Big Data Analytics

A Research and Innovation Agenda for Sweden

The Swedish Big Data Analytics Network bdainfo@sics.se https://www.sics.se/projects/big-data-analytics

Executive summary

A Swedish national agenda for the emerging field of Big Data Analytics is here presented. The purposes are chiefly to: highlight the recent and increasing importance of advanced analysis of very large data sets in society and business, and the excellent position of Sweden to potentially be at the forefront in this area by leveraging national areas of strength in research and business development; to address the limiting factors that hinder us from realising this potential; and to propose national efforts for remedying these factors and creating a fertile ground for future businesses, services, and societal applications based on Big Data Analytics.

An unprecedented growth of data constitutes a dramatic recent development in both ICT and society at large, fed by novel technology, user behavior and business models. In fact, roughly 90% of all data stored in the world has been created during the last two years[1]. Most of this data is Big Data, characterised by vast volumes, high velocities, a large variety and unknown veracity, that due to these properties poses significant challenges when collected, managed and processed.

There is an enormous commercial, societal and environmental potential in exploiting this data. Even though capacity to store, distribute and search large data sets exists today, this is not sufficient for realising the full potential in Big Data. Raw data in itself is often of little value, but often even merely a cost. It is the information the data contains that has the real value, and this value is extracted using Big Data Analytics.

The ability to refine Big Data is acknowledged to be one of the most important competitive factors during the next few years[2]. Big Data Analytics will also serve as an enabler for both smarter end-user applications and efficient management of large scale systems such as transportation networks or energy grids, and is a key component in the push towards autonomics in future large, heterogeneous, and complex information and communication technology systems.

Realising the value in data is associated with numerous challenges at different levels. We want to make use of and analyse *all* available data and therefore face new demands within storage, computation, algorithm development and networking. Big Data Analytics will happen everywhere, pushed out to our devices and into the network infrastructure, due to the fact that data volumes are too large to centralise and too sensitive to distribute in their raw form. It is also necessary to make data and analytic services available to potential users and services. For this reason, new service and exchange markets for data analytics need to be created, enabling completely new applications and businesses. This requires development of business models, incentive structures and methods for managing privacy and integrity issues. In all, it is likely that in the next few years we will change the way we look at ICT infrastructure, shifting our focus from connecting machines and computation to services and how to manage and extract information from data.

From an international perspective Sweden is in a very advantageous position to meet these challenges. There is a strong national research tradition in both basic and applied research of relevance for Big Data Analytics, for instance in critical areas such as data analysis, cloud computing and networking. Swedish companies as well as the Swedish public sector also generate and collect large data sets of high quality today. In the services area, for example, Sweden has a developing ecosystem of small to medium sized end-user service companies that generate a wealth of data. These factors in combination with a recent change in the view of the accessibility of data, lay the foundation for Swedish Big Data Analytics-based innovation.

Although Big Data Analytics has a substantial potential in creating value for Sweden, several efforts are required in order to realise this potential: strong directed support for research and infrastructure for Big Data research; activities for promoting application development, e.g. in traditional Swedish industry; development of an infrastructure for a business and service

[1] www.ibm.com/software/data/bigdata/

[2] Stefan Biesdorf et al., Big data: What's your plan?, *McKinsey Quarterly*, March 2013.

ecology around Big Data and analytics; efforts that ensure supply of key competences within Big Data Analytics; and finally, efforts to ensure collaboration and competence exchange between different actors.

Given that these areas are supported, we believe that Sweden will be in an excellent position to utilise Big Data Analytics, both for innovative societal applications, and for gaining substantial competitive advantages in traditional industry, the developing digital services area, and in emerging businesses built upon completely new value chains.

Editors

Olof Görnerup
SICS Swedish ICT
+46 70 252 10 62
olofg@sics.se

Daniel Gillblad
SICS Swedish ICT
+46 8 633 15 68
dgi@sics.se

Anders Holst
SICS Swedish ICT
+46 8 633 15 93
aho@sics.se

Björn Bjurling
SICS Swedish ICT
+46 8 633 15 89
bgb@sics.se

Table of contents

Introduction

Over the last ten years, we have seen a paradigm shift in the way information and data is generated and handled in society. The change is driven by several factors: our vastly increasing ability to store and perform computation over very large data sets; the rapid increase in sensors both in industrial systems and society at large; the introduction of Internet of Things, implying that even simple components and devices have processing power and can communicate over Internet; the mobile revolution meaning that everyone expects to be connected anytime and anywhere, and be able to both receive and send information; the appearance of cloud services and cloud computing, and other collaborative platforms and resources; and finally today's globally interconnected world, making new trends and technologies spread very fast, and giving rise to the need to handle and analyse data on a global scale. Together these dramatic changes have resulted in what is called Big Data. The amounts of data available when making decisions or keeping an overview is enormous and is produced at ever increasing rates, and the characteristics of that data produced by those heterogeneous and dynamic sources is very different from the structured data traditionally used. The overall trend is that the world is becoming more and more information centric, and the capacity to handle, process, analyse, and read value out these massive amounts of data is absolutely critical, for society, industry, and academia, as well as in individuals' daily lives. This is widely acknowledged, for example by the European Commission[3] and the National Science Foundation in the United States[4].

In order to capitalise on the potential value in data, it is of crucial importance that the data can be processed and analysed. For meeting increased demands on information extraction and for enabling analysis of increasingly complex data in increasing volumes, the processing and analysis tools have incrementally been improved over time. In competitive settings, the potential value in analysed data is relative to the competitors' capabilities of extracting value from the same or related data. Until recently, competitors have had roughly equal value extraction capabilities as data processing and analysis technologies have been scalable and adaptable to the increasing pace and volume of data flows. With the recent explosion in data availability and data collection possibilities, the competitive situation has changed. Today's processing and analysis tools may, with extensions and further improvements, still be viable options for extracting value from the data. However, the very fact that data now is in such an abundance opens up for the possibility that new processing and analysis tools are developed that are vastly superior today's technologies in terms of coping with the volumes, varieties, and the pace with which it is produced, even with respect to incremental improvements. The potential of such new tools found in Big Data Analytics diminishes, or even cripples, the potential value of information extracted using today's technologies. Thus, with the explosive developments in data availability, the data processing and analysis capabilities provided by Big Data Analytics will become one of the determining factors with respect to corporate and societal value making.

To formulate the following strategic research and innovation agenda for Big Data Analytics, several of the most active Swedish organisations within the area today, commercial, academic and in the public sector, have formed the Swedish Big Data Analytics Network. The agenda maps unique national competences and strengths, and the needs of Swedish industry and society within Big Data Analytics, as well as proposes efforts for improved national competitiveness. This would place Sweden at the forefront of the Big Data Analytics development by building on existing Swedish areas of strength. The network captures academic excellence within the central research areas, as well as industrial perspectives from both established and growth industries, and forms a basis for future cooperation in Big Data Analytics across sectors by linking stakeholders in industry, academia and the public sector.

[3] http://europa.eu/rapid/press-release_SPEECH-13-261_en.htm

[4] http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

Big Data Analytics

What is Big Data Analytics?

The term Big Data has commonly been used for any sort of data flow which is larger than usual. Here, however, Big Data refers to data sets and flows large enough to pose significant challenges when using commonly available tools and infrastructures to collect, manage and process the data within a tolerable amount of time. Big Data is not just Slightly Larger Data, nor a question of sampling in large data flows to be able to keep on as before. The point of Big Data is that it changes the way we approach data analysis, inspiring entirely new families of information services and necessitating new processing models and knowledge representations. The challenges associated with handling Big Data are broadly due to four of its characterising properties:

Volume Big Data is clearly about size - sizes that take traditional storage and computational approaches of handling data from being inconvenient to completely unfeasible. Storing indexed data from large portions of the World Wide Web, e.g. as done by Google, is an example of where new innovative methods for handling huge volumes of data by necessity had to be developed.

Velocity Big Data is generated in a very rapid pace. The Twitter fire hose (the stream of *all* tweets, globally), traffic data in mobile communication networks, and streaming video data are prime examples as this data flows at tremendous rates. The velocity of Big Data is relevant in many services where an up to date picture of information and a near real time response are prerequisites. This requires that data is processed on the fly without being stored. Financial services, for example, thrive on Big Data and utilise rapid data flows for gaining a competitive advantage, e.g. in high-frequency trading.

Variety Big Data is highly heterogenous to its nature. Traditional data analysis has to a great extent been able to rely on data being structured in tables and databases with entries of predefined types. By contrast, Big Data is typically mixtures of structured and unstructured data in various formats such as text, audio, video, sensor data and click streams. The data sources are further not necessarily fixed. With the development of Internet of Things and as a consequence of the mobile revolution, the sources may dynamically connect and disconnect in unpredictable ways and the number of sources cannot be known in advance.

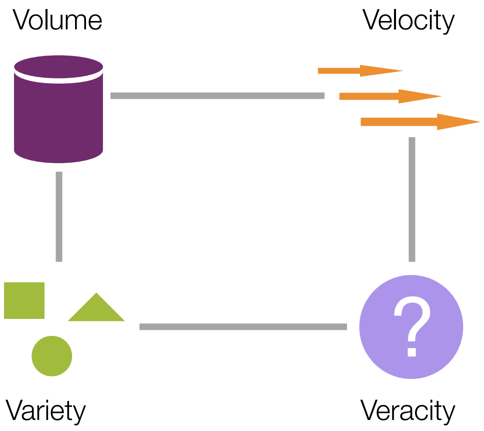


Figure 1. Big Data is characterised by the four v:s.

Veracity The quality and authenticity of Big Data (in terms of an application specific notion) are uncertain as the data may be provided by unknown and ever-changing sources. Whereas traditional approaches to data analysis rely on precise and accurate data in consistent and predictable formats, the analysis of Big Data must take this uncertainty into account to produce a viable support for decision making based on extracted information. Discussions about data quality has traditionally revolved around structured and mostly transactional data. The challenges with Big Data may well introduce a new way of looking at data quality. This will certainly require both a new mindset but also new tools and processes to handle the veracity of Big Data.

From data to information to value

Raw data is often of little value, but may even merely be a heavy monetary and environmental cost by devouring valuable resources, e.g. in infrastructure such as storage hardware, and energy for running this infrastructure. It is the *information* the data potentially carries that has the real value, and the tools for extracting this value are found in Big Data Analytics. This emerging field involves several co-existing and interdependent levels - as illustrated in Figure 2 - that each poses different sets of challenges. Firstly, the data must be collected, e.g. using sensor technology in an industrial setting or in distributed sensor networks, by extracting traffic data from cellular networks, or through embedded technology in the Internet of Things. As the data has been collected it requires scalable, robust and secure storage solutions that support huge volumes of data, e.g. using virtualised cloud environments. Processing Big Data, in turn, puts demand on computational frameworks and models that need to be fault tolerant, flexible and light weight, e.g. by supporting iterative and stream computing, as well as local processing of data. Computing and storage solutions form basis for advanced data analysis, including machine learning and statistical modeling. As analysis ideally is taking place as close to the data source as possible, simply since it is often not possible nor desirable to transfer large volumes of data, analysis is tightly integrated with the former levels. For example, by analysing data in real- or near real-time - such as in stream computing - the storage level is largely bypassed as raw data is continuously discarded after being processed, whereas the relevant information the data contains is kept and utilised. On the next level, information extracted from analysis may be conveyed to an application or human user, e.g. through visualisation, and transformed to valuable knowledge. These levels are followed by a systemic level that should support exchange of information from data between different stakeholders in cross-sectoral service and application ecologies. Such systems require well-defined interfaces that, for example, enable transfer of information without compromising privacy or exposing proprietary and business sensitive information. At the top of the stack we have an organisational level that concerns business and legal aspects of Big Data Analytics, including business model development and legislation for handling data.

Vision

We have witnessed an astounding recent development in ICT. In only a few years we have for instance gone from taking for granted an uninterrupted mobile phone call or a successful transmission of a 160 character long text message, to expecting flawless streaming video on our mobile devices, where each choice of video clip - perhaps simultaneously viewed by hundreds of thousands of other people - in an instant provides sensible recommendations of subsequent clips. Data Analytics is of great importance in ICT today, but will be paramount in the near future due to the rapid increase of data that essentially permeates all of society. We already begin to see new services and use of data - e.g. for tailoring advertisement, extracting collective sentiments from social media, and for discovering new pharmaceuticals - that will have central roles in future society and business. As Big Data Analytics extracts and utilises relevant information from data, the field will be on the center stage as the main enabler for realising the value in data as well as for allowing for a sustainable development of data intensive ICT infrastructures.

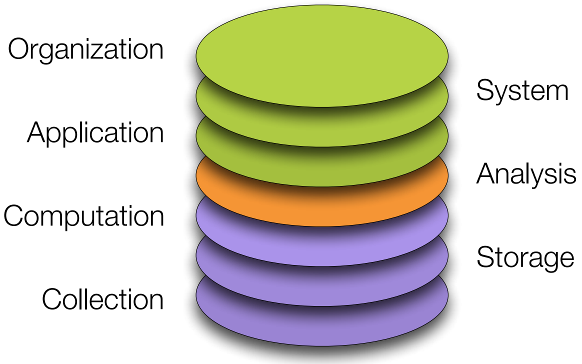


Figure 2. Levels of relevance for Big Data Analytics.

It is acknowledged that being able to utilise data is an increasingly important source of competitive advantage[5,6]. Due to new and emerging business models, pioneered e.g. by Google and Amazon, we will see a situation where companies compete with data, and the ability and knowledge to exploit this data through analytics. We are also moving toward a data and information-centric perspective of ICT, where major future research challenges and business opportunities are not primarily in linking devices per se, but in handling and utilising the data that these devices generate and exchange. We envision systems where applications can be built by third parties based on information and models provided by several

[5] Dominic Barton and David Court, Making advanced analytics work for you, *Harvard Business Review*, October 2012, Volume 90, Number 10, pp. 78-83.

[6] Stefan Biesdorf et al., Big data: What's your plan?, *McKinsey Quarterly*, March 2013.

other parties, such as network providers and social networks. Such a system has the potential to develop into a thriving market of application developers, technology suppliers and service providers. Public and commercial access to such a platform would offer countless opportunities for innovative services that reuse currently unexploited massive data sources, for instance enabling improved urban planning, epidemiological applications, traffic planning and monitoring as well as improved management and resource planning for communication networks.

Big Data Analytics will not only result in completely new value chains and business areas fed by innovative use and exchange of data. It will also be highly beneficial from societal and environmental perspectives. When the data can be collected, stored and analysed, it will be possible to optimise various processes in society. With the extensive use of smart phones, for example, motion patterns can be analysed on an aggregated level and be used in the planning process of new roads, railroads, electricity networks and public transport. By cross-examining data from the health report system with search terms used on the Internet, Big Data Analytics can be used to predict and hamper the spread of diseases. It can transform vehicle data into information that lets customers utilise their machines in better ways.

Internet of Things and cloud-based solutions will also offer services that will simplify our everyday life. The current technology evolution, where machines are connected, will transform society in many respects. It will for instance be a critical enabler for a sustainable development of energy consumption, where heat-pump suppliers, alarm companies, energy companies and service providers collaborate in order to drive standardisation allowing for sustainable IT-platforms and services. We will also see a plethora of entertainment, sports and lifestyle apps making heavy use of sensors and actuators built into mobile phones, game consoles, and other everyday objects. Added value and functionality will be provided by the patterns formed by the massive amounts of data many users generate through these devices, enabled by Big Data Analytics.

Current situation

Through the whole spectrum of society and business, large volumes of data are collected at an increasing pace. Examples include social web applications (logs, click streams, RSS streams), supply chain management (RFID data), logistics (geospatial data), from business-to-business processes, traffic planning (toll charge data) and medical applications (syndromic surveillance and patient data). In these areas, the data is utilised in an array of diverse ways: in logistic optimisation; quantification and estimation of financial or civil risk; for fraud detection; to identify trends among specific target groups; to detect pharmaceutical side effects, for tailored marketing, for realtime decision support, and for understanding customer behavior. Adding to this picture are the great number of industrial actors that are in the preparatory stages of starting to collect industrially and corporately relevant production data. For example, in the automotive industry, the new generation of trucks will be equipped with telematic gateways, allowing - in principle - online monitoring of every single truck. In home monitoring, as another example, data are envisioned to be collected either as a part of “smart homes” development or as the means of improving elderly care.

Furthermore, the Internet of Things, giving a digital presence to objects of the physical world, is becoming part of our daily lives. It is happening in different forms: smartphones and their downloadable apps, home multimedia systems, RFID identification and tracking. Many more applications are expected in the coming years, connecting our phones, cars, appliances, buildings, toys, cities, environment, and social networks. These applications will result in unprecedented amount of data, that no existing system is readily able to handle.

There is also an emerging third-party industry where service developers refine data from multiple sources for specific services, often mobile. The ongoing smart mobile explosion has been enabled not primarily by mobile network operators, but rather by actors that bring perspectives from the Internet. The range of services for mobile units has increased dramatically in recent years. A central component for network based service innovation is relative openness: an innovator can easily combine information from several different sources and create a new service, without being limited by the constraints that exist in traditional manufacturing industry.

The ongoing digitalisation has also resulted in aggregation points; services that collect a given type of data, such as photos (for example Flickr and Photobucket), documents (Dropbox, among others), music (Spotify), and geographic information (Google maps, Open Streetmap). These services have created complementary resources for other innovators in the form of digital services that in turn easily can be used by others.

The most popular framework for building data intense applications today is by far Apache Hadoop. The framework is based on MapReduce - a programming model developed by Google to manage large data sets. Examples of actors that use Hadoop for storing and processing Big Data are Amazon, Facebook and Yahoo, where the latter, for example, employes the framework in their web search system. Despite the large impact of Hadoop and similar platforms, it is widely acknowledged that these frameworks have crucial shortcomings that limit their applicability. The need for computing models that are more flexible than MapReduce is slowly being realised, e.g. with the YARN framework - that will be included in the next version of Hadoop - Storm by Twitter, S4 by Yahoo and Pregel by Google. However, new computing models will still be required to handle new data sets with different processing requirements.

Global trends

The development in Big Data Analytics can be summarised into four main trends:

From search to awareness We are beginning to see completely new services that do not, in a manner of speaking, focus on arranging documents so that we easily can find an interesting copy in our collection, but that provide context and awareness by using *all* data, structured and unstructured. Analysis components for Big Data are in this way going beyond the archival services provided by today's search engines, and will also provide services which give their users a situation awareness, a sense of what is going on, and a refined sense of what the information flow is in real-time. This calls for entirely new families of tools and knowledge representations.

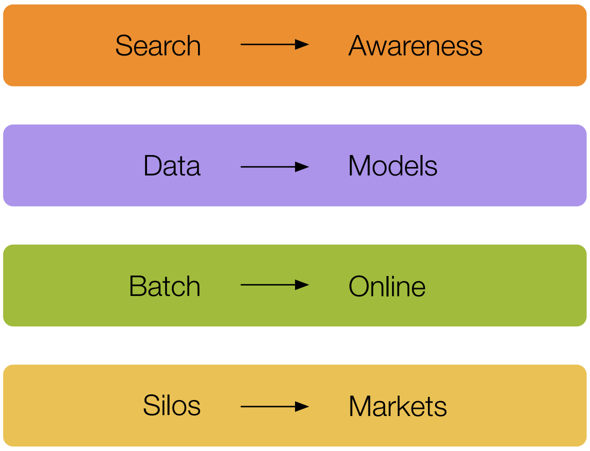


Figure 3. Major trends in Big Data Analytics.

Furthermore, where traditional systems have been focusing on the task of finding answers to specific questions, the amount of data that is available today is resulting in a shift towards more descriptive and explanatory analysis. The user is no longer required to know exactly what questions to ask, but the system is capable of highlighting interesting aspects: deviations and anomalies, relations and co-occurrences, etc. In many applications it is not obvious as to what kind of knowledge is “hidden” in the data, and thus the ability to automatically or semi-automatically discover and describe interesting patterns is becoming increasingly relevant.

From data to models For several reasons, instances of data are being replaced by models that capture the relevant information content in the data. First of all, this trend is driven by computational necessity since it is often no longer possible to store or transfer raw data. By formulating and maintaining models that, for example, capture relevant statistical properties of the data, data volumes are drastically reduced. Representing information in data using models is also done for legal reasons as it enables anonymisation of sensitive data, and for business reasons since it allows for control of which aspects of the data that can be revealed to customers or business partners. An example is cellular network traffic data that, on the one hand, carries great potential for enabling societally and commercially valuable applications, but at the same time can not be readily applied or released in its raw form for all of the above reasons.

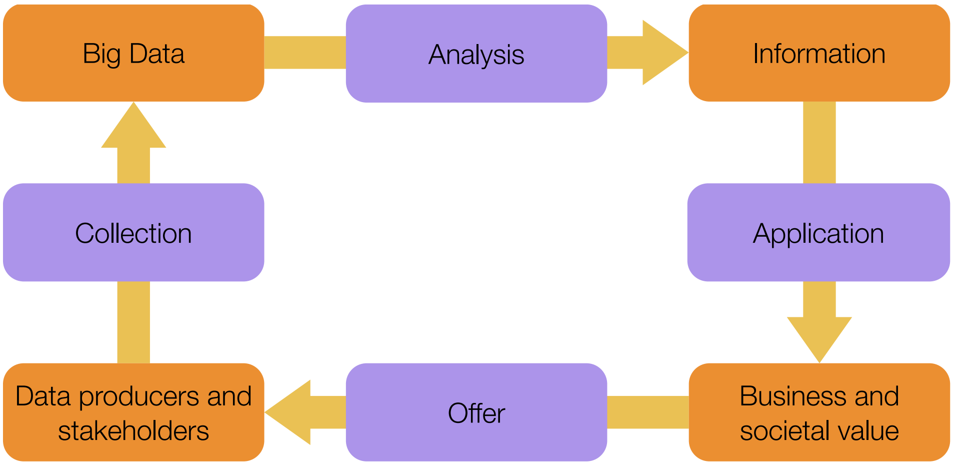


Figure 4. Cycle enabled by Big Data Analytics resulting in value creation and new markets.

From batch to online processing The traditional approach of handling data is to collect it to a central location for storage and computation. In this scenario the data is processes in a batch-based manner where one runs through the whole data set during analysis. This approach, however, is often infeasible when handling Big Data as it needs to be processed timely, even in real-time, and close to the data source. Online processing is also gaining attention as time is an increasingly important factor in understanding the value and the actionability of information. Trends, changes, dynamics of information over a time-line will determine much of what is relevant, pertinent, and interesting in an increasingly fluid information landscape. For this reason, a trend is to use incremental models of data, continuously updated, local and distributed.

From silos to markets We are going from applications based on collecting and storing large volumes of data solely for internal use - e.g. for supporting major web search engines or recommendation systems in web shops - to business models based on exchange of data and models thereof. This situation is schematically illustrated in Figure 4. For example, companies base their businesses on collecting data from social media and offer this to customers that employ it for taking the pulse on public opinion or for tracking collective reactions after a product release. Similarly, streaming micro-blog data from Twitter is piped to other actors that use this data for extracting trends, common sentiments etc, that they in turn sell to their customers. Another example that reflects this growing trend of markets based on Big Data Analytics is services, such as Google Analytics, where companies offer insights through analysis, and in exchange acquire data that can be used in other services, such as targeted advertising, that are offered to third parties. In this way an ecology of services and applications based on innovative use and refinement of data is emerging.

Challenges

The potential value in the massive data flows typical of Big Data can only be extracted under the condition that appropriate measure are taken for adapting society, business, and technical solutions to meet the corresponding and new needs. In this section we give a broad overview of the challenges that need to be addressed in order to build a national capability in Big Data Analytics with an internationally competitive edge.

Organisational Level

Business Models Big Data Analytics promises to revolutionise the society with respect to data analysis possibilities and availability of refined information. However, utilising the full business potential in Big Data Analytics presupposes a functioning service ecology, where the participating actors obtain competitive incentives to collect, share, and distribute data, information, and analyses, according to sound market principles. Doing this in a well-structured way, with clear incentives for all parties involved, poses challenges in several areas, such as legislation, policy-making, public administration, and management. It also requires development of completely new business models, or suitable adaptions of existing models.

These wider aspects of Big Data Analytics need to be addressed to the extent necessary to propose a sound technical basis for ecologies based on collecting, analysing and sharing information extracted from Big Data. There is a need for development of methods and protocols for exchange of information while supporting negotiation, payment, and verification. Guidelines for data integration between different types of sources need to be devised.

Privacy, Security and Legislation Information can be extracted from a combination of several more or less isolated systems that are managed by different stakeholders. Each such system will in general contain information that the owner of that data will not be willing, able and/or legally allowed to distribute or exchange with others. Patterns inferred from user data for instance, may reveal sensitive information, such as the whereabouts of individuals or groups of individuals.

In order to ensure data availability it is necessary that privacy issues and concerns are addressed. For ensuring a functioning data service market and the development of new business models, it is further of vital importance that data storage, transfers, and processing can be made while respecting the basic requirements of information security: *integrity*, *authenticity*, *confidentiality*, *availability*, and *non-repudiation*.

Current legislation may prevent governmental agencies to share and distribute data. It is a challenge to make it easier for governmental agencies to share and distribute data across agency boundaries or to the general public.

System Level

Service Frameworks A fast growing segment of digital services are based on large data volumes. Driving this expansion is a constant combination and recombination of diverse types of data forming ever new user services. Such services in turn drive usage of services with digital components which generate more data enabling further service innovation. The better the access to essential types of data, the better the service innovation climate and ease of access should therefore be a first priority.

In large parts of society data remain largely locked within organisational boundaries used to optimise product development, manufacturing, marketing, aftermarket services, or logistics internally or within supply chains or closed networks. This phenomenon of closed innovation has been challenged with open innovation strategies lately. However, in most industries business models rarely reflect data or analytics as tradable commodities.

While it is enticing to pursue a strategy of data stream optimisation, redundancy is at the heart of rapidly growing digital service industry. Robust supply of essential data sources should build on multiple approaches simultaneously to minimise detrimental consequences of negative developments in any one source. For instance, geographic information is available from public sources for free or a fee, from private sources for a fee, free, or through another business agreement, and via open source projects. In all, these overlapping sources maintain a steady supply to a multitude of service innovators with widely varying requirements on data quality and financial resources.

New services should use not only existing systems but take care to incorporate and cultivate new massive streams of data from customers or the public. In this sense, service innovation is not reduced to a mere analysis of “what is already there”, but rather a means to understand the limits of current information resources and to supplement missing data if not from an existing source then by cultivating new ones. The service industry will join the manufacturing industry who is already harvesting massive data volumes from their connected installed base.

Information Platforms Data infrastructures allowing several actors to collaborate and share data become increasingly important. It is a challenge to determine how to design and configure common platforms where participants in the data service market can find desired data sources and data owners can share and offer their data with other participants. In particular, this poses challenges for finding common standards and interfaces for supporting payment models and security and privacy aspects; for representing data and analysis requirements; for searching among data sources and services; for allowing fusion of data sources; and for integration of existing data sources and services into new applications and business models.

It is of further of importance, not at least for environmental concerns, to address challenges with respect to efficiency in collection and storage of data and with respect to determining where data should processed in the most efficient manner.

Presentation and Interaction The massive amount of data together with the numerous possibilities in aggregation, processing, and analysis of the data necessitates new approaches to visualisation and representation of data. It is a challenge to develop new presentation and visualisation techniques that allow users to efficiently understand the data, the analyses of the data, and the potential new information models. In particular, new presentation techniques need to allow representation of many-dimensional and temporal data as well indicating meta-data such a credibility of data sources. It is further a challenge to represent streaming data and temporal changes to streaming data where no records are or can be kept.

In order to allow users to discover relevant data and information, a challenge that need to be addressed is that of investigating both interactive and dialogue-based knowledge approaches to discovery, and ways to use available human-generated information (for example, design documents, field reports, service logs, etc.) as guidance. Dialogue-based techniques can be used for addressing the challenge of allowing users to interact with the representation of data for customising it to suit the user's information extracting needs.

Technical Level

Scalability Scalability is one of the most crucial technical challenges in Big Data and Big Data Analytics. Current analysis methodologies will only scale to certain levels of complexity and amounts of data. Beyond such levels, current methodologies will become increasingly irrelevant and likely not be suited for providing refinements and analyses of competitive value. One challenge is to deploy existing technology such as Grid and Cloud computing as well massive parallelisation through Multicore to improve the computational performance of data analysis algorithms. The computational challenges must be further addressed through new distributed algorithms. Also the increasing complexity and diversity of the data poses scalability challenges. One way of coping with the amounts of information available nowadays is to analyse different data granularities at different levels of abstraction.

Data management Managing the massive amount of data is one of the most apparent technical challenges in Big Data and Big Data Analytics. In some cases it may not be a viable option to store data. This can be either because it is physically and/or economically infeasible to store (immensely) large volumes of data or that the management overhead becomes too large (e.g. when data is updated faster than it can be stored.) Or, it could be that data may not be stored due to privacy or sensitivity concerns. A further reason for not storing data could be that the computational models or information models require on-line real-time data for their operations. Stream processing is an approach to addressing several of the challenges concerning analysis of data that cannot be, may not be, or is better not stored. In particular, for some applications may batch processing be infeasible or otherwise undesirable. It is also a challenge to define representation models for stream processed data. In the cases where data is stored, there are several challenges regarding redundancy, distributed storage, and cache distribution that need to be addressed.

A part of the Big Data revolution is the diversity of data formats, and in particular the increased number of sources of unstructured data. This poses several challenges for data indexing, representation, and retrieval in database technologies. It is noticeable that not only data in symbolic form can be analysed. For example, with speech recognition technology, it will be possible to refine live speech data into actionable information. Several related challenges concern the diversity in the ways data need to be collected and stored in order to be of value for specific applications. For example, several applications require that data is recorded as time series.

Modeling and Analysis The key for extracting value from data, is typically to build a suitable model of the interesting aspects in data, and use that model to analyse key values, detect anomalies and trends, make predictions, and perform other analysis. Because of this, statistical modeling and machine learning have received an increased interest in the context of Big Data Analytics. There are several challenges related to such modeling, including to develop new kinds of incremental models for meeting the challenges with regard to potential inadequateness of either batch processing or intermediary storage, or both.

Because of the sheer amount of data available, some of it will necessarily contain errors, or interesting parts of the data may be missing. Further, there are almost always uncertainties associated with the data, either due to noise in the sensors, uncertainties in the generation process, or uncertainties about future development of a situation. A huge challenge is therefore to develop robust methods that work in light of all these uncertainties. A special case is probabilistic computations, which represents the uncertainties in each step of the calculations. Related to this, despite the large sizes of the data sets, there is often a significant sample bias present in Big Data. How to manage this efficiently when combining different data sets is still largely a research issue.

Despite the huge amounts of raw data available in a large number of domains, only very little of it has been classified or analysed by experts. Therefore, one challenge is to develop bootstrapping or active learning techniques, where limited amount of expert knowledge is put to the best possible use. In addition, there is a need for more research in the area of interactive learning, as well as in learning from uncertain and accidental data: frequently, the data available does not correspond directly to the problem being analysed, while still being relevant enough that it should not be simply ignored.

It is a challenge to predict system behavior, especially when available data sets are incomplete or limited, or the models too complex for analytical evaluation. In such cases, it may be necessary to rely on simulation models. Simulation studies of complex systems have in addition become a key competence in the innovation process. Before large amounts of money are invested in a product, simulation studies can uncover crucial features in the design that should be improved or changed before the product is fabricated. Large computing resources have in the past belonged to academic groups and institutes, but with the development of computing clouds massive computing is available for industry.

Artificial Intelligence Techniques Several aspects of Big Data and Big Data Analytics renews the interest for, and puts focus on, several techniques from Artificial Intelligence. Machine learning has already been mentioned above as an important means to read out value from data. In addition to that, one of the most prolific sources of massive data is free text from blogs and other social media, which poses the challenge of adapting techniques from Natural Language Processing to extract valuable information from such sources of data. Further, data in the form of speech or optical sensor data puts renewed interest in Speech Recognition and Computer Vision for obtaining valuable information from such sources. Also the fields of Artificial Neural Networks and Deep Learning will draw attention from the challenges stemming from the importance of obtaining classification from unstructured and complex data.

Furthermore, the abundance of data and data sources make important to be able to discern the relevance of particular sources and data sets with respect to the characteristics of the information the user expects to extract value from. It is a challenge in the field of Decision Support to develop the methods and tools appropriate for discerning the usefulness of data, information, and knowledge with respect to the user's intentions and information needs.

Data Collection and Fusion As more and more devices become networked and individually addressable, the amount of information about locality and connectivity of devices collected and maintained by the network operators is expected to increase rapidly. As the rates of data and meta-data increase, the allocation of network and configuration of network parameters pose challenges for network operators.

Smart phones and other mobile devices produce data that can be used for extracting mobility pattern data. In particular for telecom operators, mobility pattern data can be used for addressing the challenge of efficient resource management in information intense networks on which new information intense services and applications rely.

Mobility pattern data can further be used for public policy and transportation planning.

Sensor technology is being deployed in various and an increasing number of contexts, for example in cars, containers, unmanned aerial vehicles (UAVs), adding to the abundance of data. Further, the deployment of the Internet of Things will generate large amounts of text-like communication between devices. The volume of data produced by these sources implies challenges in fields ranging from network management to energy efficiency. The data will be generated on many levels of abstraction and in very various levels of editorial quality. For any service, product, or analysis based on human- or device-generated language, there will be a challenge in representing the meaning of communication represented in the data.

Combining data from multiple sources is a challenge, especially if this data comes in very different forms, such as sensor readings and design documents. A specific challenge is to find ways to combine statistical data analysis with knowledge-based or symbolic approaches to data analysis. It is further important to be able to take into account relevant external conditions, meta-data or other forms of contextual data, when analysing and fusing data.

International efforts

Big Data and Big Data analytics development have to a large degree been driven by industrial pioneers in the area, such as Facebook, Amazon, and most importantly Google. In fact, commonly used tools such as Hadoop and HDFS are open source implementations of frameworks originally developed internally at Google to tackle its needs to perform computations over huge amounts of data. Private companies continue to be very important in the development of the field, contributing to shared and open implementations used in a multitude of companies, and companies such as IBM has pushed the analytics field forward through e.g. the Watson project. American companies continue to be at the forefront of this development, with examples such as HP, Teradata, 10gen, Microsoft, and Splunk in addition to the companies mentioned above, but new Big Data and Analytics enterprises are getting a stronger foothold in both Europe and Sweden, with companies such as Spotify and Neotechnology.

On the academic side, the US National Science Foundation recognises that Big Data Analytics technologies will be fundamental to further research and national development[7]. It will support Big Data technology development as well as development in related areas such as machine learning, data mining, and visualisation. Additionally, the President's Council of Advisors on Science and Technology (PCAST) recently acknowledged that Big Data Analytics technologies will be of utmost importance, noting that the pipeline of data to knowledge to action has tremendous potential in transforming all areas of national priority. These efforts will further strengthen an already strong US research community in the area, with excellent research initiatives such as AmpLab at UC Berkeley[8] and MIT’s Computer Science and Artificial Intelligence Laboratory (CSAIL), which in 2012 launched the big-data initiative bigdata@CSAIL[9].

On the European level, Big Data and Analytics is recognised as a critical development area[10]. The Coordination Action "Big Data Public Private Forum" (BIG) started late in 2012 within the FP7 framework. It is working towards the definition and implementation of a clear strategy that tackles the necessary efforts in terms of research and innovation in Big Data, aiming at providing a major boost to technology adoption and supporting actions for the successful implementation of the Big Data economy, particularly in health, finance & insurance, retail,

[7] http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

[8] https://amplab.cs.berkeley.edu

[9] http://bigdata.csail.mit.edu

[10] http://europa.eu/rapid/press-release_SPEECH-13-261_en.htm

manufacturing, energy, transportation, telecommunications, and media. As part of this strategy, outcomes of this project will be used as input for Horizon 2020, the next European framework program, in which Big Data and analytics will be important parts. Several national initiatives are also being formed in European countries, with examples including the Italian Trento RISE Association Open-Big Data National Initiative and the recent French allocation of 11.5M EUR to Big Data projects[11].

It is imperative that Sweden follows these developments to ensure that our position within this developing field remains strong.

Big Data Analytics innovation in Sweden

Value for Swedish industry and society

Exploiting information from data has increasing importance for gaining and maintaining a competitive advantage, where companies that do not utilise the value in data run the risk of being overrun by those who do[12]. Since this is particularly significant in data intensive sectors of strength in Sweden today; telecommunications (e.g. Ericsson), media services (e.g. Spotify), process and manufacturing industry (e.g. SKF) and the transport sector (e.g. Volvo), bringing Big data analytics forward will result in a substantial competitive advantage for Sweden.

Data as such is sector invariant, ubiqutous and to a large extent subject to the same fundamental analytics regardless of domain and branch of business. Big Data Analytics is not only applicable in different sectors, but also serves as a bridge across these sectors. It enables cross-fertilisation between otherwise separated domains such as the transport sector, media distribution area and in telecommunications. By linking domains through the cycle illustrated in Figure 4, Big Data Analytics is an enabler for development of innovative products and services that will strengthen established and developing businesses, as well as forward new start-ups. This will in turn give Sweden a clear competitive advantage internationally by staying at the very forefront in a game-changing development, resulting in numerous job opportunities as business thriving from the value in data is growing.

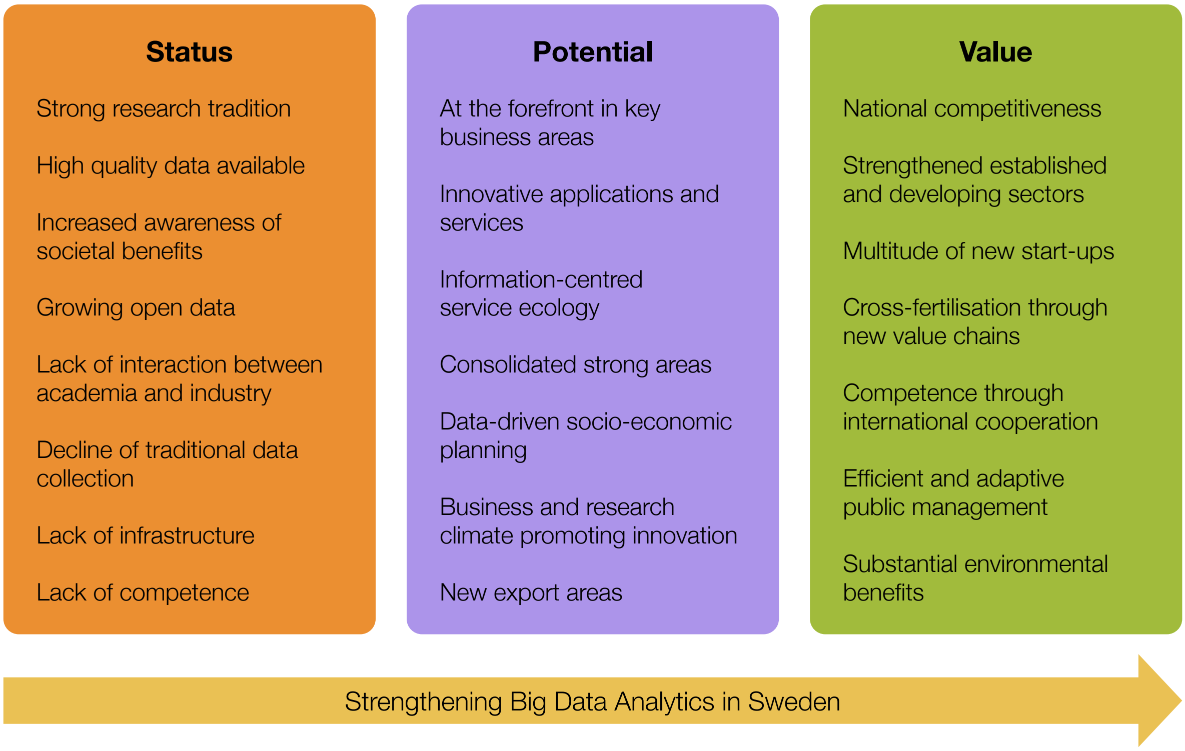
The benefits of an information advantage are not exclusive to commercial stakeholders. In addition, public actors like governmental agencies and municipalities are increasingly dependent on structured and interpreted information about the social and economic context in which they operate, to be able to dimension their commitments and activities in a more effective way. With increased globalisation and a rapid technological development, society has become more complex with fast and ambiguous changes that require a proactive approach for meeting demographic and environmental challenges. Here data analytics will play an important role by extracting relevant information, e.g. applicable in decision support.

With the proliferation of Big Data, knowing how to extract information from data sets is becoming increasingly important for assessing societal trends, and consequently to respond accurately to structural changes taking place. It is likely that the quality and suitability of the data analysis tools will become the determining factor with respect to the quality of decisions. Being able to use information extracted from very large data sources is key to successful adaption and to grasp innovation opportunities occurring with major changes in society.

In socio-economic planning, for example, models based on privacy-preserving data mining of user data can be utilised to calculate both the need for and the effect of planned infrastructure. Today, a combination of surveys, demographic databases and other register data, and stated preference studies are used to create these models. However, this data is often of low quality, low coverage and not up to date. The great potential with analysing user data, e.g. from mobile communication networks and social applications, lies in the possibility to continuously build dynamic models using very large data sets. By exploiting existing data from several sources, the accuracy and current validity of these models could be greatly improved.

[11] http://www.redressement-productif.gouv.fr/programme-investissements-davenir

[12] *Harvard Business Review*, October 2012, Volume 90, Number 10, pp. 78-83.



Status

Strong research tradition

High quality data available

Increased awareness of societal benefits

Growing open data

Lack of interaction between academia and industry

Decline of traditional data collection

Lack of infrastructure

Lack of competence

Potential

At the forefront in key business areas

Innovative applications and services

Information-centred service ecology

Consolidated strong areas

Data-driven socio-economic planning

Business and research climate promoting innovation

New export areas

Value

National competitiveness

Strengthened established and developing sectors

Multitude of new start-ups

Cross-fertilisation through new value chains

Competence through international cooperation

Efficient and adaptive public management

Substantial environmental benefits

Strengthening Big Data Analytics in Sweden

Figure 5. Big Data Analytics has the potential to enable numerous values for Swedish industry and society.

Furthermore, the environmental benefits of applying Big Data Analytics are likely to be substantial, e.g. by making widespread and heavily used ICT systems more energy efficient. For example, optimising cellular networks - used by essentially everybody today - by utilising data-driven techniques based on actual user behaviour, much energy can be saved. The same is true for heavy industry, such as in the mining and steel sectors, where even minor improvements gained by analysis of data will result in very large energy savings. Another example of the environmental benefits of Big Data Analytics can be found in the traffic and transport area, where data e.g. of collective mobility patterns revealing current bottlenecks, can be used to optimise freight traffic as well as decisions of individual commuters. With urban growth putting strain on traffic systems both in Sweden and globally, the value in making these systems much more efficient and adaptive is particularly timely. Likewise, smarter planning of flights by analysing and predicting customer needs, the airline industry - one major contributor of carbon dioxide today - as well as the environment have a lot to gain.

Current position of Sweden

Sweden is in a strong position in Big Data Analytics in several respects. Swedish companies as well as the Swedish public sector generate and collect large data sets of high quality today, e.g. in healthcare, from the Swedish Land Survey data, and Swedish Statistical Offices (SCB). In the healthcare sector, for example, there are report systems that merge data streams from lab results, ambulances and healthcare centres. When the output from these report systems is compared with information about social activities in society, such as terms used in search engine queries, the evolution and spread of contagious diseases can be monitored and predicted. Large amounts of data is also collected in traditional industries, where sensors and distributed systems are increasingly used.

However, the traditional methods used by statistics offices to collect data, e.g. surveying and using enterprise and real estate registers, are facing several crucial challenges that can be met by means of Big Data Analytics. Firstly, the traditional infrastructure of statistics is expensive and not well-suited for handling a fast changing and fluid society. Secondly, a decline in response rates in statistical surveys is becoming more apparent in recent years, actualising an urgent need to study alternative and complementary data sources and collection methods. Thirdly, lack of interaction between academia and research institutions, and the public sector prevents Swedish agencies from utilising readily available data from existing sources.

There has recently been a change in the view of the accessibility to data in Sweden, partly in alignment with the recommendations made by the Royal Society in the United Kingdom[13]. This is driven by initiatives such as öppnadata.se, trafiklabb.se, openaid.se, which offer easy access to data. Similarly, there are national initiatives for making public data more accessible through common data infrastructures, such as the Geodata portal and Artdatabanken, with input e.g. from crowd sourcing. In essence, there is a clear direction in Swedish society in making data available, which tremendously increases data sources which can be used for various purposes. Among Swedish public authorities there is also an increasing awareness of the potential and societal benefits of utilising Big Data, e.g. at the Swedish Transport Administration (Trafikverket), Mistra Urban Futures - a centre for sustainable urban development - and the Swedish Institute for Communicable Disease Control (Smittskyddsinstitutet).

Another factor that puts Sweden in a good position is that there is a very strong research tradition in fields on which Big Data Analytics relies, such as data analysis, statistics and natural language processing. High quality applied research, also of relevance for Big Data Analytics, is another national strength, enabling industrial actors to leverage on large data sets in collaboration with research institutes and academia. For example, frontline research is undertaken in health care analytics, which aims for providing efficient and effective decision support for health care and pharmaceutical research. This includes the development of techniques and tools to support decision making and discovery of drug effects by analysing structured and unstructured data. To take another example, Sweden is at the forefront in the financial sector, where investment decisions and derivative trade are based on combined information streams.

Although Sweden is in a very advantageous position in Big Data Analytics from an international perspective, there are several crucial limiting factors that must be addressed, both in short and long term, in order for a strong and sustainable competitive advantage to be realised: lack of directed support for research, infrastructure for Big data analytics, and commercialisation of current yet untapped state-of-the-art research in the field; lack of common standards for sharing data and information; and finally, insufficient competence supply. These limiting factors will be elaborated on below.

Development potential

Despite current efforts in Big data analytics, vast amounts of data generated in society is largely unexploited today due to the many challenges outlined earlier. Although today there are a large number of companies that successfully base an essential part of their activities on large-scale data analysis, valuable information is lost every day due to the lack of service platforms, considerable risks concerning privacy, and insufficient analysis methods.

Knowledge of and access to patterns mined from data have the potential of enabling a multitude of new information-intensive services and applications that result in improvements in a number of important areas, e.g. in telecommunications, transport and media. For instance, the possibility of targeting offers to consumers depending on knowledge of user data patterns will for example be worth an immense amount of money given the large sums paid in the advertising industry today. User data is also in itself of great commercial value, and providing this information to stakeholders other than the immediate providers of user data has the potential of creating a new industry and enabling start-up companies and innovators implementing services based on this information.

There is a very strong potential in consolidating Sweden's strengths in Big data analytics with respect to several application areas, such as telecommunications, biomedical applications, global systems science (within climate and energy), societal planning, eScience (e.g. fundamental physics and astronomy) and business intelligence. The ICT and business infrastructure supporting such applications is

[13] http://royalsociety.org/policy/projects/science-public-enterprise/report/

relatively mature in Sweden and Swedish industry; and - as outlined above - there is a strong R&D tradition within critical areas such as data analysis, cloud computing and networking. In general, there is an opportunity in Sweden to develop novel communication and service technologies in support of industries that become increasingly data-driven.

Within telecommunications, Sweden is in an excellent position at the forefront of cellular technologies, moving into services and cloud technologies - both areas that heavily rely on Big Data Analytics. On the networking side, there is a potential to innovate towards harnessing the large amount of operational data that is constantly produced in such systems and perform analytics, possibly in-network, both to improve real-time fault, performance, and security management as well as creating novel analytics services based on this data. On the service infrastructure side, developing platforms for Big Data processing and analytics would create a significant competitive advantage. In the services area, Sweden has a developing ecosystem of small to medium sized end-user service companies generating a wealth of data that can be harvested for more effective and efficient operation, for serving users better, e.g. by anticipating their needs, and for monetising aggregated user statistics.

With Facebook's establishment of a data center in Luleå, we are also seeing the beginning of a service ecology consisting of both smaller and larger companies establishing themselves within Big Data management and services. This development, in conjunction with the attractiveness of Sweden for data centers due to access to cheap and carbon neutral energy and cooling, that with the right support has the potential of becoming the foundation of a new Swedish strength- and export area completely reliant on Big Data and Analytics expertise.

Further, there is an untapped potential in employing Big Data Analytics in the public sector, e.g for data driven socio-economic planning. This is reflected in that the Swedish government has in its strategy for a digitally cooperating public administration declared an intention that Swedish public administration should be more open and innovative by the use of open data and open solutions within the administration and in relation to the society[14]. Big Data Analytics has a crucial role as the government wants to increase the possibilities of developing new and innovative digital services in the society. Extracted relevant information from data can provide policy makers and industry with detailed descriptions of the world in order to make informed decisions in both operational situations and at strategic crossroads. The specific role of quantitative analysis is to bring objectivity and agreed-upon and well-documented methods into the description of society. Analytics can adapt to new needs as society changes, and when the availability of data increases, completely new methods can be developed.

Strengthening Big Data Analytics in Sweden

Although Big Data Analytics has an enormous potential in creating value for Sweden, several efforts are required in order to strengthen our position and create a fertile ground for future businesses and services. We have identified five main areas of support listed below.

Big Data Analytics technology and infrastructure

Firstly, there is a need for strong directed support for research and infrastructure for Big Data research, as stated for example in the National Science Foundation call BIGDATA in the United States[15]:

> *[...] to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to: accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life. The new*

[14] http://www.regeringen.se/sb/d/16772/a/207625

[15] http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

> *knowledge, tools, practices, and infrastructures produced will enable breakthrough discoveries and innovation in science, engineering, medicine, commerce, education, and national security - laying the foundations for US competitiveness for many decades to come.*

This holds equally true for Sweden, and such support would enable applied research across several levels of Big Data Analytics, from data collection and storage, computational frameworks and advanced analysis, through development of API:s necessary for viable exchange of data, to research and implementation of business models. Such support could also include seed funding for projects involving both entrepreneurs and researchers with the explicit goal of developing business plans for new enterprises within Big Data Analytics.

Application development To both strengthen competitiveness and to develop Big Data competence in Sweden, it is imperative that real-world Big Data Analytics applications are developed. This is true both for the largely new digital services area, where this would involve development of completely new consumer and business services, and more traditional Swedish industry, where such development could generate a huge competitive advantage against other actors. In the latter case, there are often major gains to be found with relative ease, such as in increasing efficiency within forestry and mining and provide more advanced management services for telecommunication networks. In the current situation, such application development would most likely include both industry and academia to ensure competence within Big Data Analytics and industrial relevance.

Business ecology In order to build a business and service ecology around Big Data and analytics in Sweden, and to facilitate new value chains and business models based on refined data, common standards for sharing data and information in order to make it widely usable must be developed. More broadly, support of the development of a national infrastructure for sharing techniques and tools for Big Data Analytics as well as datasets would strongly benefit the field. Such an infrastructure should involve both software and hardware components, as well as platforms for support and experience sharing. There is also a need for knowledge transfer and development of open data in Sweden, including effective exchange of experiences between organisations who have, or plan to offer, open data sources. Such a strategy should be implemented in coherence with the strategy on e-infrastructures for research currently being formulated by the Council for Research Infrastructures at the Swedish Research Council[16]. Small enterprises within Big Data and analytics should be supported as well so that the area can reach critical mass in terms of related business ventures and competence within Sweden.

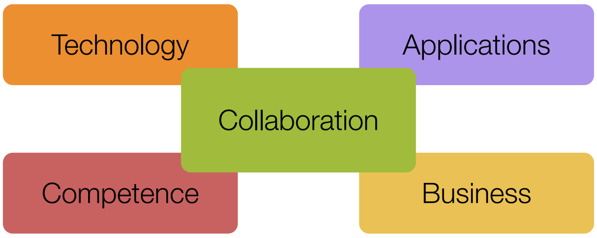


Figure 6. Big Data Analytics development areas.

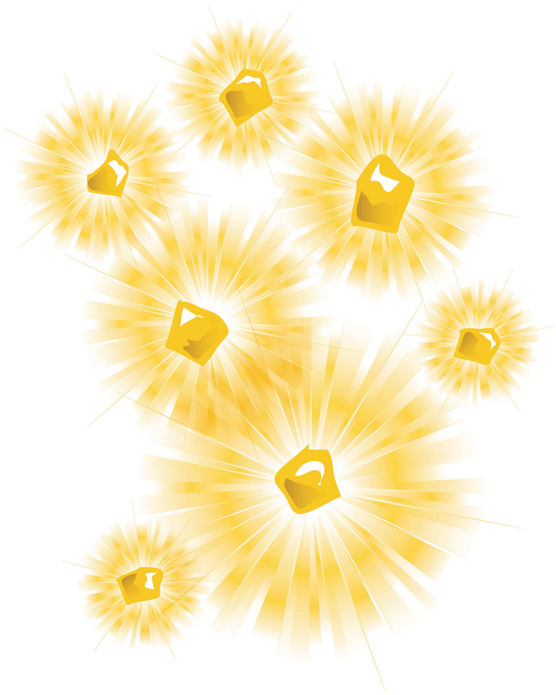
Competence supply Today, much of the key competence within Big Data Analytics comes in the form of Ph.D:s educated in related fields but with experience from advanced data analytics and hands-on experience of working with large data sets. As the area develops and gains importance, this is not going to be a tenable situation, and educational efforts for undergraduates will have to increase significantly. These efforts must combine theoretical education with experience of real world problems and business knowledge to ensure that graduates can contribute to solving practical Big Data Analytics problems. Further, educational efforts need to be directed to active professionals, both to complement the relative lack of experience with Big Data and analytics throughout industry and to ensure that they can update their skill-set to stay on top of current changes. Here, professional networks around Big Data Analytics could also have a significant impact.

[16] The Swedish Research Council's Guide to Infrastructures 2012, *Vetenskapsrådets rapportserie*, 3:2012.

Collaboration and internationalisation The full value of the support areas listed above can only be realised through efficient collaboration between parties, interests, and development areas. Technology research and development must use real applications and business cases as drivers, while new business and application opportunities depend upon the possibilities created by new infrastructure and research. None of the aforementioned areas will be able to develop without the necessary competence, the development of which need to stay on top of the latest research and industry needs. Thus, collaboration efforts are crucial to the success of Big Data Analytics in Sweden. We suggest to support the creation of a strong, open national network of both industry and academia which can serve as a contact surface for interested parties. Further, collaboration between areas should be encouraged in supported projects within the area, such as combining technology, application, and business development with educational and research efforts in universities and institutes.

Finally, Sweden will need significant international collaboration to continue to stay at the forefront of Big Data Analytics Development. We cannot lead the development in all aspects of Big Data and related areas, meaning that we have to ensure that we have efficient means of bringing competence and technology into Sweden. Here, European collaboration will play a vital role. The European Horizon 2020 initiative lists Big Data and Analytics as important innovation areas, and Sweden has several partners within EIT ICT labs which could be used to further develop and exploit results. Collaboration with the United States is also highly desirable, but although there are some good contacts within academia these should be strengthened and extended if possible.

If these development areas are appropriately supported, we believe that Sweden will be in an excellent position to gain both significant competitiveness as well as considerable societal and environmental value through Big Data Analytics.



The Swedish Big Data Analytics Network

Behind this agenda stands a consortium of partners including several of the most important organisations within Big Data Analytics and its related areas in Sweden. The consortium includes large established companies, small and medium sized companies, universities, institutes, and stakeholders in the public sector. The network is open for further partners with an interest in this important field.

SICS Swedish ICT, Swedish Institute of Computer Science, is a non-profit research organisation, and the leading research institute in computer science in Sweden. SICS' mission is to contribute to the competitive strength of Swedish industry by conducting advanced and focused research in strategic areas of computer science, and actively promoting industrial use of new research ideas and results in industry and society at large. SICS works in close collaboration with industry and the national and international research community.

SICS has substantial competence, many years of experience and several research projects within most of the Big Data Analytics research area, including: networked systems, storage and computation platforms, advanced data analysis and machine learning, visualisation and interaction, and service platforms.

Contact: Daniel Gillblad, 08-633 1568

Viktoria Swedish ICT Part of Swedish ICT, Viktoria has conducted applied research on data driven service innovation for a number of years. Viktoria Swedish ICT has built competence in research projects together with vehicle manufacturers, the transport industry, and universities nationally as well as internationally.

The institute has documented experience on both organisational land technical level research. The research groups Digitalisation Strategy and Cooperative Systems have worked with business models and open innovation as well as data mining applications in complex organisational contexts.

Contact: Magnus Andersson, 0736-457163

KTH - ICT The School for Information and Communication Technology at KTH has activities that span the whole field of information technology in its broadest sense, including basic as well as applied research in close collaboration with leading international universities. At ICT eminent international research is carried out especially in nano physics, photonics, electronic and computer systems, and communication systems and services.

Contact: Seif Haridi, 070-512 15 40

KTH - Electrical Engineering The School of Electrical Engineering at KTH has longstanding activities within the area of communication and processing of large data sets, both in research and in teaching. Communication theory provides us with the fundamental understanding to represent and efficiently store large data sets; signal processing gives us with highly effective algorithms for these tasks and also the capabilities to process data streams in real time; networking research provides us with the means to efficiently transport the data, and control theory gives us the methods to built stable and globally controllable systems.

Contact: Rolf Stadler, 08-7904250

KTH - Geoinformatics The Geoinformatics division at the Department for Urban Planning and Environment is responsible for both research and education within the broad field of geographical information technology (GeoIT). Geoinformatics or GeoIT is the science and technology for collection, management, visualisation, analysis and presentation of geospatial data. The research at the Geoinformatics Division at KTH is focused on methodology development and the applications of GeoIT for sustainable urban/regional planning, environmental monitoring, crime analysis, and health studies. The Geoinformatics Division is responsible for the GeoIT Profile in the Built Environment Program and for the International Masters of Science Program in Geodesy and Geoinformatics.

Contact: Yifang Ban

KTH - FMS The Environmental Strategies Research – FMS division is a part of the Department of Urban Planning and Environment, which in turn is a part of the School of Architecture and Built Environment. FMS' aim is to develop solutions for, knowledge on and debate around strategic environmental problems. This is primarily done through multi-disciplinary research. The area of research is the interconnections between environmental issues, technological developments and societal change.

Contact: Viveka Palm, 070-5854219

Stockholm University - DSV The research group within data and text mining at the Department of Computer and Systems Sciences consists of two professors, one associate professor, three lecturers and nine PhD students. A particular focus of the research is on predictive data mining using ensemble methods, i.e., techniques for generating sets of models that collectively form predictions by voting, and on methods for generating interpretable models, e.g., rule learning. The research on text mining focuses on efficient and resource lean methods using language technology for very large text sets.

Major application areas of the research include healthcare analytics and pharmaceutical research, and one of the projects lead by the group within these areas is "High-performance data mining for drug effect detection", which is funded with 19 MSEK by the Swedish Foundation for Strategic Research during 2012-2016. Another application area is automotive research, and the group is participating in the project "Integrated Dynamic Prognostic Maintenance Support", lead by Scania AB, and supported by 11.6 MSEK from Swedish Governmental Agency for Innovation Systems during 2012-2017.

Contact: Henrik Boström, 08-161616

Karolinska Institutet - Unit of Computational Medicine Karolinska Institutet (KI, ki.se) is one of Europe's largest medical universities. 80% the activities are devoted to research and 60% of the research is funded with external grants. The annual production from the Institute is 4000 research papers including hundreds of papers with an impact >15. As of 2007 KI coordinated 34 FP projects and was a partner in 220 projects. The mission of the Institute is to improve the health of mankind through research and education. Each year, the Nobel Assembly at Karolinska Institutet awards the Nobel Prize in Physiology or Medicine. KI has also developed a system supporting innovation with Karolinska Institute Holding governing these activities.

The Unit of Computational Medicine is headed by Professor Tegnér and houses >30 researchers (>60% PhD's) with expertise and capacity in Computational Modeling, Molecular Biology, Bioinformatics, Translational Informatics, Software engineering and Big Data Analytics, and a sequencing facility. The Unit resides in a clinical research center - The Center for Molecular Medicine (cmm.ki.se).

Resources: Storage and computing facilities, in-house LIMS and translational informatics resources, a fully equipped molecular biology laboratory, in-house Illumina HiSeq2500 sequencer in operation, several in-house large-scale clinical databases including the Patient Disease Networks with > 1 million disease pairs over 40 years including > 5 million individuals. To the best of our knowledge this is the largest comorbidity and clinical outcome database in the world.

Contact: Jesper Tegnér, 070-680 4989

Chalmers University - Department of Computer Science and Engineering The Department of Computer Science and Engineering (CSE) at Chalmers University is strongly international, with about 70 faculty and 70 PhD students from about 30 countries. CSE provides a dynamic research environment and has groups of world renown in a number of fields, and expertise within e.g. algorithms, machine learning, and software engineering. The participating group is leading a 2,5 M Euro project funded by the Swedish Foundation for Strategic Research on “Data Intensive Systems” together with an industrial consortium led by Recorded Future. This project aims at developing technologies for predicting time evolution of networked data based on collection of open source data feeds from the internet, paying attention to privacy concerns.

Contact: Devdatt Dubhashi, 031-7721046

Uppsala University - Department of Physics and Astronomy The High Energy Physics Department at Ångström Laboratory, Uppsala University is participating in the ATLAS experiment at the CERN Large Handron

Collider (LHC) in Geneva. Our specific interest in this research program is to search for the charged Higgs boson, which is predicted to exist by theories beyond the High Energy Physics Standard Model theory such as Supersymmetric theories. The ATLAS collaboration has over the last months analysed and collected a several billion collision events, each of ca 1 Megabyte size, i.e. in total of the order of several Petabyte of raw data. At the same time approximately the same number of Monte-Carlo-simulated collisions have been generated and analysed.

We have contributed to the development of the grid e-science technology since its inception 10 years ago and are currently using this technology for processing our data for the charged Higgs searches. The Grid was one of the key infrastructures leading to the discovery of the Higgs boson in 2011. In the development of the grid technology we are collaborating within the eSSENCE project. Uppsala University is together with Lund University leading the development of the Advanced Resource Connector (ARC) middleware that integrates computing resources (usually, computing clusters managed by a batch system or standalone workstations) and storage facilities, making them available via a secure common Grid layer. THe ARC middleware is deployed in all European Grid tiers running with distributed resources.

Contact: Richard Brenner, 018-471 7616

Luleå University of Technology (LTU) is renowned for applied research in close collaboration with national and international companies and stakeholders. The yearly turnover is around 160 million Euros, of which 90 million Euros is related to research activities. LTU has 1600 employees and 17000 students distributed over 4 campuses. The university has several strategic partners, both including leading universities like Stanford University and Monash University, and multi-national corporations like Ericsson, Scania, Shell, IBM, Volvo Aero, LKAB and Scania. Several research centres are located at LTU, such as the Centre for Distance-spanning Technology (CDT). LTU is also a partner in the EIT ICT Labs, which is one of three established European Knowledge and Innovation Communities (KIC) that capitalise fully on the knowledge triangle to accomplish impact through integrated innovation and that thus offer additional catalyst activities based on carrier projects to support higher impact. The Pervasive and Mobile Systems research group work with among other things how Big Data can be made simpler for individual users and on tools to empower non-programmers to create their own mobile applications. The research group also works on how to visualise data in innovative ways by building on tangible interfaces.

Contact: Peter Parnes, 0920-491033

Halmstad University, and in particular the research centre CAISR, investigates data analysis along three "awareness" themes: human-awareness, situation-awareness and self-awareness. We are especially interested in near real-time processing of data streams using methods that are suitable for embedded systems, often characterised by limited bandwidth and memory capacity. In line with the notion of "ubiquitous knowledge discovery", we are often working in settings where information cannot be saved for any significant amount of time. At the same time, this data should also be integrated with other more low-frequent but complex information, e.g. vehicle repair or usage histories.

Examples of domains where such restrictions are common include diagnostics and planning on-board vehicles, responding to various traffic situations (on roads, at crossings, etc), production facilities in distance places (e.g. windpower plants), "smart grids" for distribution of energy and "intelligent homes" for assisted living.

Focus of the research in our group is on algorithms that are scalable and cost efficient, while at the same time being transparent and producing results that are easy to interpret for users. Since the final acceptance of solutions often requires more than just algorithms, we are also studying how services can be developed around the technology.

Contact: Thorsteinn Rögnvaldsson, 035-167477

Statistics Sweden is an administrative agency. Our main task is to supply customers with statistics for decision making, debate and research. We are mainly assigned these tasks by the government and different agencies, but we also have customers in the private sector and among researchers.

Besides producing and communicating our statistical data, we are tasked with supporting and coordinating the Swedish system for official statistics. We also take part in international statistical cooperation. Statistics Sweden is responsible for maintaining several official registers. Most of our official statistical production is freely accessed through the Statistical database (SSD).

Contact: Viveka Palm, 070-5854219

Lantmäteriet is a governmental agency under the Ministry of Social Affairs. One of Lantmäteriets missions is to contribute to sustainable and economic development by providing conditions for search, find and use geographic information and property information. Lantmäteriet produces geographic information and real estate information for society - government, business and individuals. Lantmäteriet is also responsible for building a spatial data infrastructure. This infrastructure is the basis for information exchange of spatial data from a large number of authorities and is also the Swedish node to a European geodata portal. With more and more data available, there is a great potential in increased data use, where data from different sources can be combined to produce new applications and create new services and products. However there a research questions that need to be addressed and development project to be carried out in order to optimise the data exchange and use by different actors and in different areas of application. Big data analytics can provide very useful knowledge and tools in this process.

Contact: Hanna Ridefelt, 026-634605

SIS, Swedish Standards Institute, is a member-based, non-profit association specialised in national and international standards. We therefore collaborate with companies, government agencies and local authorities, research scientists and professional organisations, in Sweden and the rest of the world, to establish standards that facilitate more rational routines, cost-efficient business flows and enhanced quality. In a globalised world, there is a clear need to achieve coordination across industries, cultures and national borders.

Information Management is one of the eight business areas at SIS, Swedish Standards Institute. Both the private and public sector benefit from making effective use of a robust and future proof IT infrastructure. Standardisation within ICT aims to create methods and rules for an efficient and safe management of information, promoting business development through the use of information communication technology. Current ICT projects include document management, geographic data, e-archive, e-learning and e-health.

Big Data has developed into a discipline in its own right. As formal and commonly shared definitions emerge SIS, Swedish Standards Institute, believe that Big Data can benefit from also developing standards in order to more effectively communicate these definitions. Areas for possible standardisation include descriptions regarding quality, completeness and precision etc. Another area identified is that regarding the interpretation of data, for example modelling, coding and data specifications. There are already today several standards in place governing the use of information technology and it is possible that many of these can be applied on Big Data. Geographic data for example hosts a number of well-developed IT standards. Users of Big Data might also wish to consider how management standards can be applied and implemented in order to address issues such as legal aspects, traceability, responsibly and confidentiality.

Contact: Fredrik Fehn, 08-555 521 22

Ericsson is the world's largest communications equipment vendor and a leading provider of related professional services and service platforms to operators of mobile and fixed networks. Within Big Data, Ericsson investigate the possibilities to utilise new and emerging technologies in the data analytics area in order to cope with the ongoing data explosion. Finding value in unstructured, incomplete, heterogenous and dirty data is a corner stone in the network society. Combining Ericsson's unique knowledge of the network and proven methods for processing large scalable analytics creates very interesting opportunities for the future.

Contact: Martin Svensson, 010-7165105

IBM Big data represents a new era in data exploration and utilisation. IBM is uniquely positioned to help clients design, develop and execute a Big Data strategy that will enhance and complement existing systems and processes.

Big Data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile, and to answer questions that were previously considered beyond your reach. Until now, there was no practical way to harvest this opportunity. Today, IBM's platform for Dig Data uses state of the art technologies including patented advanced analytics to open the door to a world of possibilities.

IBM sees use cases, or Big Data Sweet Spots, in where the majority of the business benefits and technology merge together into value driven solutions. One example of such a use case is Operational Analysis where the trend revolves around integrating and analysing large amounts of machine generated data in real-time, from sources like smart meters, power stations or others kinds of sensors.

Contact: David Rådberg, 070-793 2019

Volvo Technology is the research and innovation center within the Volvo Group. We pioneer new product, services and solutions for all brands within the Volvo group. We are 500 research engineers and scientists working out of technology centers in Gothenburg, Lyon, Greensboro, Bangalore, Shanghai and Ageo.

Big Data analysis is an important technology domain for the Volvo Group. Our products are today constantly connected to the surrounding world, logging data about the vehicle and its surroundings. We see great opportunities in transforming this data into information for helping our customer utilise their machines in better ways.

Contact: Daniel Zackrisson, 031-322 04 85

SKF is represented in more than 130 countries. The company has more than 100 manufacturing sites and also sales companies supported by about 15,000 distributor locations. SKF also has a widely used e-business marketplace and an efficient global distribution system. Data analysis is an important part of day to day business. For example, the output of the business area SKF Reliability Systems is asset efficiency - optimising machine performance to enable a plant to increase production, while at the same time maintaining or even decreasing costs. Solutions include hardware and software, as well as services such as consulting, mechanical services, predictive and preventive maintenance, condition monitoring, decision-support systems and performance-based contracts.

Contact: Bengt Thulin

Sandvik is a high-technology, engineering group with advanced products and world-leading positions within selected areas. Worldwide business activities are conducted through representation in more than 130 countries. In 2012 the Group had about 49,000 employees with annual sales of approximately 99,000 MSEK. The Sandvik Group conducts operations in five business areas with responsibility for research and development, production and sales of their respective products: Sandvik Mining, a leading global supplier of equipment and tools, service and technical solutions for the mining industry; Sandvik Machining Solutions, a global market-leading manufacturer of tools and tooling systems for advanced industrial metal cutting; Sandvik Materials Technology, a world-leading manufacturer of high value-added products in advanced materials, special alloys, metallic and ceramic resistance materials; Sandvik Construction, provides solutions for virtually any construction industry application encompassing such diverse businesses as surface rock quarrying, tunneling, excavation, demolition, road building, recycling and civil engineering; and Sandvik Venture, aiming to create the best possible environment for growth and profitability in attractive and fast-growing operations.

Profiting from IT and information has always been a key success factor for Sandvik, e.g. enabling continuous improvements in production and providing value-added to their customers. Big data offers a new source process for internal process development and additional service offerings. Currently, the three main challenges are: Integration of traditionally separated data set; developing methodologies to cope with quality and authenticity of data originating from unknown and changing sources; and establishing mind-set to explore the opportunities in dynamic big data rather than traditional static rules of thumbs.

Contact: Ulf Hermansson, 026-266568

Spotify is a music streaming service that connects more than 24 millions of users in 20 different countries with the right music every moment. Spotify is one of the leaders in the music delivery business. Spotify is available everywhere: desktop clients, web client, handheld devices, embedded systems, cars, televisions and set-top devices.

Spotify is a data oriented company where decisions tend to be based on real data instead of gut feeling. AB testing and data collection is essential to achieve this. With the amount of data that 24 millions of users generate each day big data is critical. With more than 5 TB of data generated each day our focus on processing data and extracting valid information from it grows each day. Analytics and reports to music right holders is one of the first uses of data but not the only one. Studying user behavior to provide better recommendations or understanding relation between music track is critical to improve the user experience while using the service. Service metrics are also comprehensive studied to improve our service and prevent failures.

Contact: Pablo Barrera, 0735-011 722

TIBCO Spotfire Businesses collect increasingly large amounts of data through business events and customer interactions. TIBCO Spotfire develops analytics tools that can handle all this data effortlessly while enabling users to spot patterns, examine outliers, and detect trends in data that have previously remained hidden. It comfortably works with today's big data sets, helping users –- including both data analysts and business users –- find the most meaningful insights with intuitive, attractive, visual displays of information.

Contact: Olle Landström, 031-7041566

Recorded Future is a start up headquartered in Cambridge, MA and Göteborg, Sweden, with 25+ employees around the globe attacking a Big Data Analytics problem – organise the web in a radically new and useful way. The world's 24/7 media flow is filled with temporal signals, including reports of what's transpired or statements of what's expected to come. Recorded Future's linguistic and statistical algorithms extract time-related information and through temporal reasoning we structure the unstructured. In doing so, Recorded Future have formed the world's first temporal analytics engine.

Contact: Staffan Truvé, 070-5933885

Gavagai AB is a language technology company formed in 2008 in Stockholm as a spin-off from SICS in response to commercial requests to license technology for distributional semantics and large scale text analysis to industrial stakeholders. Gavagai develops and markets its proprietary text representation Ethersource, a base technology for semantic analysis of large text streams. Ethersource identifies meaningful relations between terms in text and can be used to identify attitudes or moods in text for financial, security, and market analysis purposes by commercial and public organisations for the analysis of open source intelligence in numerous languages – its main performance features are scalability, dynamic learning, and robustness in face of vast, vastly growing and heterogenous text streams on internet scale. Gavagai currently (2013) has eight full time employees in its core development team, most with PhDs and all with research background, and a number of external consultants with backgrounds ranging from linguistics and philosophy to computer science, engineering, financial analysis and military intelligence analytics. The technology developed by Gavagai provably more effective than others, and is radical departure from the standard models in use today. Gavagai will continue investing heavily in operational implementation of the latest research results in in-memory on-line fully incremental knowledge representation for human language in volumes orders of magnitude larger than competitors' models. This proposed research framework is in line with our strategic direction.

Contact: Jussi Karlgren, 070-3142564

Lincube Group, an SME, is since 2007 specialised in Big Data, Big Analytics and Data Warehousing. By partnering with the leading suppliers on the Big Data arena our +25 experts in data warehousing, large-scale data analysis, algorithm development and software engineering gives us a unique capability to help clients capture data, structure it, interact with it, understand and act upon it to improve performance.

Lincube delivered services range from development resources by the hour to complete Big Data Analytics turnkey solutions with included infrastructure.

Currently Lincube focus on Big Data research and development in the following areas:

- Analytics and data warehousing for Social Media. Plug & Play solution for integrating Social Media unstructured data with customers structured data to enable combined analytics to understand how to accelerate performance.

- Digital Analytics Warehouse - Real-time analytics of customers' digital on-line behavior to optimise design and content for maximising performance in the digital channel.

- External data integration hub - Development of interfaces for visualisation, text analytics and entity management of external unstructured data sources for data warehousing. Makes external unstructured data such as public and/or social media data accessible from any traditional data warehouse with a minimal effort.

Contact: Stefan Lavén, 0708-66 00 51

Findwise, founded in 2006, is a vendor independent SME, expert in creating solutions based on the leading enterprise search and analytic platforms: Autonomy IDOL, FAST ESP (a Microsoft subsidiary), Google GSA, IBM OmniFind, Microsoft SharePoint, LucidWorks Enterprise, Oracle, Exalead, Eurling, Splunk, Apptus and Apache Lucene/ Solr (Open Source). In addition to our technical expertise, we take pride in offering a large number of unique search related skills in areas such as business analysis, usability, computational linguistics, information management and security. Findwise is mainly a knowledge-based company with five offices, Gothenburg (HQ), Stockholm, Copenhagen, Oslo, Warsaw and Sydney. Of our 97 employees (Dec 2012) 95% hold a Masters degree from an accredited University. Our technical focused employees are developers skilled in mixed environments; integration, architecture and development of (Distributed/Cloud) search and content processing based systems. Findwise has experience from working with more than two hundred clients that are mainly (but not limited to) large companies and organisations in northern Europe. Findwise have experience of technical andmanagement work in "Better Search Engine", funded by Swedish Foundation for Strategic Research and the following European projects: FP6 RUSHES, FP6 AMIRA, FP6 DILLIGENT.

Findwise's interest in Big Data Analytics: We have for the last years seen the opportunity to apply our Language Technologies and Content/Data Processing framework within (Social) Media Monitoring to identify influencers and apply sentiment analysis of various products and services, Predictive Maintenance in the process industry, Pattern Mining in Electronic Health Records to identify relationships between diagnosis, medicine and comobirdity in Health-Care, Dialogue Systems within cars, and many other areas. But, our interest and expertise is not limited to structured data, but instead to join structured and unstructured data as well as create structured data – fact-finding – by mining a big repository of unstructured data.

Contact: Henrik Strindberg, 070-9443905

Appendix A: Application domains

The following appendix presents a sample of domains for which Big Data Analytics is important.

Industry

Process and manufacturing industry In most industrial plants and systems, the whole production line is equipped with sensors, which feed into monitoring and control systems. A single plant may have on the order of tens of thousands of sensors, often sampled at millisecond rates. Furthermore, enterprises today consist of not a single plant but several, distributed at different locations, often worldwide. The enterprises are in turn connected in supply chains, where each part depends on the others. There is a demand to use analytics of the collected data to get an overview of the production situation in the whole chain; to detect deviations and problems in time; to predict the production outcome; and to plan and dynamically adjust the production in respons to both the internal situations and external demands.

Telecom and internet Obvious enablers for the current rapid development in ICT, including Big Data, mobility, cloud services, and Internet of Things, are the telecommunication industry and the availability of internet everywhere. However, to monitor, maintain and upgrade this huge infrastructure, and to protect it from various threats on all levels, requires Big Data Analytics of the data flows, to find patterns, trends, and unexpected events in it.

Transportation industry Modern vehicles contain numerous sensors and electronic control units that generate large amounts of data. This data can be very useful for diagnostics, traffic safety, product development, etc. At the same time, the distributed and mobile nature of those systems makes them challenging to analyse. However, the benefits that can be obtained, for example by increasing fuel efficiency or by reducing the number of dangerous situations, make it an important area for both research and innovation.

Finance The role of Big Data Analytics in financial applications has gone through a generational change in which traditional analysts without a specialisation in data analytics have been urged towards modern analytics that involve network- and transactional data. The extreme constraints on financial data flows, where information value drops by the millisecond, means that any Big Data Analytics tools must rely on sampling of real-time flows. Fears and actual cases of herd behaviour and automatic triggering of orders has led to discussions on further regulating algorithmic trading. The maturity of Big Data Analytics tools and methods is key to keeping algorithmic trading safe and efficient.

Streaming media There has been a rapidly growing business around streaming media. Streaming audio services has been around for a while, and there are already several competing services for movies. Other examples include YouTube, and that e.g. SVT offers all their programs on the web. A significant part of all web traffic today consists of streaming material.

Characteristic for a provider of streaming media is the huge volumes of data to distribute, the requirement to minimise delays, and typically a very large number of customers. Together this calls for analytics both to monitor the distribution itself to detect possible problems, and to analyse user behaviour and trends, to be able to predict user demand or provide recommendations. Sometimes it is also used as a basis for caching of the distributed material.

Academia

Physics / eScience There are ever increasing sources of non-structured data from high throughput experiments in biology, large experimental facilities in Physics, energy grids, large climate studies and simulations, etc. New parallel, distributed, heterogeneous high performance computing architectures like clouds, multicores, clusters, FPGAs, as well as a new generation of algorithmic and statistical techniques is being developed to address this.

As an example, consider the search for particles within high energy physics. Searching for the charged Higgs boson, several billion collision events was analysed and collected, each of ca 1 Megabyte size, i.e. in total of the order of several Petabyte of raw data. At the same time approximately the same number of Monte-Carlo-simulated collisions have been generated and analysed. For the physics analysis of these vast amounts of data novel e-science tools and analysis methods need to be developed and employed. The extremely large data samples which need to be produced and analysed in order to find the very rare collision events in which new heavy particles are produced require data processing capacities. Currently about 250000 cores, 160 PB of disk storage and 90 PB of tape storage distributed over 150 sites around the world are used. The capacity is divided equally between data analysis and to generate and analyse Monte-Carlo-simulated events. These requirements have led to the development and build up, over the 10 passed years, of the computing data grid, which currently successfully satisfies the exceptionally high needs for compute power of the LHC experiments. The same grid technology is currently used also in a few other science areas like quantum chemistry, biophysics and astroparticle physics.

Life science Genomics research has high value to both society and industry. It is used by biomedicine researchers, hospital diagnostics, food industries, agronomy, and pharmaceutical industries. A quantum shift is happening in the area of human genomics. A huge wave of big data is approaching, driven by the decreasing cost of sequencing genomic data, which has been halving every 5 months since 2004. These improvements in both the cost and throughput of DNA sequencing machines have caused a mismatch between the increasing rate at which they can generate genomic data and the ability of our existing tools and computational infrastructure to both store and analyse this data. The scale of the storage requirements for genomic data is huge – a single human genome amounts to coping with the analysis of three billion base pairs. In addition to the storage of genomic data, its analysis will require both massive parallel computing infrastructure and data-intensive computing tools and services to perform analyses in reasonable time. Biobanks, that are used to store and catalogue human biological material, are not prepared to handle this wave of data - there is a Biobank bottleneck: a lack of platform support for the storage and analysis of the coming massive amounts of human genomic data.

Computational epidemiology In health applications leveraging on ICT developments, the possibilities for employing network data is huge. In particular, computational epidemiology – the study of all things epidemiological except the pure medical aspects – is thriving. Centres for disease control around the world are today employing ICT services and data analyses. They are also enjoying competition from non-medical services, such as Google Flu Trends. Computational epidemiology is resting on the access and analysis of massive data, and requires the amalgamation of non-medical data (obtained from so-called syndromic surveillance indicators) and medical data, e.g., lab results and health records. When combined, these two kinds of data provide for possibilities of constructing flexible and dynamic systems with attractive real-time properties; such uses include early warnings, halting or mitigation of disease spread, simulations and scenario-based reasoning relating to health policies, and real-time decision support to first responders. There are also possibilities for tracking animal activities, since strict laws regulate animal movement in many countries. Such tracking gives important clues about the movement of zoonoses, and about animal- human encounters, something that is key to understanding, for example, the origin of certain influensa virus strains.

Society

Health With the changing age profile of the society, it is becoming more and more important to provide systems that can support people in their lives in an unobtrusive but efficient way. Those systems need to “understand” humans and seamlessly adapt to their habits. An important concept is that of “aging together”, where the focus shifts from comfort and convenience for young and healthy people, towards safety and protection for elderly or sick ones. With the ubiquity of cheap sensors available today this is possible in theory, but new developments in data analysis techniques are needed in order to implement it in practice.

Transports and smart cities A large number of data is being gathered every hour in today's cities, but there is surprisingly little global analysis that is being done on it. While combining data from multiple sources needs to be done in a careful way to preserve privacy, the benefits of being able to detect abnormal situations or discover surprising relations between events definitely make it worthwhile. This area is a prime example of the need for combining very diverse types of information, and for presenting results in a flexible way.

Urban and physical planning Data for urban and physical planning is collected and produced by local, regional and national authorities, but is not generally shared and used in an efficient manner. To this data from all available sources can be added and used. An important part of this is to create work processes from the early data collection stage to the visualisation and presentation stage in order to optimise well grounded political and/or business decisions.

An example is location of preschools. An efficient preschool planning require historical and present data in order to make prognosis of future demographical development. This work is carried out most efficiently if all available data is easy accessible and handled, which rarely is the case. It is also important with cooperation between different municipality administration units. Here the work processes need to be much more lateral. The visualisation part is very important. Using data allows a much easier way to draw conclusions based on visual output, rather than long text documents with little possibility to give a large overview of a particular situation. For established preschools sensors could be used to monitor and improve indoor and outdoor environment and social activities, e.g. noise, ventilation, movement patterns. This could also be of importance to provide information of spread of diseases.

Socio-economic planning For socio-economic planning there exist data from Statistics Sweden for small areas called NYKO (similar to census districts), and the planning system itself is organised in each municipality. Today, it has become possible to build data systems on Internet that could be used by municipalities, private companies, media, and citizens. This would give a more Open Data situation. Such systems can incorporate functions like database handling, visualisation, change detection, small area mapping, forecasts, comparisons with other regions or even countries.

Spatialised Big data Big data analysis of large text volumes in Internet can be combined with geography. For example “which topics are discussed in certain areas”. Here it can be needed to aggregate geo-positions given by addresses to towns or districts, a typical geospatial task. One can also combine with background geo statistics to relate to population properties in different regions/districts, e.g. districts with large immigration or large emigration (national or international). Big data analysis with small areas, changes and relations can give a new source of information to act upon.

Integrated data initiatives Detailed geodata have been available for at least 10 years, however because of high costs most users have been restricted to very few data sources, e.g. registers from one authority or municipality, remote sensing data, or international data in outdated formats. This is now changing rapidly through Open source technologies, Open data initiatives and geodata cooperation. To some extent, it is the same type of data that was available also 5-10 years ago, but the extended access has dramatically increased the possibility to combine them for new applications. Examples can be seen in recent work by Statistics Sweden: Green areas in urban areas from satellite imagery combined with population data has provided information on green areas available by distance from person’s homes; shopping districts delineated using data on work places combined with population registers gives statistics on possible number of customers; commuting statistics, road data, bus stops, rail stations etc. yields data for planning of bus and train services.

Data can either be made available in detail to users, or can provided as services by which users can request computations from authorities or private companies. Services for computations on data with confidentiality restrictions are new methods for data production. Computation services can be from Statistics Sweden, transportation authorities and companies, mobile phone network providers, medical data etc.

Appendix B: Categories of data

Depending on the application domain, there are many different types of data involved in Big Data Analytics, each posing different challenges.

Structured data Although more and more data is labeled “unstructured”, there are still significant amounts of structured data and traditional data bases in many domains. For example in a process industry, the whole production line is equipped with sensors, collecting data about the process. Data may be missing or noisy, but it is stored in rows and columns, is predominantly numerical, and the meaning of each data item is well defined.

Another example of structured data is that found in official registers. Official register data is a key element of the infrastructure for official statistics. An official register is created and maintained by governmental institutions and typically comprises data on individuals, enterprises and real estates. If this data has a geographical dimension then it is often called geodata or spatial data. Many official registers were set up for administrative purposes, such as tax administration, but some are also created for statistical purposes. Some of the core registers for production of official statistics in Sweden is the register of total population, the real estates tax assessment register and the business register.

It is also becoming increasingly common that the interpretation of available data is known “in principle”, but in practice nobody has the full picture, with multiple actors having control over their own parts of the database. There is also the need to distinguish between human created data and automatically generated data (e.g. from sensors). Those two types have very different characteristics, the most important one being that human created data is often on a higher level of abstraction, with concepts close to what the output of knowledge discovery should be. At the same time it is often also less trustworthy, because noise that is present can be in the form of systematic bias.

Geographic data Large volumes of data from Geographic Information Systems are available today, regarding the physical landscape, e.g. roads, lakes, buildings, addresses, and also information about as persons, work places and transport routes. An important change in recent years is that producers are more willing to share data with others, given that this can be done in a safe manner. This opens up the possibility for completely new application areas and innovations, both in society (e.g. for urban planning or to monitor environmental effects) and commercially.

A branch of statistics focusing on spatial or spatiotemporal datasets is Geostatistics. As a research field it is associated with methods for interpolation and statistical models for estimation and simulation of spatial phenomena. In a more general sense, geostatistics refers to statistics aggregated (or disaggregated) to small geographical areas, typically grids or sub-municipal units, allowing studies on a fine spatial level. Geostatistics is closely related to geodata and in general some sort of geographical information system is required to make use of it. Geostatistics is generally created and provided by statistics producing institutions.

Real time media Real-time media includes streaming of both canned and live media to any number of media consumers. Today huge amount of pictures, audio and video is being created where online sites like e.g. Flickr, YouTube and Vimeo are the most prominent examples where users share media. To be able to further enhance these media much research and development is needed on how to tag people, places etc. in the media streams. A special case of media is that will come from the emerging area of LifeLogging with companies such as Memoto which produces products for capturing images “all the time” and sharing them via cloud services and automatically analysing the photo streams for event and action tagging. Another important source of real-time media is conferencing where audio and video date could be further enhanced by analysing it in real-time and in recordings. A special attribute of real-time media is that it usually generates very large amounts data compared to other types of data. The amount of real-time media data being created in the future will be staggering.

Natural language We know that the volume of human generated data, especially such data as is in verbal form, will grow manifold in the next decade. We can confidently predict that the ubiquitous availability of speech capture devices will soon be harnessed for recording and archiving purposes; we can see that the growth in telephony will be immense in regions of the world which are poor in ICT today; we can predict that the Internet of Things will generate large amounts of text-like communication between devices. All of this data will be generated on many levels of abstraction and in very various levels of editorial quality. For any service, product, or analysis based on human-generated language or even device-generated language, a processing layer or a base component which can be used for representing the meaning of communication will be necessary.

Time series The time aspect of the available data is very important in many applications, where the task is detecting trends and anomalies, comparing individual against the group or comparing individual at different times. This needs to be combined with detecting discrete events in a continuous data, as well as identifying context and external influences.

Event data Matching external events with time series and analysing their consequences is an area where there are more questions than answers. Similarly, finding causal relations and identifying important and unimportant events is going to be important aspect of Big Data Analytics solutions.

Network data An important category of data concerns very large networks, such as communication networks, biological networks (e.g. gene regulatory and metabolic nets) and social networks. Much of the interesting information is in the connections themselves, and in the structure of connectivity.

Linked data In computing, linked data describes a method of publishing structured data so that it can be interlinked and become more useful. It builds upon standard Web technologies such as HTTP and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried.

We build dataportals today, but there are indications that portals will not be necessary in the future. Instead, formats which enable intelligent search on the web might be the future. This should be explored further. How can the concept of linked data be used in society, for example to build intelligent web services using data from different public authorities both on a national and local level?

Appendix C: Related research areas

Several research areas within ICT are crucial to the development of Big Data Analytics, of which the most important ones are described briefly below.

Networked systems

The large majority of the applications and services described in the introduction of this strategic agenda that underpins the Big Data Analytics area are networked. Examples are the Internet of things, cloud services, media distribution and all kinds of mobile applications. They communicate and are part of complex *networked systems*. As such, they put requirements on the communication and networking technology that form the infrastructure for these networked systems.

The trends in the communication network area are largely overlapping with Big Data. The network traffic volume is clearly dominated by media distribution and specifically video and TV. The rise of the Internet of things means that we are getting yet an order of magnitude increase in the number of connected devices that produce data of various kinds. Mobility is the norm - people and devices roam the world. Services are increasingly provided by `the cloud', or more correctly, from large data centers run by third-party providers.

These trends together with the requirements from the applications and services lead to research issues and challenges for the development of future networking technology. Efficient mechanisms including in-network caching and information-centric networking are needed to support media and other content distribution services. New solutions are needed to reduce the latency induced from the network infrastructure and cloud service platforms. Improved mobility mechanisms are needed that can handle the large amounts of mobile devices in the Internet of things. Operation and management of networked systems needs to explicitly deal with uncertainty stemming from the nature of networked systems using probabilistic approaches and programmable networks.

Cloud computing

Cloud computing is an enabling technology for Big Data. It democratises Big Data by enabling even small companies to rent large-scale computational infrastructure to process and store large volumes of data. The recent emergence of platform-as-a-service (PaaS) for Big Data has also reduced the barrier for small companies, as they lower the cost and effort in deploying and administering large-scale clusters. Research into storage and data-intensive PaaSes for cloud computing is transforming how we will store and process Big Data.

Mobile services

The landscape in which mobility experiences thrive is messy, and likely to remain so. Our actions will be situated in a hybrid landscape of technology installations and devices and not all will talk to each other, or even be connected to the Internet. Every situation becomes reachable by and using a multitude of services at once. We are no longer just in one place at a time (as we are communicating with people and things in other places), and we are no longer doing just one thing at a time (since we are simultaneously interacting with a plethora of services). Mobility, or movement, of persons and objects is essential in modern society and a mobile lifestyle brings with it e.g. temporary relation to the locations we visit and the situation we engage in. It allows us to encounter more people and visit many places and affect the experiences of these interactions. It has consequences for the way we interact socially.

We foresee three types of technological paths through which mobility services will continue to emerge i.e. sensors, networks and software. Mobility experience services will change dependent on the availability of new sensors that are integrated in a mobile device (the phone); attached to the body or to clothes, or embedded in the environment. Apart from the already prevalent accelerometers and pedometers, we expect that other sensors will become more commonplace, such as pressure sensors, bio-sensors for measuring heart rate, galvanic skin response (GSR), sensors measuring air pollution, weather sensors, moisture sensors, motion sensors and similar. Oftentimes, the camera and microphone on the mobile can be used as sensors. For example, the microphone has been used as a stethoscope and the camera has been used as a motion sensor. Sensors are becoming embedded in textiles, can be made out of paper, or integrated in plasters. Increasingly, these sensors are integrated with wireless communication in so-called sensornodes. We will also see more actuators, such as small subtle vibrators, heat actuators, or materials that increase/decrease in size or shape.

These services are also dependent on advances in networking technologies. An important property of these services lies in how the different units are networked to one-another in various ad-hoc networks or to the Internet and the data cloud. Jointly, these ad-hoc connected devices may create for functionality arising from dynamically configured, mobile settings of many networked units. The units may produce large streams of data sensing various aspects of user movements, bodily data or interactions with other people, that can be capitalized to create services – using crowd-sourcing, recommender systems, social navigation techniques, or other machine learning algorithms.

Internet of Things

Internet of Things, giving a digital presence to objects of the physical world, is becoming part of our daily lives. It is happening in different forms: smartphones and their downloadable apps, home multimedia systems, RFID identification and tracking. Many more applications are expected in the coming years, connecting our phones, cars, appliances, buildings, toys, cities, environment, and social networks, towards a richer world. These applications will result in unprecedented amount of data, that no existing system is readily able to handle. The challenges are numerous: scalable database storage, data center management, information-centric schemes - to name a few.

Artificial Intelligence

Artificial intelligence is not a single homogenous research area, but a collection of several disciplines, loosely characterised by dealing with tasks that traditionally have been considered hard to automate and thus have required human beings to solve. Areas stemming from this ambition include: machine learning, computer vision, speech recognition, natural language processing, robotics, knowledge representation, automatic reasoning, planning, agent techniques, and many more. After a relatively long period of basic research, many of these areas have now matured, and we suddenly find ourselves surrounded by techniques ultimately coming from these areas: programs that recognises faces, telephone answering machines that you can talk to, autonomous lawn mowers and vacuum cleaners, automatic translation between different languages, gps:es that recommend the best route, etc. In industry there are even more examples, such as software for surveillance, planning, and optimisation, based on AI techniques.

The connection to Big Data Analytics is twofold: first, with the enormous amounts of data to analyse, and the rapid pace of change in all of society, much of the handling of data must be autonomous and self-adapting, which require technology from the above areas. Second, much of the recent success in the same areas, e.g. in machine translation and machine learning, is due to the actual availability of huge amounts of training data today.