Scientific Big Data Analytics by HPC

T. Lippert, D. Mallmann, M. Riedel

published in

NIC Symposium 2016
K. Binder, M. Müller, M. Kremer, A. Schnurpfeil (Editors)

Forschungszentrum Jülich GmbH,
John von Neumann Institute for Computing (NIC),
Schriften des Forschungszentrums Jülich, NIC Series, Vol. 48,
ISBN 978-3-95806-109-5, pp. 1.
http://hdl.handle.net/2128/9842

Scientific Big Data Analytics by HPC

Thomas Lippert, Daniel Mallmann, and Morris Riedel

Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany
and JARA-HPC
E-mail: {th.lippert, d.mallmann, m.riedel}@fz-juelich.de

Storing, managing, sharing, curating and especially analysing huge amounts of data face an immense visibility and importance in industry and economy as well as in science and research. Industry and economy exploit "Big Data" for predictive analysis, to increase the efficiency of infrastructures, customer segmentation, and tailored services. In science, Big Data allows for addressing problems with complexities that were impossible to deal with so far. The amounts of data are growing exponentially in many areas and are becoming a drastical challenge for infrastructures, software systems, analysis methods, and support structures, as well as for funding agencies and legislation.

In this contribution, we argue that the Helmholtz Association, with its objective to build and operate large-scale experiments, facilities, and research infrastructures, has a key role in tackling the pressing Scientific Big Data Analytics challenge. DataLabs and SimLabs, sustained on a long-term basis in Helmholtz, can bring research groups together on a synergistic level and can transcend the boundaries between different communities. This allows to translate methods and tools between different domains as well as from fundamental research to applications and industry. We present an SBDA framework concept touching its infrastructure building blocks, the targeted user groups and expected benefits, also concerning industry aspects. Finally, we give a preliminary account on the call for "Expressions of Interest" by the John von Neumann-Institute for Computing concerning Scientific Big Data Analytics by HPC.

1 Introduction

The Helmholtz Association develops and operates large-scale infrastructures and makes them available for scientists, research groups, and communities. The effective usage of these infrastructures is ensured by scientific peer-review. These science-led processes not only guarantee the most beneficial usage of the infrastructures, but also steer their evolution and focus through the involvement of research communities in key areas of science and engineering.

The John von Neumann Institute for Computing (NIC) coordinates the scientific peer-review process for the provision of supercomputing cycles at the Jülich Supercomputing Centre (JSC). Scientists and researchers who apply for computing time on these systems are supported by a continuously growing number of domain-specific Simulation Laboratories (SimLabs) at JSC. The SimLabs offer support on a high level and push forward research and development in co-design together with specific scientific communities that take advantage of High Performance Computing (HPC) techniques in massively parallel applications.

In order to address the challenges of *Scientific Big Data Analytics (SBDA)* it is essential to enhance the central elements (cf. red box in Fig. 1) as research, development of infrastructures, and especially the provisioning through a scientific peer-review process. The user support through SimLabs and DataLabs, who have an integrating role, needs to be expanded within Helmholtz as well as to communities outside Helmholtz.

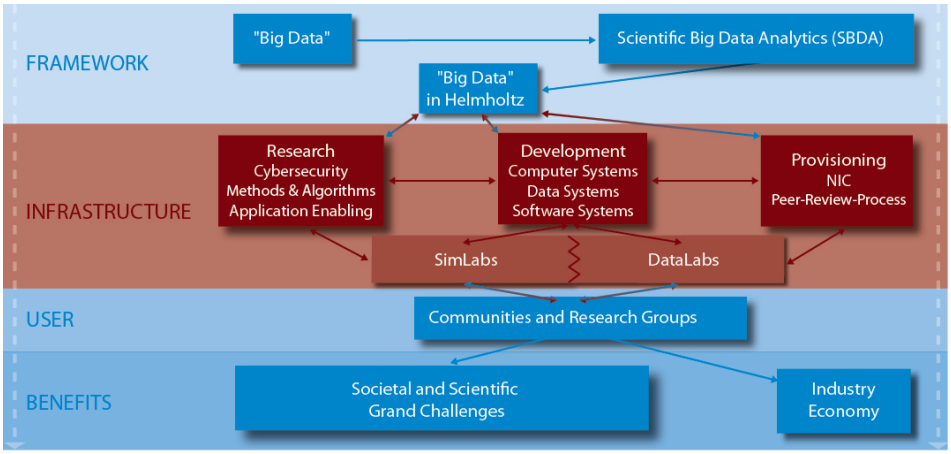

Figure 1. "Scientific Big Data Analytics" and the role of lage-scale research.

The remainder of this paper is structured as follows. After the introduction, the scene is set by clarifying the terms "Big Data" with a focus on "Analytics" as a framework concept for scientific environments in general and Helmholtz in particular. Sect. 3 describes the core building blocks of the infrastructure that are necessary to implement this framework concept, while Sect. 4 provides information on relevant communities and research groups. Sect. 5 then outlines benefits for research and industry when implementing the framework concept. Sect. 6 gives insight into the initial steps already undertaken to realise parts of the framework and its implementation using a couple of case studies as examples.

2 Big Data in Science and Engineering

2.1 On Understanding "Big Data"

Simple definitions of "Big Data" refer to data sets that become so large that it is impossible to process them using traditional methods. This definition leaves many questions unanswered such as "what are traditional methods" or "what exactly means processing in this context". This is in particular the case when one considers the difference of processing and traditional methods in science and industry. More recently, the term "Big Data" is often used to refer to data challenges with an increasing number of the (in)famous "Vs" such as "Variety", "Volume", "Velocity", "Veracity", "Variability", and most notably "Value" that altogether emphasise the fact that the "Volume" alone is not necessarily the only problem.

Economy and industry have a variety of applications for "Big Data": personalised marketing and product offerings, product development according to trends, infrastructure expansion based on usage statistics, or better logistics are only a few examples. For these purposes the customer behaviour is analysed and used to e.g. personalise advertisements, or to optimise logistics chains of supermarkets, or the infrastructure of wireless carriers. The goal is to deliver the best product or best service literally tailored for a customer at minimal cost. This is often achieved by focusing on "correlation" and related statistical

models that indicate a pattern in datasets. For product development, additional data is used, e.g. material properties, production costs, or measurement data obtained from production machines to help developing innovative products and to remain competitive in the market.

In science and research, Big Data in the context of our framework illustrated in Fig. 1 means, that based on data analysis, complex questions are dealt with, models are refined, or new models are developed. In contrast to just analysing Big Data for correlations and trends (cf. Google Flu Predictions[1]) as often done in industrial research questions, science and research need to focus on *causality* in order to prove or disprove a specific hypothesis including also the reproducibility of scientific findings (cf. "lack of Google Flu predictions"[2]).

In all scientific areas, the amount of data is increasing and the addressed problems become more complex. This is also due to the impact of "open scientific data" providing much more opportunities for data fusion by using multiple datasets in one scientific use case. Unfortunately, current state-of-the-art processing technologies can only handle a few TeraBytes. Beyond these limitations, methods of parallel data management and high performance computing need to be exploited. The largest data sets that have to be analysed in leading projects require to process several PetaBytes. It is almost impossible to handle data sets of this size in universities or research institutes. This task is reserved for large-scale research institutions and research infrastructures.

2.2 Scientific Big Data Analytics (SBDA)

The field of Big Data is strongly influenced by books such as "Big data: A revolution that will transform how we live, work, and think"[3] as well as commercial use cases such as "Industry 4.0", whereby factories of the future are automatically built and optimised by their machines that continuously measure and exchange data or industrial case studies about predictive maintenance, potentially leading to massive reductions in operating costs of large machines. The enormous momentum in media and industry leads to the development of a wide variety of technologies that constantly change and thus tends to impede stable algorithm development (cf. the relatively fast move from map-reduce[4] solutions based on Apache Hadoop[5] 1.0 , 2.0, and more recently Apache Spark[6]).

Some of our latest findings reveal that many solutions based on these recent technologies cannot offer effective algorithmic solutions required by science. Examples include classification algorithms like support vector machines (SVMs)[7] as shown in Cavallaro *et al.*[8] or clustering algorithms like density-based spatial clustering of applications with noise (DBSCAN)[9] as shown by Goetz *et al.*[10]. Due to constant change of the basis technologies and because of the significant lack of yet to be developed or commonly adopted standards for remote and distributed processing, a community agreed algorithm code (e.g. as often seen in the simulation sciences – based on the mature and strong message passing interface – such as numerical weather prediction codes[11]) is thus hard to establish. The constant changes also lead to resource usage focusing on demonstrating new technology options rather than algorithms supporting deep scientific hypothesis analysis and testing. We strongly believe that a community-led and accepted peer-review in the field of SBDA can bring a major consolidating effect to the development of this technology field in general and could push its technical maturity in particular.

Researchers are constantly distracted from exploring new and partly innovative technologies (with often rather similar functions for a variety of scientific problems) and in too

many cases lack a sound infrastructure for developing solutions beyond small-scale technology islands or simple testbeds. The notion *Scientific Big Data Analytics* comprises the work of researchers and engineers, that is based on the creation and analysis of big data sets, and thus relies on the availability of appropriately sized infrastructures, in order to be competitive in the respective domain. It is necessary to provide large-scale infrastructures to scientists and engineers of universities and research institutes, who perform projects with highest demands on storing, transfer, and processing of data sets. The provisioning should be done similar to the provisioning of HPC resources as done for the simulation sciences since many years. This constitutes a key element in the SBDA building block in our framework shown in Fig. 1.

To guarantee that the data analysis achieves the highest scientific quality, it is necessary to apply the principle of autonomous controlling of resource allocation by science in a competitive peer-review process, like it is common practice for international large-scale infrastructures. In addition, the scientific controlling of resource allocations will allow to focus on problem areas that are highly relevant for science and society. The steering process prevents that science gets lost in the details of this industry-driven topical area, with many technologies that are highly relevant only for industry (e.g. recommender engines, shopping basket association rule mining, etc.). Instead, new approaches will be developed and, subsequently have to be translated to economy and industry. Scientific approaches will mature, leading to community-approved codes to tackle an ever increasing amount of research questions.

2.3 Helmholtz-Specific Elements

The framework as outlined in Fig. 1 puts strong emphasis on the Helmholtz association being in a key position to realise SBDA. Firstly, in its role as operator of infrastructures, like supercomputers and data storages of the highest performance category, secondly, as leading organisation for research and development in this domain, and thirdly, with its internationally acknowledged competence in the scientifically controlled allocation of resources of large-scale instruments, not to forget its experts from a variety of application domains. Given these strong competences, Helmholtz is able to push the field forward and help to achieve new insights for science and society. To give an example, solving a wide variety of inverse problems from the field of Big Data as available in Helmholtz centres can actually lead to better algorithms for simulation sciences that in turn then deliver more accurate models to understand our world. In order to establish this strong productive loop between HPC simulations and more data-intensive applications, a strong foundational infrastructure is required with technologies as mature as seen today in the field of scientific computing.

Fig. 2 illustrates one of the key capabilities of Helmholtz we refer to as "full productive loop for SBDA". Big Data is already an integral part of many Helmholtz centres and most centres have demands for data storage, management, and data analysis, as it was already outlined in a Helmholtz concept paper 2012 ("Further Development of High-Performance Computing and Big Data Management in the Helmholtz Association"). The user demand is the key driver for the Helmholtz Association as a whole. But in contrast to many other scientific associations, Helmholtz can as well provide the large-scale resources that are required in order to tackle Big Data and its challenges – with sustained superior performance.

The leading organisation in the Helmholtz Association at the forefront of simulation and data intensive computing is the John von Neumann-Institute for Computing (NIC), a long-standing partnership of the research centres Jülich, GSI and DESY.

3 Infrastructure

3.1 Infrastructure Research

The analysis of scientific data sets is characterised by well established methods and algorithms from mathematics and computer science, which evolved over decades, and which now reach their limits with the exponentially growing demands of Big Data. This is especially true for the scaling of algorithms and the throughput in data analysis. Core principles, like parallelisation, can be adapted from other areas, like HPC Examples include the use of "halo/ghost areas" in smart domain decompositions or the use of parallel and scalable I/O. In this manner, several limitations can be overcome. Of course, new methods have to be developed, and in parallel, the process for scientific and engineering application enabling must be established in the infrastructures. Concrete issues will be solved in co-design with industrial partners. Further challenges in the area of data privacy and security are closely coupled to the research area cybersecurity, and SBDA will be a central use case for this. This kind of infrastructure research is a first core building block of our framework implementation shown in Fig. 1.

3.2 Infrastructure Development

The continuous development of compute and data systems is essential for the continuous advancement of the infrastructure, which simultaneously must offer stable operation and

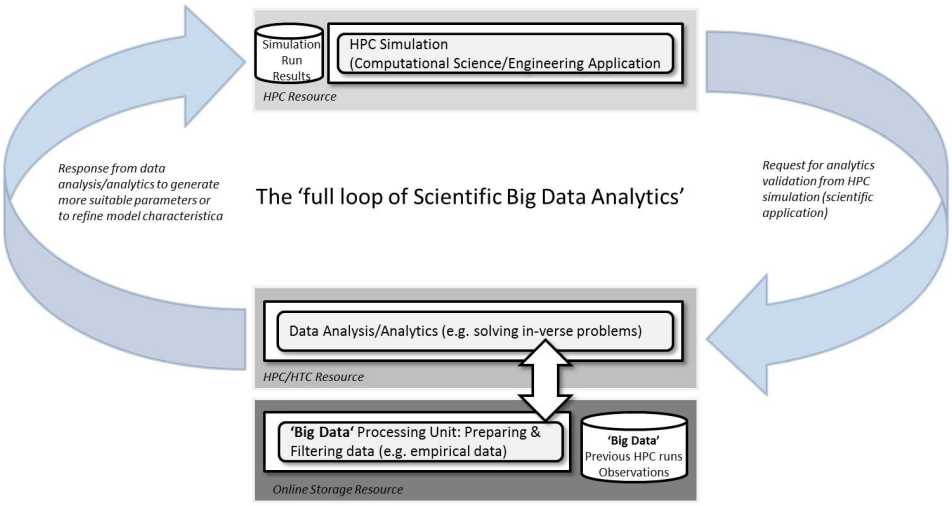

Figure 2. The "full productive loop of SBDA".

usage, in order to allow researchers and engineers to stay competitive in their domain. Big Data currently is shaped by many new approaches for software systems, which are too disjunct to offer efficient and stable usage. Several of these approaches are driven by technologies originating in the fast moving industrial sectors as outlined above (Hadoop 1.0, Hadoop 2.0, Google Dataflow, Apache Spark, or Hana and Teracotta for in-memory databases). The usability of these software systems for scientific applications needs to be evaluated and compared to well established approaches (e.g. message passing and/or parallel I/O in simulation sciences), and, if suitable, must be further developed. The development of the provided compute, data, and software systems in community and multi-disciplinary compute and data centres will be driven by the demands of the scientific applications.

3.3 Infrastructure Provisioning

In order to reach the highest scientific quality in data analysis and the greatest possible advancement in research and development in SBDA by HPC, it is crucial to establish the scientific competitive peer-review process for the allocation of resources as shown in the third core building block in Fig. 1. A well known institution for scientifically controlled allocation of resources is the Scientific Council of the John von Neumann-Institut for Computing (NIC), who allocates a major part of the computing time of the supercomputers operated by JSC to scientists in Helmholtz and throughout Germany. The NIC allocation principles served as a blue print for the allocation of computing time in the Gauss Centre for Supercomputing (GCS) as well as the Partnership for Advanced Computing in Europe (PRACE). Given NIC's strong experience over many years and its scientific advancement, it is natural to apply the NIC provisioning concept to data infrastructures and analysis resources. This kind of provisioning together with the activities in the Helmholtz programme "Supercomputing & Big Data", will promote research and development for supercomputing and data infrastructures. It is at the heart of truly innovative SBDA.

3.4 Data- and SimLabs

The efficient usage of these infrastructure will be made possible through interlocking the domain-specific SimLabs for scientific computing with domain-specific DataLabs. In this respect, it is important to realise that data analysis, and simulation are not separable. This fact is clearly outlined in one of the recent reports of US Department of Energy that suggests that "computing gets more intertwined with data analysis"[12]. Many scientific projects show that there is no border between simulation and data analysis, because the HPC-driven data modelling is mandatory for solving grand challenge problems as shown in Fig. 2 while data analysis is equally important for empirical research and theory. While numerical simulations with HPC, based on physical laws or engineering models are key of bridging experiment and theory, SBDA and its full loop provide the "breakwaters for HPC" to safeguard these bridges from collapsing while faced with an ever increasing number of big data waves.

What is more, DataLabs and SimLabs as shown in Fig. 1 have an important role in establishing SBDA as a central element with international dimension promoted by their members. In order to assemble teams with highest competence, the members need to be

recruited based on international reputation and scientific standing. More notably, they have to be enabled to link with their scientific communities in the best possible way. Ideally, as shown in SimLabs by supporting and driving community codes, key scientific and engineering data-intensive algorithms, mostly developed in international collaborations, must be supported and further developed within such DataLabs.

4 Communities and Research Groups

It is the users of the infrastructures, scientific communities and research groups, that define which methods, algorithms, and approaches, are relevant for their domain. Those methods, algorithms, and approaches are the ones to be further developed and improved and are selected as part of the infrastructure provisioning. The SimLabs and DataLabs as installed by the JSC in recent years are the liaison teams that can coordinate the development and can push the transfer of algorithms and tools into the infrastructure. In order to accomplish this objective, SimLabs and DataLabs must be integrated into the respective research communities. The latter condition is of key importance since many techniques and solutions to Big Data challenges will be developed by geographically dispersed research teams, that often lack proper communication and exchange on the synergistic level.

Analysing JSC's SimLabs activities reveals that several of them are already working on data-intensive aspects. The JSC will strongly foster data science aspects in the existing SimLabs. One can expect to see quite some impact in the following communities served by JSC's Sim/DataLabs in the next years: terrestrial systems, climate sciences, neuroscience, computational biology, molecular systems, to name a few. Examples of fields, that today demonstrate these trends, can be found in the earth sciences (e.g. hyper-spectral earth observation image data) and in the medical sciences (e.g. high-resolution medical image data).

An important driver of SBDA activities will be the availability of publicly shared datasets. Many large funding organisations (e.g. the European Commission) or major community journals are pushing in the direction to publish not only scientific findings in a journal, but also to make associated datasets publicly available. The technologies to enable full reproducibility of paper findings from such datasets are still in their infancy, but are continuously developed. These technologies and the subsequent larger public data access will profoundly influence the activities of communities and research groups in data-intensive computing in the next decades.

What is more, one can foresee an increasing number of scientific cases where the availability of data across different scientific domains will provide so far unknown research opportunities. This demands that SimLabs and DataLabs from different domains should not work in isolation from each other, but may tackle joint Big Data challenges. Examples would be in the area of earth sciences where climate challenges directly or indirectly influence modelling and data-intensive computing of terrestrial systems.

5 Benefits

5.1 Grand Challenges

The framework sketched for large-scale research including infrastructure activities and users as shown in Fig. 1 will allow to tackle societal and scientific grand challenges in

innovative ways (e.g. solving inverse problems jointly with HPC simulations). In an initial survey throughout the Helmholtz Association, 14 domains were identified including one or more Helmholtz centres each, which encounter major challenges in their scientific data analytics. Among others those domains are climate research, epidemiology, environmental diseases, water research, or polar and marine research. Additional domains and research areas will be identified in the context of research groups at universities and research institutes.

It should be emphasised that methodologies, algorithms, tools, and analysis procedures originating in fundamental research very often are the basis to support solutions of grand challenges pursued in the applied sciences. The interlinked SimLab-DataLab framework is a catalyst to realise this transfer. Examples from energy are areas like Smart Grids, electric mobility, and battery research, in health sciences the DataLab Neuroscience supports, for instance, the development of the three-dimensional model of the human brain.

5.2 Benefits for Industry and Economy

Industry and economy in general analyse data in order to solve issues different from scientific research. However, methodologically, both sides often need to use the same algorithms and tools. The activities in research and development will influence the activities in industry and economy through improved algorithms and methods developed in joint research. Small and medium enterprises can benefit, because they can overcome scaling issues they face while expanding without the need to employ their own data science teams. In addition Sim- and DataLabs are supposed to also support industry and economy through targeted contract research for a limited period, coordinated by a so-called industry hub currently put up by the JSC.

Moreover, SBDA activities have been started such as the German initiative "Smart Data Innovation Lab"[13]. Here, data-intensive computing challenges are tackled in the four domain areas energy, industry 4.0, smart cities, and personalised medicine. Research organisations in Germany with quite a few companies formed this collaboration in order to jointly tackle problems in industry and economy. While the infrastructure is just getting started to be used by selected scientific cases, one can already observe many options for improvements taking into account major arguments of this paper. Most notably, the current infrastructure is driven by High Throughput Computing (HTC) needs and technologies while solutions based on HPC are equally important to be considered in order to consistently provide benefits for industry and economy, not to forget the quality assuring principle of peer-review.

6 Scientific Big Data Analytics by HPC – Expression of Interest

In order to gain better insights into the demand by communities and requirements for SBDA, the JSC has performed an initial step towards implementing principles as proposed in this paper. The importance of data analytics, management, sharing, and preservation of very big, often heterogeneous or distributed data sets from experiments, observations and simulations is of increasing significance for science, research and industry. This development has been recognised by many research institutions, among them leading HPC centres. They want to advance their support for researchers and engineers using SBDA by HPC.

At the beginning of 2015, the John von Neumann Institute for Computing (NIC) has, for the first time, invited Expressions of Interest (EoI) for SBDA projects using HPC to identify and analyse the needs of scientific communities. The goal is to extend and optimise the HPC and data infrastructures and services in order to provide optimal methodological support. As described above, a peer-review was performed on the EoI submissions in a similar manner as known from HPC calls. The EoIs have been submitted from various domains. They clearly demonstrate in which areas SBDA by HPC is of major importance:

EoI submissions have been received in the field of Biology where HPC knowledge mining of molecular structure data was considered to be an interesting challenge for SBDA. These contributions already showed how intertwined approaches for SBDA, using machine learning and statistical data mining on the one hand, can be successfully combined with HPC technologies on the other hand. This fact is further underpinned with evidence by two more EoI submissions related to HPC simulations: One is on the statistical analysis of high-Rayleigh number turbulent convection data while the second outlines Big Data challenges and opportunities in the area of turbulence databases from direct numerical simulations.

Two EoI submissions in the area of earth sciences are related to each other. Big Data challenges from the SimLab Climate Science raises the demand for a joint atmospheric data repository and processing unit in order to advance in SBDA while the SimLab Terrestrial Systems outlines challenges for the implementation of a high-performance big data storage and analysis framework for large-scale earth science simulations.

A further EoI submission in the field of neurosciences raised the request for the availability of more parallel and scalable SBDA algorithms such as deep learning and unsupervised clustering for analysis of cellular cortical structures in the human brain. While this contribution was focused on statistical data mining aspects, a further EoI contribution described SBDA challenges to solve inverse problems by big data evaluations for a high complexity energy meteorological *in situ* analytics application. Finally, turning to particle physics, a lattice QCD submission is highlighting several Big Data challenges in the post-processing steps of a large-scale simulation project.

7 Conclusions

The first experiences with SBDA by HPC in the EoI call of the John von Neumann Institute for Computing tell that there is an urgent demand for an SBDA framework concept as provided in this paper. The Helmholtz infrastructure as outlined promises to implement SBDA by HPC in a sustainable manner, with SimLabs and DataLabs as well as autonomous scientific peer-review by community representatives as a cornerstone.

Acknowledgements

We commemorate the late Walter Nadler who was invaluable in forming the EoI call for SBDA by HPC in the John von Neumann Institute for Computing. Walter, we will not forget you!

References

1. J. Ginsburg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinksi, L. Brilliant, *Detecting influenza epidemics using search engine query data*, Nature 457, 2009.
2. D. Lazer, R. Kennedy, G. King, and A. Vespignani, *The Parable of Google Flu: Traps in Big Data Analysis*, Science Vol (343), 2014.
3. V. Mayer-Schoenberger and C. Kenneth, *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt, 2013.
4. J. Dean and G. Sanjay, *MapReduce: simplified data processing on large clusters*, Communications of the ACM 51.1 (2008): 107-113.
5. D. Borthakur, J. Sen Sarma, J. Gray, *Apache Hadoop goes realtime at Facebook*, Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, ACM, 2011.
6. S. Ryza, U. Laserson, S. Owen, J. Wills, *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*, O'Reilly Media, ISBN 1491912766, 2015.
7. C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning, vol. 20(3), pp. 273297, 1995.
8. G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, A. Plaza, *On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Issue 99, pp. 1-13, 2015.
9. M. Ester *et al.*, *A density-based algorithm for discovering clusters in large spatial databases with noise*, Kdd. Vol. 96, 1996.
10. M. Goetz, M. Richerzhagen, G. Cavallaro, C. Bodenstein, P. Glock, M. Riedel, J. A. Benediktsson, *On Parallel and Scalable Classification and Clustering Techniques for Earth Science Datasets*, Sixth Workshop on Data Mining in Earth System Science (DMESS 2015), International Conference on Computational Science (ICCS), Reykjavik, 2015.
11. J. Dudhia *et al.*, *The weather research and forecast model: software architecture and performance*, Proceedings of the Eleventh ECMWF Workshop on the Use of High Performance Computing in Meteorology, 2005.
12. US DOE ASCAC Report, *Synergistic Challenges in Data-Intensive Science and Exascale Computing*, 2013.
13. Smart Data Innovation Lab (SDIL), German Initiative,
    Online: `http://www.sdil.de`