A background chapter in preparation of Norrbotten's Regional Digital Agenda 2017

Big Data Sovereignty in Norrbotten

Ahmed Elragal* – Luleå University of Technology

March 2017

> "Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway"
> Geoffrey Moore

1. INTRODUCTION

Gleaned from various sources such as emails, sensors, social media, etc., big data continues to be the topic of much discussion and companies that have pioneered to analyze big data and integrate it with traditional data are finding that the analytics benefits are real.

Despite its long history, recently data is described as a new type of asset and is increasingly recognized as a necessity for the digital economy. Recent research shows that the top 100 EU manufacturers, even with limited use of big data analytics, are capable of furthering EU economic growth by an added 1.9%, by 2020 [6]. For both public and private sectors, alike, big data analytics has become essential to economic activities and the fact-based decision making process. Therefore, it is inevitable that governments ought to have a big data strategy, if they opt to tap into such lucrative market.

This document is setout to be part of a revision of the digital agenda for Norrbotten, and aims to identify the key role of big data analytics in such agenda. The chapter begins by introducing the concept of big data, and presenting different techniques for how it can be analyzed. It then explores the challenges and opportunities of big data analytics followed by a discussion on how to address these challenges. Finally, potential benefits of developing and implementing a big data analytics strategy for Norrbotten is presented and an action plan is provided. The chapter is relevant for academia, public sector and the industry in Norrbotten, whereas a big data analytics strategy needs to be addresses by these three actors in order to be successful.

2. WHAT IS BIG DATA?

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Four main features characterize big data: volume, variety, velocity, and veracity. The volume of the data refers to its size, and how the exponential growth of data requires new approaches for analytics. Variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data [5]. Veracity, however, deals with the correctness of the data. See Figure 1.

Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it is coming from a greater variety of sources than ever before, including Internet of Thins (IoT) data, logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now

* E-mail: ahmed.elragal@ltu.se

joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML), JSON or Rich Site Summary (RSS) feeds. There is also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume.

Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites [13].

Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [14].

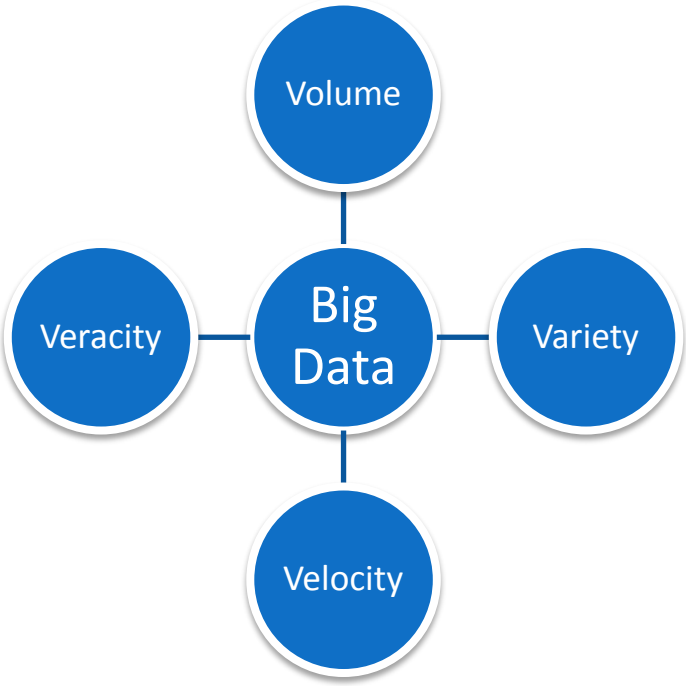

Figure.1: Big Data Characteristics

3. BIG DATA ANALYTICS

It is difficult nowadays to open a popular publication, online or in the physical world, and not run into a reference to data science, analytics, big data, or some combination thereof [1]. The interest in big data analytics (BDA) is on the increase. Google's adoption of the MapReduce was definitely a catalyst, which has led to a lot of developments in the area of BDA. Further, the development and deployment of Apache Hadoop, SPARK, and Mahout has also opened the doors for organizations to process extremely large datasets. BDA is the use of advanced techniques, mostly data mining and statistical, to find (hidden) patterns in (big) data. BDA is where advanced techniques operate on big datasets [13]. The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems [4]. A significant amount of these techniques rely on commercial tools such as relational DBMS, data warehousing, ETL, OLAP, and business analytics tools. During the IEEE 2006 International Conference on Data Mining (ICDM), the top-ten data mining algorithms were defined based on expert nominations, citation counts, and a community survey. In order, those algorithms are: C4.5, k-means, SVM (support vector machine), Apriori, EM (expectation maximization), PageRank, AdaBoost, kNN (k-nearest neighbors), Naïve Bayes, and CART. They cover classification, clustering, regression, association analysis, and network analysis. Actually,

not only organizations and governments generate data; each and every one of us now is a data generator [9]. Humans produce data using our mobile phones, social networks interactions, GPS, etc. Most of such data, however, is not structured in a way so as to be stored and/or processed in traditional DBMS. This calls for BDA techniques in order to make sense out of such data.

BDA is inherently related to data mining, a term that has often been used interchangeably with knowledge discovery in database (KDD). However, we see data mining as a step towards knowledge discovery. The term KDD was coined in 1989 to point to the process of finding knowledge in data [7]. KDD is also defined as the process of finding patterns hidden information or unknown facts in the database. Traditionally the notion of finding useful unknown patterns and hidden information in raw data has been given many titles including knowledge discovery in database, data mining, data archaeology, information discovery, knowledge discovery or extraction, and information harvesting. The lack of consensus on the term is attributable to the relative novelty as well as the multi-disciplinary nature of KDD. Multi-disciplinary means that KDD belongs to many disciplines like statistics and computer (machine learning, artificial intelligence (AI), databases, data warehousing, expert systems, knowledge acquisition and data visualization. Data mining is considered a step in the KDD process of discovering useful knowledge from data, where data mining points to the application algorithm or technique used for extracting patters and unknown information from the raw data.

Big data analytics is mostly used with the intention to predict. Prediction is the ability to foresee the future, based on applying certain techniques on datasets. Predictive analytics is a process whereby information extracted from various data sources is utilized to elucidate patterns as well as predict the future. Predictive analytics has the potentials to bring great business value to organizations and individuals, alike. Added to that, prediction has been identified as a key research area of the future.

On the other hand, predictive analytics is differentiated from prescriptive analytics which refers to the determination of a course of actions or decisions. In other words, the focus of prediction is on what will happen, whereas the focus of prescription is on how to make it happen [10]. For example, in a telecommunications operator content, predicting works to identify which customer will churn, while prescription works in ways to avoid it from happening via say simulation models.

Big data analytics ecosystem is represented in figure 2. That is, collecting and storing big data on its own is never the objective of any business. Therefore, in order to extract value from big data, it must be analyzed in a timely manner, and the results need to be available to decision makers, enabling the so-called fact-based decision making. The value realized by an organization is a function of having the appropriate combination of people, process and technology. Fact-based decision making is the extensive use of data, statistical, data mining and machine learning techniques to predictive models to support business decisions. Analytics enable organizations to enhance performance via turning data into intelligence.

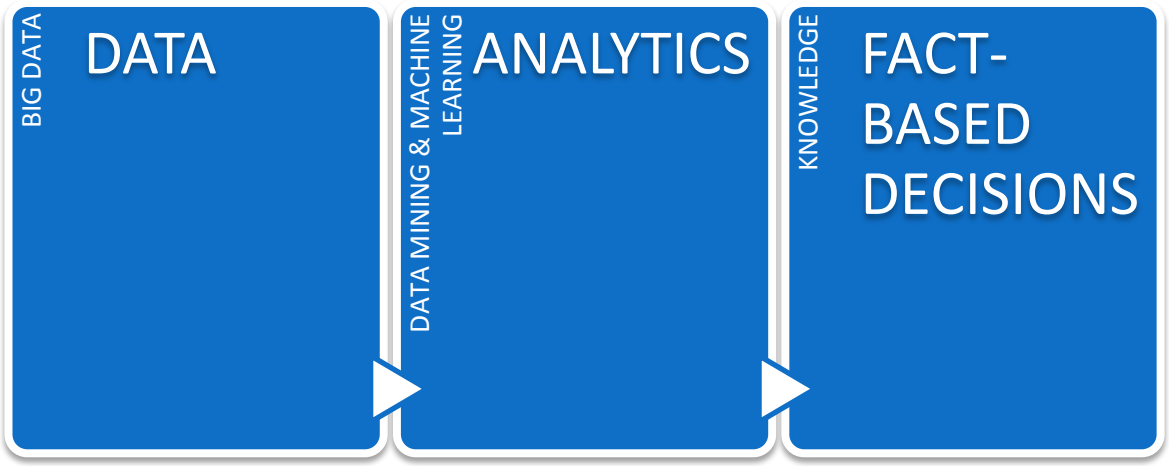

Figure.2: Big Data Analytics

4. NORRBOTTEN TO TAP INTO DATA-RELATED LUCRATIVE MARKET

The global market for big data related hardware, software and professional services (such as data center computing, networking, storage, or analytics) is booming and is predicted to reach EUR 43.7 B by 2019 – 10 folds more than 2010 [8]. Relative to that market, we strongly believe that Norrbotten has a degree of attractiveness. That is, the area is witnessing growth in the data center business. Added to that, the potential for innovative research in various areas, such as space research. Key attractiveness points are discussed in the next paragraphs. See below.

I. *Cloud* computing: Today, the role cloud computing is playing in the ICT industry is critical. Recently, Norrbotten in particular has been at the forefront of data center and cloud investments. For instance, Luleå University of Technology, in cooperation with SICS, have started the first phase of a national large-scale data center for testing and experimentation of cloud and big data projects. Also, Hydro66 is another shining example. Indeed, the data center growth fuels the digital economy in numerous ways. For example, it brings international business to the region – e.g., Facebook - as well as facilitates research and innovation. Lastly, cloud computing contributed to deeply change the way business and society think about technology;

II. *Space & big data research*: the presence of space research at LTU together with IRF and SSC represents an interesting area for potential future research. The digital revolution, combined with decreasing costs of manufacturing and launching satellites, means space is becoming a powerful tool for collecting data at global and local scales. The everyday life of Europeans has become unimaginable without satellites orbiting in space, as people across the continent make use of satellite, navigation and timing services provided by Europe Galileo satellites. The future potential is huge as it coincides with other advances for which navigation is a key, such as connected devices – aka IoT, self-driving vehicles – aka IoV, etc.;

III. *Location*: A recent survey with companies, when asked to rate specific drivers of location relating to data operations, the top five most important issues include tax rate, ease of doing business, legal framework, talent and data-related regulations. We believe those factors need to be studied in order to increase location attractiveness in such lucrative data-related market.

These developments offer opportunities to Norrbotten region; however, enablers need to be present in order to tap into such lucrative data-related market and being able to utilize big data analytics in a way to address emerging societal challenges and contribute to better policymaking. That is as-if we are saying that there exists a reciprocal effect between big data and policymaking. See figure 3.

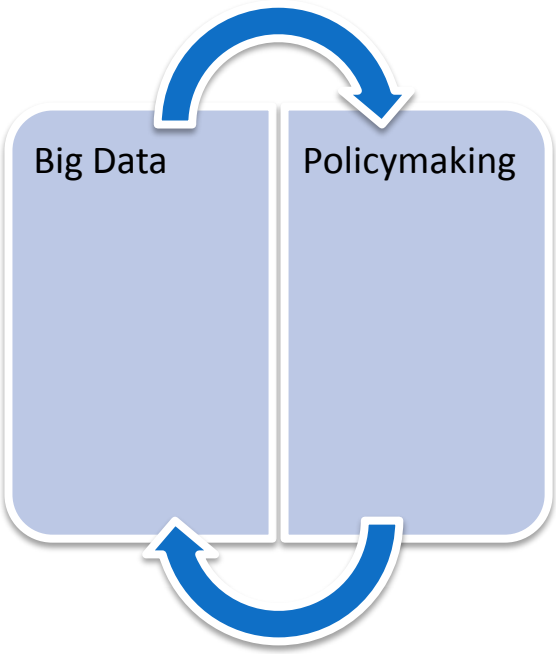

Figure.3: Big Data & Policymaking

5. CHALLENGES FACING THE BIG DATA-DRIVEN ECONOMY

The digital economy, see previous section, is one that is based primarily on data and analytics. Hence, it is worth noting that in such data-driven economy, organizations and governments alike are facing challenges which need to be addressed in order to reap-up its tremendous values. Below are some key challenges.

I. *The scope of big data processing*: there is inclination amongst data scientists and analysts to include as much data as possible in the discovery process. In one hand, this increase possibility of results comprehensiveness and accuracy, but on the other hand puts lots of pressures on the discovery process in terms of the hardware and tools required;
II. *Value-realization*: big data is preprocessed and then analyzed using analytics techniques. The objective is to elucidate knowledge which has the potential to render value to companies in a form of better pattern recognition as well as predictions.
III. *Data is the new oil:* data is being collected by the digital platforms for three reasons: - to provide and enhance the services such platforms offer for their users; - for the purpose of marketing; & - for monetization purposes. The challenge, however, is to decide what data to collect? What data to share?;
IV. *The limit on permissible data processing*: unfortunately, data protection laws and legislations do not address the issue of how to deal with the uncontrolled power in the hands of companies processing and analyzing users' data;
V. *Privacy and profiling users*: even though users have not setup their own profile, an accurate profile, however, could be gleaned from the data exhaust which we leave in various online platforms;

VI. *Data Governance:* data governance is a prerequisite for a successful big data analytics project. Governance ensures that the analytics does not violate and legal framework and ensures trust for all stakeholders;

VII. *Scarcity of the data scientist*: worldwide, there exist a need for data scientists whereby the demand exceeds the supply. See also section 7;

VIII. *Streetlight projects*: Big data is being passively created and continuously collected, and this has opened the door for plenty of research to be conducted. However, big data projects should be formulated around important problems [11]. Yet, it has been noticed recently that big data projects may have suffered from the so-called 'streetlight' effect. That is, the tendency of organizations and researchers to study phenomena for which there exist plethora of data, instead of studying relevant problems. Hence, such data may be biased towards solving local problems, and not necessarily the grand problems; &

IX. *Data monetization*: it is the ability of a company to generate money from its available datasets. In today's environment, companies have become aware of the meaning of the term "data is the new oil". Accordingly, each company is sitting on sheer amounts of datasets that need to be utilized towards value creation. The way data monetization is implemented at companies could either be direct or indirect. Direct data monetization means selling (part of) the dataset of a company. Indirect monetization uses the dataset to create new products and services, such as Amazon is using its customer records to suggest other products or Alibaba via its targeted finance. Another form of indirect monetization takes place whereby a company is bartering its datasets. The challenge with monetization is that sometimes a certain project is not able to get the dataset it needs to fulfill its objectives, due to finance shortage. Therefore, we end up with incomplete knowledge, and that definitely affect fact-based decision quality!

6. BIG DATA SOVEREIGNTY IN NORRBOTTEN: THE PRINCIPLES

There exist a number of principles which are required to facilitate big data utilization, growth, and innovation in Norrbotten.

I. *Data for public good:* datasets that both public and private sectors hold are considered to be national assets. Therefore, data should be used for public good. Sharing of data, in accordance with the Open Data in Sweden (visit http://oppnadata.se), will enhance the culture of engagement;

II. *Skillsets and capabilities sharing:* skillsets in big data analytics should be shared amongst Norrbotten government agencies in order to facilitate speed lane to face-based decision making and innovation; whenever appropriate. Capabilities such as datasets, analytics models, and infrastructure necessary to perform these computations, would be shared amongst agencies whenever applicable;

III. *Crossing the gap between academia & the industry:* recently, academia and the industry have been working on big data analytics projects. Relevant government agencies in Norrbotten are required to engage with industry and academia, as well as SME's and international relevant bodies in order to work on big data analytics projects and innovations. Such engagement will leverage expertise in big data analytics and increase knowledge in this area, in Norrbotten;

IV. *Regulatory framework:* the EU's General Data Protection Regulation (GDPR), regulates the processing and use of personal data in the EU [6][12]. It represents a basic landmark to creating a data-friendly atmosphere whereby citizens and companies feel confident that their privacy preferences are protected, while also safeguarding economic interests and innovation. Personal data represents one part of valuable business data; and businesses are now looking towards the EU to also

ensure a level playing field and legal certainty with regard to the use of non-personal data[†] so that they can release the potentials of the digital economy;

V. *PPP:* The European Commission Communication "*Towards a thriving data-driven economy*", paves the road for actions to support the emergence of the data economy [3]. Among the measures suggested was the development of a "big data" community, based on public-private partnerships, as well as "*excellence networks*"; support for the deployment of necessary ICT infrastructure, including cloud computing; and an initiative to promote the adoption of open data and big data analysis within public administrations [2]. PPP should also be promoted in Norrbotten.

7. DATA SCIENTIST: HAVE WE GOT TALENT?

The scarcity of high quality big data resources is a challenge, not only in Norrbotten or Sweden, but also in Europe[‡]. Norrbotten needs to address the need for educating learners with the appropriate big data skillsets. Education as well as training programs in the region are the core and crux in creating big data innovations. There exist a number of European initiatives. For example, the European Data Science Academy (EDSA) is analyzing the required sector specific skillsets for data analysts across the main industrial sectors in Europe to develop a data science curricula which meet these needs. Added to that, the EIT-Digital Master on Innovation recently launched a major on Data Science as a joint initiative of 6 European Universities (Universidad Politecnica de Madrid (UPM), Eindhoven University of Technology (TU/e), and Universite Nice Sophia Antipolis (UNS), KTH, TUB). It is a two-year master in which the 1st year includes foundations courses (data handling, data analysis and data management, visualization) and the 2nd year focuses on application on various domains. Having that being said, most of the programs ignores the storage, while focusing more on analytics and preprocessing.

Data Scientists require knowledge in key areas such as statistics, machine learning and data mining, programming, and visualization. They also are assumed to know data management tools. Data scientists are expected to develop novel algorithms and approaches for big data. Examples include: learning algorithms, predictive analytics mechanisms, etc. furthermore, data scientists should be able to evaluate the analytics results obtained. Education programs should focus on these skillsets.

Norrbotten should address, in collaboration with higher education institutes and education providers, the following:

I. Big data related *new educational programs* based on interdisciplinary curricula;
II. *Professional courses* to educate and leverage the skillsets of current workforce with the specialized big data skills;
III. A *forum* between scientists and industry experts;
IV. *MOOC*s in the area of big data analytics.

8. ACTION PLAN

So what? In this section, we introduce a number of actions to be planned and executed in order to leverage Norrbotten's potentials, capabilities, and attractiveness in the big data analytics domain. See below.

[†] Distinguishing personal and non-personal information is harder than ever before. To that end, there exist a lack of a universally accepted definition for what belongs to each of the two categories.
[‡] There is an unbelievable shortage of data experts globally, and in the EU. Some reports project the number needed in the EU to reach 500,000. Visit https://joinup.ec.europa.eu/community/opengov/news/500000-data-scientists-needed-european-open-research-data

A. *The establishment of a Big Data Working Group (**BDWG**) at Norrbotten level in order to follow-up on various issues pertaining to big data. Members should come from Region Norrbotten representing government, academia, and the industry;*
B. *The **BDWG** is to identify and report on the barriers to big data analytics. Report is produced twice a year detailing* barriers and providing mechanisms to address them;
C. *The **BDWG** is to work together in order to design and promote education & training programs that enhance skillsets and experience in big data analytics in Norrbotten;*
D. *The **BDWG** is responsible for design and publication of a guide to the successful data analytics projects within Norrbotten. The objective is not only reporting, but also elucidating knowledge and guidelines for successful endeavors. Such reporting is profiled in an annual networking event;*
E. *The **BDWG** is responsible to develop a database of skillsets, tools, datasets, programs, and projects for local use and promotion.*

By continuing to place ample efforts on big data analytics in Norrbotten, in collaboration between government agencies, academia and the industry, we will be in a position to creating a nurturing ground for big data international players to enter our region and stimulate this growing data-based digital economy.

REFERENCES

[1]. Agarwal, R., & Dhar, V. (2014). Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research, 25 (3), 443–448.
[2]. COM (2015)0192.
[3]. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: 'Towards a thriving data economy', COM (2014) 0442 final, 2 July 2014.
[4]. Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. The 14th Industrial Conference on Data Mining (ICDM). Petersburg: Springer-LNCS.
[5]. EMC (2012), Data Science and Big Data Analytics. In EMC Education Services, pp. 1-508.
[6]. European Commission, 'The EU Data Protection Reform and Big Data: Factsheet', March 2016.
[7]. Fayyad, U., Piatetsky, G., and Padharic Smyth, (1996). From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U., Piatetsky, G., and Padharic Smyth. Advances in Knowledge discovery and Data Mining. AAAI Press/ The MIT Press, pp.1-34.
[8]. International Data Corporation (IDC) Research: 'Worldwide Big Data Technology and Services Forecast, 2015–2019', October 2015.
[9]. McAfee, A., & Brynjolfsson, E. (2012, October). Big Data: The Management Revolution. HBR, 3-9.
[10]. Provost, F., & Fawcett, T. (2013). Data Science for Business. CA: O'Reilly.
[11]. Rai, A. (2016). Synergies Between Big Data and Theory. MIS Quarterly, 40 (2), iii-ix.
[12]. Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016/L 119/1, 27 April 2016.
[13]. Russom, P. (2011), Big Data Analytics. In TDWI Best Practices Report, pp. 1-40.
[14]. TechAmerica (2012), Demystifying Big Data: A Practical Guide to Transforming the Business of Government. In TechAmerica Reports, pp. 1-40.

"Data is the new oil. No, data is the new soil"
David McCandless