Statistics and big data: Different perspectives

Cite as: AIP Conference Proceedings **2293**, 420108 (2020); https://doi.org/10.1063/5.0026847
Published Online: 25 November 2020

Sandra Nunes, Teresa A. Oliveira, and Amílcar Oliveira

Statistics and Big Data: Different Perspectives

Sandra Nunes[1, a)], Teresa A. Oliveira[2, b)] Amílcar Oliveira[2, c)]

[1]*School of Business Administration, Polytechnic Institute of Setúbal, Campus do IPS – Estefanilha, Setúbal, Portugal and Centre for Mathematics and Applications, Faculty of Sciences and Technology, New University of Lisbon, Portugal*

[2]*Universidade Aberta, Palácio Ceia, Rua da Escola Politécnica and Center of Statistics and Applications, University of Lisbon, Lisboa, Portugal*

[a)]Corresponding author: sandra.nunes@esce.ips.pt
[b)]Teresa.Oliveira@uab.pt
[c)]Amílcar.Oliveira@uab.pt

Abstract. Big Data has become the new slang in the world of information collection and analysis. The researches we conduct and the data we collect continue to grow, due to rapidly expansion of technology. Disciplines such as Computer Science, Engineering, and Statistics play a key role in the analysis of big data, each with its specificity but all equally important, an opinion that is not shared by all, being the Statistic considered the weakest link. This work attempts to show that Statistics have a distinct and essential role in this new world of Big Data, showing that Statistics and Big Data denote a crucial union. We will start with a brief introduction to Big Data and the several existing definitions.

INTRODUCTION

Every day we are bombarded with data. Most recently we witnessed an explosion of data, and it is nearly impossible to ignore the increasing and the potential use for Big Data. Our society is increasing in complexity. We are extremely mobile, technological advances are quickly changing the way we live, data users want more data in more detail. Digital transaction data are increasing substantially, online news and blogs are replacing newspapers, smart phone GPS data offer traffic data, e-commerce transactions provide signals as to what items cost, local governments are making data available for public access, internet use is growing exponentially, and the tendencies are not expected to reverse. No one can ignore this growing ocean of digital data mainly the statisticians. We can read in McKinsey Big Data report, 2011 that "*A significant constraint on realizing value from Big Data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning ... we project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions*".

Big data problems require multidisciplinary teams by their very nature. At the very least, they typically require theme area experts, computational experts, machine learning experts, data miners, and statisticians. According Roger Peng from Johns Hopkins School of Public Health, "*In Big Data, statistical sciences and domain sciences are more intertwined than ever before, and statistical methodology is absolutely critical to making inferences*".

Previously to Big Data development, corporations could not store all their archives for long periods nor efficiently manage massive data sets. Nowadays a large volume of data is generated at unparalleled rates. This is due to many technological trends, including the Internet of Things, the proliferation of the Cloud Computing (Botta et al., 2016). Behind the scene, powerful systems and distributed applications are supporting such multiple connections systems (e.g., smart grid systems (Chen et al., 2014), healthcare systems (Kankanhalli et al., 2016), retailing systems (Schmarzo, 2013), government systems (Stoianov et al., 2013), etc. Collecting, storing, processing and analysing this volume of data turns out to be increasingly challenging. Organizations that are able to overcome these challenges and

International Conference of Numerical Analysis and Applied Mathematics ICNAAM 2019
AIP Conf. Proc. 2293, 420108-1–420108-4; https://doi.org/10.1063/5.0026847
Published by AIP Publishing. 978-0-7354-4025-8/$30.00

extract business value from Big Data, will have substantial competitive advantages. Big Data is often seen as a "fancy word" for a more insightful data analyses, but Big Data is much more than a that. Big Data management requires significant resources, new methods and powerful technologies. Big Data require to clean, process, analyse, secure and provide a granular access to massive evolving data sets. Companies and industries are more aware that data analysis is increasingly becoming a vital factor to be competitive, to discover new insight, and to personalize services (Davenport et al., 2012). Big Data is sometimes highlighted as fundamental for productivity growth, innovation and customer relationship, benefiting business areas like healthcare, public sector, retail, manufacturing and modern cities (Manyika et al, 2011; Chen et al, 2014).

Countless projects in the Big Data area have led to the development of many new models, frameworks and technologies to provide more storage capacity, parallel processing and real time analysis of different heterogeneous sources. Additionally, new solutions have been established to ensure data confidentiality and security, and simultaneously offer more flexibility and performance. Moreover, the cost of most hardware storage and processing solutions is continuously dropping due to the sustainable technological advance (Purcell, 2013).

To extract knowledge from Big Data, numerous models, technologies, software and hardware have been designed, always with the purpose of ensuring more precise and reliable results for Big Data applications. One problem lies in choosing among the numerous new technologies. Actually, many parameters should be considered: technological compatibility, deployment complexity, cost, efficiency, performance, reliability, support and security risks. There exist many Big Data surveys in the literature but most of them tend to focus on algorithms and approaches used to process Big Data rather than technologies (Ali et al., 2016; Chen and Zhang, 2014; Chen et al., 2014).

The leap in Big Data speech to more popular outlets implies that a coherent understanding of the concept and its terminology is yet to develop. For example, there is little consensus about the fundamental question of how big the data has to be to qualify as "Big Data" (Gandomi and Haider, 2015).

DEFINING BIG DATA

There is no widely accepted threshold for which data becomes "Big Data". Ward and Barker (2013), attempt to clearly define Big Data by presenting several definitions highlighting that "Big Data" is predominantly and "anecdotally" associated with data storage and data analysis, terms dating back to distant times. They also argue that the adjective "Big" implies significance, complexity and challenge, but also makes difficult to quantitatively define "Big Data". In this work Ward and Baker presents several definitions, some defining Big Data by its characteristics, others based on the expansion of traditional data with more unstructured data sources, and others trying to quantify it. They also present definitions that rely on the inadequacy of traditional technologies to deal with this new type of data, presenting several perspectives from the industry. At the end Ward and Barker conclude that all definitions include at least one of the following aspects: size; complexity; and technologies to process huge and complex datasets. They conclude by stating that "*the concept of Big Data includes storage and analysis of large and complex datasets, using a set of novel techniques*".

According Dumbill (2013), "Big Data is data that exceeds the processing capacity of conventional *database systems. The data is too big, moves too fast, or does not fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it".* Chen et al. (2014) support this definition by focusing on the fact that traditional software and hardware cannot recognize, collect, manage or process this new type of data in reasonable time. Krishnan (2013) also agrees with these perspectives, defining Big Data by "*its complexity, speed and several degrees of ambiguity, whose processing is inadequate for traditional methods, algorithms and technologies*".

According Gandomi and Haider (2015), state that "*size is the characteristic that first stands out, but other characteristics have become usual to define Big Data*".

Several data scientists and experts define Big Data by the following three main characteristics, called the 3Vs, (Furht and Villanustre, 2016):

Volume - Large volumes of digital data are generated continuously from millions of devices and applications. According to McAfee and Brynjolfsson (2012), it is estimated that about 2.5 exabytes were generated each day in 2012. This amount is doubling every 40 months approximately. By 2015, digital data grew to 8 ZB (Rajaraman, 2016). According to IDC report, the volume of data will reach to 40 Zeta bytes by 2020 and increase of 400 times by now (Kune et al., 2016).

Velocity - Data are generated in a fast way and should be processed rapidly to extract useful information and relevant insights. For instance, YouTube is a good example that illustrates the fast speed of Big Data.

Variety: Big Data are generated from numerous sources and in multiple formats. Large data sets consist of structured and unstructured data, public or private, local or distant, shared or confidential, complete or incomplete, according Emani et al. (2015) and Gandomi and Haider (2015) more Vs and other characteristics have been added by some actors to better define Big Data: Vision (a purpose), Verification (processed data conform to some specifications), Validation (the purpose is fulfilled), Value (pertinent information can be extracted for many sectors), Complexity (it is difficult to organize and analyse Big data because of evolving data relationships) and Immutability (collected and stored Big data can be permanent if well managed).

WHY IS IT IMPORTANT FOR STATISTICS TO BE ONE OF THE KEY DISCIPLINES FOR BIG DATA?

Statistics is fundamental to ensuring meaningful and accurate information is extracted from Big Data. The following issues are crucial and are only intensified by Big Data:

- Data quality and missing data;
- Observational nature of data;
- Quantification of the uncertainty of predictions, forecasts and models.

Like in any data, through Big Data analysis we will find bias, false positives and uncertainty. Statistics brings sophisticated techniques and models to bear on these issues. Statisticians help translate the scientific question into a statistical question, which includes carefully describing data structure; the underlying system that generated the data (the model); and what we are trying to assess (the parameter or parameters we wish to estimate) or predict.

In a Working Group of the American Statistical Association (2014) entitled "Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society" we may read:

"*Big Data will often not be served well by "off the shelf" methods or black box computational tools that work in low-dimensional and less complicated settings, and therefore require tailored statistical methods. Statisticians are skilful at assessing and correcting for bias; measuring uncertainty; designing studies and sampling strategies; assessing the quality of data; enumerating limitations of studies; dealing with issues such as missing data and other sources of non-sampling error; developing models for the analysis of complex data structures; creating methods for causal inference and comparative effectiveness; eliminating redundant and uninformative variables; combining information from multiple sources; and determining effective data visualization techniques*".

And:

"*With a major Big Data objective of turning data into knowledge, statistics is an essential scientific discipline because of its sophisticated methods for statistical inference, prediction, quantification of uncertainty, and experimental design. Such methods have helped and will continue to enable researchers to make discoveries in science, government, and industry.*"

And yet:

"The age of Big Data will be a golden era for statistics. Scientific fields are transitioning from data-poor to data-rich and—across industries, science, and government—methods for making decisions are becoming more data-driven as large amounts of data are being harvested and stored. However, alone, data are not useful for knowledge discovery. Insight is required to distinguish meaningful signals from noise. The ability to explore data with skepticism is required to determine when systematic error is masquerading as a pattern of interest. The keys to such skeptical insight are rigorous data exploration, statistical inference, and the understanding of variability and uncertainty. These keys are the heart of statistics and remain to be used to their full potential in Big Data research".

There is no question Big Data has hit the business, government and scientific sectors and we have no doubt that statistical thinking will be essential to solve numerous Big Data challenges. Unfortunately, the role of statistics seems too often to be undervalued. Instead, computer science, applied math and other fields are frequently mentioned as the pertinent scientific discipline while statistics is often left out. Numerous papers concerning Big Data neglected to mention statistics as a discipline important in this area. Sometimes implicitly but often explicitly authors claimed in much of the Big Data literature that statistical thinking is no longer relevant in the petabyte age. We believe just the opposite. Fundamentals of good modelling and statistical thinking are crucial for the success of Big Data projects. Thorough statistical practices, such as ensuring high-quality data, incorporating thorough domain knowledge, and developing an overall strategy for large modelling problems, are even more important for Big Data problems than small data problems.

CONCLUSIONS

In this work we try to show a different perspective from the ones found in most papers related to Big Data. We try to show that statistical thinking is required when we work with Big Data and that multidisciplinary teams involving statisticians are crucial to solve problems in this area. In our opinion statistical thinking feeds the intersection of ideas between scientific fields (such as biological, physical, and social sciences), industry, and government. We believe that further engagement of statisticians and cutting-edge statistics (as one of the core data science disciplines) will help advance the aims of Big Data challenges.

ACKNOWLEDGMENTS

This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the projects UID/MAT/00297/2013 (CMA/NOVA.ID.FCT) and UID/MAT/00006/2019 (CEA/UL).

REFERENCES

1. Ali, A., Qadir, J., urRasool, R., urRasool, R., Sathiaseelan, A., Zwitter, A., Crowcroft, J. Big data for development: applications and techniques. Big Data Anal. 1, 2, 2016.
2. Botta, A., de Donato, W., Persico, V., Pescapé, A. Integration of cloud computing and internet of things: a survey. Future Gener. Comput. Syst., 2016 56, pp. 684–700.
3. Chen, M., Mao, S., and Liu, Y. "Big Data: A Survey," Mobile Networks and Applications, 2014, vol. 19, no. 2, pp. 171–209.
4. Chen, C.P., Zhang, C.-Y. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci., 2014, 275, pp. 314–347.
5. Costa, C. and Santos, M.Y. Big Data: State-of-the-art Concepts, Techniques, Technologies, Modeling Approaches and Research Challenges. IAENG International Journal of Computer Science, 2017, vol. 44, no. 3, pp. 285-301.
6. Davenport, T. H., Barth, P., and Bean, R. "How big data is different." MIT Sloan Management Review, 2012, vol. 54, no. 1, pp. 43–46.
7. Dumbill, E. "Making sense of big data," Big Data, 2013, vol. 1, no. 1, pp. 1–2.
8. Emani, C.K., Cullot, N., Nicolle, C. Understandable big data: a survey. Comput. Sci. Rev., 2015, 17, pp. 70–81.
9. Furht, B., Villanustre, F. Introduction to big data. In: Big Data Technol. App. Springer International Publishing, Cham, 2016, pp. 3–11.
10. Gandomi, A. and Haider, M. "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, 2015, vol. 35, no. 2, pp. 137–144.
11. Google Trends, "Interest in Big Data over time", 2016. Available: https://www.google.pt/trends/explore#q=big%20data.
12. Kankanhalli, A., Hahn, J., Tan, S., Gao, G. Big data and analytics in healthcare: introduction to the special section. Inf. Syst. Front., 2016, 18, pp. 233–235.
13. Krishnan, K. Data Warehousing in the Age of Big Data, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2013.
14. Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R., Buyya, R. The anatomy of big data computing. Software: Pract. Experience, 2016, 46, 79–105.
15. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh C. and Byers, A.H. Big Data: The next frontier for innovation, competition and productivity. Mckinsey Global Institute, 2011.
16. McAfee, A. and Brynjolfsson, E. Big Data: the management revolution. Harvard Bus. Rev., 2012 90, pp. 60–68.
17. Oussous, A., Benjelloun, F., Lahcen, A. and Belfkih, S. Big Data technologies: A survey. Journal of King Saud University – Computer and Information Sciences, 2017.
18. Purcell, B.M. Big Data using cloud computing. Holy Family Univ. J. Technol. Res, 2013.
19. Rajaraman, V. Big data analytics. Resonance, 2016 21, pp. 695–716.
20. Rudin, C., Dunson, D., Irizarry, R., Ji, H., Laber, E., Leek, J., McCormick, T., Rose, S., Schafer, C., van der Laan, M., Wasserman, L. and Xue, L. Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society. A Working Group of the American Statistical Association, 2014. (www.amstat.org/asa/files/pdfs/POL-BigDataStatisticsJune2014.pdf).
21. Schmarzo, B. Big Data: Understanding How Data Powers Big Business. John Wiley & Sons, 2013.
22. Stoianov, N., Urueña, M., Niemiec, M., Machnik, P., Maestro, G. Integrated security infrastructures for law enforcement agencies. Multimedia Tools App.,2013, pp. 1–16.
23. Ward, J. S. and A. Barker, A. "Undefined by Data: A Survey of Big Data Definitions," arXiv:1309.5821 [cs.DB], 2013.