A roadmap for Big Data at KTH

The KTH Big Data Working Group

December 19, 2014

Executive summary and Conclusions

Scope The *KTH Working Group on Big Data* was formed in 2012, and has been supported by the KTH ICT and Life Science platforms.

The term Big Data can mean different things in different contexts and communities. We have chosen to be inclusive, and to include in the description KTH activities that can be called either (*i*) Data Analytics, or (*ii*) Data Engineering or (*iii*) Data Generation (or some of each). Recommendations and conclusions of this report are however primarily written from the perspective of (*i*) Data Analytics. Wherever possible we have referred to the recent report *Frontiers in Massive Data Analysis* from the National Research Council of the National Academies USA (2013) which is becoming the international benchmark [48].

Conclusions KTH, as many research-oriented universities faces Big Data challenges. Many KTH research units generate Data which are Big, and for many there are problems to analyze and fully make use of the data.

KTH has great potential strengths in fundamental research of relevance to Data Analytics, represented at several KTH schools.

KTH has great strengths in Data Engineering and High-Performance Computing-related aspects of Big Data, well integrated into ongoing high-profile projects such as the ongoing upgrade of the KTH supercomputing resources, work done on massively large data resources at KTH ICT, work done in the Swedish e-Science Center (SeRC), and other initiatives and efforts.

KTH lacks integration between the above three aspects of Big Data (Data Analytics, Data Engineering, Data Generation). There is also a lack of educational offers such as e.g. MSc programs.

Background and Previous work The *KTH Working Group on Big Data*, in this document for short also referred to as "the Big Data group" or the "the group", currently consists of Erik Aurell (CSC, coordinating), Gunnar Karlsson (EES), Timo Koski (SCI), Mikael Skoglund (EES) and Ozan Öktem (SCI). The group has been awarded two grants (2013, 2014) from the KTH ICT Platform to prepare a Whitepaper and a Roadmap, and to organize activities such as seminars, workshops, etc. The group has also been supported by the KTH Life

Science Technologies platform.

The Whitepaper, prepared in collaboration with Prof Scott Kirkpatrick, KTH Dr HC, was delivered on January 22, 2013, and has been widely circulated inside KTH. Progress towards the project objectives have been reported to the KTH ICT and Life Science Technologies platforms continuously, and are summarized in Appendix to this document.

Preparations of the Roadmap The Big Data group has based its work on the Roadmap on participation in workshops and conferences (detailed below), on an on-line questionnaire, and one separate interviews in several of KTH Schools. Preliminary results were presented by M Skoglund to the KTH ICT platform Steering Group on May 9, 2014.

Organization of the Roadmap document Overall, this Roadmap is structured into an introduction to Big Data, an inventory of Big Data research currently done at KTH, a SWOT analysis of future Big Data at KTH, the recommendations of the group, and Appendices. The detailed lay-out of the document is given below.

Contents

1 Introduction to Big Data

The last decade has seen an unprecedented acceleration in the amount and complexity of data collected and stored, in all spheres of human activity. The total world production of digital data will soon reach Avogadro's number (6×10^{23}) bytes per year, while several large projects in science by themselves already produce petabytes (10^{15}). To reduce such Big Data to information which is actionable, and which can be used to advance modern societies, and modern science and technology, poses fundamental challenges to data analysis and statistical methodologies [48].

Data Analytics means to address the problems Big Data poses to statistics, the science of how to analyze and draw conclusions from data. A main assumption of classical statistics that the number of observations n is sufficiently large, and in particular larger than the data dimension p. In Big Data p is quite to the contrary often as large, larger, or even much larger than n. For example, in a global gene expression measurement comparing two biological samples n is two, but the number of (human) genes is about $20,000$. A naïve use of this data would be to take all the gene variations as meaningful (it is well understood by practitioners that this is not the case), while a caricature of conservative statistics practice would be to use the data to gain information about only one gene, because restricted to one gene indeed $n > p$ $(2 > 1)$. In Big Data practice one often faces the unsatisfactory consequence that while the total number of data items (np) may have grown enormously, the amount of new reliable information has grown only modestly. Closing this gap is the main problem of Data Analytics.

During the past 50 or so years, several new academic disciplines that can be said to also deal with problems of data analytics have emerged, ultimately tracking their origins to applied mathematics and mathematical statistics. When sorting under Electrical Engineering, relevant sub-fields include statistical signal processing and information theory. Similarly, the term machine learning is often used for related activities sorting under Computer Science. In this report, when we say 'statistics' we also mean to cover such derived fields.

Data Engineering comprises all aspects of the physical systems that support the storage, handling and communication of the data and the computations carried out based on the data. In academia, such activities usually sort under Electrical and Computer Engineering. At KTH, the Schools primarily involved would be EES, ICT and CSC (PDC).

Data Generation (and or Acquisition) is the process of measuring and collecting data to be engineered and analyzed. Elsewhere in this report we highlight KTH work on problems related to instrumentation, sensoring and measurements. Activities at KTH also depend on externally generated or collected data. Exampled include car traffic data collected from Stockholm taxi cabs, or various biological data accessed via collaboration with Karolinska.

Complexity of data Data can be complex in a variety of ways: they can be voluminous, noisy, heterogeneous, indirect, multi-source, and collected over a range of temporal and spatial scales. Harnessing the power of such data through new methods that combine data and simulations offers tremendous opportunities to gain understanding, optimize performance, and optimally control dynamic systems.

In all of these settings, complex data provide a new opportunity to create predictive models that embody both physical principles and rich data. For example, data may be used to adapt the model, or to characterize uncertainties due to a lack of fundamental understanding as in *e.g.* underlying physics. An equally important aspect is the use of models for adaptive sensing and experimental design. Critical challenges include: how to characterize the uncertainty in large and complex data sets, particularly when their provenance is unknown or when the data pass through a number of processing steps (typically involving additional models)? How do we reconcile information from multiple observational modalities? How do we extract pertinent information from the large volumes of data?

Complexity of models Conventional sampling methods have been able to solve statistical inverse problems for simple models. It is a real challenge to extend these methods, and develop entirely new ones, to solve statistical inverse problems governed by large-scale complex models, for instance in the form of Partial Differential Equations (PDEs), derived from physics, population genetics and many other fields.

Complexity stems from a wide range of spatial and temporal scales represented in the model; the coupling of multiple mechanisms; the hierarchical nature of a model; heterogeneity of models (e.g., continuum and atomistic, discrete and continuous, structured and unstructured); stochastic model components; severe nonlinearities; and so on. Extremely large model state dimensions (to billions of unknowns and beyond) are a ubiquitous feature of such problems, and massive parallelism is therefore essential. To allow for storage and processing of raw data, various compression mechanisms may be necessary, including lossy processing (involving e.g. image compression).

Despite these enormous challenges, the critical need to quantify uncertainties in realistic models of natural and engineered systems from observational data requires the development of a new generation of statistical inverse methods able to cope with model complexity. How can we construct reduced models that can capture the pertinent statistical features but are much cheaper than the original high-fidelity models? How can we characterize uncertainty in model structure?

High dimensionality of model input Many frontier statistical inverse problems are characterized by uncertain fields for model input, which when discretized give rise to very high-dimensional parameter spaces. Examples include uncertain initial or boundary conditions, heterogeneous sources, heterogeneous model coefficients, and domain shapes.

As with expensive computational models, high-dimensionality is prohibitive with conventional statistical computing methods. Still, uncertain fields abound in models for complex phenomena so we must develop methods to address the underlying challenges. These include how to devise Markov-Chain Monte-Carlo (MCMC) (and other) sampling strategies that scale to very large parameter dimensions; how to specify priors that are consistent with the discretization; how to adapt parameter fields to the information contained in the data; and how to take advantage of hierarchies of discretization.

A common feature of such problems is that, despite the increasingly large volumes of data available for inference of models, the data are often informa-

tive about a low-dimensional subspace of the parameters: can methods be developed that invoke the expensive computational models only in those subspaces? Principal Component Analysis (PCA) and Classical Multi-Dimensional scaling (CMDS) are two traditional methods based on linear data models. However, they are not effective if data do not well reside on sub-planes. Recently, many methods of nonlinear dimensionality reduction, also called manifold learning, have been developed based on nonlinear data models, which assume each observed high-dimensional data resides on a low-dimensional manifold, see Appendix B, Section B.4 for a brief survey. In general, a key challenge is the analysis and implementation of how to preserve crucial information when a process of dimensionality reduction is carried out.

Model learning One recently popular approach is to exploit sparsity, that is an a priori assumption that in some basis, though not known at the outset of the analysis, data dimension is actually not high [compressed sensing]. The opposite approach is to assume that the data, while intrinsically high-dimensional, is in some sense typical. This notion can be made precise by positing that the data has been generated as random draws from an unknown probability distribution. To understand the data then means to deduce the probabilistic model by which they have been generated, and the information to be gained are the relations between the variables encoded in the model. For data analyzed in batch two simple ideas have here recently shown to be very powerful: (i) learning models with many more than n parameters and (ii) learning such models approximately, as maximum likelihood is computationally unfeasible. The first idea calls for regularization, as otherwise the problem is ill-defined, which can be given a Bayesian interpretation as learning with weakly informative priors [29], an a priori assumption that the world, in the large, is somehow simple. The second idea has already spawned a large literature surveyed, among other places, in the monograph by M. Jordan and M. Wainwright [55].

Errors There are several sources of error in the process from data generation to inference. Errors and noise in the measurements themselves, lossy compression of data, mistakes in model assumptions, and (unknown) failures in algorithm executions. It is important to realize that 'more data' will often lead to 'more errors.' That is, rather than capitalizing on having more than enough data to perform successful model fitting, there will often be expectations to reach additional conclusions based on the data. It is quite possible to turn data into something resembling knowledge, and often only a human can decide whether the result of the inference is reasonable, see [48] and Appendix B below.

2 Current Big Data research at KTH

The inventory of this section has the objective to provide a background to the discussion of future research questions at KTH in the Big Data area given in Section 3 below. The inventory is based on the workshops and conferences on organized by the Big Data group and documented below in Appendix A, on direct contacts with researchers at the different schools, institutes and centres known to the group, in a few cases on direct contacts with the respective school management, and in the case of CSC school by an on-line questionnaire.

2.1 Big Data activities at KTH Schools

ABE The School of Architecture and the Built Environment addresses matters concerning how cities, buildings and infrastructure will be designed and built. The school has several application domains where advanced data analysis is used, such as traffic and transportation studies, environmental monitoring and energy consumption. The School hosts a research center with direct use of big data, namely the Centre for Traffic Research(CTR). The Department of Transport Science is central research unit for analysis and modeling of traffic data to monitor, control and facilitate the use of the transport system. Further application areas are found in geodesy and geoinformatics pertaining to collection, management, visualization, analysis and presentation of remote-sensed geospatial data. The Department of Civil and Architectural Engineering use and develop computational science methods.

BIO KTH School of Biotechnology is in at forefront of modern high-throughput biology with flagship projects of national importance such as *The Protein Atlas*. Science for Life Laboratory, presented separately below, is primarily, as a KTH unit, associated to BIO. BIO is one of the main centres of Data Generation at KTH, and can be expected to be so for many years to come. BIO has also recently hired competence in Bioinformatics covering aspects of Data Analytics.

CHE The KTH school of Chemical Science and Engineering (CHE) has both applied and theoretical material-related research. Several of the current research projects are related to energy (solar cells, fuel cells, carbon capture, insulating nano-composites), environment (biodegradable plastics, biofuel, cellulose-based materials) and nano-technology (graphene, nano-composites, surface modifications). Micro- and macroscopic phenomena like crystallization, electrical conduction, polymer/solvent interactions, porous media diffusion, nano-particle dispersion, heat transfer, material degradation and mechanical deformations are simulated at the school.

CSC Research at KTH School of Computer Science and Communication ranges from theoretical computer science to man-machine interaction. The KTH supercomputer centre PDC is described separately below. CSC has leading competence in complexity, security and privacy-preserving computation, of foundational importance to Data Engineering. CSC addresses aspects of Data Engineering and Data Analytics in many several application areas such as robotics, audio-visual processing, visualization, neuroinformatics and computational biology, and runs a master course on machine learning. CSC has ongoing research using deep learning of representations for image classification based on large scale labeled or unlabeled sets of visual data.

EES The School of Electrical Engineering (EES) hosts extensive activities related to methods and systems for data analytics. One core discipline is statistical signal processing, with basic research on statistical inference, algorithms and machine learning, with applications to, e.g., positioning, wireless networks and multimedia. A related activity is that in information and communication theory, with basic research on transmission, networking, coding and compression. The

School also has a very strong environment in optimization, systems and control, with applications in wireless systems, process control and intelligent transportation. The School's general expertise in computer and communication networks covers queuing theory, network optimization and resource allocation, management, and network security. Big data analytics is also crucial for the School's activities in electric power systems and the smart grid, and in software-based enterprise engineering.

ICT The School of Information and Communication Technology has a broad research charter that ranges from material and nano physics to software and computer systems. The systems-oriented departments are Communication Systems and Software and Computer Systems. The latter is the KTH stronghold for research in fundamental aspects of software and computer systems with a focus on cloud computing, service computing, social networks, time aware systems, data science, and applied AI. The department collaborates closely with SICS Swedish ICT in SCALE to address the handling of vast amounts of data from large number of connected users and devices by scalable processing in the cloud and scalable algorithms. Big data sets from genomics and the Internet of things require secure, scalable storage infrastructure as well as secure data-intensive computing support. Novel programming technologies are needed to manage the increased system complexity.

ITM The School of Industrial Engineering and Management (ITM) focuses on developing new products, material and production processes in a sustainable fashion as concerns technical management, financial profitability and the environment. The core knowledge areas of the School – industrial design and innovation, product and production development, materials development, micro- and nano-technology, industrial economics, organisation and management – are all actual or potential application domains of Big Data. Materials science in particular is data-rich involving *e.g* optimization of parameters depending on chemical composition and structures on many scales. The School hosts a large number (14) of competence centers and is strongly represented in the strategic research area XPRES – Initiative for excellence in production research.

SCI The School of Engineering Sciences (SCI) is the largest school at KTH, with research areas stretching from fundamental mathematics to theoretical and experimental physics and to applied mechanics. Data Acquisition and Data Engineering issues are at the forefront of several research directions at the school, often pursued in collaboration with international universities, research institutes and industry. An example from Nuclear Physics is detailed elsewhere in this report.

SCI Department of Mathematics is the largest KTH resource for Data Analytics and its supporting sciences, with research and education connected to the topic is carried out in imaging, financial mathematics, statistics (high-dimensional classification, Bayesian machine learning and biostatistics), topological data analysis, dynamical systems, scientific computation (numerical analysis) and optimization. The Department also has a wide range of external industrial contacts with companies such as RaySearch, Scania CV, Ericsson, Folksam, Skandia Liv and Algorithmica. The Department's MSc programme in

Computational and Applied Mathematics is one of the largest at KTH in terms of number of enrolled students. Data Analytics is also carried out elsewhere at SCI, such as using information theoretical tools and other methods to extract information from very large biological data sets.

STH Research at the School of Technology and Health concerns the interaction between human activities, technology and environment for medical technology, logistics and ergonomics in health care. Data analysis methods and systems are primarily relevant to research in the following units: Structural Biotechnology, Medical Imaging as well as Medical sensors, signals and systems.

2.2 Activities at KTH Institutes and Centres

We here briefly describe activities at KTH Institutes and Centres, with an emphasis on Data Analytics, that were prominently visible at the events organized by the group.

ACCESS The ACCESS Linnaeus Center coordinates research on a broad front to facilitate tomorrow's networked systems, with applications in intelligent transportation systems, wireless communications, active buildings, and the smart power grid. The Center advocates a holistic approach with a foundation based on signals, systems and control theory, optimization, networking and theoretical computer science.

Nordita The Nordic Institute for Theoretical Physics (Nordita) is an international organisation under the Nordic Council of Ministers, since 2005 is located in Stockholm and where the permanent faculty is affiliated with KTH, Stockholm University or Uppsala University. All current Nordita permanent faculty are recipients of ERC Advanced Grants and incoming professor John Wettlaufer is the recipient of a VR Excellence Grant, as is recently appointed Director Prof K Freese.

Nordita conducts a large amount of interdisciplinary research on the interface between Theoretical Physics and other fields, both directly and through its visiting programs. Recent areas of of relevance in Data Analytics include financial and social time series and the analysis of the causal structure. Finding efficient ways of how merge and analyze such data is a big issue in further development of more stable financial and economic systems on different time scales.

PDC PDC provides large scale computing and storage resources to Swedish academia and has a dedicated team working on cloud computing and Big Data Engineering. This team provides access to cloud computing resources and investigates programming models and architectures for the use of massive data, including developing access models towards high-performance computing environments for analysis of Big Data, data analysis as a service, privacy thread modeling and security frameworks. PDC is leading partner in a number of Swedish, Nordic and European cloud projects, and works closely with the Swedish e-Science Research Centre (SeRC) on data-intensive problems, particularly in life science and in traffic analysis.

SciLifeLab Science for Life Labs in Stockholm and Uppsala is a national resource dedicated to high-throughput biology, from genomics and proteomics data to high-throughput 3D/4D fluorescence microscopy imaging data.

SciLifeLab is a major Data Generation centre, both nationally and internationally. With the imminent addition of new cutting-edge sequencing equipment, SciLifeLab will have a theoretical capacity of ca 1.7 PB of sequence data per year, corresponding to 17,000 whole human genomes. SciLifeLab has a very significant effort in Data Engineering, and a recently established facility for Bioinformatics long-term support provides significant resources in terms of Data Analytics.

Although DNA and RNA sequence data dominates the output of SciLifeLab in sheer bytes, the different facilities produce large and extremely heterogeneous data sets ranging from molecular dynamics simulations, protein sequence and levels, chemical screening to microscopy imaging and many other techniques. In addition, there is a vast array of publicly available databases that can provide essential contextual information. A significant amount of analysis at SciLifeLab is performed on different kinds of environmental samples, where different analysis and integration issues come into focus.

SeRC The Swedish e-Science Research Centre (SeRC) is a strategic research area funded by VR and formed around the two largest high performance computing centers in Sweden, the National Supercomputer Centre (NSC) at LiU and PDC at KTH, as well as the visualization centers Visualization Center C in Norrköping and the VIC studio at KTH. The university partners of SeRC are KTH, Stockholm University, Karolinska Institute and Linköping University.

SeRC is one of the main Swedish centers for Data Engineering as the leading provider of high-performance computing infrastructures, especially in the area of data management, visualization, and hardware for distributed and parallel computing. SeRC foresees a continued expansion into Data Engineering with a focus on support for extremely large data sets. SeRC has strong expertise in equation based modeling centered around computational methods with an emphasis on simulation.

2.3 Other stakeholders

We here describe important stakeholders which are external to KTH, or which are internal but do not (mainly) associate to one institute or centre, or to one KTH School as a whole. The list is not intended to be exhaustive.

Financial and social time series: Financial markets generate huge amount of data, e.g. stock prices and trading volumes on the minute scale. Being in essence probabilistic objects, financial markets reflect investors' moods and expectations, hence they are tightly connected to the news. Social networks provide enormous amount of data per day which can be used for related analysis, e.g. mood analysis basing on posts. Finding efficient ways of how merge and analyze these data is a big issue in further development of more stable financial and economic systems on different time scales.

Data from epidemiology, clinical medicine and genomics: Recent biotechnological advances in next-generation sequencing, tandem mass-spectrometry, high-throughput screening, cellular imaging, etc, are enabling more in-depth insights into the individual disease processes by generating millions of data points for each subject under analysis. Efficient integration of complementary information sources from multiple levels, including tissue characteristics from cellular imaging, the genome, transcriptome, proteome, metabolome and interactome, have the potential to greatly facilitate the discovery of true causes and states of disease in specific sub-groups of patients sharing a common genetic background. The behavior of big data versus classical correlational effects, such as Simpson's paradox, is important, in particular if they can be transferred to causal effects. New systematic methods to infer causal connections over time by means of massive data sets will be a major focus of research.

Imaging: For medical imaging there is Center for Medical Image Science and Visualization (CMIV) in Linköping that also works with visualization. Another stakeholder are the county councils (landstigen) and hospitals. These handle big data, *e.g.*, in the form of hospital records. For physics and material sciences we have Max IV and the future spallation neutron source, both in Lund. These are expected to generate big data in the context of material characterization.

Telecommunication network traffic patterns: In wireless telecommunication networks, the real time traffic patterns can have a major impact on the end user quality of experience and on the utilization of system resources. Understanding, modeling and predicting real time traffic patterns is key for creating next generation wireless communication networks - using less energy, having higher resource utilization and providing better end user performance. Telecommunications networks also pose industrial grade requirements on algorithm design, computational complexity, and system robustness. There is also an increasing business for companies like Ericsson in developing and providing services that capitalize on the availability of massive data collected in the use of the communication network.

Intelligent transportation systems: Monitoring and directing the flow of vehicles in and around a city like Stockholm is a challenging task that will benefit greatly from upcoming big data principles. There are also related applications to the transportation of gods via trucks, on much larger-scale road networks. Here new principles supported by data analytics include platooning of trucks for reduced traffic congestion and fuel drain. Other examples are monitoring of driver's behavior including distractive driving patterns in swarms of (road) vehicles benefits from big data analytics for increased traffic safety and reduced number of accidents. Risk assessments based on data analytics can here improve the accuracy in *e.g.* insurance policy calculations.

Data from Nuclear Physics: Big Data first appeared in the physical sciences dealing with large questions such as the origin and nature of the universe and its constituents. In the 2013 Whitepaper the group highlighted Big Data from Astronomy & Astrophysics. Another area, much more broadly represented at KTH, is Nuclear Physics, the primary goal of which is to understand the origin,

evolution and structure of matter that interacts via the strong force, constituting more than 99% of the visible matter in the universe. KTH Nuclear Physics is strongly involved in several large international projects to prepare experiments that will generate enormous amounts of data. One example is AGATA (Advanced Gamma Tracking Array)1, a European project with large KTH involvement to develop and operate the next generation gamma-ray spectrometer for nuclear physics experiments at the new accelerator infrastructures, based on the technique of gamma-ray energy tracking in electrically segmented high-purity germanium crystals. This system relies on technical advances such as the development of encapsulated highly segmented germanium detectors assembled in a triple cluster detector cryostat, an electronics system with fast digital sampling and will require data processing at the high rate of more that 1Tbit/s in the full system. The first phase of AGATA is operational, and already today AGATA uses high-performance computing (HPC) resources at the level of 0.7 Mcoreh/y, 0.1 Pbyte/y and handles data at the rate of 1 Gbit/s. AGATA also has a Data Analytics aspect as a technical key ingredient is to determine the locations of gamma-ray interactions inside germanium crystals from observations of electron/hole pairs arriving at electrodes on the crystal surface. This can be formulated as an on-line linear inverse problem on streaming data, $Me(t) = s(t)$, where $e(t)$ is the desired result, a vector consisting of all interaction points and their interaction energies, $s(t)$ is a composite signal meta-vector sampled with 14-bit resolution at a rate of 100 MHz and containing all non-zero signals from the affected electrodes, and M a transformation matrix. The problem is difficult because one needs to have both a good understanding of M, which depends on crystal and detector geometry and many other properties, and because the data to be treated arrive at such high speeds. Many other current and future projects in Nuclear and Particle Physics can be expected to pose analogous challenges to both Data Engineering and Data Analytics.

3 Future Big Data at KTH

The purpose of this Section is not to propose or put forward concrete research questions – obviously a futile exercise, in any quickly developing field – but to discuss the Big Data on a strategic level for KTH. The presentation is for clarity organized as a traditional SWOT analysis (strengths, weaknesses, opportunities and threats), where all concepts are to be understood relative to KTH's peers and potential competitors.

The objectives of KTH are to carry out research, education and interaction with society. The discussion under each heading is therefore taken in this order. Recommendations are given separately in Section 4 below.

3.1 Strengths

KTH has strengths in research on all three levels of Data Generation, Data Engineering and Data Analytics. In Data Generation these can typically be identified with areas of research excellence of KTH, such as, but not only, high-throughput Biology (as highlighted in several places in this report); in Data Engineering they can be identified by structured and well-funded long-term

efforts, often on the national level beyond KTH, and tied to important infrastructures as represented by the supercomputing centre PDC. The strengths in Data Analytics are more broadly distributed reflecting the fact that Big Data research is carried out in many of the schools. Two long-time focal points are Big Data Analytics research is the Dept of Mathematics, with the broadest base of fundamental research, and Nordita, with the strongest concentration of excellence in mathematical and adjoining sciences in the country.

3.2 Weaknesses

While KTH has great strengths in Data Generation in some areas, this is not uniformly so, and in some other areas Big Data is not available, or is not being used. As remarked elsewhere in this report, too often research, particularly in theoretical fields, is based on standard data used in a community for a long time (perhaps already outdated), and sometimes it is based only on simulated data. In the Big Data age this is a competitive disadvantage.

Data Engineering requires very large data storage facilities as well as other super-computing resources to compete on an international level and the situation can change quickly. KTH has potential weaknesses if the facilities are not kept at the state-of-the-art levels.

The main weakness in Data Analytics research is organizational fragmentation. KTH does not have a Department of Statistics where Big Data research would belong, seen in an international perspective [1]. Instead the activities are dispersed over the Schools, as surveyed above, and the effort to coordinate is larger and it is more difficult to obtain critical mass for the research to have impact.

Our education programs are organized by the Schools, which is a problem for new areas, whether they stem from cross-disciplinary research topics or from new technologies. Many groups could contribute, but in each separate School the mass of the new area is (or could be perceived to be) sub-critical. Big Data is here a case in point. As detailed elsewhere in this report, the first MSc courses in Big Data have already been launched in the US, and European alternatives have been announced this year. There is nothing similar at KTH, which is a weakness.

For interaction with society, see above.

3.3 Opportunities

KTH has strengths in all three aspects of Big Data covered in this report, and Sweden has a industrial base for Big Data applications, with many large industrial firms in the Stockholm area that have close links to KTH (Ericsson, Scania, other). Sweden also already houses one of the largest commercial data centres in the world (Facebook in Luleå). We also have successful companies, such as Klarna and Spotify, that rely on Data Analytics for their businesses. For these and other reasons, KTH has an opportunity to establish itself as

[1] Out of the top ten universities world-wide in the 2014 Shanghai ranking six have a Department of Statistics (Harvard, Stanford, Berkeley, Columbia, University of Chicago, Oxford). Cambridge has a joint department of " Pure Mathematics and Mathematical Statistics" (like KTH), while MIT, Princeton and Caltech do not have Statistics in their line organizations.

the natural center for Big Data research in Sweden, and an important center world-wide.

KTH is the natural candidate to organize undergraduate and graduate-level education in Sweden, owing to the breadth of research pertaining to Big Data as well as its size and international renown. It is also well positioned to take the lead in interacting with society. KTH has excellent vehicles to gain additional leverage on the European level through its partnership in the EIT KICs, especially InnoLife, where Big Data for health and aging is a priority.

3.4 Threats

A main threat in general is that the weaknesses will not be duly addressed and thus putting KTH at a risk of falling behind on developments. With the tremendous pace, it might then be hard to catch up later. A consequence might be that Swedish industry would not see KTH as a strong partner and hence seek collaboration elsewhere.

As in any fast-moving area there is a first mover advantage. For instance, if other Swedish universities launch educational initiatives earlier than KTH, the playing field will be leveled with respect to visibility and KTH might miss out on domestic funding opportunities.

4 Recommendations

4.1 Research and Outreach

In the line-organization at KTH, activities are organized in Schools and Departments, and research activities are in the main funded by external competitive grants. The recommendations of the group are therefore limited to the organizational question, which naturally impacts also on education.

We recommend that KTH creates a Center for Big Data to integrate the activities in different levels and parts of KTH. One important objective of such a center is outreach and to raise awareness inside and outside the organization, in continuation of the activities of the group to date. Alternatively, on a more ambitious level, one could also consider a separate KTH platform on Big Data, as the area in general cuts across all current KTH platforms.

Furthermore, as this strongly impacts the educational offering, it is the view of the group that KTH would also benefit from having a unit in the line-organization being the natural home for Big Data questions. In many leading universities around the world this unit is the Department of Statistics. At KTH, an alternative could be a Department of Data Engineering and Data Analytics. In Appendix B to this document we include an extended survey of Data Analytics, its current state and main methods.

4.2 Education

The demand for data scientists today only in Stockholm can be assessed, e.g., by a glance at the the web page

`http://se.indeed.com/Data-Scientist-jobb`

At the date of writing this report there are 10 announcements in data science. Let us quote ad verbatim a section in one of these:

> As a Data Scientist your main responsibility is using available data to improve our products and the customer experience, both driven by the business needs and self-driven research and data mining.
>
> It is important that you have good statistical understanding, strong analytical insight and some programming background. Previous experience and/or proven interest in artificial intelligence, machine learning, data mining, natural language processing, predictive analysis etc. are a plus. You will generally be free to use tools you see fit, but your colleagues will have a preference for Perl, R and Hadoop.

The eligibility for this is precised as *´BSc, MSc or PhD in Math, Statistics, Computer Science, Artificial Intelligence or comparable'.* Even the rest of these announcements gives more or less identical description of the profile and tasks of a data analyst.

We can outline the requirements of a data scientist as follows:

- Mathematics: The coming data engineers must master complex mathematical models and sophisticated data processing techniques. This should be reflected in the educational curricula. Data Scientists must be competent mathematicians. They must be able to understand computational linear algebra, a.k.a. matrix computations, numerical analysis, and matrix analysis. Most data mining applications use matrix computations as their fundamental algorithms.

  The mathematical skills include fields of mathematics not necessarily associated with applications, like algebra, combinatorics, and geometry. These are needed to provide students with the necessary mathematical tools for properly formalize data mining and inference problems, to let students focus both on how to compute and also what to compute.

- Probability theory and statistics, signal processing, data mining: Machine learning and statistical algorithms are powerful but often require considerable sophistication from the side of the human user, especially with regard to selecting features for training and choosing model structure (e.g., for regression or in graphical models). Many of students today learn the mathematics behind these models but do not develop the skills required to invoke them effectively in practice on large and big data sets. Clearly, statistics and machine learning algorithms must be improved to be more robust and easier for general unsophisticated users to apply. But at the same time there must be a rigorous training of students to master the intricacies of the methods as experts.

- Computing:

  – Programming/Scripting Languages: Next, computer architectures are continuing to become more complex with multicore CPUs, GPUs and memory hierarchies that require novel programming paradigms to be fully utilized. The computing platforms consist of thousands of interconnected processors with supervisory management of data flows and processor loads. Training students to work in massive data analysis will require experience with handling vast amounts of data

data and with computational infrastructure that permits real problems associated with massive data to be revealed. It is hence central to work on real data sets and systems to train the students. The availability of benchmarks, repositories (of data and software), and computational infrastructure will be a necessity in training the next generation of data engineers.

  - Relational Databases: A solid understanding of SQL-based systems is a central requirement. To learn the fundamentals of database design and management is another.
  - Distributed Computing Systems and Tools: solid background in a range of the systems/tools such as Hadoop/HBase, Cassandra, Hive, Pig, MapReduce et al.

- Visualization: Data scientists must be able to take the hard data from within the data warehouse and other storage facilities, mine it for the most important and present it in graphics so that users can understand and employ.

Therefore it is clear that capitalizing on the strengths in computational mathematics, optimization, discrete mathematics, probability theory, statistics, machine learning, signal processing, networking, algorithmics and high-performance computing, KTH has all the educational resources and research-wise firmament needed to create and start an education of data analyst.

KTH should consider developing a masters program in data analytics that gives students the technical training, as outlined above, needed to succeed in high-tech jobs in data analysis and computing exemplified above. The program should draw on competence in several schools of KTH and could seek application areas in yet a wider set of fields.

Finally, it is important for students to be trained in an interdisciplinary environment that blends mathematics, systems engineering and computer science with real application areas. The programs should, e.g., also include training on ethical issues associated with data management and analysis, and in societal issues. The discussion of these is found in the paper by Microsoft researchers Danah Boyd and Kate Crawford: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, Communication & Society*, 15, 5, p. 662–679, 2012.
Jake Porway, most recently the data scientist in the New York Times R&D lab, writes:

> A data scientist is a rare hybrid, a computer scientist with the programming abilities to build software to scrape, combine, and manage data from a variety of sources and a statistician who knows how to derive insights from the information within. S/he combines the skills to create new prototypes with the creativity and thoroughness to ask and answer the deepest questions about the data and what secrets it holds.

Thomas Davenport and D.J. Patil summarize in Harvard Business Review (Oct. 2012) (pp. 70–76: *Data scientists are the new sexy and they are needed now and into the future in ever-increasing numbers.*

Appendices

A Big Data group progress report

This appendix contains a comprehensive report on activities carried out by the Big Data group after the delivery of of the Whitepaper on 2013-01-22. Most of this material has been reported to the KTH ICT and Life Science Technologies platform at status meetings and on an ongoing basis.

A.1 Workshops and conferences

"Digital Horizons" June 13, 2013 The group hosted Prof Scott Kirkpatrick, The Hebrew University of Jerusalem, Israel, who was an invited panelist in the Big Data session at the KTH-The Economist-summit on "Digital Horizons".

Big Data Day with Scott Kirkpatrick June 14, 2013. This whole-day event on Big Data in the Life Sciences included presentations by Dr Mikael Huss, SciLifeLab, Dr Andrey Alexeyenko, SciLifeLab, and Big Data group member Dr Ozan Öktem, KTH Mathematics, and a discussion session with Scott Kirkpatrick.

KTH Symposium on Big Data in the Life Sciences December 12, 2013. This half-day event was held in lecture hall FB42 in the main building on the AlbaNova campus, and included lunch, a poster session, and a buffet dinner.

Speakers:

- Prof Mikael Skoglund, EES
- Dr Jim Dowling, ICT & SICS
- Prof Jan-Olov Strömberg, SCI
- Dr Lukas Käll, BIO & SciLifeLab
- Prof Tony Lindeberg, CSC

Discussion session: The event included a lively discussion session with input from, among several others, Prof Stefan Carlsson, CSC, Dr Jussi Karlgren, SICS, and Torbjörn Hägglöf, IBM Sweden.

KTH Symposium on Big Data May 26, 2014. This full-day event was held in the Oskar Klein Lecture Hall, AlbaNova, Stockholm. Speakers and program (including abstracts and slides) have been reported previously.

Speakers:

- Dr Abdel Labbi, IBM Research Zurich, Switzerland, *From Big Data to Smarter Decisions*
- Prof Jussi Taipale, Karolinska Institutet, Sweden *Genome-wide analysis of protein-DNA interactions*
- Mattias Lidström, Ericsson, Sweden *Real Time Big Data Analytics Challenges*

- Prof Holger Rootzén, Chalmers University of Technology, Sweden *How not to waste effort in big screening experiments: tail methods for estimating the number of false positives*
- Prof Erkki Oja, Aalto University, Finland *Big Data Analysis by Machine Learning*
- Prof Erwin Laure, KTH-Royal Institute of Technology, Sweden *d-Science – the Big and Not-so-Big Data*
- Prof Søren Brunak, Technical University of Denmark & University of Copenhagen, Denmark, *Fine-grained phenotypes, comorbidities and disease trajectories from data mining of electronic patient records.*

A.2 Attendance at workshops and conferences (participant lists)

Participation in Big Data Day, June 14, 2013

1. Yucheng Hu, Tsinghua University, Beijing, China
2. Nicolas Innocenti, Computational Biology, KTH
3. Andrey Alexeyenko, Karolinska Institute, Stockholm
4. Astrid de Wijn, Physics, Stockholm University
5. Ralf Eichhorn, NORDITA, Stockholm
6. Mikael Skoglund, Communication Theory, KTH
7. Timo Koski, Mathematics, KTH
8. Mikko Vehkaperä, Communication Theory, KTH
9. Ozan Öktem, Mathematics, KTH
10. Mikael Huss, SciLifeLab, Stockholm
11. Marcin Skwark, SciLifeLab, Stockholm
12. Scott Kirkpatrick, The Hebrew University of Jerusalem, Israel
13. Erik Aurell, Computational Biology, KTH
14. Jim Dowling, SICS, Stockholm
15. Dhribaditya Mitra, NORDITA, Stockholm
16. Carlo Fischione, Control Theory, KTH
17. Mikael Johansson, Control Theory, KTH
18. Sarunas Girdzijauskas, Communication Networks, KTH
19. Darya Goryanskaya, SciLifeLab, Stockholm
20. Pan Lu, SciLifeLab, Stockholm

21. Carolyn L Beck, U of Illinois, USA
22. Saikat Chatterjee, Communication Theory, KTH
23. Rolf Stadler, Communication Networks, KTH
24. Seif Haridi, SICS, Stockholm
25. Martin Nilsson, SICS, Stockholm
26. Per Brand, SICS, Stockholm
27. Maria Papadopouli, Communication Networks, KTH
28. Carl Gustaf Jansson, ICT, KTH

Participation in Symposium on Big Data in the Life Sciences December 12, 2013

1. Laeeq Ahmed HPCViz,CSC, KTH
2. Tahmina Akhter, KI
3. Rossen Apostolov PDC, KTH
4. Lars Arvestad, CSC and SciLifeLab
5. George Athanasiou, KTH
6. Erik Aurell, CSC
7. Mauricio Barrientos, KI
8. Brynjar Smári Bjarnason, Stockholm University
9. Petter Brodin SciLifeLab
10. Stefan Carlsson, CSC
11. Frida Danielsson, SciLifeLab
12. Jim Dowling, ICT and SICS
13. Åke Edlund, PDC
14. Örjan Ekeberg, CSC
15. Arne Elofsson, Stockholm University and SciLifeLab
16. Olivia Eriksson, Mechanics
17. Hossein Farahani, SciLifeLab
18. Hamed Farhadi, KTH
19. Christoph Feinauer, Politecnico di Torino, Italy
20. Anders Forsgren, Mathematics
21. Jesper Gantelius KTH

22. Stefania Giacomello, SciLifeLab
23. Mario Giovacchini, SciLifeLab
24. Maksym Girnyk, KTH Communication Theory
25. Dariya Goranskaya, SciLifeLab
26. Viktor Granholm, SciLifeLab
27. Björn Hallström, Hjärtkliniken, Danderyds Sjukhus
28. Henrik Hellqvist, KTH
29. Dan Henningson, KTH
30. Pawel Herman, KTH
31. Yue Hu, SciLifeLab
32. Mikael Huss, SciLifeLab
33. Anne Håkansson, SciLifeLab
34. Nicolas Innocenti, CSC
35. Sverker Janson, SICS
36. Magnus Jansson, KTH Signal Processing
37. Jussi Karlgren, SICS
38. Gunnar Karlsson, EES
39. Timo Koski, Mathematics
40. Per Kraulis, SciLifeLab
41. Peter Larsson, KTH
42. Erwin Laure, PDC
43. Sara Light, SciLIfeLab
44. Tony Lindeberg, CSC
45. Ulrika Ljungman, KTH
46. Daniel Lundin, SciLifeLab
47. Atsuto Maki, KTH
48. Mirco Michel, SciLifeLab
49. Luminita Moruz, SciLifeLab
50. Dinesh Natesan, KTH
51. Tobias Oechtering, KTH Communication Theory

52. Ljubica Pajevic
53. Elena Panizza, SciLifeLab
54. Evangelia Papadaki, KI
55. Lukas Persson, SciLifeLab
56. Otto Pintor, KTH
57. Florian Pokorny, KTH
58. Guillermo Rodriguez Cano, KTH
59. Oxana Sachenkova
60. Zeeshan Shah, PDC
61. Hossein Shokri
62. Joel Sjöstrand, CSC
63. Mikael Skoglund, KTH Communication Theory
64. Yajing Song, SciLifeLab
65. Josephine Sullivan, CSC
66. Dennis Sundman, KTH
67. Ragnar Thobaben, KTH Communication Theory
68. Richard Tjörnhammar, KTH
69. Daniel Trpevski
70. Izhar ul Hassan, PDC
71. Per Unneberg, SciLifeLab
72. Chathuranga Weeraddana, KTH
73. Hugo Wefer, SciLifeLab
74. Björn Winckler, KI
75. Dan Wu, KI
76. Ming Xiao, KTH Communication Theory
77. Duo Xu
78. Rerngvit Yanggratoke, KTH
79. Yan Zhou, KI
80. Ozan Öktem, KTH Mathematics

Participation in KTH Symposium on Big Data, May 26, 2014

1. Ilgar Abdullayev, Karolinska Institutet
2. Magnus Adolfson, Scania CV AB
3. Kunal Aggarwal, KTH student
4. Altug Akay, KTH
5. Andrey Alexeyenko, SciLifeLab & MTC, KI
6. Tom Andersson, MSB
7. Frej Andreassen, General Electric – GE Money Bank
8. Mariette Annergren, KTH
9. Lars Arvestad, CSC and SciLifeLab
10. Erik Aurell, KTH
11. Marcus Berg, Statistiska institutionen, Stockholms universitet
12. Daniel Berglund, IMM
13. Brynjar Smari Bjarnason, Örebro universitet
14. Johan Blaus, KTH Näringslivssamverkan
15. Jorrit Boekel, ScilifeLab Stockholm, BILS
16. Magnus Boman, KTH & SICS
17. Mikael Borg, BILS – Bioinformatics Infrastructure for Life Sciences
18. Søren Brunak, Center for Biologisk Sekvensanalyse, DTU
19. Stefan Carlsson, KTH
20. Wojtek Chachólski, KTH
21. Georg Chambert
22. Saikat Chatterjee, KTH
23. Veronique Chotteau, KTH/BIO School
24. György Dán, KTH
25. Hatef Darabi, Karolinska Institutet
26. Ino de Bruijn, SciLifeLab
27. Pol del Aguila Pla, KTH – Signalbehandling
28. Jim Dowling, KTH and SICS
29. Rong Du, KTH Automatic Control Lab

30. Daniel Edsgärd, Karolinska Institutet
31. Ariel Ekgren, KTH
32. Maria Ekholm, Karolinskas sjukhus
33. Arne Elofsson, Department of Biochemistry and Biophysics, Stockholm University
34. Pär Engström, Science for Life Laboratory, Stockholm University
35. Martin Eriksson, Myndigheten för samhällsskydd och beredskap
36. Olivia Eriksson, SeRC, KTH
37. Carlo Fischione, KTH Automatic Control
38. Stefan Fleischmann, DBB
39. Viktoria Fodor, KTH/EES
40. Mattias Frånberg, Stockholms Universitet
41. Babbi Fröding, KTH, Filosofi
42. Andrea Ganna, Karolinska Instituet
43. Benjamin Garzon, KI
44. Daniel Gillblad, SICS
45. Maksym Girnyk, KTH Kommunikationsteori
46. Ulrika Gunnarsson Östling, KTH
47. Omar Gutierrez Arenas, KTH
48. Mikael Haglund, IBM
49. Seif Haridi, KTH
50. Liqun He, Uppsala University
51. Ivone Herre, Scilifelab
52. Håkan Hjalmarsson, KTH/EE School
53. Olga Hrydziuszko, Karolinska Institutet
54. Yue Hu, KTH/SciLifeLab
55. Torbjörn Hägglöf, IBM Svenska AB
56. Nicolas Innocenti KTH, Royal Institute of Technology
57. Elling Jacobsen, KTH
58. Magnus Jansson, KTH
59. Ashwini Priya Jeggari, Karolinska Instituet

60. Ioannis Kalfas, KTH
61. Gunnar Karlsson, KTH EE
62. Johan Karlsson, KTH, Matematik
63. Mehmood Khan, KTH
64. Mohammad Khodaei, KTH, EE, LCN
65. Ali Khorshidi, KTH
66. Kittipong Kittichokechai, KTH
67. Timo Koski, Institutionen för matematik/KTH
68. Per Kraulis, NGI Sweden, SciLifeLab, DBB, SU
69. Lukas Käll, KTH BIO
70. Erik Lampa, Uppsala universitet
71. Wilhelm Landerholm, Queue AB
72. Peter Larsson, KTH
73. Tore Johan Larsson, KTH
74. Haibo Li, CSC/MID
75. Erik Lindahl, SciLifeLab
76. Tony Lindeberg, KTH
77. Thomas Lindh, KTH
78. Anders Lindquist, KTH
79. Sven Lindqvist, KTH Industriell Ekonomi
80. Filip Lindskog, KTH Matematik
81. Agneta Lissmats, RCC
82. Chaoren Lu, Karlstad University
83. Stefan Magureanu, KTH
84. Atsuto Maki, CSC/KTH
85. Raffaele Marino, Ph.D Student Nordita/KTH
86. Majid Nasiri Khormuji, HUAWEI
87. David Nilsson, KTH
88. Torbjörn Nordling, Uppsala Universitet / SciLifeLab
89. Malin Nordström, Karolinska sjukhuset

90. Anton Osika, KTH
91. Arash Owrang, KTH
92. SenthilKumar PaneerSelvam, Stockholm University
93. Oleksii Pasichnyi, SEED, ABE, KTH
94. Bengt Persson, Uppsala University, SciLifeLab and BILS
95. Staffan Persson, Scania CV
96. Markus Persson, Aon Hewitt
97. Valeria Pestana Gianvittorio, Science for Life Laboratory
98. Christoph Peters, Stockholm University, DBB
99. Xiaoyan Qian, Uppsala University
100. Ryan Ramanujam, Karolinska Institutet
101. Olof Ramström, KTH
102. Lars Rasmussen, KTH
103. Martin Ribe, SCB
104. Oxana Sachenkova, Stockholm University
105. Hossein Shahrokni, KTH
106. Mauricio Sodré Ribeiro, Industrial Ecology / KTH
107. Ann-Charlotte Sonnhammer, SNIC
108. Erik Sonnhmmer, SU
109. Ola Spjuth, Karolinska Institutet
110. Rolf Stadler, KTH
111. Ying Sun, AstraZeneca
112. Martin Sundin, Avdelningen för Signalbehandling
113. Thomas Svensson, SciLifeLab
114. Oleg Sysoev, Linköpings Universitet
115. Håkan Terelius, KTH
116. Ragnar Thobaben, KTH/EES/CommTH
117. Johan Thunberg, KTH Royal Institute of Technology
118. Martin Tran, Karolinska Institutet
119. Kostas Tsirigos, Department of Biochemistry & Biophysics, SU

120. Misbah Uddin, KTH
121. Bo Wahlberg, KTH
122. Liping Wang, KTH/EE
123. Per Warholm, Stockholm University
124. Hugo Wefer, KI
125. Francesco Vezzi, SciLifeLab
126. Madeleine Winberg, Diabetes Tools
127. Björn Winckler, KI
128. Valtteri Wirta, KTH / Scilifelab
129. Stefan Vlachos, Karolinska University Hospital
130. Rerngvit Yanggratoke, KTH
131. Stavros Yika, KTH
132. Yan Zhou, Karolinska Institutet

A.3 Research visits and lectures

The following research visits have been organized or co-organized:

1. Prof Scott Kirkpatrick, HUJI, Israel, June 2013. Prof Kirkpatrick participated as a panelist in the Big Data session of the KTH-The Economist summit Digital Horizons. He also was the key panelist for the event on Big Data in the Life Sciences held on June 14, 2013.

2. Prof Nicolas Macris, EPFL, visited for a week in January 2014, and gave a talk on in the seminar series at AlbaNova organized by Aurell. Prof Macris represents the group at EPFL headed by Ruediger Urbanke who is an important contact for the Aurell and Skoglund groups.

3. Prof Osamu Watanabe, Prof Yoshiyuki Kabashima, and Dr Haiping Huang, all from Tokyo Institute of Technology, Tokyo, Japan, visited for a week in March 2014 (16th March to 21st March). Profs Watanabe and Kabashima are respectively PI and co-PI of two large research Japanese (MEXT-funded) centres in the area. Prof Watanabe gave a talk at KTH Dept Mathematics, Prof Kabashima at EES School, and Dr Huang in the seminar series at AlbaNova organized by Aurell. Only the visit of by Prof Kabashima was funded by the project; the others were self-funded.

4. Dr. David Koslicki from Oregon State University visited KTH EE School in April 2014 to discuss large scale data analysis in life sciences, hosted by Mikael Skoglund and Dr. Saikat Chatterjee. These discussions resulted in a joint paper [17].

5. Dr Danny Bickson, co-founder of GraphLab, a very visible Big Data start-up based in Seattle US visited Stockholm June 3-5 2014 in the context of an EIT-supported workshop organized by Dr Sarunas Girdzijauskas, KTH EES. We primarily discussed educational issues with Dr Bickson such as the structure of a possible MSc course. Dr Bickson also provided us with a list of US-centered benchmark examples given below.

Information about Big Data MSc programs obtained from Dr Danny Bickson: First, Dr Bickson pointed to an on-line list of programs available at `http://whatsthebigdata.com/2012/08/09/graduate-programs-in-big-data-and-data-science/`. This list, though long, is however not complete, and some of the most relevant and likely ones do not appear.

Dr Bickson suggests that the following shorter list of programs is the most relevant. Not all of these programs have the words "big data" or "analytics" in the title. Where available the approximate tuition in stated, generally somewhat higher than the tuition fees charged for non-European students at KTH. Order between these programs not important.

- Data Science at Columbia U (`http://idse.columbia.edu/`) - which was eventually delayed for next year. Cost should have been approx $50kUSD$ (tuition in this program is paid by course/points not by semester)
- Data Science at NYU (`http://cds.nyu.edu/`) – a Big Data science center has opened in NYU attracting big name as well as industry collaboration.
- MITS at CMU (`http://www.cmu.edu/mits/index.html`) – computer science + software engineering program. 3-4 semesters, $25kUSD$ per semester.
- MCDS @ CMU (http://mcds.cs.cmu.edu/) - very intense program. emphasizes practical aspects of Big Data. 3 semesters ($20kUSD$ each) + summer internship
- Data Science Track inside Computer Science at Stanford (`http://icme.stanford.edu/academic-programs/ms/data-science`)
- Master of Science in Analytics at Northwestern (`http://www.analytics.northwestern.edu/index.html`)
- Computational Science and Engineering (CSE) at Gtech (`http://www.cseprograms.gatech.edu/csems`) approximate cost $45kUSD$
- Applied and Computational Mathematics at Johns Hopkins (`http://ep.jhu.edu/graduate-programs/applied-and-computational-mathematics`)
- Master of Science in Statistics: Analytics Concentration at University of Illinois at Urbana-Champaign (`http://www.stat.illinois.edu/degrees/msanalytics.shtml`)

A.4 Outreach activities

Seminar at Sveriges Riksdag "Big Data – Hot eller nytta?", October 16, 2014 Mikael Skoglund, together with Gösta Lemne, VP Strategy Development at Ericsson and member of IVA, presented at the lunch seminar "Big Data –

Hot eller nytta?" organized by IVA and "Sällskapet Riksdagsledamöter och Forskare" (RIFO). The seminar was held in the Parliament Building, Stockholm (Riksdagshuset, Partimatsalen) on 2014-10-16, and gathered about 30 Members of Parliament, including some on the ministerial level.

Attendance list of "Big Data – Hot eller nytta?": The following attendance list is as attached to the invitation letter signed by Yvonne Andersson, chairperson of Rifo and Johan Weigelt, IVA. In our estimate about 30 of the listed attendees participated (listed participant 34/35 was likely a double booking):

Order	First name	Family name	Affiliation
1	Yvonne	Andersson	KD
2	Maria	Andersson Willner	S
3	Roger	Aule	UbUs kansli
4	Finn	Bengtsson	M
5	Helena	Bouveng	M
6	Örjan	Carlborg	Rifo
7	Sotiris	Delis	M
8	Susanne	Eberstein	S
9	Annika	Eclund	KD
10	Ingrid	Edmar	UbUs kansli
11	Lars	Eriksson	Rifo
12	Olle	Felten	SD
13	Emma	Henriksson	KD
14	Hans	Hertz	Rifo
15	Per-Ingvar	Johnsson	C
16	Fredrik	Lagergren	Rifo
17	Mimmi	Lapadatovic	UbUs kansli
18	Gösta	Lemne	IVA
19	Kerstin	Lidén	Rifo
20	Betty	Malmberg	M
21	Catarina	Molin	UbUs kansli
22	Lars	Nilsson	IVA
23	Cecilia	Nordling	UbUs kansli
24	Lotta	Olsson	M
25	Johan	Pehrsson	FP
26	Johan	Persson	IVA
27	Leif	Pettersson	S
28	Anna	Sjöström Dou	Rifo
29	Mikael	Skoglund	KTH
30	Lars-Arne	Staxäng	M
31	Thomas	Strand	S
32	Per-Anders	Strandberg	UbUs kansli
33	Elin	Sundin	S
34	Gunilla	Svantorp	S
35	XX	Svantorp	

(continued)

Order	First name	Family name	Affiliation
36	Anna	Söby	S
37	Rasmus	Weibull	
38	Johan	Weigelt	IVA
39	Per	Westerberg	M
40	Annette	Åkesson	M

Round Table Conference on Data-analysis, KTH & Elisa, September 29, 2014 Group member Koski organized a one-day conference with Mr Kimmo Pentikäinen, Vice President for Research & Development at Elisa Corporation, Finland, the largest mobile operator in Finland.

Round table conference part 1.

10.00−10.15 Coffee and refreshments

10.15−10.25 Presentation of KTH. Timo Koski, KTH

10.30−10.40 Presentation of Elisa. Kimmo Pentikäinen, Elisa

10.45−11.30 Presentations of data analysis at KTH. Erik Aurell KTH, Wojciech Chachólski KTH

11.30−12.30 Lunch, Restaurang Syster o Bror; Drottning Kristinas väg 24

Round table conference part 2.

12.35−12.45 Presentation of Department of Mathematics. Boualem Djehiche

12.45−12.55 Presentation of Elisa. Kimmo Pentikäinen Elisa

13.00−13.20 Presentation of data analysis & communication networks. Rolf Stadler, School of Electrical Engineering

13.25−13.40 Presentation of data analysis and statistics at KTH. Timo Koski KTH

13.45−14.00 Big Data and mathematics. Ozan Öktem KTH

14.00−14.30 Discussion

14.30−15.00 Coffee and refreshments

Round table conference part 3.

15.00−17.00 General discussion.

The list of participants of this event was as follows:

Name	e-mail	Affiliation
Erik Aurell	eaurell@kth.se	KTH Biological Physics
Wojciech Chachólski	wojtek@kth.se	KTH Mathematics
Boualem Djehiche	boualem@kth.se	KTH Mathematics
Timo Koski	tjtkoski@kth.se	KTH Mathematics
Tatjana Pavlenko	pavlenko@math.kth.se	KTH Mathematics
Kimmo Pentikäinen	kimmo.pentikainen@elisa.fi	Elisa Corp, VP
Ryan Ramanujam	ryan@ramanujam.org	KTH Mathematics
Martina Scolamiero	scola@kth.se	KTH Mathematics
Rolf Stadler	rolf.stadler@ee.kth.se	KTH EES/LCN
Ozan Öktem	ozan@kth.se	KTH Mathematics
Jonas Hallgren	jonas@math.kth.se	KTH Mathematics
Felix Rios	flrios@kth.se	KTH Mathematics

Other outreach: Erik Aurell and Timo Koski were members of the Senior Program Committee of the AI & Statistics 2014 conference, a high-profile venue for Data Analytics.

One member of the Big Data working group (Skoglund) is also the KTH coordinator for the KTH-Ericsson framework agreement. The steering committee for those activities has appointed a formal contact person at Ericsson (R. Cöster) with whom we have met and discussed opportunities for joint activities. These discussions are still ongoing.

B Data Analytics: the concept, its potential, and current limitations

The concept and its potential

Data analytics is the discovery of meaningful patterns in data. The generic goal is thereby to discover useful information, suggest conclusions, and support various forms of decision-making. Data analytics develops models that explain the past and predict the future [25]. In this it relies on hybrid applications of statistical learning theory, computer programming, signal processing, algorithms on graphs, operations research, visualization and predictive analytics. Predictive analytics focuses on application of statistical and machine learning models for predictive forecasting or classification. In predictive analytics one uses regression techniques [36] like linear (multiple) regression model, discrete choice models, logistic regression, probit regression, time series models, survival or duration analysis [2], classification and regression trees and multivariate adaptive regression splines, as well as machine learning techniques like neural networks, support vector machines, naive Bayes and k-nearest neighbours [32]. Data analytics seems in other words to be emerging as a new focus for organizations and universities promoting a fusion of related fields.

Data analytics needs tools of *data analysis* for inspecting, cleaning, and transforming data. Data integration is a precedent to data analysis. Data mining [31] is one technique of data analysis for modeling and knowledge discovery for predictive rather than purely descriptive purposes. In a more statistical view, data analysis is separated into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA) and prescriptive analysis. EDA

focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. The third field of data analysis is prescriptive (optimization and simulation).

It has been estimated [33] that in 2007, it was possible for the global society to store 2.9×10^{20} optimally compressed bytes, and to communicate almost 2×10^{21} bytes (10^{21} bytes = zettabyte). Organizations must deal with data in petabyte-scale (10^{15} bytes). Every day, Google alone processes about 24 petabytes of data [24]. The Americans are estimated to have consumed in 2008 a flow of 3.6 zettabytes of media information [9].

Data may come in click streams, transaction histories, and sensors. The data may have social media content (tweets, blogs, Facebook) and video data from retail, from video entertainment (games, movies). Big data also encompasses everything from call center voice data to genomic and proteomic data from biological research (c.f. the details in section B.6) and medicine [5]. Not only is data big, but it has often to be processed quickly. E.g., fraud detection is to be done at a point of commercial transaction.

The massive complex data sets as commonly seen today are in the petabyte and terabyte (10^{15} bytes) scales, which might serve as a crude and preliminary delineation of the notion of big data. The challenges of such data, which is frequently in a constant state of change (streaming), are ubiquitous in science, medicine [5], technology and business. One can in particular note five specific types of business that are visible in this, see e.g. [47], - cybersecurity, financial analytics, health analytics, new media and smart cities. The rise of massive, complex data sets has the potential to transform the understanding of phenomena ranging from physical and biological systems to social and economic behavioral patterns, e.g., human cognitive learning about causality.

The concept and its current limitations

Due to advances in computer hardware– faster CPUs, cheaper memory, and (forthcoming advances [41]) in new technologies such as Hadoop, MapReduce, and text analytics for processing big data, it seems now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for data analytics. But it could be asked whether 'big data is driven more by storage capabilities than superior ways to ascertain new knowledge' [11].

In this respect we must note that traditional 'small data' can often offer information that is not contained (or containable) in big data. It should therefore be recognized that the factors described above are ipso facto enabling the more traditional small data collection. The same factors are thus enhancing the scope of time-honoured applied statistical data analysis of small or large data sets, too. One can think of future forms of data analytics that use data from all sources, traditional and new ones. In fact, the questions in the best selling popular book [6] are imaginative applications of traditional statistical models in this genre: How can a football coach evaluate a player without ever seeing him play? Or, how can a formula out-predict wine experts in determining the best vintages?

There is a place for a word of caution: We quote Prof. Michael I. Jordan, one of the main authors of the Report [48], from an interview in The IEEE Spectrum on 3 October 2014[2]

[2] http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan

> ... the analogy of building bridges. If I have no principles, and I build thousands of bridges without any actual science, lots of them will fall down, and great disasters will occur.
>
> Similarly (in Big Data), if people use data and inferences they can make with the data without any concern about error bars, about heterogeneity, about noisy data, about the sampling pattern, about all the kinds of things that you have to be serious about if you are an engineer and a statistician − then you will make lots of predictions, and there is a good chance that you will occasionally solve some real interesting problems. But you will occasionally have some disastrously bad decisions. And you will not know the difference a priori.

The emphasis in this report/survey is to summarize the pieces of work in mathematics and engineering that are judged to be reliable and promising points of departure in development of big data analytics supported by scientific principles. In this respect the current report differs appreciably from the unrestrained promotion and boosting of the benefits of big data (and its analytics) as, e.g., [44].

B.1 Pitfalls of Data Analytics

- All empirical scientific research stands on a groundwork of measurement. The standard of scientific research is, that we want to guarantee that the instrumentation actually is capturing the theoretical concept targeted at. Furthermore, the measurements are to be replicable and comparable across cases and over time. And we need to ascertain whether measurement errors are systematic or not.

  In today's organizations the engineers are incessantly changing the algorithms to improve the service. As pointed out in [39], platforms such as Twitter and Facebook are always being re-engineered. Whether studies conducted even a year ago on data collected from these platforms can be replicated in later or earlier periods is an open question.

- Due to huge amount of data analyzed, it is very easy to find spurious dependencies in projects with Big Data. With large data sets the appetite for hypotheses tends to get even larger. And if it is growing faster than the statistical strength of the data, then many of the inferences are likely to be false. This state of affairs seems to become neglected by the phantasmagorical praise of correlation studies over big data, c.f. [44]. There are statistical methods to deal with these problems, like family-wise error statistical tests, but many of them have not been studied computationally and it requires hard mathematics and engineering. A report of a round table discussion of data correlations vs. scientific models is found in [11].

- Data analysis can deliver inferred data at a certain level of quality and there must be explicit considerations about it. It is needed to add error bars to the inferred analytics. This point of view is missing in much of the current machine learning literature.

-on-the-delusions-of-big-data-and-other-huge-engineering-efforts#qaTopicThree

- Hal Varian from Google is quoted in [11] stating that *the point of big data is ... to be able to pick a random sample and to analyze it.* One gets a result from the random sample as good as looking at everything. But the difficulty is to make sure that it is really a random sample.
- An addendum to the preceding: it is difficult to determine what is an outlier in big data.
- There may be a tendency to overlook the old and elementary pieces of statistical wisdom [12] calling attention to issues like selection bias, endogeneity/exogeneity and confounding. There is in addition the phenomenon of *dark data*. By this one means [10] that most data is is created, used and thrown away without any person being aware of its existence. This seems like a digital technology version of sampling bias masking possible confounders.
- The reliance on time-honoured statistical concepts like sufficiency [46] and coefficient of correlation [51] can turn out to be obsolete, when dealing with massive data sets.
- There must be an awareness of *over-fitting* a small number of cases to a huge number of models, c.f. the difficulties in Google Flu Trends predictor, as reported in [39].

The mere 'access to data, processing power and a certain amount of statistical know-how', the triumvirate celebrated in [44], is thus not sufficient for big data analytics [9]. The analysts must with some assurance be giving out reasonable answers and are quantifying the likelihood of error. It will take decades, at least according to M.I. Jordan (interview, loc.cit), to get a real engineering approach that contains understanding of where results came from and/or why models are not working and also of necessary exploratory tools for visualizing data and models. Ergo, data analytics needs a foundational corpus.

B.2 Foundations of Data Analytics

The need for Foundations

Research on foundations (principles) on data analytics is a strategic priority for the academia. This means research on formal and mathematical models for data processing, as well as on issues concerning the engineering of large-scale data processing systems. The foundations on data analytics is a new discipline that has emerged to address the need for professionals and researchers to deal with massive data. Its object is to provide the underlying theory and methods of the data tidal wave.

Massive data sets involve phenomena that are too complex to permit intuitive predictions and simple statistical summaries. The main tool of conventional data analysis by any single human, or any group of humans, is to process statistical summaries like, e.g., sufficient statistics. The sets of big data are also too voluminous to fit on a single computer, to manipulate with traditional databases or statistical tools, or to represent using standard graphics software. The data is also heterogeneous, not resembling the highly curated data analyzed in the

recent past. Working with big data analytics requires distinctive new foundational research. The blend of statistical physics, computational theory and information theory, as in [45], can turn out to be promising in this. Massive data sets require the support of mathematical and computational modelling for quantitative assessments combined with tools of theoretical computer science to check the tractability of the any proposed method.

Data Base Management Systems and Data Analysis Tools

Existing data base management systems (DBMS) and data analysis tools of do not always lend themselves to data analytics at the scale required by massive data sets. In fact, data does not necessarily fit nicely into an existing processing tool. In other words, big data necessitates an analysis that existing data base management systems can not readily provide [41] and [42].

Tools such as SAS, R, and Matlab support relatively sophisticated analysis, but are not designed to scale to data sets that exceed the memory of a single machine [42]. Additionally, neither DBMSs nor MapReduce are good at handling data arriving at high rates, providing little support for techniques such as approximation, sub-linear algorithms, or sampling that might help users ingest massive data volumes. Research is needed to bridge the gap between large-scale data processing platforms such as DBMSs and MapReduce, and analysis packages such as SAS, R, and Matlab.

Modelling paradigms

The phenomena encountered in big data analytics are often not derivable from a few simple principles. This means that foundational data analytics needs to work on *model identification* and *model reduction* [21]. The former is the problem of determining a model that replicates data or qualitative properties thereof with sufficient degree of accuracy, the latter is to find the simplest model with these properties. Models also need to integrate and model multiple types of data (*data/model fusion*). These issues, which are *the* challenges in data analytics are handled differently depending on the modelling paradigm one uses.

Most approaches to modelling are based on explicitly describing the system in a deterministic or stochastic setting. The aim is not only to describe the system, but also to provide a conceptual simplification of it. Mathematical analysis offers powerful tools to describe phenomena that display continuous behaviour. Advanced data analytics will make use of tools, in addition to those mentioned above, from discrete mathematics (graph theory), algebra, geometry, and theoretical computer science, signal processing, control theory, and optimization.

A highly relevant modelling paradigm is *data-driven modelling* where techniques from analysis of data are used to infer the model directly from data without explicitly describing the system. Most methods stem from statistical learning, but dynamical systems theory and information theory and statistical physics [45] are also applicable. These modelling approaches take into account all the data but do not offer any conceptual simplification of the phenomena that are studied.

Model identification

Model identification is recast as a parameter identification problem, which is handled using tools from *inverse problems*. These also offer means for model reduction and assessment of modelling error. Next, many challenges evolve around model identification in phenomena that manifest discrete behaviour, *e.g.* inferring the regulatory network from genomics data. Statistical methods that account for the combinatorial structure of the logical connections, such as Bayesian networks [38] and Markov models, are applicable to address such model identification issues. For hybrid models, model identification for continuous and discrete phenomena is handled separately and there are clear advantages in handling these simultaneously. For data-driven modelling, the data provides the model thereby taking care of model identification and model reduction. Methods for assessing the model error are however limited.

Multiscale methods

Tools from *multiscale methods* are needed when phenomena that are studied involve multiple temporal and/or spatial scales that are not separable. The challenge is to capture small-scale effects on the large-scales without resolving all small-scale features. Most multiscale methods are developed for addressing scientific applications in engineering or physics, *e.g.* simulation of high frequency wave propagation, statistical mechanics, computational material sciences.

Network/differential equation models can handle multiscale phenomena of massive data sets, where spatial scales are separable. They also offer a framework for data/model fusion in that data related to some specific part of the system can be associated to the node(s) relevant for that part.

Statistical learning and inverse problems: Connecting models to data

Inverse problems are central in analysis of massive data and arise whenever one seeks to fit theory to such data [13], [30]. Measured data often has significant amount of noise, is sampled non-uniformly, and/or a significant amount of it is missing (incomplete data), leading to very challenging inverse problems.

The last decade has witnessed significant progress in addressing these challenges with the advent of *sparse signal processing* methods and *sparsity promoting regularisation*. Sparse signal processing is a general mathematical framework for modelling and extraction of patterns from data. This general framework not only yields algorithms that compete favourably with application specific state-of-the-art approaches, but also prescribes optimal data collection protocols (compressed sensing).

These methods have had tremendous impact on compression and reconstruction problems, especially in situations with incomplete data and/or highly noisy data, cases that previously were considered impossible to handle.

Modelling of complex systems

Qualitative inverse problems is a general framework where model parameters are matched to observed qualitative behaviour of data, a task that among others involves sparse signal processing, dynamical systems theory, and regularisation theory. For network/differential equation models, the model identification can

be addressed using combinatorics to link properties of network structure with dynamic systems theory describing the behaviour of the system. Dynamical systems theory offers several tools (approximate entropy, delay embedding theorems, Conley index theory) relevant for model identification/reduction that are yet to be utilised [20].

An important tool is sparsity promoting regularisation, which until now is mostly used for imaging. It has great unexplored potential in other inverse problems, such as in qualitative inverse problems. The theory of sparse signal processing and sparsity promoting regularisation remains to be exploited in the context of massive data sets.

B.3 Clustering and Classification

Approaches for clustering, classification, and regression of high dimensional complex data can be divided into two categories, *universal* and *domain specific* [36], [32], [25]. Domain specific approaches utilise invariances and/or structural constraints specific for the data at hand and are suitable for analysing data sets where data is of the same kind, such as images. Universal approaches do not use such information and are suitable for analysing heterogeneous data sets.

Universal approaches

Among the universal approaches, the most common methods are based on machine learning theory. These are well suited for discovering hidden linear relationships in low noise data, but data in life sciences is often highly noisy and relationships are highly nonlinear. Other approaches have appeared that offer methods better suited for such data.

First we have *diffusion geometry* which, based on tools from harmonic analysis, provides means for elucidating the large-scale geometry of a manifold or a graph representing the data set. Diffusion geometry has been very successful in a variety of applications with a performance at state-of-the-art levels or better on standard community benchmarks [19]. Second, there is *topological data analysis* which provides a mathematical framework and techniques necessary to determine the fundamental geometric structures underlying massive data sets, *e.g.* to infer how discrete points assemble into a global structure [16]. These methods have the capability to recognise the number of parameters required to describe the global structure, without actually parametrizing. Third, we have *sparse signal processing* that, in the context of data analysis, provides a mathematical framework for compressing high dimensional data while preserving important features. Finally, *dynamical systems* and *algorithmic information theory* offer complexity measures on information that can be used to identify patterns in data without knowing the type of pattern.

Domain specific approaches

Methods of mathematical statistics are currently used in an ad hoc manner without careful consideration of their applicability to the situation of data analysis at hand. Data analytics needs a unified, consistent and mathematical theory for model identification in probabilistic models with computationally tractable methods taking into account the complexity of the estimators (several classical

methods of data reduction turn out to be intractable for complex systems). One example is the need to develop computationally tractable Bayesian networks for large and complex systems

Image data represents a class of data that is still mainly analysed manually. Extracting objects from a noisy background and recognition and classification of shapes of biological objects from images are perhaps among the most challenging problems in image analysis.

B.4 Mathematical structures

A key step in modelling a phenomena is to assign proper mathematical structures, often called *spaces*, to encode those aspects of the phenomena and data that are relevant for the problem at hand. In fact, all types of data analysis is based on some underlying choice of space.

For example, the problem of searching the elements of a data set that are close to a some query under some similarity criterion has a vast number of applications from pattern recognition to textual and multimedia information retrieval. One may consider the cases where the similarity criterion defines a metric space, instead of the more restricted case of a vector space. If the data set has a metric (the data consists of points embedded in some metric space), then one can talk about a distance between its elements.

A *space* is the abstract representation of the concrete situation being modelled. It is not just a universal set, it carries with it a number of useful concepts and determines what kind of mathematical operations one can perform. If the space has a topology, then one has a notion of how close two elements are. The algorithms currently invoked in data analytics will encounter difficulties (the curse of dimensionality) on high dimensional metric spaces [18], that is, those whose histogram of distances has a large mean and/or a small variance. This so-called curse of dimensionality, well known in vector spaces, is also observed in metric spaces. In high dimensional spaces even big data may become sparse. The research summarized in [4] shows that the notion of a distance in high dimensions becomes problematic as an approach to clustering. This problem is more fundamental than the performance degradation of high dimensional algorithms.

Many solutions have been proposed in different areas, in many cases without interdisciplinary knowledge of efforts by others. Some basic results that explain the quantitative definition of the concept of intrinsic dimensionality are needed. From the data modeling point of view, the shift towards the setting of large $p =$ dimension of the space), small $n =$ the number of sample points, and where $\lim_{n\to\infty} p(n)/n$ is a kind of intrinsic dimensionality, holds a great promise for data analytics, but it also implies previously unseen unique modeling challenges [15], [40]. In particular, the traditional analytic modeling frameworks that were developed typically under settings in which the number of study samples (n) exceeds the number of study variables (p) will not be ideally suited for data analytics with massive data analytics.

Nonlinear dimensionality reduction (manifold learning)

In general, dimensionality reduction (DR) [36] means the following problem: Given an observed high-dimensional data set (e.g., in a metric space), find a

low-dimensional data set such that certain tasks of data analysis of the original high-dimensional data can be realized on the low dimensional data within a tolerable error. The DR problems are usually classified into two categories.

Hard dimensionality reduction problems: In hard dimensionality reduction problems, the extrinsic dimensions of the data usually range from hundreds to hundreds of thousands of components, and usually a drastic reduction (possibly of orders of magnitude) is sought out. The components are often repeated measures of a certain magnitude in different points of space or in different instants of time.

In this category we would find pattern recognition and classification problems involving images (*e.g.*, face recognition, character recognition, etc.) or speech (*e.g.*, auditory models).

Soft dimensionality reduction problems: In soft dimensionality reduction problems, the extrinsic dimensions of the data usually are not too high–less than a few tens, for example, and the reduction is not very drastic. Typically, the components are observed or measured values of different variables, which have a straightforward interpretation.

Most cases of statistical analysis in fields like social sciences, psychology, etc., fall in this category. Multivariate analysis is a main tool for soft DR.

Usually, the main focus of soft DR is not on data reduction, but on data analysis. The dividing line between these two categories is not very strict. Some techniques, such as PCA (= principal component analysis), may be used in both of them.

Visualization often needs to make use of DR, where the high-dimensional data need to be reduced to 2- or 3-dimensions in order to display it. Several presentation techniques allow one to visualize up to about 5-dimensional data, using colors, rotation, stereography, glyph or other devices, but they lack the appeal of a simple display. Both soft DR and hard DR methods may reduce high-dimensional data to visible data. To choose a better DR method for the visualization of a particular data is quite dependent on the purpose of the visualization.

Below we survey the various nonlinear manifold learning models that have emerged during the recent two decades.

Minimum Variance Unfolding MVU Maximum variance unfolding was formerly known as semi-definite embedding or semi-definite programming [56]. The intuition for this algorithm is based on that when a manifold is properly unfolded, the variance over the points is maximized.

This algorithm begins by finding the k-nearest neighbors of each point, then maximizes the variance over the whole data set, preserving the distances between all pairs of neighboring points. MVU preserves the local data geometry very well. It can also be applied to similarity configuration, serving as a nonlinear MDS. However, MVU kernel is dense too. Hence it has a high computational cost. When the dimension of input data is too high, landmark technique is often adopted to reduce the computational time. Some other methods adopt the similar idea as MVU. For example, manifold sculpting [28] uses graduated optimization to find an embedding. It also computes k-nearest neighbors and tries

to seek an embedding that preserves relationships in local neighborhoods. It slowly scales variance out of higher dimensions, while simultaneously adjusting points in lower dimensions to preserve those relationships. If the rate of scaling is small, it can find very precise embeddings. The similar idea is also adopted by locally multidimensional scaling [54], which performs multidimensional scaling in local regions, and then uses convex optimization to fit all the pieces together.

LLE This is a method to embed linearly each neighborhood on the data graph to a low-dimensional space so that the manifold the data resides on is mapped to the same space. LLE, which is described in [52], begins by constructing k-nearest neighbors of each point, and then computes the barycentric coordinates of the point with respect to its neighbors to describe the local similarity. Finally, the DR data is obtained by preserving the similarity weights.

LLE produces sparse DR kernel. Taking the advantage of sparsity, it implements faster than Isomaps and MUV. LLE tends to handle non-uniform sample densities poorly because there is no fixed unit to prevent the weights from drifting as various regions differ in sample densities.

LTSA This method, first descried in [57], is based on the intuition that when a manifold is correctly unfolded, all of the tangent hyperplanes to the manifold will become aligned. The method is very similar to LLE. It begins by computing k-nearest neighbors of every point, then computes a. basis for the tangent space at every point. The local coordinates associated with the basis produce the weights between the point and its k-nearest neighbors. The weights then derive a sparse DR kernel. LTSA finds an embedding that aligns the tangent spaces by using the eigenvector optimization of the DR kernel.

Laplacian Eigenmaps Laplacian eigenmaps [7] use spectral technique in DR processing. In Laplacian eigenmaps, the high-dimensional data is still assumed to be laid on a low-dimensional manifold. The method first builds a graph that defines a neighborhood system on the data set, then constructs a kernel that approximates the Laplace-Beltrami operator on the manifold. The near-zero eigenfunctions of the Laplace-Beltrami operator are used as the embedding.

Hessian locally linear embedding Like LLE and Laplacian eigenmaps, Hessian locally linear embedding [26] also produces a sparse kernel. It tends to construct the approximate Hessian operator on the manifold on which the given data resides so that the DR data is obtained from the near-zero eigenvectors of the operator. It often yields results of higher quality compared with LLE and Laplacian eigenmaps, when the given data is smooth. As a trade-off, it has a more costly computational complexity. Besides, it is relatively sensitive to data noise.

Diffusion maps The idea here is that the Neumann heat diffusion operator on a manifold has the same set of eigenvectors as the Laplace-Beltrami operator, but the small eigenvalues of the Laplace-Beltrami operator becomes the largest ones. The method, described in [19], constructs DR kernel that approximates the heat diffusion operator. The DR data then is provided by several leading eigenvectors of the kernel (*i.e.*, the eigenvectors achieving several largest eigenvalues).

In many applications, the dimensions of DR kernels can be very large. That will cause both memory and computational cost problems, and the computational stability problems. To overcome the difficulty caused by the large sizes of DR kernels, one has used some fast algorithms for spectral decompositions of DR kernels.

B.5 Bayesian, signal processing and neural networks techniques

Bayesian Machine Learning

The Bayesian approach rests on prolific thinking rooted in an astoundingly simple formula [12]. As discussed above, it is already well understood that the massively high-dimensional data, combined with small sample sizes, bring new statistical challenges [13], [50]. With multiple data sources the challenges are even harder but, on the other hand, new opportunities become simultaneously available for using advanced data integration methods, which will be developed in foundational data analytics research.

Machine learning provides in general a powerful approach to predictive models for identification, diagnosis and prognosis, complementing mechanistic modeling. Bayesian machine learning approaches enable computing data-driven models which combine prior knowledge and several data sources, producing predictions and hypotheses on action mechanisms [23]. However, acquiring machine learning solutions on data is in many contexts a high-performance computing challenge.

Machine learning based data integration methods can be divided in two categories: unsupervised and supervised. Supervised methods are more common; their goal is simple in principle: form any combination of the data sources which results in maximal regression or classification performance. In practice, in particular for very high-dimensional data, this goal is far from trivial both statistically and computationally. Unsupervised methods find and characterize relationships between the data sources. They are needed for complementing the supervised methods in cases, where it is important to understand the relationships in more detail.

A fully Bayesian approach is particularly well suited for addressing complex questions in integrative biology regarding structural links and information synthesis between different sources of data, for building in hierarchical relationships and informative priors based on substantive knowledge and for assessing the relative weight of alternative strategies in translational genomics. It facilitates the propagation of uncertainty, and therefore leads to realistic estimates of the statistical reliability of discoveries, which is of paramount importance in the health sciences.

Causality and complex data

Massive data is being collected in medical registries and databases: these are prime examples of heterogeneous data. Due to the unique personal identification numbers used in the Nordic countries, these countries are in an exceptional position to follow up individuals and to connect various registries and cohorts.

A view of causality that is very natural in medicine is the counter-factual one which focuses on the question: What would have happened if the intervention had not been made (or if a different intervention had been made)? So one compares the actual situation (with intervention) with the counter-factual one (without intervention) to see what is the causal effect of the intervention. New causal methods are particularly needed when there is time-dependent confounding, meaning that confounders are processes developing over time as opposed to simple variables measured at a given time, e.g. when dynamic treatment regimes are considered. The concept of a directed acyclic graph (DAG) is central in the causality calculus. Such graphs are describing Bayesian networks, but they are given a causal interpretation by an ingenuous definition of intervention [49].

The analysis of data in health science is presently undergoing fundamental changes. The main connecting idea is an attempt to systematically look at causal connections [2], [3]. Causal modeling has over recent years become a major topic, with emphasis on graphical models, networks, Bayesian nets [38] and counter-factual models. New methods for causal analysis can elucidate the effects of changes and developments over time. The new ideas on causality are not a panacea, solving all problems of causality, but they do represent a more systematic approach to this difficult field. Hence we end up in conclusions about causal research and big data antithetic to those in [44].

Data compression

Data compression has a close relation with DR. It is another technique often used in the high-dimensional data acquisition and processing. Data compression or source coding is the process of encoding data down to size smaller than their normal presentation. Because of the huge size of high-dimensional data, it is necessary to transmit the compressed data instead. There are two compression schemes: loss-less and lossy. Lossy compression scheme can reach a higher compression rate than the loss-less one, but, as a trade-off, will lose some information. Hence, the quality control for a lossy compression scheme is crucial. There is a close relation between DR and compression: a high-dimensional data can be first reduced to a low-dimensional data, and then compressed further. Since the compressed data still need to be restored to the original one (within a tolerance), usually only linear DR methods can be used as a pre-process for data compression. The main difference between DR and data compression is that the main purpose of DR is efficiency of data processing while data compression seeks efficiency behavior of data transmission and of automated learning. The compressed data cannot be used in data processing unless it is restored, whereas DR data can.

Recently, compressive sensing has motivated intensive research activities. This new methodology integrates the compression schemes into the data acquisition so that no additional compression is necessary for the acquired data. The readers interested in compressive sensing may refer to [14, 27].

SOM

Also called self-organizing feature map, or Kohonen maps [37], SOM is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional) discretization of the input space

of the training samples. Self-organizing maps are different from other artificial neural networks in that they use a neighborhood function to preserve the topological properties of the input space [53].

ICA

Independent Component Analysis (ICA) is a computational method for separating a multivariate signal into components that they are statistically independent from each other [35, 22]. In linear ICA the data are represented by the random vector $x = (x_1, \ldots, x_m)^T$ and the components as the random vector $s = (s_1, \ldots, s_n)^T$, and the task is to transform the observed data x, using a linear static transformation W as $s = Wx$ into maximally independent components s measured by some score function $F(s_1, \ldots, s_n)$.

Deep learning and similar approaches

Deep learning (also called deep structural learning or hierarchical learning) is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations [3]. While the potential advantage of such architectures has been appreciated for a long time, see *e.g.* [8], the recent interest stems from new algorithms to (approximately) learn such models incrementally, a prominent being (fast learning of) restricted Boltzmann machines [34]. Deep learning has attracted considerable attention in the public media as some of the most visible researchers have been hired by large companies to set up labs (Yann LeCun by Facebook and Geoffrey Hinton by Google, both in 2013), and by the success of IBM's DeepQA to win the US television quiz show *Jeopardy!* [1]. Dr Abdel Labbi from IBM Research Zurich spoke about DeepQA and future applications at the KTH Symposium on Big Data on May 26, 2014. A related approach to nonlinear dimensionality reduction is autoencoders, a special kind of feed-forward neural networks. Related to autoencoders is the NeuroScale algorithm, which uses stress functions inspired by multidimensional scaling and Sammon mappings to learn a non-linear mapping from the high-dimensional space to the embedded space. The mappings in NeuroScale are based on radial basis function networks.

B.6 A Case: Complex and/or large data sets in Life Sciences

Recent biotechnological advances in next-generation sequencing, tandem mass-spectrometry, high-throughput screening, cellular imaging, etc, are enabling more in-depth insights into the individual disease processes by generating millions of data points for each subject under analysis. One important development is that whole-genome sequencing is becoming accessible to small laboratories and its price is transformed into a low cost biological assay.

However, the large-scale profiling experiments are notoriously prone to technical variability and instrument-specific biases; therefore, looking at any single data source alone will lead to limited and potentially biased views and predictions. For instance, the massive number of genetic variants found in genome-wide association studies or in exome/whole-genome sequencing makes it very

[3] The source of this concise definition is Wikipedia.

challenging to distinguish between the variants truly associated with a disease and those originating from technical or disease-independent variability, leading to frequent false positive and negative findings [43]. Moreover, the exponentially increasing number of potential interactions between the variants makes the pure experimental approach quickly unpowered, and translates into a need for integrated approaches that are scalable, robust and economical.

New technologies, capable of transforming the economics and speed of data acquisition, also provide an unparalleled opportunity to collect large scale data from multiple scales, thereby further driving the usage of large-scale complex data sets.

Efficient integration of complementary information sources from multiple levels can greatly facilitate the discovery of true causes and states of various phenomena. Recent activities in, e.g., biocomputing demonstrate a vivid activity among these themes.

B.7 Conclusions

There is a big scientific gain to be expected in the development of the potential and depth of big data analytics. The conclusion of the round-table conference [11] is that the findings of data analytics will be *generators of new theories* in different domains of science. The direct societal and economic potential lies in the production of more information out of projects in big data that can be acted upon.

The analytical techniques surveyed above and those perhaps overlooked, for one reason or other, tend to improve incrementally. Therefore they do not provide in the short run the quick spectacular breakthroughs improving the business opportunities and the profit chances (targeted by CEOs of corporations and their ilks).

There should exist principles underlying the analytical techniques that link measures of inferential accuracy with intrinsic characteristics of the data generating process and with computational resources such as time and space [48]. These principles can, however, be searched for only after experience with concrete and specific problems of Big Data analytics.

References

[1] The deepqa research team.

[2] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer, 2008.

[3] Odd O Aalen, Kjetil Røysland, Jon Michael Gran, and Bruno Ledergerber. Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175:831–861, 2012.

[4] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *ICDT'2001 Lecture Notes in Computer Science, pp. 420–434*, 2001.

[5] Charles Auffray, Timothy Caulfield, Muin J Khoury, James R Lupski, Matthias Schwab, and Timoth Veenstra. Genome medicine: past, present and future. *Genome Med*, 3, 2011.

[6] Ian Ayres. *Super crunchers: How anything can be predicted*. John Murray, 2007.

[7] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[8] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2:1–127, 2009.

[9] Roger Bohn and James Short. Measuring consumer information. *International Journal of Communication*, 6:1000–10026, 2012.

[10] Roger E. Bohn. How much information. 2009 report on arnerit consumers. Technical report, University of Califomia, San Diego, Retrieved from httpi. hmi. ucsd. edu/howmuchinfomresearch_, 2010.

[11] David Bollier and Charles M Firestone. *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC, USA, 2010.

[12] Uri Bram. *Thinking Statistically. 2nd Edition*. Kuri Books, 2012.

[13] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. A computational framework for infinite-dimensional Bayesian inverse problems. part I: The linearized case, with applications to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013.

[14] E. Candes and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[15] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[16] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009.

[17] Saikat Chatterjee, David Koslicki, Siyuan Dong, Nicolas Innocenti, Lu Cheng, Yueheng Lan, Mikko Vehkaperä, Mikael Skoglund, Lars K. Rasmussen, Erik Aurell, and Jukka Corander. SEK: sparsity exploiting k-mer-based estimation of bacterial community composition. *Bioinformatics*, 30(17):2423–2431, 2014.

[18] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM computing surveys (CSUR)*, 273–321, 2001.

[19] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21:5–3, 20060.

[20] Validation Committee on Mathematical Foundations of Verification, Uncertainty Quantification; Board on Mathematical Sciences, Division on Engineering Their Applications, and National Research Council Physical Sciences. *Assessing the Reliability of Complex Models: Mathematical and*

Statistical Foundations of Verification, Validation, and Uncertainty Quantification. The National Academies Press, 2012.

[21] National Research Council Committee on Mathematical Sciences Research for DOE's Computational Biology. *Mathematics and 21st Century Biology*. The National Academies Press, 2005.

[22] P. Comon and Jutten C. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[23] J. Corander, T. Koski, and M. Ekdahl. Parallell interacting MCMC for learning of topologies of graphical models. *Data Mining and Knowledge Discovery*, 17:431—456, 2008.

[24] Thomas H Davenport, Paul Barth, and Randy Bean. How 'big data' is different. *MIT Sloan Management Review*, 54, 2013.

[25] Luc Devroye. *A probabilistic theory of pattern recognition*. Springer, 1996.

[26] D. L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100:5591—5596, 2003.

[27] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer Verlag, Basel, 2013.

[28] M. Gashler, D. Ventura, and T. Martinez. Iterative non-linear dimensionality reduction with manifold sculpting. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 513–520, Cambridge, 2008. MIT Press.

[29] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[30] O. Ghattas, T. Isaac, J. Martin, N. Petra, and G. Stadler. Big data meets big models: Large-scale bayesian inverse, with applications to inverse modeling of antartic ice sheet dynamics. In *11th World Congress on Computational Mechanics (WCCM XI), 5th European Conference on Computational Mechanics (ECCM V), 6th European Conference on Computational Fluid Dynamics (ECFD VI), July 20–25, 2014, Barcelona, Spain*, 2014.

[31] David J Hand, Heikki Mannila, and Padhraic Smyth. *Principles of data mining*. MIT press, 2001.

[32] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*. Springer, 2009.

[33] Martin Hilbert and Priscila López. The world's technological capacity to store, communicate, and compute information. *Science*, 332:60–65, 2011.

[34] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

[35] Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*. J. Wiley, 2001.

[36] Alan J Izenman. *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer, 2009.

[37] Teuvo Kohonen. *Self-organization and associative memory*. Springer, 1988.

[38] T. Koski and J. M. Noble. *Bayesian Networks. An Introduction*. Wiley, 2009.

[39] David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343:1203–1205, 2014.

[40] Woojoo Lee, Donghwan Lee, Youngjo Lee, and Yudi Pawitan. Sparse canonical covariance analysis for high-throughput data. *Statistical Applications in Genetics and Molecular Biology*, 10:1–24, 2011.

[41] Sam Madden. From databases to big data,. *IEEE Internet Computing*, 16:0004–6, 2012.

[42] Volker Markl. Breaking the chains: On declarative data analysis and data independence in the big data era. In *Proceedings of the VLDB Endowment*, 2014.

[43] Pekka Marttinen, Jussi Gillberg, Aki Havulinna, Jukka Corander, and Samuel Kaski. Genome-wide association studies with high-dimensional phenotypes. *Statistical applications in genetics and molecular biology*, 12:413–431, 2013.

[44] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. John Murray (Publishers), 2013.

[45] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[46] Andrea Montanari. Computational implications of reducing data to sufficient statistics. arXiv preprint arXiv:1409.3821.

[47] The Swedish Big Data Analytics Network. The big data analytics. the research and innovation agenda for sweden. Technical report, SICS bdainfo@sics.se https://www.sics.se/projects/big-data-analytics, 2013.

[48] Committee on the Analysis of Massive Data; Committee on Applied, Theoretical Statistics; Board on Mathematical Sciences, Their Applications; Division on Engineering, and Physical Sciences; National Research Council. *Frontiers in Massive Data Analysis*. The National Academies Press, 2013.

[49] Jude Pearl. *Causality: models, reasoning and inference*. Cambridge Univ Press, 2000.

[50] N. Petra, J. Martin, G. Stadler, and O. Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems. part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM Journal on Scientific Computing*, 2014. Accepted.

[51] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *Science*, 334:1518–1524, 2011.

[52] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323—2326, 2000.

[53] A. Ultsch. Emergence in self-organizing feature maps. In *Workshop on Self-Organizing Maps (WSOM 07). Bielefeld (2007)*, 2007.

[54] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19(6):889—899, 2006.

[55] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[56] K. Q. Weinberger, B. D. Packer, and L. K. Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *Proceedings of the 10th International Workshop on AI and Statistics*, 2005.

[57] Z. Y. Zhang and H. Y. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1):313—338, 2004.