KUNGLIGA TEKNISKA HÖGSKOLAN

KTH-Aalto initiative on Big Data to Small Information

ICT platform white paper

Erik Aurell (CSC, Chair), Scott Kirkpatrick(KTH Dr HonC), Timo Koski (SCI), Mikael Skoglund (EES), Ozan Öktem (SCI)

1/22/2013

This document surveys the field of Big Data and the new scientific challenges and opportunities posed by Big Data. Big Data is today a commercial reality, a result of the success of data aggregators such as Amazon and Google, and it is supported by advances in flexible cloud computing for storage, maintenance and data delivery. These developments will have a profound impact on the conduct of science in fast-moving areas because of their generality. While in the past, Big Science such as the LHC and the confirmation of the existence of the Higgs Boson has involved vast amounts of data distributed to a data grid with enormous computing power, these have served one experiment, and a small group of people, answering only one question. Astronomy is an example in which vast amounts of data serve a larger group of people, ranging from "black belt" experts in a single observational mode, to astrophysicists studying the data in multiple spectral bands, to school children around the world. Biology is the next frontier for Big Data enabling discoveries by many scientists to be extracted from the flood of DNA, RNA and protein sequence data now available. In all of these areas, the key to unlocking these secrets is concepts, methods and analytical tools to do more and to do different things with Big Data. The document argues that Big Data is a potential area of strength of a KTH-Aalto University collaboration which should also be of interest to many of KTH's industrial partners. The document was commissioned by the KTH ICT platform and has been produced by a committee of Erik Aurell (CSC, Chair), Scott Kirkpatrick (KTH Dr HonC), Timo Koski (SCI), Mikael Skoglund (EES) and Ozan Öktem (SCI), and with the participation of Dr Danny Bickson (Carnegie Mellon University, USA), and Academy of Finland Centre of Excellence on Computational Inference Research (Aalto University, Finland). The opinions and recommendations expressed are in all cases those of the authors and do not necessarily reflect those of their respective units or organizational affiliations.

Extended executive summary

Background and Objective: 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 ($6x10^{23}$) bytes per year, while several large projects in science by themselves already produce petabytes (10^{15}). Big Data on such a scale no longer by itself confers information on which humans can act. From Big Data we need to extract Small Information which is actionable, and which can be used to advance modern societies, and modern science and technology. The objective is to make KTH and Aalto world-wide leaders in this endeavor.

Challenges and Methods: Big Data can be used in simple ways such as computing averages, correlations, etc. The challenges are to use Big Data also in more intelligent ways which scale to very large data sizes. This translates to developing tools and theories for inference from Big Data for basic tasks such as estimation, prediction, clustering and model learning. The underlying assumption is that this can be done if and when Information is much smaller than Data. A confluence of interests in many different fields such as sparse signal processing, Bayesian hierarchical modeling and statistical physics and is making this a very active area.

Opportunities and Timeliness: Sweden and Finland are world leaders in telecommunication, and play leading roles in the European Institute for Innovation and Technology (EIT) KIC ICTlabs. KTH hosts Swedish Science Council (VR) Centre ACCESS on complex networked communication systems, the largest research effort in this area in Europe, recently very successfully evaluated by VR. Many groups in ACCESS have recently moved into sparse signal processing problems. KTH and partners have created SciLifeLab, a center combining advanced technical know-how and state-of-the-art equipment generating many different types of Big Data in the life sciences. Finnish research is a long-term world leader in data analysis. Aalto University and University of Helsinki host the recently created Academy of Finland Centre of Excellence in Computational Inference Research (COIN). In conclusion, there is a window of opportunity to establish KTH and Aalto and Sweden and Finland as leaders in a Big Data effort, in EIT, in applications to life sciences and health, and more generally.

Table of Contents

What is Big Data?

In this report we take Big Data to mean data so large that it cannot be comprehended and directly used by humans. It is today a commercial reality, directly with data aggregators such as Amazon and Google, but also with credit cards, cell phones, and everywhere else, and is supported by advances in flexible cloud computing for storage, maintenance and data delivery. We will argue that these developments will have a profound impact on the conduct of science and society because of their generality. We will also argue that a main focus should be to develop theories, methods and tools to use and to exploit Big Data in novel ways which scale to very large data sizes.

While we will be vague on precisely what (present or future) data is or is not "big", we note that petabytes (PB, 10^{15} bytes) of data are reached by today's largest commercial actors and single research projects[1]. Let us recall the storage units of megabyte (MB, 10^6 bytes), gigabyte (GB, 10^9 bytes), terabyte (TB, 10^{12} bytes) and, on the other side, exabyte (EB, 10^{18} bytes), zettabyte (ZB, 10^{21} bytes) and yottabytes (YB, 10^{24} bytes). According to well-known estimates the world's annual production of digital data reached zettabytes in 2011[2], and could reach yottabytes in the next ten years. To put this in perspective, one YB would weigh about *two hundred million tons* if stored in current top-of-the-line hard disks which you, gentle reader, may have in your computer[3]. Whatever the accuracy of these projections, in many situations and in many applications data much smaller than one PB will for all purposes be "big". We can therefore assume that Big Data is here, and that there will only be more of it.

Big Data has been surveyed and high-lighted in many contexts in the last years. A 2010 White Paper from TIVIT – the Finnish Strategic Centre for Science, Technology and Innovation in the Field of ICT[4] – focuses on data reserves as a potentially strategically important resource, *i.e.* on the direct ownership of the data as well as of the physical devices where it is stored:

> *Finland is a small, safe country at the very center of current western civilization. Finland is also a neutral and technically advanced country – with a good or arguably leading position in many global rankings [...] Internet world allows us to build as many bridges as we need to serve Europe, Russia, USA or China or any other global market. Finland's strong bedrock, clean water, secured energy availability and even empty factories are also assets in the data reserves business, where safety and stability are important* (Paajanen and Kuosmanen 2010, Sect 9*).*

The above scenario is already a reality as shown by the Google Hamina data center in Summa, Finland[5] and the new server farm to be built by Facebook in Luleå, Sweden[6]. The

[1] Google had about 1 PB under active management in 2006. The LHC at CERN produced about 25 PB per year in the search for the Higgs boson.
[2] Wikipedia, Zettabyte, http://en.wikipedia.org/wiki/Zettabyte
[3] A Seagate Barracuda model ST3000DM001 stores 3 TB and weighs 626 g. One would need 3 x 10^{11} such devices to store 10^{24} bytes which would weigh about 2 x 10^{11} kg. One YB would also weigh *about four kilo* if stored in double-stranded DNA, the most compact, scalable and stable storage medium currently known. Sources: Wikipedia, http://www.invitrogen.com/site/us/en/home/References/Ambion-Tech-Support/rna-tools-and-calculators/dna-and-rna-molecular-weights-and-conversions.html, and elementary estimates.
[4] Reijo Paajanen and Pauli Kuosmanen "White Paper: Finland and Data Reserves", *Tieto- ja viestintäteollisuuden tutkimus TIVIT Oy*, 15.9.2010.
[5] Helsingin Sanomat, 2010-15-1 "Google raotti ovea hakukoneen uumeniin Summassa" and http://www.google.com/about/datacenters/locations/hamina/

Hamina data centre currently employs around 100 people, mostly Finnish and from the local community, working on site as computer technicians, water/air engineers, in catering, security and in other roles; the Facebook Luleå server farm is expected to create at least about the same number of jobs when in operation.

The perspective of this white paper is that data storage, data management and data ownership are only aspects of one side of the issue, and that an equally important challenge is what to do with the Big Data one has access to. Much of it will always be publically available, either as produced by large research projects, or as generated in government or by citizens. This evaluation is supported by the conclusions of a 2011 McKinsey report[7] making the following (strong) recommendation:

> *Following a comprehensive study, we note that recent big data technological trends are to meet huge computational requirements, by using distributed systems and cloud computing, ultimately requiring high energy consumption and data communication bottlenecks. We note there is no concerted large scale effort to develop new analysis and inference tools [...] we believe that the glaring lack in developing new analysis and inferring tools will soon bring saturation in big data handling performance.*

Ericsson has launched the vision of "50 billion devices" or "everything that can benefit from a network connection will have one", stressing that in the future most digital data will be actually or potentially linked to the Internet and thus actually or potentially accessible by anyone (or anything) and from anywhere, allowing for many new types of Big Data. While a recent Ericsson White Paper[8] focuses on the concomitant business, policy and technological challenges, it is also pointed out that

> *The human aspect of more than 50 billion connected devices goes way beyond smart living and new gaming devices. Ubiquitous connectivity is about being able to control things in a way that saves time and simplifies life [...] Consumers would prefer not to be bombarded with generic, irrelevant information and messages. If consumers choose to opt in to advertising- or promotion-funded services, the messages and information they are sent needs to align with their interests, whether they are out shopping or visiting a new area. Such capabilities are already about to be realized, but can be enhanced through service enablement and data analyses, for example of location, interests and history* (Ericsson White Paper *op cit*, page 9).

In other words, learning to do more with Big Data one has, as opposed to getting more Big Data to what one already knows to do, is also part of Ericsson's "50 billion devices" vision. In addition, Ericsson has clearly pointed out Big Data as one main area for new collaboration with KTH, in the framework of the newly established KTH—Ericsson strategic partnership. A similar view is offered by The World Economic Forum, which in 2012 conducted several

[6] *Invest in Sweden* http://www.investsweden.se/Sverige/Press-information/Facebook-till-Lulea/ and e.g. *Ny Teknik*, November 7, 2012 as well as continuous coverage in *Norrbottenskuriren* and *Norrländska Socialdemokraten*.

[7] J. Manyika, M. Chui, J. Bughin, B. Brown, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*, 2011.

[8] "More than 50 billion connected devices", Ericsson white paper 284 23-3149 Uen | February 2011

panel discussions on Big Data, producing a report from which one can quote the following summary[9]:

> *A flood of data is created every day by the interactions of billions of people using computers, GPS devices, cell phones, and medical devices. Many of these interactions occur through the use of mobile devices being used by people in the developing world, people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realise the potential for channelling these torrents of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises for the benefit of low-income populations. Concerted action is needed by governments, development organisations, and companies to ensure that this data helps the individuals and communities who create it.*

On the research funding side, in March 2012, the US federal government announced a national "Big Data Research and Development Initiative"[10] with a funding budget of USD 200 million and an additional USD 250 million investment by the Department of Defense (USA) to *improve the tools and techniques needed to access, organize, and clean discoveries from huge volumes of digital data*. In Finland the Academy of Finland finances a Centre of Excellence in Computational Inference Research (COIN, Aalto University and University of Helsinki) while in Sweden there is as of today no directed research funding to Big Data[11].

Finally, in June 2012 the Royal Society (UK) released "Science as an open enterprise"[12] with a number of recommendations for the generation, preservation and dissemination of data to facilitate maximal impact from research data, and enhance its reproducibility in an age of Big Data. In particular, this report identifies six main areas where change is needed:

> *(1) developing greater openness in data sharing; (2) developing appropriate reward mechanisms for data generation, analysis and dissemination; (3) developing data standards to enable interoperability; (4) making data associated with published papers accessible, and amenable to assessment and reuse; (5) developing a cadre of 'data scientists'; (6) developing new tools for data analysis.*

While first five areas are more policy requirements, the sixth area requires new technological breakthroughs and speaks directly to the focus of this white paper.

Case studies of Big Data and its uses

Astronomy

[9] "Big data, big impact: New possibilities for international development," *A report by World Economic Forum*, 2012.

[10] "Obama administration unveils "Big Data" initiative: Announces $200 million in new R&D investments", *Executive Office of the President of United States*, 2012.

[11] Some support may be obtained if calls will be launched within the areas "Data mining" and "Streaming data" listed in *Swedish Foundation for Strategic Research* Research Strategy 2012-17, Section 5.5 "Data-X" , from the Swedish Science Council currently running program "Statistics in the Empirical Sciences", as well as from the Wallenberg foundation and other national and international research funding bodies.

[12] "Science as an open enterprise," *The Royal Society Science Policy Centre report 02/12*, 2012.

The UK Royal Society's report "Science as an open enterprise" (Ref 12, *op cit*) presents numerous examples of the social structures and infrastructure that enable fields of science such as biology to advance faster through widely shared access to and reuse of experimental data. Astronomy figures prominently in this as it was one of the first sciences to face Big Data issues and as it lacks the complication of commercial applications. New astronomical data needs to be kept private only long enough to allow careful analysis before it is first published, and publication usually requires that the data be placed in a public archive, a service that is often provided by the telescope facilities at which the data is gathered. Since telescopes are expensive and long-lived, and serve many astronomers, access to them is competitive. Making the eventual data public is typically one of the requirements that every proposed measurement program in astronomy must meet.

There are many specialties within astronomy, distinguished principally by the spectral band within which the observations are made. The newer modes, such as x-ray (done only from satellites), and radio (which may use antenna arrays) provide data which is very useful in combination with more conventional observations, but require special skills to calibrate and assess. As a result, the issues of metadata standards are now being addressed by international organizations that have been created from the bottom up by astronomers to permit "virtual astronomy (VA)." A very recent trend is telescopes that are operated as survey instruments, scanning the whole sky and placing their data immediately into a public archive for analysis. The Large Synoptic Survey Telescope (LSST) is a very high technology example of this. When completed in 2022 it will be producing 100s of PB of data per day. Many of the currently most active questions explored in astronomy require comparing observations of the same parts of the sky that are made by these very different techniques, so means of finding the highly distributed data held at different facilities' archives, and of detecting anomalous or "interesting" combinations of characteristics needed for a particular study are critical. Registries that locate data at various frequencies from a particular object are being created by the US (NSF-sponsored) VAO and the Euro-VO and standardized though international organizations such as the IVAO. But the newest observational facilities, for example, the ALMA radio telescope array in Chile, will deliver data not as brightness as a function of spherical angle (the inverse of looking down at the earth and recording its color) but in "hypercubes" that lump together observations made continuously as a function of energy and time at each position. This richer data allows chemical analysis of the molecules found in space ("origins of life?") and treatment of the many interesting non-stellar phenomena now seen, such as jets of material entering black holes, or interstellar dust out of which stars are forming, or the many planets now characterized around nearby stars. The recent Nobel Prize-winning discoveries of dark energy and the ongoing search for dark matter have benefitted from the availability of multi-spectral data from the sky.

Tools for reducing the raw observational data to permit reuse and for analysis of the calibrated results to reach scientific conclusions have evolved along a path that applies to other fields besides astronomy. In each observational mode, the first experiments required developing and sharing calibration methods so that experimental reproducibility could be achieved. Subsequent analyses were first done by "black belt" astronomers (and their close disciples) using tools that few others could master. As the analyses have become better understood, a wider group of astronomers and cosmologists can take advantage of larger amounts of data, or propose useful experiments without the same deep knowledge of instrument and technique that the "black belt" group had. Visualization tools and statistical packages also become shared by different types of astronomers.

The above developments can be seen as templates more or less shared by many other data-rich sciences. A prime example of making data public is the publication (and hence sharing) of sequence data in the Life Sciences. In large targeted research projects it is inevitable that the discoveries primarily go to the groups developing and/or having first access to the latest instruments. Two of the most important examples would be the sequencing of the human genome[13] and the recent discovery of the Higgs boson[14]. Big Data however also allows for serendipitous discoveries such as Dark Energy and the accelerated expansion of the Universe – arguably the most important discovery in fundamental Physics since quantum mechanics and relativity theory – re-using data and/or re-using existing massive data-generating equipment in novel ways. An analogous example in the Life Sciences would be the discovery that Neanderthals interbred with modern humans[15] – one of the most important advances of all time in the understanding of human origins and pre-history.

Life Sciences

The stunning development of life sciences has to a large extent revolved around the ability to acquire data on a variety of scales of biological organization. Such data can be grouped into three kinds: image data, sequence data, and text data. Medical imaging of anatomical structures and microscopy imaging of tissues have for a long time been an indispensable part of healthcare, and microscopy imaging of cells and sub-cellular structures is central in modern molecular and cellular biology. Sequence data are the result of fundamental breakthroughs in DNA sequencing with applications to transcriptomics (RNA data) and epigenomics (DNA modification and structure data), and many other fields. Text data are medical records and biomedical scientific publications and any other large data source in text format, and will in this document be discussed below in *Commercial Big Data, Social networks, and Health care*. Due to their greater chemical diversity and available techniques, large-scale data on proteins (proteomics) and of small metabolites (metabolomics) are typically acquired by methods relying on imaging and can (in the present context) be considered image data. 3D structures of macromolecules and macromolecular assemblies have also mainly been acquired by imaging, but sequence data (as outlined below) and text data as documented in the scientific literature are also very important.

[13] "Initial sequencing and analysis of the human genome" and "Finishing the euchromatic sequence of the human genome", International Human Genome Sequencing Consortium, Nature **409** 860–921 (2001); **431** 931–45 (2004).

[14] ”A new boson with a mass of 125 GeV observed with the CMS experiment at the Large Hadron Collider”, The CMS Collaboration, *Science* **338** 1569-1575 (21 December 2012); “A particle consistent with the Higgs boson observed with the ATLAS detector at the Large Hadron Collider”, The ATLAS Collaboration, *Science* **338** 1576-1582 (21 December 2012).

[15] “A Draft Sequence of the Neandertal Genome”, R Green *et al*, *Science* **328** 710-722 (7 May 2010).

We first note that not long ago, all these data were not “Big”. Microscopic images were inspected by eye by the microscopist, genes were sequenced one at a time in the “one PhD student, one gene” mode, and medical records and data were not digitalized. Today, all of these data are however naturally “Big”: microscopic images can be obtained automatically in large quantities and analysed by software, and whole genomes can be sequenced in one batch.

One characteristic aspect of Big Data of the image kind is that they immediately call for advanced data processing, the prime example being various tomography techniques in medical imaging (CT, PET/SPECT, MRI/fMRI, etc.) which rely on inverse techniques. We will discuss inverse problems and inference below. Other examples from structural biology are to understand the structural 3D/4D conformation of proteins and macromolecular assemblies to understand their function in biological processes in time and space. X-ray crystallography[16], NMR and 3D Electron Microscopy[17] are all examples of methods that combine data collection strategies with methods for inference to provide such information. The main usage of inference in imaging has mostly been relegated to situations where measured data are not interpretable by humans, e.g., a human cannot by inspecting tomographic data easily visualize the 3D structures that give rise to this data. The potential for inference of Big Data is however yet to be fully exploited in the analysis of images and this is currently one of the main issues in imaging, i.e., to automate the imaging pipeline.

GenBank, the central repository of gene sequences, today contains about 37 GB annotated data[18]. By comparison, this is only about 50 times the human genome ($3 \cdot 10^9$ nucleotides, 0.75 GB) reflecting the fact that most DNA in humans and other higher organisms does not code for annotated genes. A single modern high through-put experiment can on the other hand produce of hundreds of millions of short DNA sequences called “reads”, which is also in the tens to hundreds GB depending on the platform. Recently individual human cells where sequenced[19]. Such experiments rely (at least at present) on the same modern whole-genome technologies which can be used to sequence environmentally collected samples, profile RNA and epigenetic changes and anything else. As the costs of such technologies continue to fall one can envisage sequencing thousands of individual cells. This would give TBs or even PBs of data on e.g. genomic diversity in tumors. One can also envisage sequencing millions of patients and/or samples of the bacterial flora from millions patients which would give PBs or even EBs of data. It is clear that the economic basis for such massive data gathering will be present as the cost is already today relatively low.

One should acknowledge that the discrepancy between the amount of data generated and the amount of useful information obtained has often been criticized in Biology, perhaps most famously by Sidney Brenner

[16] de La Fortelle & Bircogne, Methods Enzymol. 276, 1997

[17] Förster et al, Methods Enzymol. 483 2010

[18] NCBI-GenBank Release 193.0 (December 15 2012) held 161140325 loci, 148 390 863 904 bases, from 161140325 reported sequences.

[19] “Genome-wide detection of single-nucleotide and copy-number variations in a Single Human Cell”, C Zong et al, *Science* **338** 1622-1626 (21 December 2012); “Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing”, S Lu et al, *Science* **338** 1627-1630 (21 December 2012);

> *It is what I call 'low input, high throughput, no output science'!*[20]

A recent report in "Issues Online: In Science and Technology" highlights the challenges of Big Data in the Life sciences from the viewpoint of the Human Microbiome Project (HMP)[21]. Although this very large sequencing project has been the subject of high-profile recent publications[22], in an interview in Science, June 6, 2012, George Weinstock, one of the principal investigators for the HMP, also stated

> *"despite the huge amount of the work that has been done on the human microbiome, the number of rigorously proved connections between disease and microbiome are few to none"*

What is missing in the above is *a*) the prosaic point that a "a few" rigorously proved connections could well be worth all the effort from the societal point of view, even if the scientists involved had hoped for much more, and *b*) the methodological point that most likely not all that could be done with the data has been done. A somewhat related example are genome-wide association studies (GWAS) linking risk of disease to personal biomedical data and which use single-nucleotide polymorphisms (SNP) arrays to probe for hundreds of thousands to millions of genetic variants. The ultimate goal of such research is to create a database of disease signatures that can be used to predict the risk of disease in an individual, and then customize appropriate prevention or therapeutic efforts for personalized medicine. While there have been many reports that GWAS studies have not delivered all the information initially hoped for – perhaps because the analysis tools available have not been powerful enough – there is no question that single results in this direction, even if they are relatively rare, can be crucially important for single patients.

We will end this brief overview by mentioning a high-profile serendipitous discovery using Big Data accumulated by many research teams and using sequence data combined with inference. It has recently been realized that 3D protein structure can reliably be inferred from the statistics of thousands of similar but not identical protein sequences[23] in some sense solving the long-standing *protein structure prediction problem*. Let us recall that life tends to re-use the same building blocks, such that similar proteins performing similar functions can be found throughout all forms of life. A protein's 3D-structure leaves an echo of correlations in the evolutionary record, and in current sequencing projects many thousands of new genomes are produced every year. This *direct coupling analysis* to determine 3D-structure from many protein sequences is a paradigm for intelligent use of Big Data from the ongoing sequencing revolution, and will likely have far-reaching implications.

[20] Interview in *Nature Reviews Molecular Cell Biology* **9**, 8-9 (January 2008)

[21] M. Sagoff, "Data deluge and the human microbiome project," *Issues Online: In Science and Technology (National Academy of Sciences (US) , National Academy of Engineering (US), Institute of Medicine (US), and University of Texas at Dallas)*, 2012.

[22] The Human Microbiome Project Consortium "Structure, function and diversity of the healthy human microbiome",   *Nature* **486**, 207–214 (14 June 2012)

[23] F. Morcos *et al*, "Direct-coupling analysis of residue co-evolution captures native contacts across many protein families", *PNAS* **108**: E1293-1301 (2011) & T Hopf *et al*, "Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing", *Cell* **149**, 1607-1621 (2012).

Commercial Big Data, Social networks, and Health care

A fast growing area in practical deployment of Big Data use today is personalization and recommendations. In the past, companies focused their marketing efforts towards advertising popular items. With the growth of data collected about user history and preferences, inference methods can be used to personalize user experience by potentially computing for each user a personalized set of recommendations. The problem is that computing for each user a recommendation is communication intensive task (vs. the task of computing a single recommendation for the whole population of users).

The Netflix contest gave a significant boost to this area, proposing a one million dollar prize for researchers who can optimize their recommendation engine in 10%[24]. While in theory the winning method achieved the requested improvement, it was too difficult to implement on the full dataset because of algorithm complexity and data magnitude, and thus only a subset is actually deployed in Netflix today[25].

To emphasize how demanding is the personalization task we give some example related to Facebook data magnitude[26]. One billion Facebook users exchange 6 billion messages a day and upload 300 million photos a day. It is clear that computing personalization for one billion different users which have rich historical, geographical and social information is a computation intensive task.

Unfortunately, there is a big gap between the computation framework and tools for managing big data and the increasing magnitude of the data. There is still a significant amount of work needed for continuing developing algorithms and tools to catch up the growing data magnitude. Furthermore, according to a recent McKinsey report[27]:

> *The study projects there will be approximately 140,000 to 190,000 unfilled positions of data analytics experts in the U.S. by 2018 and a shortage of 1.5 million managers and analysts who have the ability to understand and make decisions using big data.*

The basis of advanced training is research, and academic research in Big Data issues is therefore indirectly also highly relevant to the commercial sector. As we will argue below, it is of particular importance to KTH and Aalto and Sweden and Finland that Big Data research and then advanced training of Big Data professionals on levels of second cycle (MSc) third cycle (PhD) and beyond are carried out here.

One sector where large data of the social network type is and will be available to academic researchers (at least in countries with largely socialized medicine) is however medicine and health care. Traditionally, drug discovery/development as well as medical therapies has been oriented toward diagnosing and treating individual organ systems, focusing on one disease at a time. Hence, current treatment guidelines are geared toward treating a "standard" patient

[24] http://www.netflixprize.com/
[25] http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html
[26] http://www-conf.slac.stanford.edu/xldb2012/talks/xldb2012_wed_1105_DhrubaBorthakur.pdf
[27] http://spotfire.tibco.com/blog/?p=6886

with a single illness. Most of us patients do not fit these narrow profiles, especially as we grow older and things get complicated. We (patients) might display symptoms common to a variety of illnesses, or we may already be suffering from multiple diseases while treatment guidelines are sometimes vague and may not exist at all when a patient has multiple diseases or is at risk for developing them. Treating patients with multiple conditions is also costly. To reduce costs, doctors need ways to identify early intervention opportunities that address not only the primary disease but also any additional conditions that a patient might develop. Consequently, health care providers are forced to adopt ad hoc strategies that include relying on their own personal experiences (and knowledge), among other approaches. Straying from those guidelines (where available) may not deliver the best outcomes.

Inference from Big Data should enable healthcare providers to use individual patient data (including both structured and unstructured data) as well as insights from a similar patient population to enhance clinical decision-making. Healthcare providers and payers should also be able to move beyond a one-size-fits-all approach to deliver data-driven, personalized care that helps improve outcomes, increase the quality of care and reduce costs. An example of usefulness of inference of Big Data analytics in the health care domain is IBM's DeepQA[28], a technical platform for analysis of textual data that can be used for analysis of hospital records.

Where are the research challenges?

The title of this white paper is "Big Data to Small Information." The overall research challenge we address is how to take Big Data whatever its size and shape and transform it to actionable information, in new ways. This holds as much in science, where information can be the basis of a hypothesis or the means to falsify a hypothesis, as in the commercial world, where information underlies business decisions.

Let us set the stage by outlining two types of research challenges *not* considered in this document.

First, Big Data needs to be stored, indexed, delivered, curated and otherwise managed, and for petabyte scale and up this is very challenging. Referring to the 2010 TIVIT white paper (Paajanen & Kuosmanen, *op cit*) such questions underlie – and are crucial to – functioning Data Reserves. For this cloud computing and suitable infrastructures and advances in hardware, distributed computing and related fields will be needed. Such research challenges are not covered here.

Second, referring to the discovery of the Higgs boson, it is well known that out of all the proton-proton collisions in the LHC only a very tiny fraction produced Higgs bosons, and only some of these decayed in ways that could be observed. For each observable channel *e.g.* $H\rightarrow\gamma\gamma$ (Higgs boson decaying into two photons) enormous efforts were needed to find such

[28] "IBM computer Watson wins Jeopardy clash", Adam Gabbatt in *The Guardian*, Thursday 17 February 2011; "Building Watson: An Overview of the DeepQA Project ", David Ferrucci *et al*, *AI Magazine* Fall, 2010; http://www.research.ibm.com/deepqa/deepqa.shtml

events in the data, and also to determine the expected background *i.e.* how many similar photon pairs would be produced in the experiment by other physical processes. The discovery then appears as a relatively small deviation above the background curve at a definite energy[29]. Such results can only be conceived if and when scientists know exactly what to look for. Such research challenges are not covered here.

Turning now to our goal of how to turn Big Data into Small Information, we can frame it as how to perform the basic signal processing tasks of estimation, prediction, clustering and model learning – in the Big Data context.

Estimation

Estimation is the task of compressing data into a small set of meaningful numbers (what is the average? what is the largest and the smallest? how do these change with time? are there correlations in the data?). In network management such a task could also be called aggregation. The challenges are how to find the small sets of aggregates, how to visualize them to make them meaningful to humans, and how to compute such numbers. Swedish statistician Hans Rosling (KTH Prize Winner 2010) has shown how data on global health can be distilled and displayed giving surprising new insights[30]. This work uses innovative information design, is grounded in an intimate knowledge of the domain, and is positioned in the world-wide debate of the most important policy questions of our time[31]. We expect that similar bottom-up efforts, starting from an understanding of the phenomenon and proceeding through innovative visualization, may similarly transform the exploitation of Big Data in many other fields.

Estimation from Big Data serves several purposes. From a stream of data too large to keep we need to filter out, summarize or extract the salient events, just as CERN had to with all of the mostly irrelevant data collected by the LHC. In general, we are still in the early stages of learning how to do that. A special case where one can efficiently learn complex properties only looking once at the data as it arrives is if the data can be described by Bayesian models with conjugate priors[32]. As the data streams in, one can then recursively apply Bayes' theorem to update the posterior distribution. On a more general note, C.R. Rao, one of the leading statisticians of the past century, has pointed out that the emergence of computer intensive practices of data analysis will probably require a complete re-thinking of the theories of model based hypothesis testing and estimation theory[33].

Prediction

[29] CMS Collaboration, Fig 2, *Science* **338** p. 1570 (2012), ATLAS Collaboration Fig 6 *Science* **338** p. 1579 (2012) . In both cases the signal is about 10% of the background at the Higgs boson energy (125 GeV).

[30] http://www.gapminder.org/world

[31] See e.g. “Mortality in Iraq”, J von Schreeb, H Rosling, R Garfield, Letter to *The Lancet*, **369**, 101 (2007).

[32] http://en.wikipedia.org/wiki/Conjugate_prior

[33] “Statistics: reflections on the past and visions for the future”, C.R. Rao: *Communications in Statistics , Theory and Methods*, **30**, pp. 2235-2257 (2001).

A prediction or forecast is a statement about the way things will happen in the future, often but not always based on experience or knowledge[34]. Prediction is typically connected to estimation, where estimated values are needed for the prediction, and prediction is used to improve the estimates. The Kalman filter and its generalizations have been mainstays of control theory for over half a century.

Big Data calls for on-line prediction algorithms of which the Kalman filter is an important example. However, for large data dimensionalities the standard approach may be computationally unfeasible or unstable. For instance, if an algorithms needs to invert a matrix of correlations, this is ill-defined if the number of dimensions of the data is larger than the number of samples (the "high-dimensional statistics" setting, see e.g.[35]). Many Big Data sources will also be far from Kalman assumptions of a linear dynamical system where all error terms and measurements have a Gaussian distribution. These are very active research questions, pursued by various techniques ranging from robust control, various sparsity techniques, iterative methods such as particle filtering, re-sampling, Expectation-Maximization, etc.

Major goals of Big Data estimation and prediction are to transform research by making data readily available and manipulable by the people who will frame the next important questions. In the Life Sciences a primary challenge is to integrate diverse large-scale data sets to construct models that can predict complex phenotypes such as disease *e.g.* multiple sclerosis. As the scales and diversity of the data grow, this type of modeling will become increasingly important for representing complex systems and predicting their behavior. The exabyte scale of data will require the development of tools and software platforms that enable the integration of large-scale diverse data into models that can be operated on and refined by researchers in an interactive and iterative fashion. This is one of the central goals to meet to get impact in molecular medicine, biomedicine and the life sciences.

Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)[36]. Clustering is central to many commercial applications of Big Data including search (Google), social network modeling (Facebook), and to pattern recognition, image analysis, information retrieval, and bioinformatics, among other fields.

To cluster data of high dimensionality and/or different data types is not obvious (what is similar? what is the distance between objects? is it even meaningful to measure similarity between two objects by one number?). By itself clustering is an ill-posed problem since what

[34] The source of this admirably concise definition is Wikipedia.

[35] "Spectrum estimation for large dimensional covariance matrices using random matrix theory", N El Karoui, *Ann. Statist.* **36**, 2757-2790 (2008).

[36] Source: wikipedia.

constitutes a good cluster is at least partly in the eye of the beholder[37]. Preferences can be mathematically formalized as clustering criteria, but the choice of which criterion to use in a specific application is still with the user. On the other hand, if a criterion is agreed upon, different algorithms can be compared as to accuracy and computational aspects, and trade-offs will typically depend on the size and type of data. Many methods will also involve optimization and calculation of integrals. Alternatively, if a method is based on an iterative procedure it would need Monte Carlo sampling. A specific challenge in clustering Big Data is therefore how to handle the large-scale computational issues that arise.

Computer scientists[38] and mathematicians [39]have developed methods of spectral clustering which have shown surprisingly wide applicability. In this context, data objects are points in a space, an abstract representation of the concrete situation being formalized. It is not just an ensemble; it carries with it a number of useful concepts and determines what kind of mathematical operations one can perform, how to interpret results, and how to relate results back to the original problem. This story encourages the belief that advanced concepts from mathematics – and perhaps also from information theory and physics and other fields – will find new fields of applicability in Big Data. The challenge is here to align the interests and desired outcomes of data stake-holders with the rigor and power of the mathematical sciences. For this institutional and/or financial support is required, to fund the necessary "face time" of actors from both sides.

Model learning

Estimation, Prediction and Clustering are all connected under the heading of Model Learning. To estimate properties of the data means to estimate parameters of a (perhaps simple) model, to predict future data is one main reason to constructing a model, and to cluster the data is often the first step in estimation, prediction and model learning.

Learning the "truth" from data alone is impossible, at least in the Popperian framework adopted in the physical sciences. Learning takes place in a conceptual environment, often a formulated as a class of statistical models describing the data. Surveying all (or even a fraction of) methods developed for model learning in statistics, information theory and many other fields is out of bounds of this document, and we will only point to currently topical research directions.

First, the benchmark of principled model learning is the maximum likelihood principle of statistics. This may be extended to Bayesian learning if a prior distribution is taken into account. In both cases one aims to maximize the posterior probability of the data over the parameters in a model class. For discrete spaces maximizing likelihood amounts to a problem

[37] "Why so many clustering algorithms: a position paper", V Estivill-Castro, ACM SIGKDD Explorations **4** 65-75 (2002)

[38] "On Spectral Clustering: Analysis and an algorithm", A Ng, M. Jordan & Y Weiss, In *Advances in Neural Iinformation Processing Systems* 849--856 (2001), MIT Press.

[39] Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps
R. R. Coifman et al, *PNAS* **102** 7426-7431 (2005).

of combinatorial optimization which is typically computationally unfeasible for large data sizes.

Exponential families are an important class of statistical models characterized by smaller set of functions of the data known as sufficient statistics. All models in equilibrium statistical mechanics are examples of exponential families, where the sufficient statistics are the terms in the energy function, and the normalization constant is the free energy. In exponential families maximum likelihood models are equivalent to maximum entropy distributions on the (empirically measured) sufficient statistics. This allows for a host of variational (approximate) approaches which scale to very large data sizes, recently reviewed in[40]. The recent progress on the protein structure prediction problem (Morcos et al 2011, Hopf et al 2012, *op cit*) relied on precisely such an approach, as do the famous coding/decoding algorithms implemented in hardware in every cell phone and every mobile base station. This is a very active field with many open directions *e.g.* how to extend them to data which arrives incrementally, or how to go beyond "Belief Propagation", or how to combine discrete and continuous data more effectively then at present.

Second, it has been realized that sparse models (described by few parameters) can be efficiently determined from data even if is unknown before seeing the data which parameters to use[41]. These methods have been the focus of enormous attention in recent years and have been found to compete favorably with application specific state-of-the-art approaches in imaging and many other fields[42]. The development and the use of such "compressed sensing" or "sparse signal processing" techniques will likely only continue to grow.

Yet another way to reduce the data, in addition to compression which decreases the dimensionality of the data items, is *active learning* which uses only a subset of the data items and discards the non-informative ones[43]. By choosing examples the number of data items to be used to learn a concept can often be much lower than the number required in normal supervised learning. Finally, *Approximate Bayesian computation* (ABC) can be used when the likelihood function itself is computationally costly to evaluate. ABC has gained popularity over the last years for the analysis of complex problems arising in e.g., population genetics, ecology, epidemiology, and systems biology[44], illustrating that "Bigness" can arise as much from the (required) model as from the data itself.

New Theory for Big Data

One main thrust where a Sweden—Finland collaboration will be able to make serious impact is in establishing Big Data as its own field, on solid scientific grounds. As argued by many,

[40] "Graphical models, exponential families, and variational inference", M Wainwright & M Jordan, Foundations and Trends in Machine Learning **1** 1-305 (2008).

[41] "Near-optimal signal recovery from random projections: universal encoding strategies?" T Tao & E Candès, E *IEEE Transactions on Information Theory* **52** (12): 5406–5425, (2006)

[42] "Compressed sensing makes every pixel count", Mackenzie, Dana (2009), *What's Happening in the Mathematical Sciences*, American Mathematical Society, pp. 114–127,

[43] "Active Learning Literature Survey", Burr Settles. *Computer Sciences Technical Report* 1648, University of Wisconsin–Madison. 2009.

[44] "Approximate Bayesian Computation " M Sunnåker *et al., PLoS Comput Biol* 9(1): e1002803 (2013).

e.g. by Noam Chomsky[45], drawing conclusions from data is sometimes carried out on weak scientific grounds. While the explosion in the availability of data has opened up for ample opportunity, it is also allowing opportunistic and not always well founded use of the data.

Information theory is a branch of applied mathematics and CS/EE engineering that builds theory for asking and answering questions like “what is information?” - “how much information can be stored/conveyed given a set of physical constraints?” - “how can data be reduced/compressed and still contain the same amount of pure information?” The new age of Big Data challenges the existing theory in several different ways. For example, information theory has so far assumed that the raw data itself is available for block-based handling, that is, that the data can be stored for processing. However, as outlined before in this document, we are now facing scenarios where the data can be so huge that it is impossible to access it all. This challenges earlier definitions of “information” – what is the information content in data that cannot be accessed?

Another challenge for information theory is to guide the development of new compression schemes. It is well-known that in most existing practical applications compression/encoding is fundamentally a non-linear process, and to approach fundamental bounds on performance as predicted by information theory, non-linearity cannot be escaped. On the other hand, compressed sensing has emerged as a very promising concept for reducing the amount of data, based on linear encoding (and non-linear decoding). To increase the amount of compression beyond what is presently promised by compressed sensing, it will be necessary to consider non-linear schemes. Here guidance from information theoretic achievability proofs will serve as an important tool.

Finally, fruitful interaction has recently emerged between mathematics, statistical physics and information theory. So far results in the intersection between these fields have contributed to understanding both scaling laws in large wireless networks as well as fundamental aspects of decoding error-correcting codes. Based on experience from modeling large-scale phenomena and networks, there is ample opportunity for new contributions in the direction of Big Data.

Why KTH and Aalto?

KTH and Aalto University are the two leading technical universities in Sweden and Finland, and have formed a large fraction of the scientific-technical elite in the two countries. They have formed a strategic partnership and are in many ways each other’s closest partners on the world-wide stage. A cross-the-Baltic collaboration on Big Data represented by KTH and Aalto can be predicted to be fruitful on several different grounds:

Scandinavia is getting known as a geographically suitability location for setting up data centers, that is, vast amounts of data are actually stored in this region, and there is a need for new expertise in the general area.

[45] “Noam Chomsky on where artificial intelligence went wrong,” *The Atlantic*, Nov. 1, 2012.

Sweden and Finland are known to be at the forefront in the deployment of IT infrastructure (optical and wireless), which means that here people and machines can be connected in ways that are not possible in most other parts of the world.

The Region in general hosts world-leading institutions in health-care and biology, and in Stockholm the development of the Science for Life Lab represents a unique effort for strengthening the area. Hence, there will also be a surging need for know-how in the area of Big Data in the Life Sciences.

The Region also represents academic and industrial expertise at the highest-level ICT systems and infrastructure, as developed through the years because of the importance of the Swedish and Finish telecommunications industry.

This document has surveyed the field of Big Data and the new scientific challenges and opportunities it poses. We have argued that Big Data will have a profound impact on the conduct of science in fast-moving areas and that the key to unlocking new discoveries is concepts, methods and analytical tools to do more and to do different things with Big Data. We have shown that Big Data poses challenges for *estimation*, where there is an urgent need for on-line algorithms that analyze and learn from the data as it arrives. We have discussed *clustering* in the Big Data context, and have argued that *estimation, prediction and clustering* are all different aspects of *model learning* and have highlighted promising recent directions.

We emphasize the need for *new theory* for Big Data. Without turning data (big or small) into information we face the risk eloquently expressed by Nobel Prize Winner Piotr Kapitza more than 30 years ago

> *The biggest pollution problem in the world today is brain pollution. From all the missions we sent to the moon we got enormous data, but we do not know what to do with it.*[46]

KTH and Aalto already have a strong basis for research in Big Data. The KTH Linnaeus Centre ACCESS and the Academy of Finland Centre of Excellence COIN are both among the largest and most high-profile research efforts in Europe in the general area of ICT, and Big Data issues are addressed by many other groups as well.

The conclusion of this report is Big Data is an important frontier. We recommend the management of KTH and Aalto to take steps to have Big Data issues on the agenda in the context of EIT and elsewhere with the aim to establish KTH and Aalto as leaders in the field. We also recommend that the effort be as broad and inclusive as possible. Big Data involves many research issues not covered in the report. The issues covered in this report can on the other hand be addressed by the methods of mathematics, statistics, information theory, physics, electrical engineering and computer science, to just name the most obvious stake-holders. To potential for success is largest if all of these communities can be involved in a comprehensive and meaningful manner. We strongly recommend that such an effort is made.

[46] Piotr Kapitza on “Snillen spekulerar”, Sveriges Television (1978).