# BIG DATA LITERACY

Karen François[1] and Carlos Monteiro[2]
[1]Vrije Universiteit Brussel / Free University Brussels (VUB) - Belgium
[2]Universidade Federal de Pernambuco / Federal University of Pernambuco (UFPE) – Brazil
Karen.Francois@vub.be

*In the contemporary society a massive amount of data is generated continuously by various means, and they are called Big Data sets. Big Data has potential and limits which need to be understood by statistics consumers, therefore it is a challenge to develop Big Data Literacy to support the needs of constructive, concerned, and reflective citizens. However, the parallel developments of the concepts of statistical and mathematical literacy mirror the current gap between purely technical and socio-political characterizations of Big Data. In this paper, we review the recent history of the concepts of mathematical and statistical literacy and we will highlight the need to integrate the new challenges and critical issues from data-science associated with Big Data, including e.g. ethics, integrity, misconduct, agency, and mathwashing.*

## INTRODUCTION

The aim of this paper is to contribute to the construction of the idea of Big Data Literacy. The discussion is based on a literature review, on our previous research in mathematics and statistic education (François, Monteiro & Vanhoof, 2013; Francois, Monteiro, Carvalho, & Vandendriessche, 2015; Queiroz, Monteiro, Carvalho & François, 2017), and on the educational experience with the organization of a graduate training-network for methodology and statistics (FLAMES—Flanders training network for Methodology and Statistics (http://www.flames-statistics.eu/ ). Therefore, in the following sections we review the recent history of the concept of statistical literacy, and highlight how the meaning of the concept shifted from the basic need to understand and be able to apply statistical techniques, to a broader conception that is explicitly connected with ethical and political aspects. This broader concept includes the skills citizens need to interpret and criticize statistical information and statistical reasoning. Interestingly, the more general concept of mathematical literacy—as used in the context of PISA (OECD, 1999, 2010)—underwent a similar evolution, and is now customarily related to the needs of constructive, concerned, and reflective citizens. These parallel evolutions mirror the current gap between purely technical and socio-political characterizations of Big Data.

In the next section we present different aspects to characterize Big Data before we will analyze the new challenges and the critical issues.

## WHAT IS BIG DATA ABOUT

Caldas and Silva (2016) explain that Big Data comes on the basis of information that is generated continuously by various means and becomes a wide range of jumbled data, but that can be analyzed, processed and used in the solution of various problems. For instance, huge datasets are generated by social media users who share information, write their opinions, and contact friends. Big data also are produced from a quite wide range of social situations, which include web scraping, mobile-phone use, online shopping, and banking transactions.

Big Data allows the use of different management technologies related to the generation of data ranging from paid packages and specific data analysis software, as well as, for example, available tools in social networks. These explorations are also associated with the analytic process of *data mining* that is the search for consistent patterns and/or systematic relationships between variables to validate them by applying the detected patterns to new subsets of data (Caldas and Silva, 2016). Generally, data mining is utilized as part of business, market or scientific research, thus favoring data-based decision-making.

Big Data is often characterized from a technical point of view by the *three V's*: volume - high amount of data; velocity - high speed of data in and out; and variety - great range of data types and sources (Laney, 2001). However, Rieder and Simon (2016) argue that it is important to consider the complexity of Big Data, because it involves not only technological aspects, but also scientific and cultural factors. In this sense, boyd and Crawford (2012) argue that Big Data is not only a social phenomenon related to new technological process, but it is also influenced by widespread belief that large data sets can offer a higher form of knowledge which provides insights with the aura of truth, objectivity, and accuracy.

Zeelenberg & Braaksma (2017) state that Big Data can be classified as large datasets related to social activities which are not covered by official statistics. Therefore, Big Data comes from sources in which the populations are not well-defined as well as from sources based on surveys, census or administrative data. These authors also emphasize that Big Data may be highly volatile and selective because the population to which it refers may change from day to another, which produce non- regular time-series. Another very frequent issue is associated with the fact that some Big Data sets do not have linking variables which would allow to be related to other datasets or population frames. These limitations might increase the possibility of error on the statistical results, although it important to minimize possible bias. For example, Big Data can be combined with data from sources which utilize more standardized statistical methods.

Therefore, in order to use Big Data it is necessary to evaluate its quality under certain principles. For instance, regulations from the European Union (2009) prescribe that it is necessary to evaluate the statistics according to some fundamental ideas, such as: impartiality, objectivity, reliability, relevance, accuracy, timeliness, accessibility, comparability, and coherence.

The official and academic statistics databases continue to play important roles; however, it can be predicted that in the near future many new Big Data sources will be available at higher speeds. The statistics educators need to be aware about the limits and opportunities in utilize those emerging Big Data sets.

If we consider that statistics education need to include new perspectives and uses of data, the idea of Big Data should be approached at school. However, in school contexts, concerns for Big Data are not based on how to manage a lot of data, but in finding ways to teach students on how to deal with this data. Ainley, Gould and Pratt (2015: 409) state that "data are not big because of the size of the data file, but because they belong to a new class of data that differ in structure and source from traditional data that have inspired institutional changes in how we learn from data". These authors also suggest that it is important in the era of Big Data to understand the limits of data if they learn concepts associated with representativity of sample, the relationship between sample and population, and the sampling variability. Therefore, in order to help students learn from Big Data sources, educators have an important role to choose most relevant strategies and contents.

Ben-Zvi and Friedlander (1996) argue that the large databases are generally used to represent real-world situations, but they are difficult to handle without a technological basis. In order to approach Big Data in school settings, educators can utilize technological artifact with a more accessible language, enabling students to experience tools that lead them to explore data processing, and to experience the use of artifacts that lead them to think about data.

## LITERATURE REVIEW OF STATISTICAL AND MATHEMATICAL LITERACY

In this section we investigate the development of the concept of statistical literacy based on a literature review, arguing that the concept developed from a rather small and technical perspective in the late seventies of the twentieth century. The concept was broadened by the American Statistical Association (ASA) in the late nineties, now also emphasizing the critical aspect of statistical literacy –besides its technical components. Research from the first decade of the twentieth century confirmed this broader description of the concept and they investigated its different layers. Statistical literacy became a broad and complex concept that tries to cover knowledge elements, dispositional elements and societal responsibilities (Gal, 2002; 2004).

Initially the term *statistical literacy* was used to describe the knowledge that people need to technically understand statistics, and to make decisions based on the analysis of data. These technical aspects of the concept were studied by Haack (1979) who analysed the concept in a technical way, based on what people need to deal with statistics. He considered certain technical aspects which include the source, the type of data, the definition and measurement problems, and finally certain considerations concerning the survey sample. From this description it is clear that only the technical dimension of the concept was emphasized.

The meaning of the concept was broadened by the ASA, the world's largest community of statisticians and a representative organization of researchers in the field of statistics and statistics education. Wallman (1993) states in her presidential address to the ASA "statistical literacy is the ability to understand and critically evaluate statistical results that permeate our daily lives –coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions." (Wallman, 1993: 1).

The concept was further developed by Watson (1997) who presented a sophisticated framework of statistical literacy comprised of three tiers: (i) a technical one, (ii) a societal one and (iii) a critical one. The technical one comprises the basis understanding of statistical terminology; the societal one is related to the understanding of statistics when embedded in a wider and societal context; and the critical tier relates to the questioning of claims. This is the highest level of statistical thinking, if one can challenge statistical information in cases that claims are made without proper statistical foundation.

Gal (2002, 2004) elaborated further the concept of statistical literacy by presenting a model that comprises both knowledge elements (or cognitive elements) and a cluster of supporting dispositional elements (critical stance, and beliefs and attitudes) (Gal, 2002: 4). He portrayed statistical literacy as "the ability to interpret, critically evaluate, and communicate about statistical information and messages" (Gal, 2002: 1). This is a key ability expected of all citizens in an information-laden society.

More recently, Garfield & Ben-Zvi (2008) described the connection between the purely technical aspects of statistics and its ethical-political project. They distinguish between: (i) statistical literacy; (ii) statistical reasoning; and (iii) statistical thinking. In this model, statistical literacy provides the foundation for reasoning and thinking. Statistical knowledge or *literacy* makes it possible to reason with statistical ideas and to make sense of statistical information. To connect one concept to another and to combine ideas about data and chance is called statistical *reasoning*. The final stage of statistical *thinking* includes a deep understanding of the theories underlying statistical processes and methods. It also includes the critical competence of understanding the constraints and limitations of statistics and statistical inferences. This stage of statistical thinking is called "the normative use of statistical models" by Garfield & Ben-Zvi (2008), emphasizing that values are at work here.

Looking at the shifted meaning of statistical literacy we can observe an evolution from a pure technical meaning of the concept to a broader meaning including critical, ethical-political aspects (François, Monteiro & Vanhoof, 2013).

If we compare this sophisticated framework and description of statistical literacy with the concept of mathematical literacy one can observe a quite similar attention to the societal and critical attitude. Already at the late nineties the concept was developed as follows. "Mathematical literacy is an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen" (OECD 1999). The description was developed by the OECD in 1999 and applied to the international comparative survey PISA 2003. In 2010 the concept was reformulated to emphasize the different components and competences of mathematical literacy (e.g. concepts, procedures, facts, tools; and formulate, employ, interpret, …). The new concept was applied to the international comparative survey PISA 2012, and is defined as follows: "Mathematical literacy is an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain, and predict phenomena. It assists individuals to recognize the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens." (OECD 2010)

In the following section we evaluate what we can learn from both concepts of literacy and how they developed from a pure technical meaning to a broader meaning including critical, ethical, and political aspects.

## NEW CHALLENGES AND CRITICAL ISSUES

One thing we can learn from the overview of the development of the concepts of statistical and mathematical literacy is at least that they both evolved from a pure technical description to the implementation of a social and critical perspective.

The same layers can be observed when it comes to sophisticated algorithms. Barocas, Hood & Ziewitz (2013) suggest that algorithms should be studied from an interdisciplinary perspective considering four layers. (i) The first could be a technical approach that studies algorithms as computer sciences. (ii) The second one could be a sociological approach that studies the interaction among programmers and designers. (iii) A third one could be a legal approach that studies mathematical algorithms as a figure and as an agent in law. (iv) Finally, a fourth approach could be a philosophical one that studies the ethics of algorithms.

Algorithms and Big Data are new forms of data that are changing in a rapid and radical ways. Practitioners advocating technological change tend to have an optimistic belief in the rationalizing force of Big Data. Digital technologies and the use of algorithms are said to be value-free and thus more objective in helping people making rational decisions. In contrast to human beings, algorithms seem to be free from a variety of social factors like gender, race, class, politics, etc. They seem to have the power to analyses data in a most accurate way and maximize the amount of variance explained by the models. Therefore, Big Data are often described as the cure for inefficient, biased and discriminatory systems we had to deal with for a long time (Barocas, Hood & Ziewitz (2013). Moreover, technologists, producers and reporters tend to use the power of mathematics and mathematical algorithms to paper over more subjective choices that are made during the processing of Big Data.

If we consider the other layers of the study of algorithms (ii to iv) -besides the technical one, critical voices are currently emerging. Social media and the way Big Data are produced, used, and co-constructed, can give a good insight in how people can be misled and how mathematics can be used in an ambivalent way. Fred Benenson (Woods, 2016) coined the concept of ‘mathwashing’ to explain the complexities of data and to “describe the tendency by technologists (and reporters!) to use the objective connotations of math terms to describe products and features that are probably more subjective than their users might think.” (Woods, 2016: 2) Here, mathematics is used to paper over a more subjective reality that is behind mathematical terms like algorithms or models. Technologists (and reporters) are using the power (the certainty, the objectivity, the truth) of mathematics to inform or mislead people. A nice example Benenson gave is the Facebook trending topics that show up on the sidebar on your Facebook. It seems value-neutral but reflecting the behavior of the Facebook user.

Mathematical algorithms behind can be understood as having agency power. Algorithms shaping the way we life, act, consume, and think. The massive production and availability of digital data also changing the production of scientific knowledge (Christin, 2016). New data and information are co-constructed based on the data production and the information people are talking about or using in their daily digital practice. New questions raise about agency, accountability, authority, responsibility and control. They should be studied from the sociological (ii) and legal (iii) approach. Other questions that relate to the use of data, the collection of data, the way how information flows in the public sphere, and the privacy issue should be studied from the philosophical and ethical (iv) approach. Transparency of data became a central issue since the increasing attention to scientific integrity and scientific ethics to avoid scientific misconduct and questionable research practices (QRP). Research integrity regulations are formulated in the European Code of Conduct for Research Integrity (ESF/ALLEA, 2011) and revised in 2017 (ALLEA, 2017). They provide principles on data practices and management as follows:

“Researchers, research institutions and organisations:

• ensure appropriate stewardship and curation of all data and research materials, including unpublished ones, with secure preservation for a reasonable period.

• ensure access to data is as open as possible, as closed as necessary, and where appropriate in line with the FAIR Principles (Findable, Accessible, Interoperable and Re-usable) for data management.

• provide transparency about how to access or make use of their data and research materials.

• acknowledge data as legitimate and citable products of research.

•ensure that any contracts or agreements relating to research outputs include equitable and fair provision for the management of their use, ownership, and/or their protection under intellectual property rights.” (ALLEA, 2017: 6).

In line with these principles General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) were formulated and will become applicable by May 2018. By these regulations, the

European Parliament, the Council of the European Union and the European Commission intend to strengthen and unify data protection for all individuals within the European Union (EU) and universities work hard to implement the data management plan in the research design to meet the expectations and the regulations of the national and European government.

It seems clear that both the agency and the ambivalent role of mathematics gives rise to a data literacy that is changing. People need to be aware that they are co-constructing Big Data. They are not only passive recipients of Big Data based reports, advertisements or news bubbles. They are active participants who must be able to understand the processes behind and to value the powers and limitations of Big Data. The ambivalent role of mathematics can be understood as both critical and anti-critical. The example of an anti-critical role is mathwashing because of the use of powerful mathematics to mislead people. The critical role is exactly, to expose these practices. The double role of mathematics gives us a strong argument to look for a broad description of Big Data Literacy.

## FINAL CONSIDERATIONS

In this article we have discussed the complexities of Big Data and how people are part of the co-construction of this kind of data. People are not only exposed to these data in their daily life, they also co-produce the data by doing their daily practices. Based on the analysis of the development of both statistical and mathematical literacy we argued for a broad description of Big Data Literacy considering four levels as discussed by Barocas, Hood & Ziewitz (2013). Big Data Literacy needs to include (i) a technical approach that studies algorithms as computer sciences; (ii) a sociological approach that studies the interaction among programmers and designers; (iii) a legal approach that studies mathematical algorithms as a figure and as an agent in law; and finally (iv) a philosophical approach that studies the ethics of algorithms. In line with Ben-Zvi (2017: 32) we will conclude that "Understanding big data and its powers and limitations is important to active citizenship and to the prosperity of democratic societies. Today's students therefore need to learn to work and think with data from an early age, so they are prepared for the data-driven society in which they live.".

## ACKNOWLEDGEMENT

Thanks to Dr. Patrick Allo to invite Karen François for the panel on *Critical perspectives on the role of mathematics in data-science. Society for Philosophy and Technology – SPT,* Darmstadt, 14-17 June, to give the presentation on 'The ambivalent role of mathematics and the need for a broad description of big data literacy'.

## REFERENCES

Ainley, J, Gould, R, & Pratt, D. (2015). Learning to reason from samples: commentary from the perspectives of task design and the emergence of "big data". *Educational Studies in Mathematics 88*(3), 405-412.

ALLEA (ALL European Academies) (2017). *European Code of Conduct for Research Integrity*. Revised Edition. Berlin: ALLEA.

Barocas, S., Hood S. & Ziewitz, M. (2013). Governing algorithms. A provocative piece. In *Proceedings of the 'Governing Algorithms' conference* (pp. 1-12). New York: New York University.

Ben-Zvi, D. & Friedlander, A. (1996). Statistical thinking in a technological environment. In J. B. Garfield & G. Burrill (eds.) *Research in the role of technology in teaching and learning statistics: Proceedings of the 1996 International Association for Statistical Education Round Table Conference* (pp. 45-55). University of Granada: International Statistics Institute.

Ben-Zvi, D. (2017). Big-Data inquiry: thinking with data. In R. Ferguson et al. (Eds.), *Innovating pedagogy 2017. Exploring new forms of teaching, learning and assessment, to guide educators and policy makers* (pp. 32-36). United Kingdom: The Open University.

boyd, d. & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society, 15*(5), 662-679.

Caldas, M.S. & Silva, E.C. (2016). Fundamentos e aplicação do Big Data: como tratar informações em uma sociedade de yottabytes. *Bibl. Univ., Belo Horizonte, 3*(1), 65-85.

Christin, A. (2016). From daguerreotypes to algorithms. Machines, expertise, and three forms of objectivity. *ACM Computers & Society, 46*(1), 27-32.

ESF/ALLEA (European Science Foundation & ALL European Academies) (2011). *European Code of Conduct for Research Integrity*. Strasbourg: ESF/ALLEA.

EU (European Union). (2009, March31). Regulation on European statistics, 2009. *Official Journal of the European Union, L, 87*, 164–173. accessed February 17th, 2018 from http://data.europa.eu/eli/reg/2009/223/2015-06-08.

François, K., Monteiro, C. & Vanhoof, S. (2013). Mathematical and statistical literacy. An analysis based on PISA results. *Revista de Educação Matemática e Tecnológica Iberoamericana, 4*(1), 1-16.

François, K., Monteiro, C., Carvalho, L. & Vandendriessche, E. (2015). Politics of ethnomathematics: An epistemological, political, and educational perspective. *Proceedings of the Eight International Mathematics Education and Society Conference –MES– Vol 2 (pp. 492-504). Book Series: Mathematics Education and Society*. Portland, Oregon, USA, June 21-26, 2015. ISSN: 2077-9933.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1-51.

Gal, I. (2004). Statistical literacy. Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (Eds.). *The challenge of developing statistical literacy, reasoning and thinking* (pp. 47-78). Dordrecht: Kluwer Academic Publishers.

Garfield, J.B. & Ben-Zvi, D. (2008). *Developing students' statistical reasoning. Connecting research and teaching practice*. UK: Springer

Haack, D. (1979). *Statistical literacy: A guide to interpretation*, Duxbury Press, North Scituate, Massachusetts.

Laney, D. (2001). 3-D Data management: Controlling data volume, velocity and variety. META Group Research Note, February 6. accessed February 17th, 2018 from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Organisation for Economic Co-operation and Development (OECD) (1999) Measuring student knowledge and skills: A new framework for assessment. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2010). PISA 2012 mathematical framework. Paris: OECD.

Queiroz, T., Monteiro, C., Carvalho, L. & François, K. (2017). Interpretation of statistical data: the importance of affective expressions. *SERJ - Statistics Education Research Journal, 16*(1), 163-180.

Rieder, G. & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society,3*(1), 1-6.

Wallman, K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*(421), 1-8.

Watson, J.M. (1997). Assessing statistical literacy using the media. In I. Gal & J.B. Garfield (Eds.). *The assessment challenge in statistics education* (pp. 107-121). Amsterdam: IOS Press and The International Statistics Institute.

Woods, T. (2016). 'Mathwashing,' Facebook and the zeitgeist of data worship. *Technically Brooklyn*. Accessed online https://technical.ly/brooklyn/2016/06/08/fred-benenson-mathwashing-facebook-data-worship/ February 16, 2018.

Zeelenberg, K. & Braaksma, B. (2017). Big Data in official statistics. In T. Prodromou (Ed.), Data *visualization and Statistical Literacy for Open and Big Data* (pp. 274-296). Hershey: IGI Global.