Data Science in the Big Data era

Matrícula: 20171003230

Index Terms—**Data science, big data.**

I. Introduction

THERE is no unanimous definition about the term "Data Science". In 1997, C. F. Jeff Wu said that Data Science could be considered the natural evolution of Statistics [1], although already recognize that statistics (or data science) can be described by: (i) data collection; (ii) data modeling and analysis; and (iii) problem understanding and decision making. William S. Cleveland, in 2001, defend that the view of data science to be practiced only by the statisticians is very limited [2], because the data has to be analyzed with the support of a theory, so data science is a far more larger field, including several disciplines. Finally, the Gartner IT Glossary says that the role of the data scientist requires a broad combination of skills to understand the business issues, to discover relationships within data and detect patterns for decision making and to build the relevant dataset used for the analysis [3]. So, even though there is no unique definition of Data Science, it can be simplified by the science of analyze the data an extract information of it, and this can be done by several ways.

With this simplified definition is true to say that data science exist for a very long time, at least since the census in the ancient Rome, as it was used to determine taxes based on person's properties. So, why this field of science became so popular nowadays? This can be answered because we are living in the Big Data era. The world never produced so many data per day, is hard to say how much data are in fact produced, but the 2,5 exabytes per day is going around since 2013 [4][5][6]. Also, the amount of data is set to rise from 4.4 zettabytes in 2013 to 44 zettabytes by 2020 [5]. This grow is very feasible as we are seeing the development of new technologies and almost all of then involves the creation or manipulation of very large amount of data.

The Large Synoptic Survey Telescope will acquire 140 terabytes of information every five days, this was the same amount that the Sloan Digital Sky Survey took to archive in one decade [7]. One single autonomous car is expected to generate approximately 4 terabytes of data a day [8], estimating about 1 million autonomous vehicles worldwide gives 4 exabytes of data per day. For conclude the examples of the emerging technologies, we have the Internet of Things growing rapidly. Gartner said that in 2015, 4,9 billions "things" were connected to the Internet and estimated that this number will reach 20,8 billions in 2020 [9], and its obvious to say that each "thing" usually create some data.

Then, its easy to state that we are already living in the Big Data era and the days to come will ensure that. Big Data is often described by the following characteristics[10][11][12], as we are putting it as challenges:

Volume - Huge amount of data. Implies the ability to process these data.
Variety - Deal with different forms of data and sometimes with no structure at all.
Velocity - Real time data are often found. In this case, the data must be processed and sometimes understood fast enough to meet the real time demand.
Veracity - Big data needs to handle with "noise" found in the middle of the data or even with intentionally false data.

Considering the availability of so much data, the world found in the Data Science a way for improve his Business Intelligence, helping them to make better decisions. It realized that, discovering patterns could lead to clarify social demands that the market couldn't see yet. Or even predict some kind of behavior, leading one to be ahead of his concurrents. That's why in 2010 this business was estimated to be worth more than $100 billions and growing at almost 10% a year, twice as fast as the software business as a whole [7]. That's why Walmart is building the world's biggest private cloud, to process 2,5 petabytes of data every hour [13]. That's why Netflix thinks his recommendation engine worth $1 billion a year [14]. And that's was how the Target company could predict that a woman was pregnant [15]. But even thought this much was already done with the help of Data Science, it needs to be coined that much have do be developed/improves, specially because most of these technologies are expanding and is uncertainly what kind and how much data they will actually provide. So, Data Science have to grow along side with these technologies.

In special, it can be noted that most of the examples of this document is about companies using the Data Science to improve his results, therefore his profits. But, with Data Science, the governments can do much to benefit the population. Lets take the three main aspects of a society: health care, education and security. For the first one, the medical history of a person could be monitored to predict the development of a illness, and act to prevent that, or to right diagnosis some disease. In [16] it can be seen much more examples and initiatives that are already in progress. In education, it could be analyzed the aspects of the personal life of the students and find some patterns on his success or failure in his academics results, with that information, the government could elaborate public campaigns to maximize the success. As an example, research's in Pennsylvania did that [17]. At last, at public security, it could be studied the crimes data and see what kind of crime is likely to occur in a specific range of time and location, so it could manage the police officers in a more efficient manner. This is done by Los Angeles Police Department, and according to [18] is a success in reducing the crimes. And that's is how Data Science can grow in importance in the world.

References

[1] C. F. J. Wu, “Statistics = Data Science?” Ann Arbor, p. 13, 1997, accessed:2017-08-18. [Online]. Available: http://www2.isye.gatech.edu/ jeffwu/presentations/datascience.pdf

[2] W. S. Cleveland, “Data science: an action plan for expanding the technical areas of the field of statistics,” *International statistical review*, vol. 69, no. 1, pp. 21–26, 2001.

[3] Gartner, “Data Scientist - Gartner IT Glossary,” accessed:2017-08-18. [Online]. Available: http://www.gartner.com/it-glossary/data-scientist/

[4] IBM, “2.5 quintillion bytes of data created every day. How does CPG & Retail manage it? - IBM Consumer Products Industry Blog,” 2013, accessed:2017-08-18. [Online]. Available: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/

[5] Northeastern University, “How Much Data is Produced Every Day? - Level Blog,” 2016, accessed:2017-08-18. [Online]. Available: http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/

[6] Vcloudnews, “Every Day Big Data Statistics – 2.5 Quintillion Bytes of Data Created Daily,” 2015, accessed:2017-08-18. [Online]. Available: http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/

[7] The Economist, “Data, data everywhere | The Economist,” feb 2010. [Online]. Available: http://www.economist.com/node/15557443

[8] B. Krzanich, “Data is the new oil in the future of automated driving,” *Online: https://newsroom. intel. com/editorials/krzanich-the-future-of-automateddriving, access*, vol. 14, 2017.

[9] R. van der Meulen, “Gartner Says 6.4 Billion Connected "Things" Will Be in Use in 2016, Up 30 Percent From 2015,” http://www.gartner.com/newsroom/id/3165317, 2017, accessed:2017-08-18.

[10] IBM, “Big Data Analytics | IBM Analytics,” accessed:2017-08-18. [Online]. Available: https://www.ibm.com/analytics/us/en/big-data/

[11] M. Beyer, “Gartner says solving’big data’challenge involves more than just managing volumes of data,” *Gartner. Archived from the original on*, vol. 10, 2011.

[12] M. Hilbert, “Big data for development: A review of promises and challenges,” *Development Policy Review*, vol. 34, no. 1, pp. 135–174, 2016.

[13] Forbes, “Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud,” 2017, accessed:2017-08-18. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/#6eb5b0b6c105

[14] C. A. Gomez-Uribe and N. Hunt, “The netflix recommender system: Algorithms, business value, and innovation,” *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 4, p. 13, 2016.

[15] C. Duhigg, “How companies learn your secrets,” *The New York Times*, vol. 16, p. 2012, 2012.

[16] T. O’Reilly, M. Loukides, J. Steele, and C. Hill, *How data science is transforming health care*. " O’Reilly Media, Inc.", 2012.

[17] D. F. Perkins, A. K. Syvertsen, C. Mincemoyer, S. M. Chilenski, J. R. Olson, E. Berrena, M. Greenberg, and R. Spoth, “Thriving in school: The role of sixth-grade adolescent–parent–school relationships in predicting eighth-grade academic outcomes,” *Youth & society*, vol. 48, no. 6, pp. 739–762, 2016.

[18] M. Van Rijmenam, “The los angeles police department is predicting and fighting crime with big data,” *DataFloq. com*, 2015.