derived, where we do not only strive for theoretical results but intend to follow the whole algorithm engineering development cycle. The collaborative research center (in German: Sonderforschungsbereich) SFB876 is motivated by the fact that networked devices and sensors enable accessing the data independently of time and location: Highly distributed data, accessible only on devices with limited processing capability, versus high dimensional data and large data volumes. Both ends of the spectrum share the limitation brought by constrained resources. Classical machine learning algorithms can typically not be applied here without adjustments to incorporate these limitations. To provide solutions, the available data needs interpretation. The ubiquity of data access faces the demand for similar ubiquity in information access. Intelligent processing near to the data generation side (the small side) eases the analysis of aggregated data due to reduced complexity (the large side).

2 Content of the Special Issue

This special issue is composed of an editorial including results of an online questionnaire as well as technical contributions and research projects, and a Doctoral Thesis report. Contributing labs encompass departments of computer science, sociology, business and economics. All contributions were peer-reviewed.

One of the research project reviews concerns the *DFG Priority Programme SPP 1736: Algorithms for Big Data*, which has been established in 2013 and just entered its second funding phase. The article by Mahyar Behdju and Ulrich Meyer gives a short overview of the research topics represented in the priority programme and highlights sample results obtained by the individual projects during the first funding phase.

Two projects from SPP 1736 contributed separate technical papers, thus providing a more detailed treatment of their research areas: In the first article on *Big Data algorithms beyond machine learning*, Matthias Mnich surveys various Big Data techniques. While the main focus is on fixed-parameter tractable algorithms, he also treats topics such as sublinear sampling, parallel external-memory algorithms, and compression of uncertain data.

The second article by Hannah Bast, Björn Buchhold, and Elmar Haussmann features *a quality evaluation of combined search on a knowledge base and text*. While knowledge based search uses a knowledge base in form of 'semantic triples' of the form subject-predicate-object, in full text-search the data is given as a set of text documents. The proposed combination extends knowledge based search by a special predicate 'occurs-with' which holds if an entity of the knowledge base occurs with words in the text. After briefly describing the method, an overview of related benchmarks and analyses is given, followed by the quality evaluation for the system constructed by the authors.

Randomized primitives for Big Data processing is the topic of Morton Stöckel's PhD thesis. The thesis report provides an overview of new developments in randomized algorithms and data structures in the context of data similarity. With a focus on hashing based methods Stöckel improves the computation of intersection sizes in several areas: set intersection, sparse matrix-multiplication, and similarity joins.

Two other contributions concern the *DFG Collaborative Research Center SFB 876: Providing Information by Resource-Constrained Data Analysis*, which has been established in 2011 and is currently in its second phase. In *Big Data Science*, Katharina Morik, Christian Bockermann und Sebastian Buschjäger provide an overview of the research topics represented in the SFB and highlight the usefulness and the challenges of Big Data analytics on two challenging real-world data analytics applications in astroparticle physics within the SFB. As the authors illustrate, Big Data is not just algorithms but also architecture, i.e., it is important to develop algorithms that take the architecture of the machines into account that analyse the data in order to speed up the data analytic processing.

In *Coresets—Methods and History A Theoreticians Design Pattern for Approximation and Streaming Algorithms*, Alexander Munteanu and Chris Schwiegelshohn present a technical survey on the state of the art approaches in data reduction and the coreset framework. These include geometric decompositions, gradient methods, random sampling, sketching and random projections. The authors further outline the importance for the design of streaming algorithms and give a brief overview of lower bounding techniques.

Finally, *societal implications of Big Data* are discussed by Karolin Kappler, Jan-Felix Schrape, Lena Ulbricht, and Johannes Weyer. These are results from the German BMBF[8] interdisciplinary research cluster ABIDA[9] on *a*ssessing *Bi*g *Da*ta. Initiated in 2015, the cluster aims at monitoring and assessing current developments regarding Big Data, taking into account public opinion and bringing together expert knowledge. By discussing recent practices of self-tracking as well of real-time control of complex systems, the authors show that real-time analysis and feedback loops increasingly foster a society of (self-)control and argue that data and social scientists should work together to develop concepts of regulation, which is needed to cope with the challenges and risks of Big Data.

[8] Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)

[9] http://www.abida.de/en

3 Does Big Data converge wit AI?

Indeed, having a special issue on Big Data in an Artificial Intelligence journal raises some questions. What does Big Data really mean for AI? Is Big Data converging with AI, as suggested by some news, blogs and other media? Is Big Data really the answer to open questions in AI?

To understand this, we invited several people from academia and industry around the globe to complete a short online questionnaire on "Big Data and AI". We received feedback from 19 (with permission to use use the feedback in this editorial, but sometimes only in an anonymous fashion—attributed to "Anonymous" later) with mixed backgrounds: 11× Europe, 7× North America, and 1× Asia; 15× professors at universities and 4× industry; 4× Algorithms, 10× Artificial Intelligence, 18× Machine Learning/Data Mining, and 5× Theoretical Computer Science (multiple answers were allowed). The results are certainly not representative but give (in our opinion) a good idea of why AI should care about Big Data, in particular, given that some renown CS researchers were among the participants.

We started off by asking how to define Big Data. As illustrated in Tab. 1, the responses essentially define Big Data as data sets that are too voluminous and complex to be processed by traditional algorithms and architectures.

Next, we asked about the biggest opportunities and risks for using Big Data. The answers, see Table 2, illustrate that Big Data makes statistical machine learning more robust, solves low-level language and image perception tasks, and enables real-time control within many scientific, business and social areas. However, interestingly, only two mentioned "deeper" tasks of (artificial) intelligence such as reasoning, scheduling, and planning. This is in line with the risks expressed. Participants wondered about how to verify and reproduce results obtained using Big Data. They actually seem to fear to get "stupid" machines and people that do not aim at understanding complex problems anymore. Nevertheless, when asking "Is Big Data helping AI?", 18 participants said "yes".

To get hands on the most promising AI areas to benefit from Big Data, we explicitly asked for them. The answers, see Table 3, directly reflect the pros and cons of Big Data. They are a mixture of applications coming from biology, transportation, industry 4.0, eCommerce, finance, education, NLP, and computer vision, among others, and focus on rather low-level AI tasks, mainly using deep learning. Higher cognitive tasks such as reasoning, planning and acting in complex environments were only seldom touched. Interestingly, new research questions such as interactive machine learning and AI also popped up.

Table 1 How do you define Big Data? Some illustrative parts of answers from the online questionnaire

Definition of big data
The ability to collect huge amounts of data (i.e., database scale) about many different phenomenon (Anonymous)
Large volumes of interrelated data with multiple modalities (Sriraam Natarajan, UT Dallas, USA)
"Big Data" is the body of technology—algorithms, programming systems, and hardware, that stress our abilities to handle the data. A related, and apparently more modern term is "Data Science", which really means the same thing, but includes the application areas to which "big-data" technology is applied. What it DOESN'T mean is "AI" or "Machine Learning" or "Statistics done right", as some have claimed (Jeffrey D. Ullman, Stanford, USA)
I define it by size but also by the fact that it is typically collected as a side-effect of some other process (Anonymous)
One or more of: petabytes, high velocity, high complexity—high velocity (more data coming at you faster) predominant (Sofus Macskassy, Branch Metrics, USA)
Large volumes of data from diverse sources that can be used to learn new facts, or predict future events (Anonymous)
When it is faster to transport it on tape than to transfer it via satellite (Anonymous)
Big Data is a collection of data instances which is significantly larger than the data that has existed/been used in a given area before (Anonymous)
Computational approaches to solving computational problems lie on a spectrum of model complexity. Big data is at one end of the spectrum, effectively attempting to solve problems without explicit models. I would therefore define Big Data as the set of methods and efforts that attempt to solve problems without the necessity of explanation or understanding through explicit models (Oliver Brock, TU Berlin, Germany)
Enough data so that only very basic/generic priors (such as in NNs or CNNs) are sufficient ensure generalization to test data (Marc Toussaint, MIT, USA)
Mainly high dimensions. In fact rich SVD spectrum (Nikolaos Vasiloglou, MLtrain, USA)
"Big" is a relative term. A couple of decades ago, datasets with a million elements were considered big. But now, one routinely comes across datasets with trillions of elements (Anonymous)

The main take-away message—a 15 to 4 vote—for the AI community is: Big Data is not converging with AI!

4 Related Scientific Forums

To a significant extent papers dealing with Big Data algorithms appear in major general algorithms conferences such as SODA (ACM-SIAM Symposium on Discrete Algorithms) and ESA (European Symposium on Algorithms), data centric conferences like IEEE International Conference on Data Engineering (ICDE), or algorithm engineering conferences like Algorithm Engineering and Experiments (ALENEX) or SEA (International Symposium on Experimental Algorithms, formerly WEA). More specialized conferences for parallel and distributed algorithms include ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), IEEE International Parallel and Distributed Processing Symposium (IPDPS), and ACM Symposium on Principles of Distributed Computing (PODC). Top conferences for information retrieval and string processing are ACM Special Interest Group on Information Retrieval (SIGIR), Web Search and Data Mining (WSDM), String Processing and Information Retrieval (SPIRE), and Combinatorial Pattern Matching (CPM). For operations research and mathematical programming applications, two of the most important conferences are IPCO (Conference on Integer Programming and Combinatorial Optimization) and the triennial world congress ISMP (International Symposium on Mathematical Programming). International Conference on Intelligent Systems in Molecular Biology (ISCB), European Conference on Computation Biology (ECCB), and Conference on Research in Computational Molecular Biology (RECOMB) are the major events in bioinformatics. The top journals for machine learning and data mining are Artificial Intelligence Journal (AIJ), Journal of Artificial Intelligence Research (JAIR), Machine Learning Journal (MLJ), JMLR (Journal of Machine Learning Research), Data Mining and Knowledge Discovery (DAMI/DMKD) Journal, TKDD (ACM Transactions on Knowledge Discovery from Data), TPAMI (IEEE Transaction on Pattern Recognition and Machine Intelligence), and the International Journal of Data Science and Analytics. The main conferences are IJCAI (International Joint Conference on Artificial Intelligence), AAAI (AAAI Conference on Artificial Intelligence), ECAI (European Conference on Artificial Intelligence), NIPS (Annual Conference on Neural Information Processing Systems), ICLR (International Conference on Learning Representations), ICML (International Machine Learning Conference), European Conference on Machine Learning and Principles and

Table 2 What do you think are the biggest opportunities (left) and risks (right) for using Big Data? Some illustrative parts of answers from the online questionnaire

Opportunities of using Big Data	Risks of using Big Data
High robustness in statistical learning (Christian Bauckhage, U. Bonn, Germany)	Privacy concerns when dealing with personal data, particularly for healthcare. Inability to verify the correctness of learned models or discovered patterns when analyzing large datasets, particularly when using "black-box" models. Security issues, dealing with adversaries, information overload, etc. (Anonymous)
Almost every field of human endeavor. I see lots of activity and promise in biology and biomedicine. Lots of commercial applications too, as companies like Google or Facebook use their data gathered from billions of people to understand much about human activity, from detecting spam, to knowing what I really mean when I misspell something. I don't see how this question can be answered in a few lines. My favorite example is providing driving directions based on real-time traffic estimates obtained from cell phone location data (Jeffrey D. Ullman, Stanford, USA)	Hype and expectation (Sriraam Natarajan, UT Dallas, USA)
Deeper holistic understanding of complex data (Sofus Macskassy, Branch Metrics, USA)	too high reliance on what is learned as truth (bad models, bad assumptions, not understanding the data). biased models in ways we do not understand (e.g., gender biases) (Sofus Macskassy, Branch Metrics, USA)
Quick solutions for problems that we do not understand but for which abundant data exists (Oliver Brock, TU Berlin, Germany)	If the learning systems leveraging these Big Data sets are not designed carefully, they may end up codifying our biases and stereotypes (e.g., possible racial bias in automated airport profiling) and this may result in these biases getting even more deeply ingrained (Deepak Ajwani, Nokia Bell Labs, Ireland)
Understanding and getting the data is the key to solving many problems. It may not sound very complex or challenging on the surface, however is the key to be successful in AI / ML initiatives (Anonymous)	Frankly, I think the biggest risk is that governments, especially in Europe, will worry more about privacy than about the advantages that can come from exploiting "big data". Privacy is a modern invention. 200 years ago, before the anonymity of cities was available, people didn't imagine that they could keep their lives secret from those around them (Jeffrey D. Ullman, Stanford, USA)
To solve low-level language and image perception tasks. Discover correlations for rare medical conditions (Guy Van den Broeck, UCLA, USA)	That science in its current form ceases to exist because people just believe in the power of data but not in the power of understanding (Oliver Brock, TU Berlin, Germany)
There are two very different instances of Big Data: the one makes modelling very easy, because the large data show underlying regularities (all models are wrong and we we don't need them any more). This is a huge opportunity for analysis. The other are of the needle in the haystack type and make analysts' work harder (Anonymous)	How to verify analysis methods and results? How to reproduce results? (Anonymous)
The opportunity to analyse data streams in real-time moves machine learning into the direction of real-time control. This opens up tremendous applications.(Anonymous)	Algorithms that require large amounts of data will be less useful compared to algorithms that can work on smaller amounts of rare occurrence data (Anonymous)
With the design of more scalable learning algorithms and better parallel hardware, vast amount of collected data can be leveraged for myriad applications, particularly in augmented intelligence (Deepak Ajwani, Nokia Bell Labs, Ireland)	To get 'stupid' machines that don't reason to yield an answer, that cannot learn from few data (e.g., in personal conversations) (Marc Toussaint, MIT, USA)

Practice of Knowledge Discovery in Databases (ECML PKDD), ACM conference on Knowledge Discovery and Data Mining (KDD), SIAM Conference on Data Mining (SDM), and IEEE International Conference on Data Mining (ICDM).

Recently, there has been a growing list of new venues on Big Data such as the BigData (IEEE Conference on Big Data), the IEEE Transactions on Big Data, the Big Data Research Journal, Big Data at Frontiers in ICT, and many more.

Table 3 What are the most promising AI areas to benefit from Big Data? Some illustrative parts of answers from the online questionnaire

Most promising AI areas to benefit from Big Data
Computer vision, predictive maintenance, analyzing sensor data, biology, healthcare, etc. (Anonymous)
Deep networks (Sriraam Natarajan, UT Dallas, USA)
AI techniques based on machine learning can reach new levels of performance (Christian Bauckhage, U. Bonn, Germany)
First of all, I do not believe "AI" is really a thing. What is referred to as "AI" is really the algorithms invented for some very sophisticated applications, like speech, vision, robotics, etc. It seems, for example, that big-data approaches to translation of natural languages is a big win. It may be the same for many other applications that are typically referred to as "AI" (Jeffrey D. Ullman, Stanford, USA)
Eventually for better reasoning and explanation. Then real-world interactions and a good feedback loop (Sofus Macskassy, Branch Metrics, USA)
eCommerce, medicine, finance, education (as applications), data analysis areas / learning such as computer vision, NLP, robotics, making predictions of all sorts (Anonymous)
Those areas for which a lot of data exists (there are very few) and for which we have basically no understanding (there are very many) (Oliver Brock, TU Berlin, Germany)
Healthcare and transportation (driverless cars) (Nikolaos Vasiloglou, MLtrain, USA)
Customer behavior, New innovative ideas in solving the most simple to complex problems (Praveen Sameneni, Industry, USA)
There is no AI. Machine learning based on optimisation algorithms is not AI. At most we can train an optimisation model with Big Data (Anonymous)
Vision, NLP, Computational Biology (Guy Van den Broeck, UCLA, USA)
Machine translation, Cognitive systems, Recommendation systems (Deepak Ajwani, Nokia Bell Labs, Ireland)
Natural language processing has received a push by machine learning and then by the availability of really large language sources. Robotics certainly needs real-time processing of data streams. What happened to Planning, once a large part of AI? (Anonymous)

Acknowledgements The guest editors gratefully acknowledge the KI Journal editorial board for their support of this special issue; we particularly thank Christian Igel for his advice, authors for contributing to the special issue as well as the reviewers for their feedback. We also thank all participants of the online questionnaire for their feedback and acknowledge the use of Google Forms for running it. The work compiled in this special issue was partly support by the German Science Foundation (DFG), the Priority Programme SPP 1736 "Algorithms for Big Data" and the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis". Kristian Kersting was with the TU Dortmund University and the SFB 876 at the time the present special issue was initiated.