Linnæus University
Sweden

Title

Subtitle

What different kinds of Big Data might be there? What kind data does Supervised Machine Learning most rely on?

Name: **BABAK RAHIMI**
Date: 28th October 2020
Course: Contemporary Issues in IS Research & Development
Course Code: 5IK502, 7.5 credits
Department of Informatics

Table of Contents

1. Introduction

It is evident that we are in the era where vast volume of information and data is generated, gathered and available. As a result, the term “Big Data” is not only appearing in all sort of media, but has become the center of attention for many scientists, authors and industry people. Consequently, the need for data processing and instructing computers to use data with the aim of problem solving has become the core objective of Machine Learning approaches.

In this write up, after introducing different kinds of big data including Structured, Semi-structured, and Unstructured, the challenges and issues associated with big data is outlined looking at five characteristics of big data. Before the concluding remarks, the type of data that Supervised Machine Learning relies on as well as challenges linked to this method are briefly discussed.

2. Defining Big Data

Big Data can be defined as a collection of datasets that are large in size and volume. TechAmedica Foundation (2012) defines “Big data as a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information” (cited in Gandomi and Haider 2014, p. 138).

It is also need to be mentioned that data can be gathered from various sources such as Internet Data from social media for example or Open Data, for instance data published by government. Regardless of the source, this large volume of data is generated in *Structured*, *Semi-structured* or *Unstructured* form. As such, SAS defines Big Data “as a popular term used to describe the exponential growth, availability and use of information, both structured and unstructured” (cited in Michalik et al. 2014, p. 332). According to Fan and Bifet (2013), there are some characteristics and features of big data including *Volume*, *Velocity*, *Variety*, *Variability*, and *Value* (Figure 1), which will be looked at in the following paragraphs to critically analyze the challenges of big data in different domains.

3. Different Kind of Big Data

As mentioned earlier, big data can be categorized into three different forms namely structured, semi-structured and unstructured. On the one hand, it is argued that a large set and component of data is considered unstructured (Kwon et al. 2014). For example, data captured in forms of audio, images, or text. On the other hand, only a small portion of data is categorized under structured or semi-structured data format. In the following paragraphs different types of data based on the ways it captured will be outlined;

3.1. Structured and Semi-Structured Data

Structured and Semi-Structured Data are gathered, analyzed, and stored in a stable and secure layout (Baars and Kemper, 2008). Such datasets are well organized and accessible within the database.

Structured data can be *Created* through survey or registration forms (Wong, 2012). For instance, business data gathered through market research are generated purposely for marketing analysis or business intelligence. This can be achieved by conducting customer surveys traditionally, or loyalty programs occasionally.

Provoked data is another type of structured data captured through providing individual the opportunity to reflect on their opinion (Koutroumpis and Leiponen, 2013). Data collected through rating given by customers in restaurants or through feedback system in companies with regards to office condition that employees have are examples of this type of structured data.

Transacted or *Compiled* data, which is recorded from business transactions or online shopping for example, or cash register when customers make payment in person at the cashier counter are another sort of structured data collection method. The American company Acxion which, collecs information on credit scores, location or demographics is another example of compiled data collection method (Koutroumpis and Leiponen, 2013).

Structured data can be *Experimental* as well, where the information and data are assembled through business experiments in various marketing segments, produces or suppliers as well as clinical trial data for instance.

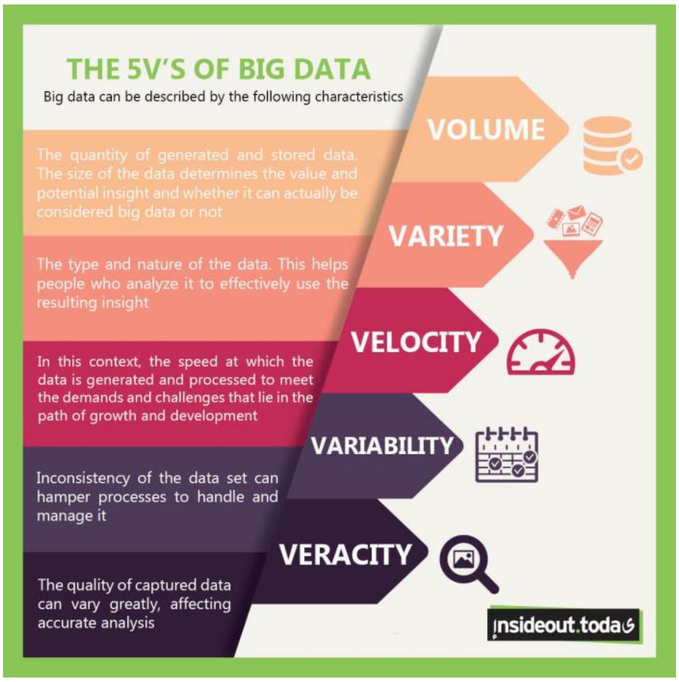
THE 5V'S OF BIG DATA
Big data can be described by the following characteristics
VOLUME
The quantity of generated and stored data. The size of the data determines the value and potential insight and whether it can actually be considered big data or not
VARIETY
The type and nature of the data. This helps people who analyze it to effectively use the resulting insight
VELOCITY
In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development
VARIABILITY
Inconsistency of the data set can hamper processes to handle and manage it
VERACITY
The quality of captured data can vary greatly, affecting accurate analysis
insideout.today

Figure 1: The 5V's of Big Data (Adel, 2017)

3.2. Unstructured Data

In the absence of organized, systematic or structured way in which the dataset is presented, it is referred to the data being unstructured (Kumar et al. 2014). Such unstructured data is difficult to be managed, analyzed and presented. A very good example of unstructured data is the one generated from e-mails, which is in various format including text, video, audio photo, and such.

Captured data is one type of unstructured data, which is generated inactively as a results of individual behavior. Examples are data captured through Google search, GPS information on smart phones or data gathered through RFID technologies.

User-generated data is another type of unstructured data (Chen and Zhang, 2014), which is created by individual inserting information in different internet platforms such as Facebook, Tweeter, commenting on news stories or Youtube posts.

4. Big Data Opportunities and Challenges

The prospective and forthcoming benefits and opportunities that the big data offers can differ from one sector to another. As such, big data and data analytics might offer different level of benefit and value to finance, insurance or banking industry to compare with government to to the accissibility of data and information. According to Yin and Kaynak (2015), higher customer satisfaction, operational optimization, effective financial risk analysis, and new product and business model developments can be achieved through effective use of big data and data analytics. Other opportunities such as availability of data in real-time (Bhimani and Willcocks, 2014), improved decision support for senior management (Davenport, 2014), improved operations (Wimmes et al. 2015) and strategic planning (Davenport, 2014) offered through use of big data can eventually lead to effective resource utilization, organizational performance and maximized profitability. Nonetheless, there are challenges and issues associated with big data and in this write-up some of these problems are elaborated by exploring the 5Vs characteristics of Big Data, as some of the challenges liaise in attributes linked to satasets (Figure 1).

4.1. Volume

This characteristic of big data is referred to the size of data that is generated from various sources (Chen et al. 2012). Depending on which source the datasets created, the volume of data may differ and require different data management techniques. Due to the large size and continuity as well as complexity of unstructured data, the storage and data analysis become a challenge leading to difficulties associated with big data management.

According to McAfee and Brynjlfsson (2012), the amount of data exchanged every second in internet equals to the total volume of data stored online 20 Years ago. The lack of managing such large volume of data can eventually lead to security and privacy concerns linked to big data. This means datasets may include personal and private information about people and government which needs to be protected through proper data management system and adoption of appropriate techniques. Moreover, there is a raising concern on data protection and individual awareness of what companies and organizations doing with personal data, sometimes without people consent.

Another challenge related to managing the large volume of data is the high cost involved to adopt technologies and necessary infrastructure to store, analyze and ensure the security and data protection from treat of unauthorized users and system attackers.

4.2. Velocity

Velocity refers the speed of data created, processed, stored and analyzed from different and all sources (Ishwarappa and Anuradha, 2015). An example is the amount and speed of data exchanged through e-mail or uploaded hours of videos in Youtube. The challenge with data velocity is related to the speed at which the data requires to be analyzed and act upon using relevant technologies in real-time.

Effectiveness of data and information sharing among and between data sources can be a challenge not only due to the high speed of data processed and analyzed, but possibility of different data warehouses or entities being reluctant to share information with one another. This problem can also exist within a company between different departments.

The fact that datasets are updated quickly and faster than the databases performance, it is difficult to analyze data at consistent level of quality and accuracy where data crunching and visualization can be done with the aim of extracting new knowledge based on the real-time data. Additionally, with regards to the real-time data processing and analysis, the lack of availability of right people with relevant skills and computational background, who might be difficult to find or maybe from certain gender can be another challenge.

4.3. Variety

When characteristics such as different types of data are taken into account, it is referred to variety of big data. In many cases big data is not structured making it difficult to be stored and organized in database in relational manner (Pantelis and Aija, 2013). The variety of big data being unstructured significantly increases the complexity of both recording and processing of datasets. This in turn will require people dealing with big data to spend considerable amount of time to clean the data before processing and analyzing.

The challenge with regards to big data variety is that usually the datasets are generated from various sources such as internet, smartphones, social media or Global Positioning System (GPS) in many different formats. Such datasets need to be sorted, managed and classified into structured format using advance data management systems and analytical tools, which can be expensive, not accessible or require highly skilled experts. In addition, there are always a challenge with ensuring the right question being asked from right people when collecting from different sources.

4.4. Variability

The next feature of big data is its variability and refers to the way in which the structure and meaning of data almost always changes (Chen et al. 2012). Ishwarappa and Anuradha (2015, p. 1171) argue that “when dealing with high volume, velocity and variety of data, it is impossible that all of the data is going to be 100% correct”. The high veracity of data can directly or indirectly influence the quality of the data generated, which can of course be affected due to unreliable, incomplete and uncertain environment and sources that data is extracted from.

The necessity to assess data veracity and verify captured data to ensure the data quality is one of the challenges of big data variability. Datasets gathered by different people and sorted in various databases is usually stored in an unstructured format (Bendler et al. 2014). The uncertain data and datasets that cannot be trusted will lead to issues with data quality and consequently result in faulty and wrong data analytics outputs and reports.

The critical challenge with data quality upon analyzing uncertain variety of datasets received from different sources can have undesirable impact on how data and information interpreted and commonly leads to misinterpretation of results and false decisions. The identification of right technology, known as visualization techniques, and relevant data processing platforms to ensure data quality and make data meaningful will help to overcome such obstacles.

4.5. Value

According to Al Nuaimi et al. (2015, p. 4), “Big data value refers to possible advantage big data can offer a business based on “good” big data collection, “management” and analysis”. Although the value of big data is evident, but collection of good data, which can turn into desired value would be a challenge. As the quality of dataset is important for adequate reporting and effective top management decision support, businesses are required to invest heavily in information technology infrastructure and systems that are usually costly. Besides reasons such as cost as well as justification of return on investment, most of businesses and managers tend to show resistance to implement relevant infrastructure as validating useful and reliable data from various datasets is challenging.

It was mentioned earlier that for gaining the maximum benefit from big data, availability of people with relevant skills and expertise is essential. Such talents of course can be developed within the organizations and businesses by investing on employee training and upgrading their knowledge and skills(Michalik et al. 2014). However, the challenge is that considering the large amount of investment and time on such infrastructure and talent development that companies need to allocate, only rich businesses are having the privilege to benefit from the big data value and those who do not have the capital or resources are less likely to gain value from the big data advantages offered.

5. Supervised Machine Learning

Machine Learning which is an application of Artificial Intelligence (AI) is defined by Mitchell (1997) as “a branch of computer science that aims to learn from data in order to improve performance at various tasks” (Cited in Jiang et al. 2020, p. 675). Machine Learning has influenced different applications such as image and speech processing or Internet of Things (IoT). ML which includes processing of data, learning and evaluation of final output can be categorized into Unsupervised Learning, Reinforced Learning and Supervised Learning (Russell and Norvig, 2010). When the system does not provide the desired output and the aim is to find input patterns, it is referred to *Unsupervised Machine Learning*. Parveena and Jaiganesh (2017, p. 32) argue that “in the *Reinforced Machine Learning* the algorithm learns to react to an environment” (Figure 2).

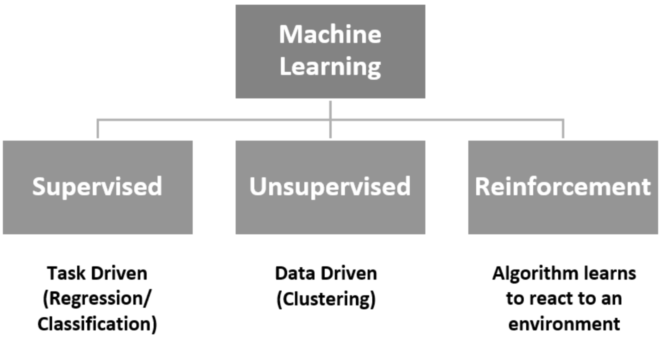
Machine Learning
Supervised
Unsupervised
Reinforcement
Task Driven (Regression/ Classification)
Data Driven (Clustering)
Algorithm learns to react to an environment

Figure 2. Machine Learning and its Types
(Parveena and Jaiganesh 2017, p. 32)

Supervised Machine Learning is a task driven process which labelled datasets are used to process meaningful outcome (Parveena and Jaiganesh, 2017). Therefore, it relies on structured and established dataset collection where the required data are identified and pre-processed. As the structured data are used in supervised machine learning, the input can be adjusted if the ultimate performance is not satisfactory. According to Wuest et al. (2016), Supervised machine learning also relies on data which is labeled. The specific dataset used for supervised learning (trained dataset) aim to address a defined and explicit problem with expected quality outcome. Further, this type of machine learning highly depends on knowledgeable external supervisor that can provide constant expert feedback to handle and process data, deliver necessary training and effectively test dataset so that the targeted task is performed.

5.1. Supervised Machine Learning Opportunists and Challenges

In general, supervised machine learning has been successfully employed in different processes including data analysis, outcome prediction, optimization, control application and many more. For instance, this type of machine learning technique is used to improve quality control in manufacturing systems, where the environment is complex at certain level (Harding et al. 2006). Supervised machine learning provides the ability to identify implicit relationship among large and complex dataset as well as providing solution for high dimensional problems to understand a particular domain. By determining unknown knowledge, supervised machine learning provides the opportunity to learn from dynamic systems and adapt to the changing environment. According to Nilsson (2005), the adoption of supervised machine learning offers a reasonable fast data analysis within almost all domains to compare with traditional techniques.

Nevertheless, there are number of challenges associated with supervised machine learning. The first challenge is with regards to *incomplete* supervision. The effectiveness of supervised machine learning is limited in situation where smaller portion of datasets are labeled to compare with unlabeled (Zhou, 2018). Insufficient quantity of labeled data leads to incomplete supervision and incomplete input value, which this eventually leads to inaccurate results generated.

Another challenge is associated with supervised machine learning is related to *inadequate pre-processing* of data. The lack of adequate preparation and pre-processing of data can have critical impact on the final results produced. According to Sun et al. (2015), the pre-processing of data can be difficult if the data is unlabeled, for instance data is redundant, inconsistent or noisy. In data preparation and pre-processing the lack of expert for data training can be another obstacle related to the supervised machine learning.

The other challenge linked to supervised machine learning is with regards to *inexact* supervision, which according to Zhou (2018) is a situation where the given data does not have the exact criteria required. Datasets that are not precise and have irrelevant input features can result in inaccurate output. To avoid inexact supervision, examination of where the data is generated from and appropriate data cleaning is necessary.

Inaccurate supervision, which deals with situation where information and data are not always true is another challenge linked to supervised machine learning. Frenay and Verleysen (2014, p. 855) state that “some labeled information may suffer from errors”. Data inaccuracy resulting the unexpected output can be due to many reasons including incorrect data capturing and storing in database, using faulty instruments, dealing with sensitive issues or ignoring human behavior and people desire factors.

Data privacy and *trust* is another challenge related to supervised machine learning. The example would be healthcare data which can be collected from multiple hospitals or entities that each may have various data protection policies. Provided that most of the time sharing data is necessary for operationalizing supervised machine learning, then they way in which distribution and sharing datasets taken place becomes a challenging problem.

6. Concluding Remarks

Although the notion of big data is nothing new, but the concept has been given major attention by both academia and industry. In this write up different kinds of big data including structured, semi-structured and unstructured were looked at and both opportunities and challenges involved in using data in different context critically discussed by looking at characteristics of big data. While big data provides opportunities such as operational optimization, availability of real-time data and resource utilization or improved performance, the fact that many challenging problems are linked to big data was also highlighted. It was argued that there are many concerns with data security, and privacy, problems with data and information sharing, issues associated with unstructured and unlabeled data as well as challenges with regards to the cost involved and management resistance.

It was also concluded that besides advantages that supervised machine learning has to offer including real-time data analysis, quality and control application, opportunities to learn from dynamic systems or ability to adapt with changing environment, adopting supervised machine learning can be challenging at the same time. Problems with incomplete, inexact, and inaccurate supervision, issues related to inadequate pre-processing of data or obstacles with data privacy and trust are some of the challenges discussed.

Therefore, there are many considerations to be taken into account when adopting supervised machine learning and making use of big data in any particular framework. Lack of big data training and expertise, sometimes inappropriate techniques to capture and store data, lack of proper policy to protect privacy of information and lack of supervision may lead to possibility of making wrong decisions based on information received from supervised machine learning and data analytics.

7. References

Adel, N. 2017. *How Open Data Can Leverage your Content Marketing Efforts; Big Data provides a new paradigm for Data-Driven marketing* [Online]. Insideout.toda. Available at: https://blog.insideout.io/en/tag/the-adoption-of-big-data/ [Accessed: 5 October 2020].

Al Nuaimi, E. et al. 2015. Applications of big data to smart cities. *Journal of Internet Services and Application* 6(25), pp. 1-15

Baars, H. and Kemper, H. 2008. Management Support with Structured and Unstructured Data – An Integrated Business Intelligence Framework. *Information Systems Management* 25, pp. 132-148.

Bendler, J. et al 2014. Taming uncertainty in Big Data: Evidence from Social Media in Urban Areas. *Business & Information Systems Engineering* 6(5), pp. 279-88.

Bhimani, A. and Willcocks, L. 2014. Digitisation ,Big Data and the transformation of accounting information. *Accounting and Business Research* 44(4), pp. 469-90.

Chen, C.P. and Zhang, C.Y. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275, pp. 314-47.

Chen, H. et al. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly* 36(4), pp. 1165-1188.

Davenport, T.H. 2014. *Big Data at Work: Chancen Erkennen, Risiken Verstehen*. Boston: Vahlen.

Fan, W. and Bifet, A. 2013. Mining big data: current status, and forecast to the future. *ACM SIGKDD Explor Newsl* 14(2), pp. 1-5.

Frenay, B. and Verleysen, M. 2014. Classification in the presence of label noise: a survey. *IEEE Trans Neural Network Learn Syst* 25, pp. 845-869.

Gandomi, A. and Haider, M. 2014. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35, pp. 137-144.

Harding, J. A. et al 2006. Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering* 128, pp. 969-976. doi:http://dx.doi.org/10.1115/1.2194554.

Ishwarappa, K. and Anuradha, J. 2015. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. *International Conference on Intelligent Computing, Communication & Convergence*. Interscience Institute of Management and Technology. 2015. India: Elsevier, pp. 319-324.

Jiang, T. et al 2020. Supervised Machine Learning: A Brief Primer. *Behavior Therapy* 51, pp. 675-687.

Koutroumpis. P. and Leiponen, A. 2013. Understanding the Value of (Big) Data. *IEEE Conference on Big Data*. October 2013, Silicon Valley, CA.

Kumar, R. et al. 2014. Apache Hadoop, NoSQL and NewSQL Solution for Big Data. *International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE)* 1(6), pp. 28-36.

Kwon, O. et al. 2014. Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management* 34(3), pp. 387-394.

McAfee, A. and Brynjolfsson, E. 2012. Big Data; The Management Revolution. *Harvard Business Review* 90(10), pp. 60-69.

Michalik, P. et al 2014. Concept definition for Big Data architecture in the education system. In Applied Machine Intelligence and Informatics (SAMI), *IEEE 12th International Symposium*, 2014. pp. 331–334.

Nilsson, N. J. 2005. Introduction to machine learning. *Theoretical Computer Science* 298, pp. 207-233. doi:http://dx.doi.org/10.1016/S0304-3975(02)00424-3.

Pantelis, K. and Aija, L. 2013. Understanding the value of (big) data. *2013 IEEE International Conference on Big Data*. 23 December, 2013. Silicon Valley, CA, USA: IEEE, pp. 38-42.

Parveen, M. and Jaiganesh, V. 2017. A Literature on Supervised Machine Learning Algorithms and Boosting Process. *International Journal of Computer Applications* 169(8), pp. 32-35.

Russell, S. and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. New Jersey: Prentice Hall.

Sun, S. et al. 2015. A review of Nyström methods for large-scale machine learning, *Inf. Fusion* 26, pp.36-48.

Wimmes, C. et al. 2015. Die Bedeutung von Big Data im Controlling – Eine empirische Studie. *Controlling*, 27(4/5), pp. 256-262.

Wong, P.C. et al 2012. The top 10 challenges in extreme-scale visual analytics. Computer Graphics and Applications. *IEEE* 32(4), pp. 63-67.

Wuest, T. et al 2016. Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research* 4(1), pp. 23-45.

Yin, S. and Kaynak, O. 2015. Bog Data for Modern Industry: Challenges and Trends. *IEEE*, 103(1), pp. 143-146.

Zhou, Z. 2018, A brief introduction to weakly supervised learning. *National Science Review* 5, pp. 44-53.

Lnu.se

Linnæus University
Sweden

Faculty of Technology
SE-391 82 Kalmar | SE-351 95 Växjö
Phone +46 (0)772-28 80 00
teknik@lnu.se
Lnu.se/fakulteten-for-teknik