BIG-DATA LITERACY AS A NEW VOCATION FOR STATISTICAL LITERACY

KAREN FRANÇOIS
Vrije Universiteit Brussel, Brussel
karen.francois@vub.be

CARLOS MONTEIRO
Universidade Federal de Pernambuco, Recife
carlos.monteiro@campus.ul.pt

PATRICK ALLO
Vrije Universiteit Brussel, Brussel
patrick.allo@vub.be

ABSTRACT

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 statisticians and statistics consumers, therefore it is a challenge to develop Big-Data Literacy to support the needs of constructive, concerned, and reflective citizens. However, the development of the concept of statistical literacy mirrors the current gap between purely technical and socio-political characterizations of Big Data. In this paper, we review the recent history of the concept of statistical literacy and highlight the need to integrate the new challenges and critical issues from data science associated with Big Data, including ethics, epistemology, mathematical justification, and math washing.

Keywords: *Statistics education research; Big-Data literacy; Statistical literacy; Justification; Ethics*

1. 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 previous research in mathematics and statistics education (François, Monteiro, & Vanhoof, 2013; François, 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, n. d.). We review the history of the concept of statistical literacy and highlight how its meaning 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 criticise statistical information and reasoning. Without going into a detailed analysis of the concept of mathematical literacy, it is interesting to see how the concept – as used in the context of PISA (OECD, 1999, 2010) – underwent a similar evolution, and is now related to the needs of constructive, concerned, and reflective citizens. The parallel between statistical and mathematical literacy mirrors the current gap between purely technical and socio-political characterizations of Big Data.

Statistics Education Research Journal, 19(1), 194–205, http://iase-web.org/Publications.php?p=SERJ

2. WHAT IS BIG DATA ABOUT

In this section we present different aspects to characterise Big Data before we analyse the new challenges and the critical issues.

Generally, Big Data refers to large amounts of information that is generated continuously by various means from a quite wide range of social situations, which include web scraping, mobile-phone use, online shopping, and banking transactions. For instance, huge datasets are generated by social media users who share information, write their opinions, and contact friends.

Big Data allows the use of different management technologies related to the generation of data ranging from specific data analysis software to tools available in social networks (Martins, Monteiro, & Prodromou, 2017). These explorations are also associated with the analytic process of *data mining* that is the search for consistent patterns or systematic relationships between variables to validate them by applying the detected patterns to new subsets of data (Caldas & Silva, 2016). Generally, data mining is utilised as part of business, market or scientific research, thus favouring data-based decision-making.

Big Data is often characterised 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 also its complexity 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. In recent work (Shmueli, Bruce, Stephens, & Patel, 2017), a fourth *V* of *Veracity* was added to emphasise the ethical and epistemological concern of the trustworthiness of data. Even a fifth *V* of *Value* is added to highlight the way Big Data is impacting business, companies and society (De Mauro, Greco, & Grimaldi, 2015).

Zeelenberg and Braaksma (2017) state that Big Data can be classified as large datasets related to social activities which are not covered by official statistics. Such 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 emphasise that Big Data may be highly volatile and selective because the population to which it refers may change from day to another, which produces non-stationary time-series. Another very frequent issue is associated with the fact that some Big-Data sets do not have linking variables which would allow relating them to other datasets or population frames. These limitations might increase the possibility of error on the statistical results. However, it is important to minimise possible bias, which may be done by combining Big Data with data from sources, which utilise more standardised 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 (EU, 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 utilise those emerging data sets. If we consider that statistics education needs 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, p. 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".

They suggest that it is important to understand the limits of data and re-consider concepts associated with representativity of samples, the relationship between sample and population, and sampling variability. In order to help students learn from Big-Data sources, educators have a key role to choose 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 utilise technological artefact with a more accessible language, enabling students to experience tools that lead them to explore data processing, and to experience the use of artefacts that lead them to think about data.

3. REVIEW OF LITERATURE ON STATISTICAL LITERACY

In this section, we present the development of the concept of statistical literacy to mirror the current gap between purely technical and socio-political characterizations of Big Data. We review the recent history of the concept of statistical literacy and we compare this concept with the concept of mathematical literacy as it is developed by the OECD (1999, 2010) to highlight the need to integrate the new challenges and critical issues from data science associated with Big Data.

Based on a literature review, we argue that the concept of statistical literacy developed from a rather small and technical perspective in the late 1970s. The concept was broadened by the American Statistical Association (ASA) in the late 1990s, now also emphasizing the critical aspect of statistical literacy – besides its technical components. Research from the 2000s confirmed this broader description of the concept. Thus, statistical literacy becomes a broad and complex concept that covers 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 and the type of data, the definition and measurement problems, and finally certain considerations concerning the design of the survey sample. From this description it is clear that only the technical dimension of the concept was emphasised. 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. In her presidential address to the ASA, Wallman (1993, p. 1) states that

> "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."

The concept was further developed by Watson (1997) who presented a sophisticated framework of statistical literacy comprised of three tiers:

- a technical one,
- a societal one, and
- a critical one.

The technical one comprises the basic 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 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). He

portrayed statistical literacy as "the ability to interpret, critically evaluate, and communicate about statistical information and messages" (Gal, 2002, p. 1). This is a key ability expected of all citizens in an information-laden society.

More recently, Garfield and Ben-Zvi (2008) described the connection between the purely technical aspects of statistics and its ethical-political significance. They distinguish between: statistical literacy; statistical reasoning; and 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. Garfield and Ben-Zvi call this stage of statistical thinking "the normative use of statistical models", 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 now compare this sophisticated framework and description of statistical literacy with the concept of mathematical literacy as it was developed by OECD (1999), one can observe a quite similar attention to the societal relevance and critical attitude:

> "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."

This conception was applied to the international comparative survey PISA 2003. In 2010, it was reformulated to emphasise the different components and competences of mathematical literacy. There was a new emphasis on the technical skills to link the critical literacy with mathematical skills (concepts, procedures, facts, tools; and formulate, employ, interpret, etc.). These technical and mathematical skills will serve and assist pupils to become critical and constructive citizens. The new conception was applied to the international comparative survey PISA 2012 and it is described as follows (OECD, 2010):

> "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 recognise the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens."

4. STATISTICAL LITERACY AND BIG DATA

In this section, we evaluate recent work on the discussion of statistical literacy related to Big data and we bring in two more issues that seem relevant in the debate on the revision of the statistical literacy concept. The relevance of statistical literacy for the proper appraisal of the promises of Big Data has been raised before (e.g., Horton, 2015; Ridgway, 2016; Wild, 2017). Ridgway (2016) highlights the inadequacies of the current statistics curriculum in view of the data revolution; he stresses

> "a need to create curricula that devote more attention to the interpretation of large scale data sets" (p. 529), reminds us that "statistics is about solving real problems" (p. 531), confirms that the "idea of prediction is conceptually simpler than hypothesis testing and should be introduced early into the curriculum" (p. 534), and asks for and broadening the scope of the curriculum "ensuring that students

> engage with every phase of statistical problem solving and making students aware of current uses of statistical methods that affect their lives directly." (p. 536).

The question of how the field of statistics should change or the question of how our conception of statistical literacy should be revised in view of the rise of novel data practices is typically developed to address two specific concerns. The first is that Big Data leads to the redundancy or irrelevance of statistical reasoning. The second is that, as a profession and as a field of expertise, statistics has to take action to remain relevant and to ensure that future statisticians have the right data skills. We first review current efforts to affirm the relevance of sound statistical reasoning and to adapt the scope and ambition of statistics education. In a second move, we argue that while such efforts are indispensable, they aim at a form of statistical literacy that is narrower than Big-Data literacy as a new vocation for statistical literacy – which is exactly what we have in mind.

As a reply to the recurrent claim that Big Data makes statistical reasoning and careful thinking about sampling redundant, Tim Harford, in his 2014 Significance lecture at the Royal Statistical Society International Conference, shows how statistical lessons of the past are the perfect antidote to the overly optimistic promises of Big Data. Statistical reasoning, then, retains its value because it demystifies Big Data and replaces it with more realistic expectations about what one can achieve with data. Harford's (2014) argument challenges the four promises or *articles of faith* of Big Data, namely

(i) highly accurate results;
(ii) the irrelevance of (thinking about) sampling because N = all;
(iii) the replacement of causal knowledge with detection of correlation; and
(iv) the redundancy of statistical or scientific models.

He argues that many of the risks that careful statistical reasoning, and especially the interaction between statistical reasoning and data collection, are meant to avoid, are still present in the context of Big Data. Bias remains a problem that does not disappear as N increases. Even more: it is harder to control when so-called opportunistic data is used. Similarly, whereas taking action in view of detected regularities, we do not fully understand, can often be profitable, actionable knowledge based on correlations remains very brittle. The inevitability of such risks reaffirms, in Harford's view, the relevance of the critical reflexes built into statistical reasoning for Big Data.

Wild (2017, p. 34) develops an analogous argument, reiterates the risks associated with the use of found data, and stresses that "[Big data] can tell more lies and we are more likely to believe them." In addition, he explicitly highlights the relevance of statistical literacy as a means to assess specific statistical results and to understand where the claims about the unprecedented benefits of Big Data break down. He does not relate this need to reinforce statistical literacy to the needs of data subjects (the permanent datafication of our life online), but ties it to the fact that the production of statistics and the processing of data is no longer the exclusive responsibility of experts.

This last insight is also central for Horton (2015), who argues for a revision of how introductory statistics courses are taught. His goal is to train statisticians (even at the undergraduate level) to enable them to tackle complex real-world data problems and to equip them with skills that let them compete with computer scientists – another dimension along which the relevance of statistics needs to be asserted – on the data-science job market. In his view, statistics should offer data science the prospect to become "more rigorous, scientific, and reproducible". The general moral that arises from Harford (2014), Wild (2017) and Horton's (2015) attempts to make statistics and statistical literacy relevant to the rise of Big Data is that statisticians should become more sensitive to the needs and challenges of contemporary data science; they might even be able to safeguard data science from predictable disappointments.

Our goal in the present paper is complementary to theirs. We explicitly argue for the further expansion of statistical literacy as a form of Big-Data literacy. Their arguments and the proposals formulated by Horton (2015) provide only a partial answer to the challenge we set

ourselves. As we see it, there is more to critical data skills than the recognition of the limits of what we can learn from Big Data or the ability to detect specific classes of errors. Input from other disciplines is needed as well. We conclude this section with three interdisciplinary rather than purely statistical lessons. Each of these lessons introduces a concern for Big-Data literacy that is not, or is only insufficiently addressed by current conceptions of statistical literacy.

Lesson 1: Understanding the complexity of how knowledge is created and decisions are justified Statistics views the construction of knowledge from the standpoint of how we deal with uncertainty (Kline, 1967/1985); we acquire knowledge by reducing uncertainty and by quantifying that uncertainty. As such, it builds on a narrow conception of justification that cannot readily take into account the role of trust, accountability, and responsibility in justifying decisions based on data. If, however, we want to understand and evaluate what is at stake epistemologically in the context of Big Data, we cannot restrict our attention to the types of justification that good statistical reasoning supplies. Especially in the context of decision-making based on the algorithmic processing of Big Data, we also want to avoid decisions that are unaccountable or that are based on processes that are secret, opaque, or hard to comprehend (see the next section). The ability to detect typical statistical mistakes does not help to address, or even to become aware of such concerns. In our view, such abilities are essential for the development of the critical data skills of experts as well as lay people. They should therefore be integrated in the broader conception of Big-Data literacy we envisage.

Lesson 2: Making room for ethical norms We need critical data skills to address the specific risks associated with Big Data (Mittelstadt & Floridi, 2016; Floridi & Taddeo, 2016; Ethics Advisory Group, 2018). We can think of these risks as ethical (Ridgway, 2016, §6.8; Vayena & Tasioulas, 2016) as well as epistemological risks: the risk of doing something wrong or the risk of being too confident. Often, ethical risks can be reduced by avoiding epistemological risks (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016): if we properly assess uncertainty (Dunson, 2018, §3), we avoid mistaken beliefs and our actions are better informed. In this sense, careful statistical reasoning serves an ethical purpose. Yet, for Big Data, the ethical challenges vastly outstrip the epistemological challenges. Non-discrimination and fairness, but also privacy, data protection and security stand out in this context. We would like to stress this point in a more general fashion by drawing attention to the value of uncertainty and explicitly including awareness of what one ought not to know as a precondition for data literacy.

Lesson 3: The perspective of the data subject Traditional conceptions of statistical literacy take the perspectives of the producer and the consumer of statistics as their point of reference. Being statistically literate, in the technical but also in the critical sense, is valuable because it allows one to make sense of data and to avoid being misled by statistical claims. When Harford (2014) and Wild (2017) underscore the value of statistical thinking in the context of Big Data, they still do this from the same perspective. In a context of massive datafication, however, where in the first place we are the subject of statistical analyses (i.e., data subjects), we need to be able to critically engage with Big Data even when we are neither the producer nor the consumer of statistical claims. Using the framework outlined in Zwitter (2014), we may say that statistical literacy traditionally caters for the needs of the Big-Data collectors and utilisers but not for the needs of the Big-Data generators. If we think of risks of Big Data in terms of the power imbalances it creates, then seeing Big-Data literacy as an attempt to redress these imbalances means that we need to think more explicitly about the kind of statistical and Big-Data literacy that can benefit all data subjects; even those who are not aware that they are being counted, quantified, and datafied.

5. NEW CHALLENGES AND CRITICAL ISSUES

Based on the overview of the development of the concept of statistical literacy and on the three lessons learned on justification and ethics, we further investigate the question on how the concept of Big-Data literacy should be developed. One aspect we take from the development of the concept of statistical literacy is the importance of integrating the social and critical perspective which we need even more in the context of Big Data. Another aspect we learned from our discussion on knowledge building and justification, and the relation between justification and ethics, is the need for an interdisciplinary approach to deal with the complex processes on data handling. In this section, we illustrate some of the issues that make the context of Big Data more complex.

The same layers as discussed by Garfield and Ben-Zvi (2008) can be observed when it comes to sophisticated algorithms. Barocas, Hood, and Ziewitz (2013) sketch the different approaches in studying algorithms:

> "[Someone] suggested that there could be a technical approach that studies algorithms as computer science; a sociological approach that studies algorithms as the product of interactions among programmers and designers; a legal approach that studies algorithms as a figure and agent in law; and a philosophical approach that studies the ethics of algorithms. [But then they resume asking for an interdisciplinary approach:] Compartmentalizing the study of algorithms into disciplines may reduce complexity and, in fact, discipline the discourse. But what are the risks involved in this 'division of labour'? Would an interdisciplinary study of algorithms be fruitful? How would it work?"

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 analyse data in a most accurate way and maximise 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 non-technical layers of the study of algorithms from above, 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, p. 2)

> "[coined the concept of 'math washing' 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."

Here, mathematics is used to paper over a more subjective reality that is behind mathematical terms like algorithms or models. Technologists (and media) are using the power (the certainty, the objectivity, the truth) of mathematics to inform or mislead people. A nice example Benenson gave is the trending topics that show up on the sidebar on your Facebook. It seems value-neutral but reflecting the behaviour of the 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 are 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 rise about agency, accountability, authority, responsibility, and control. They should be studied from the

sociological and legal 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 approach. Transparency of data became a central issue since the increasing attention to scientific integrity and ethics to avoid misconduct and questionable research practices. 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 (p. 6):

> "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."

In line with these principles, general data protection regulation rules (EU, 2016) were formulated and are applicable since May 2018. By these regulations, the European Parliament, the Council of the EU and the European Commission intend to strengthen and unify data protection for all individuals; universities work hard to implement the data management plan in the research design to meet the new regulations.

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. Rather than passively receiving 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 math washing because of the use of powerful mathematics to mislead people. The critical role is, to expose these practices. The double role of mathematics gives us a strong argument to look for a broad description of Big-Data literacy.

6. FINAL CONSIDERATIONS

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 statistical literacy and the comparison with the notion of mathematical literacy we argued for a broad description of Big-Data literacy considering four levels as discussed by Barocas, Hood, and Ziewitz (2013). Big-Data literacy needs to include a technical approach that studies algorithms as computer sciences; a sociological approach that studies the interaction among programmers and designers; a legal approach that studies mathematical algorithms as a figure and as an agent in law; and finally, a philosophical approach that studies the ethics of algorithms. And all that done in an interdisciplinary approach.

We argue that the conceptions of statistical literacy should be revised to address two concerns: that Big Data leads to the irrelevance of statistical reasoning and that statistics has to take action to remain relevant. Related to statistical literacy and the teaching of statistics, teachers must ensure that future statisticians have the right data skills and interdisciplinary competences to collaborate with people engaged in data (statisticians, mathematicians,

computer scientists, machine learning experts, and domain experts). The scope and the ambition of statistics education have to be adapted to this Big-Data move; we aim at a form of Big-Data literacy as a new vocation for statistical literacy, in line with Ben-Zvi (2017, p. 32):

> "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."

Big-Data literacy is a wider concept than both statistical and mathematical literacy. It holds the aim of a critical and reflective citizenship based on interdisciplinary skills and competences. It is interdisciplinary in nature, bringing not only statisticians and mathematicians together but also computer scientists, sociologists, lawyers, and philosophers. It has to bridge the gap between a different subject-matter (Big Data) and a need for different skills and attitudes.

It is different from statistical literacy in that knowledge creation based on Big Data is generated from data that are of a different nature – not (representatively) collected but already there and in progress while analysing – and processes are more complex. It is thus a different way of making and justifying claims on the basis of data that must be understood. It is different from statistical literacy because there are different risks one has to be aware of. Dig Data is dealing with the ethical challenges in terms of non-discrimination, fairness, privacy, data protection, and security; and Big Data will significantly outstrip the epistemological challenges in creating robust knowledge. It finally differs from statistical literacy because there is a new perspective and a new audience to be taken into account. The perspective of the data subject is new for Big-data literacy. Especially (young) pupils may misunderstand the impact their data trails may have on their life. They are unaware of the commercial practices they are subjected to and probably do not understand the power imbalances that data processing creates. In a context where there is a growing attention to the human right not to be measured, we need a concept of literacy that empowers data subjects.

The implementation of an interdisciplinary field in a traditional curriculum that is mainly build upon discrete educational fields such as mathematics, statistics, computer sciences, but also sociology, law, and philosophy, challenges contemporary teaching practices. Big-Data literacy generates not only big responsibilities but also big challenges for rethinking statistical literacy as big-data literacy, which has to be further investigated.

ACKNOWLEDGEMENT

The research is partially funded by the Research Foundation – Flanders (Project G083620N).

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 (2017). *The European code of conduct for research integrity*. Revised Edition. Berlin: ALLEA – ALL European Academies.

Barocas, S., Hood S., & Ziewitz, M. (2013). Governing algorithms. A provocation piece. In *Proceedings of Governing Algorithms. A conference on computation, automation, and control* (pp. 1–12). New York: New York University.
[Online: governingalgorithms.org/resources/provocation-piece/]

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). Milton Keynes, UK: The Open 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 IASE Round Table Conference* (pp. 45–55). Voorburg: International Statistical Institute.
[Online: iase-web.org/Conference_Proceedings.php?p=Role_of_Technology_1996]

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. *Bibliotecas Universitárias: pesquisas, experiências e perspectivas, 3*(1), 65–83.

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

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP conference proceedings, 1644*, 97.
[Online: doi.org/10.1063/1.4907823]

Dunson, D. B. (2018). Statistics in the Big Data era: Failures of the machine. *Statistics and Probability Letters, 136*, 4–9. [Online: doi.org/10.1016/j.spl.2018.02.028]

Ethics Advisory Group (2018). *Towards a digital ethics*. Report 2018. Brussels: European Data Protection Supervisor.
[Online: edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf]

ESF/ALLEA (2011). *European Code of Conduct for Research Integrity*. Strasbourg: European Science Foundation & ALL European Academies.
[Online: allea.org/portfolio-item/the-european-code-of-conduct-for-research-integrity-2/]

EU (2009). Regulation on European statistics, 2009. *Official Journal of the European Union, L*, *87*, 164–173. [Online: data.europa.eu/eli/reg/2009/223/2015-06-08]

EU (2016). General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council. [Online: eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e3265-1-1]

FLAMES (n. d.). Flanders Training Network for Methodology and Statistics.
[Online: www.flames-statistics.eu/]

Floridi, L., & Taddeo, M. (2016). What is data-ethics? *Philosophical Transactions of the Royal Society A*, *374*(2083), 1–5.

François, K., Monteiro, C., Carvalho, L., & Vandendriessche, E. (2015). Politics of ethnomathematics: An epistemological, political, and educational perspective. In S. Mukhopadhyay & B. Greer (Eds.), *Proceedings of the Eight International Mathematics Education and Society Conference, Vol 2* (pp. 492–504).
[Online: mescommunity.info/MES8ProceedingsVol2.pdf]

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.
[Online: periodicos.ufpe.br/revistas/emteia/article/view/2240/0]

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities (with discussion and a rejoinder by the author). *International Statistical Review, 70*(1), 1–51.
[Online: onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00336.x]

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*. Dordrecht: Springer

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

Harford, T. (2014). Big data: Are we making a big mistake. *Significance, 11*(5), 14–19. [Online: doi.org/10.1111/j.1740-9713.2014.00778.x]

Horton, N. J. (2015). Challenges and opportunities for statistics and statistical education : looking back , looking forward. *The American Statistician*, *69*(2), 138–145.

Kline, M. (1967/1985). *Mathematics for the Nonmathematician.* New York: Dover Publications.

Kline, M. (1980). *Mathematics. The Loss of Certainty*. Oxford: Oxford University Press.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note. [Online: blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf]

Martins, M. N. P., Monteiro, C. E. F., & Prodromou, T. (2017). Teachers analyzing sampling with TinkerPlots: Insights for teacher education. In T. Prodromou (Ed.), *Data visualization and statistical literacy for Open and Big Data* (pp. 194–222). Hershey, PA: IGI Global.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2). [Online: doi.org/10.1177/2053951716679679]

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of Big Data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, *22*(2), 303–341. [Online: doi.org/10.1007/s11948-015-9652-2]

OECD (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: Organisation for Economic Co-operation and Development (OECD). [Online: www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33693997.pdf]

OECD (2010). PISA 2012 Assessment and analytical framework. mathematics, reading, science, problem solving and financial literacy. Paris: Organisation for Economic Co-operation and Development (OECD). [Online: www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf]

Queiroz, T., Monteiro, C., Carvalho, L., & François, K. (2017). Interpretation of statistical data: The importance of affective expressions. *Statistics Education Research Journal, 16*(1), 163–180. [Online: iase-web.org/Publications.php?p=SERJ_issues]

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education. *International Statistical Review*, *84*(3), 528–549. [Online: doi.org/10.1111/insr.12110]

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.

Shmueli, G., Bruce, P. C., Stephens, M. L., & Patel, N. R. (2017). *Data mining for business analytics: Concepts, techniques, and applications with JMP Pro*. Hoboken, NJ: Wiley.

Vayena, E. & Tasioulas, J. (2016). The dynamics of big data and human rights: The case of scientific research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160129.

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 thinking using the media. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). Voorburg: International Statistics Institute. [Online: iase-web.org/Books.php?p=book1]

Wild, C. J. (2017). Statistical literacy as the earth moves. *Statistics Education Research Journal*, *16*(1), 31–37. [Online: iase-web.org/Publications.php?p=SERJ_issues]

Woods, T. (2016). 'Mathwashing,' Facebook and the zeitgeist of data worship. *Technically Brooklyn*. [Online: technical.ly/brooklyn/2016/06/08/fred-benenson-mathwashing-facebook-data-worship/]

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, PA: IGI Global.

Zwitter, A. (2014). Big Data ethics. Commentary. *Big Data & Society, 1*(2), 1–6. [Online: doi.org/10.1177/2053951714559253]

KAREN FRANÇOIS
Department of Philosophy
Vrije Universiteit Brussel (VUB)
Pleinlaan 2, BE-1050 Brussel, Belgium