ISR Italian Sociological Review
University of Verona, Italy
ISSN 2239-8589 | Editor in Chief Paola Di Nicola
Managing Editor Debora Viviani

From a Theory-Centric View of Social Research to a Data-Centric Approach

Sonia Stefanizzi

How to cite
Stefanizzi, S. (2022). From a Theory-Centric View of Social Research to a Data-Centric Approach. [Italian Sociological Review, 12 (7S), 651-663]
Retrieved from [http://dx.doi.org/10.13136/isr.v12i7s.575]

[DOI: 10.13136/isr.v12i7S.575]

1. Author information
Sonia Stefanizzi
Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy

2. Author e-mail address
Sonia Stefanizzi
E-mail: sonia.stefanizzi@unimib.it

3. Article accepted for publication
Date: May 2022

From a Theory-Centric View of Social Research to a Data-Centric Approach

Sonia Stefanizzi*

Corresponding author:
Sonia Stefanizzi
E-mail: sonia.stefanizzi@unimib.it

Abstract

Digital data have become so pervasive that they are the main resource used in scientific research, business and society. With reference to scientific research, a data-centric approach is emerging in all scientific disciplines. The article argues that data centrism does not mean the absence of theory. In fact, as will be shown, data are always produced in relation to and to precise expectations and conceptual schemes. Just as the algorithms that are used to analyze huge amounts of data are "material executors" of actions predetermined beforehand.

A relevant aspect of data centrism is that the large amount of data available is closely linked to a revolution in the communication of research results that becomes Open Science. The link between Big Data and Open Data is, therefore, the true revolutionary contribution of data-centrism. In fact, the dissemination of data using open formats (Open Data) is able to provide new opportunities in terms of greater transparency on the part of the producers of the data. The data-centric approach allows us to tackle problems that in the past were considered difficult or impossible to analyze, it also raises other questions related to data quality and ethics such as the right to privacy, transparency in accessing public data, and fair treatment of predictive techniques. These aspects are part of the so-called data colonialism, which is lack of transparency in the collection and use of data, an increase in dehumanization at work and a potential violation of privacy.

Keywords: Big Data, data-centric approach, data colonialism.

* Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy.

1. Theory-centric versus data-centric approach

The emergence of the Internet and global connectivity has given rise to an accumulation of huge amounts of data stored in digital databases, the amount of which is constantly doubling. This amount of data is called Big Data, not only because of its quantity, but also because of the possibility for researchers to analyze it with increasingly automatic and fast procedures (algorithms).

Also, in the field of sociological research, in recent years we have witnessed a vertiginous technological innovation in the production, communication and analysis of data used for research purposes. As has been said, thanks to digital technologies we have at our disposal huge amounts of data that open up a radical change in the way research is done and how scientific knowledge is produced.

This mass of data has different types and origins and is put together to better describe and understand certain social phenomena, so as to produce new forms of analysis and knowledge.

The availability of Big Data and the ease with which it is produced represents a major challenge and opportunity for social research. Access to and analysis of data becomes the driving force of research: we are witnessing a shift from a theory-centric approach to a model of innovation that is called data-centric. Entering into the merits of the two approaches to research, in the theory-centric perspective the usefulness of data consists essentially as evidence of the falsifiability of a hypothesis in order to confirm or not the empirical validity of a theory. Theoretical knowledge constitutes the framework of the investigation that serves to formulate hypotheses from empirical observations that must then be confirmed or falsified by empirical evidence. In other words, it offers a perspective, a conceptual scheme drawn up with reference to the specific problem under investigation, which can therefore orient the collection and analysis of data[1]. It often follows that data which do not work well as evidence in this sense or which could even be interpreted very differently, are removed and typically not exposed to further analysis by researchers.

In the theory-centric view what is important is the corroboration or otherwise of a theory, the formulation of new theoretical frameworks, all other components of the research, from the data to the techniques of data collection and analysis are seen as secondary to the creation of theories and strongly oriented by the theory itself. Scientific knowledge is thus based not only on respect for what we call scientific evidence, but also on the flexibility of the scientist, who is able to change his/her mind when confronted with new results

[1] Popper, 1934 wrote that "Theories are nets cast to catch what we call -the world-: to rationalize it, to explain it, to dominate it".

that dismantle his/her previous interpretation of reality. The analysis of Big Data clearly indicates a different and probably Baconian understanding of the role of hypotheses in science. Theoretical expectations are no longer seen as guiding the process of enquiry and empirical input is recognized as primary in determining the direction of research, the phenomena and related hypotheses considered by researchers.

According to some scholars, the rise of Big Data represents a sort of reclaiming of inductivism in the face of the countless criticisms directed at theory-free reasoning (Williamson, 2004). Advances in automation, combined with the exponential rise of Big Data would contribute to the value of the inductive philosophy of science (Williamson, 2004; Pietsch, 2015).

Much of the recent philosophy of science, and in particular modelling and experimentation, has challenged theory-centrism by highlighting the role of models, methods and modes of action as research outputs rather than mere tools, and emphasizing the importance of expanding the philosophical understanding of scientific knowledge to include these elements alongside propositional statements. The rise of Big Data offers another opportunity to reframe the understanding of scientific knowledge as not necessarily theory-centred. Referring to the vast literature on perspectivism and experimentation (Gooding, 1990; Giere, 2006; Radder, 2006; Massimi, 2012), Werner Callebaut has argued forcefully that the most sophisticated and standardized measurements embody a specific theoretical perspective, and this is no less true for Big Data (Callebaut, 2012). Elliott and colleagues point out that conceptualizing Big Data analysis as atheoretical risks encouraging unsophisticated attitudes towards empirical enquiry as a "fishing expedition", having a high probability of leading to meaningless results or spurious correlations, relying on scientists who lack adequate expertise in data analysis, and producing data distorted by the way it is collected (Elliott et al., 2016: 880).

The data-centrism view of research can be considered as "a particular model of attention within research, within which concerns around data handling take precedence over theoretical questions" (Leonelli, 2016: 178). Data-centric social sciences have been recently developing based on ICT. A large amount of data on socioeconomic-technological systems has slowly accumulated in several institutions and fields. The data is generated and collected in some type of data-generating mechanism and then stored in a database or computer storage. The data is eventually distributed or analyzed for a purpose, which is to interpret the world from the data and to make decisions. We can also start to formulate a model of a specific phenomenon from the data since the data may provide us with useful insights to construct a model. In this perspective, data are seen as public entities that have scientific value independently of their role in testing a given hypothesis and that can be interpreted in different ways depending on the

skills and interests of the researchers analyzing them. We are therefore witnessing a radical reassessment of the potential of data in generating knowledge (Leonelli, 2018). One of the criticisms levelled at the data-centric approach is the assumption that Big Data analysis implies the death of theory, i.e. that the theoretical assumptions used to produce and manage the data are not relevant to their interpretation. This interpretation caused a stir, especially following an editorial by Chris Anderson published in Wired magazine in 2008, called "The end of theory" (Anderson 2008). Anderson, in fact, claimed that Big Data would make the scientific method obsolete, since there would no longer be any need for any kind of theoretical-interpretative mediation, there would no longer be any need to read the data in a certain way rather than another: the data are so many that they directly tell the reality for what it is [2]. Hence the death of theory, of the interpretation of the social phenomena we are interested in understanding. This kind of process would represent a sort of Copernican counter-revolution on the scientific method, in which theory is no longer the starting point, but, on the contrary, the final achievement of the process of data analysis. This epistemological shift reduces the importance of causality to a meaningful correlation, sufficient to formulate a theory in a big data-driven approach.

But data-centrism does not necessarily mean "the death of theory". Even in the case of Big Data, the "value reference" of which Weber talks is an ineradicable dimension. In fact, data are always produced in relation to precise expectations and conceptual schemes. Not only that, but all the operations within the different research phases carried out by the researcher to acquire information (from indicators to variables to techniques used, data collected, etc.) are elements of mediation between us and the world. It is, therefore, plausible to assume that data come from certain conceptual perspectives, just as the ways in which they are produced, organized and analyzed derive from theoretical frameworks. In particular, as far as objectivity and incontrovertibility are concerned, contemporary epistemology is now too disenchanted to still believe in the neo-positivist myth of the neutrality of empirical data (such as, for example, those recorded by quantitative measurements), which are instead considered a variable sensitive to the context and theoretical paradigm in which they are inscribed.

[2] Anderson's view that data speaks for itself led in a number of cases to Big Data being used with a fideistic certainty that everything it produced in terms of patterns was of better quality than traditional methods. This led to the generation of what has been called Data Hubris, an arrogance of data as a tool for asserting truth. The example that has been studied more than any other as a phenomenon in this direction is Google Flu Trends (Butler, 2013) and (Lazer, 2014).

Ken Waters has provided the useful characterization of theory-informed enquiry (Waters 2007), which can be invoked to emphasize how theory informs the methods used to extract meaningful patterns from Big Data, and yet does not necessarily determine either the starting point or the results of data-intensive science. This does not resolve the question of the role that theory actually plays. Rob Kitchin (2014) has proposed to consider Big Data as related to a new mode of hypothesis generation in a hypothetico-deductive framework. Leonelli is more skeptical about attempts to match approaches to Big Data, which are many and diverse, with a specific type of inferential logic. Rather, she has focused on the extent to which the theoretical apparatus at work within big data analysis rests on conceptual decisions about how to order and classify data - and has proposed that such decisions may give rise to a particular form of theorizing, which she calls classificatory theory (Leonelli, 2016).

These disagreements indicate that Big Data elicits different interpretations of the nature of knowledge and enquiry, and the complex iterations through which different inferential methods build on each other. Once again, in the words of Elliot and colleagues, attempting to draw a sharp distinction between hypothesis-driven and data-intensive science is misleading; indeed, these modes of enquiry are not orthogonal and often intertwine in actual scientific practice (Elliott et al. 2016: 881, O'Malley et al. 2009; Elliott, 2012).

In the view of these considerations we assume that the different classificatory and data collection systems do not merely provide an empirical basis for measuring reality and perception of phenomena, but have a strong economic potential and acquire a different value depending on who uses them and their relevance is never contained to a single domain (Leonelli, 2018). One of the elements characterizing the logic underlying Big Data is its predictive effectiveness: the possibility of cross-referencing and processing such a huge amount of data in real time would make it possible to make largely reliable predictions, precisely because they result from the use of such vast and varied information. This means that, unlike the logic of the Small Data, with the Big Data a quantitative, rather than qualitative, analysis is conducted: whereby the data is not used for a specific purpose and reconnected to the information contained in it, but a huge quantity of data is collected to arrive at the most disparate information and, therefore, with different purposes.

In this regard, it is useful to remember that the algorithms that are used to analyze huge masses of data often coming from heterogeneous sources turn out to be only "material executors" of actions predetermined upstream. In fact, behind each algorithm there is a mathematical or statistical model that determines the results. It will be executed within a protected environment that often conceals mechanisms that are opaque and invisible to the outside world, except to their direct creators (O'Neil, 2016). Algorithms generate models that

very often define an internal reality of their own which they then use to justify their results, sometimes even losing sight of the external context. As O'Neil argues, in order to determine whether a model works, they usually try to perform a comparison with an expected result derived from a widely accepted informal model, which does not guarantee full objectivity (O'Neil, 2017). Indeed, they not infrequently need feedback from outside the system in order to try to adapt and improve their performance, for example through the analysis of errors and exceptions[3]. Often the opacity of an algorithm, i.e. the ignorance of the assumptions on which it is based and the failure to adequately consider the limits of its applicability, leads to its incorrect use or to interpreting the predictions provided by the algorithm in terms that are not always scientifically founded, and leads to making decisions without the necessary reasons and explanations. If, therefore, it becomes difficult to know and explain how certain algorithms manage to arrive at a certain outcome, the results obtained lose their relevance in the face of the impossibility of demonstrating why we have achieved them, and consequently knowledge is not increased.

A relevant aspect of data-centrism is that the large amount of available data is closely linked to a revolution in the communication of research results that becomes open science. According to this paradigm, Open science is "doing science" in a way that other researchers can also collaborate and contribute, where research data and other processes are made available making them reusable and redistributable thus encouraging the reproduction of research and its data. Big Data opens us, therefore, to the possibility of exploring, aggregating and relating vast sets of data, but it changes the status acquired by the data itself whose production implies an obligation to distribution so that the data is used by more people making it possible for different components of society to communicate and work together (Leonelli 2018).

The link between Big Data and Open Data is, therefore, the true revolutionary contribution of data-centrism. In fact, the dissemination of data using open formats (Open Data) is able to provide new opportunities in terms of greater transparency on the part of the producers of the data, as, for example, in the case of public administration; better access to information by citizens; and the creation of new products and new services by businesses. Open Data is accessible public data that people, companies, and organizations can use to launch new ventures, analyze patterns and trends, make data-driven decisions, and solve complex problems. All definitions of Open Data include two basic features: the data must be publicly available for anyone to use, and it must be

[3] A striking example was the one that hit Google with its algorithm for automatic image recognition and cataloguing, which in 2015 had mistakenly catalogued a photo with two black people as a "Gorilla", generating deep general indignation.

licensed in a way that allows for its reuse. Open Data should also be relatively easy to use, although there are gradations of "openness". And there's general agreement that Open Data should be available free of charge or at minimal cost.

The distinctive feature of Open Data, therefore, and which identifies, we might say, its specific DNA, is the possibility of exploiting and reusing the information contained in these data, including its economic potential. However, it must be specified that the notion of Open Data must include not only "Open Government data" (i.e. data relative to the work of the public administrations), but also all that "other" information, which is modifiable and accessible to all without being subject to any control that could block or limit its reproduction.

Open Data is a subset of Big Data, but the purposes and uses of these two broad categorizations of data are profoundly different. With Big Data, we collect massive amounts of information to fuel discovery science. With Open Science, we make that data and the tools to analyze it free and accessible to researchers around the world. Taken together, this free sharing of data with researchers across disciplines is advancing the scientific research.

2. Some critical aspects of the data-centric approach

The use of this source of knowledge poses some problems, especially with respect to data quality and reliability. Indeed, quantity does not necessarily imply quality. In this regard, the questions that arise concern the dependence of data analysis on the context in which the data are extracted and used, which can vary immensely depending on the situation and the questions posed by the researcher. A further critical issue is the fact that it is not possible to analyze Big Data without having access to the so-called metadata, i.e. information on its provenance (i.e. how it was generated, with respect to what and under what circumstances) that allows researchers to assess whether the data is reliable and what interpretations are plausible (Leonelli, 2018). The case of Cambridge Analytica also exemplifies another problem that refers to the social role of research and has to do with the problem of sensitive data, i.e. data that can be used to represent characteristics of individuals or groups of people[4]. The increasingly sophisticated analysis of these data, and the opportunity to relate them to each other offered by Big Data, opens the door to an ever better

[4] Cambridge Analytica was founded in 2013 by Robert Mercer and specializes in collecting huge amounts of data from social networks about their users. This information is then processed by models and algorithms to create profiles of each individual user, with an approach similar to that of "psychometrics", the field of psychology that deals with measuring abilities, behaviors and more generally personality characteristics.

understanding of the real needs of citizens, and thus to better informed and more efficient political, social and environmental decisions. At the same time, the mishandling of such data, or the adoption of ethically and socially problematic research methods or purposes, can easily generate enormous harm to the individuals concerned - for instance by making them vulnerable to surveillance and manipulation by malicious actors, or by generating unreliable or biased knowledge about them, which is then used by social, commercial, medical or insurance services.

The problem of how to use sensitive data is, therefore, both an epistemic and an ethical one, where it is not possible to distinguish criteria used to produce reliable knowledge from those used to ensure that the methods used do not reinforce unjust and arbitrary social discrimination. The lack of a clear separation between scientifically and ethically correct conduct is particularly relevant in the case of Big Data. Reflection on the social consequences of using old, biased, unreliable and corrupt data is always inexorably linked to an assessment of the ethical value of the choices made in the selection, management and interpretation of the data. In this sense, not only are the scientific and ethical values of data not necessarily in conflict, but they are typically associated with each other.

The phenomenon of datafication also introduces other issues and side effects such as: data colonialism, lack of transparency in the collection and use of data, an increase in dehumanization at work and a potential violation of privacy.

In fact, in a logical parallel to the historical term colonialism, in which land and resources were appropriated for profit, the term "data colonialism" was created to express the exploitation of human beings through their data, paving the way for the "capitalization of life without limits" (Couldry, Mejias, 2019). In a modern declination, this phenomenon combines predatory extractive practices with the abstract quantification methods of modern information systems (Couldry, Mejias, 2019). For this reason, the term data colonialism has been coined to express the exploitation of human beings through their own data (which then becomes a raw material from which to extract value). It is not possible to delimit this trend into specific geographical or thematic areas as it is a global phenomenon that potentially involves any person or thing that is within today's connected infrastructures. In this virtual environment, our social life becomes an easily accessible and potentially exploitable resource for extracting value from it. Ricaurte (2019) has shown how the process of data colonialism can be placed in a scenario divided into three distinct but complementary phases: the extraction of data from sources, the processing of data through the use of increasingly complex algorithms, and the exploitation of the information obtained by market agents. The parallelism that has been made by the authors

(Couldry, Mejias, 2019) of this neologism is intended to emphasize the aspect of resource appropriation with a predatory connotation. This data collection is also carried out by trying to shape aspects of our lives, such as social relations, making "this process of extracting value from them as natural as possible" (Couldry, Mejias, 2019). Digital platforms have, in fact, made it possible to create an ideal medium for the emergence of this new social paradigm, for which every aspect can be continuously tracked, structured and analyzed in the form of data. Arvidsson (2016) had already analyzed these techniques of value extraction by digital platforms of their contents as a form of self-financing, which makes the mechanism of appropriation of user data opaque and not very transparent; counterbalanced by the fact that the use of these services is often made free of charge. But this principle of appropriation of social aspects as elements of value is central, also encouraged by the continuous sharing and exposure of our thoughts and activities (Couldry, Mejias, 2019). To counterbalance these negative aspects, Human Data Science, a multidisciplinary model that develops innovative approaches to reading and interpreting data while keeping the person at the centre, has been developing in recent years. In fact, if it is true that today's world population produces an almost infinite amount of data, in order to capitalize on this wealth of information it is increasingly important to transform data into value, adopting models that also consider the human factor.

In the age of post-truth, one of the most relevant effects of the growth of Big Data and, in particular, of data from social networks, is their reliability, as the line between opinions and information is often blurred[5]. Even if working on Big Data could help to solve some methodological problems that characterize quantitative studies, such as social desirability, the interviewer effect, or the economic sustainability of the surveys, as already mentioned, questions remain about the reliability of the data. In other words, who judges the quality of the data and the technologies used to sort them? Indeed, there is a difference between the knowledge and interests of those who produce the data and those who re-interpret them for new purposes. Even if the re-interpretation may lead to new discoveries, there remains the risk of relying on data generated and evaluated by the producers on the basis of different criteria than those of the secondary analyst. The risk is that the researcher using Big Data tends to forget that these data bases do not represent a representative

[5] We refer to the Oxford Dictionary definition of post-truth, where post-truth is viewed as referring to circumstances in which the objectivity of facts is less influential in shaping public opinion than emotionality and personal beliefs. It therefore indicates a rhetorical-persuasive strategy in which the subjective and passionate component prevails over the referential one.

sample of reality, but a selection made on the basis of conceptual reasons and practical limitations whose interpretation instead of helping to understand social phenomena leads to mystifications (Leonelli, 2018). The moment data are deposited in a database they lose their original meaning. A gap is thus created, because the audience is different from the one for which they were intended. By losing their meaning, they acquire another meaning. This is precisely what is known as the decontextualization of databases. It follows that if one does not have documentation that reflects the circumstances and limitations under which they were produced (so-called metadata), it becomes difficult to recontextualise the data.

Another problem concerns the issue of scientific interpretability and explanation. The text entitled "The Book of Why", written by Judea Pearl and Dana McKenzie, deals in an interesting way with the difference between a causal model and a data-driven approach, although the term data-driven is not explicitly used in the text. The text highlights what a model is able to do, and what a data-driven approach is not able to achieve. According to the authors there are a number of obstacles that separate a data-driven approach from the current state of traditional empirical research. The first obstacle is the adaptability or robustness of a model. Researchers have encountered and recognized that current machine learning systems are unable to recognize or react adequately to new circumstances for which they were not explicitly programmed or trained. The second obstacle relates to the issue of interpretability. At present, in fact, the machine learning models remain unable to explain and motivate their predictions. The third and last obstacle is due to the lack of understanding of the connection between cause and effect. The user of digital data should, for example, be able to answer questions of a manipulative nature such as "What would happen if?", "What would happen if we induced a certain event in the analysis?", and of a retrospective or explanatory nature based on counterfactuals such as "What would have happened if the individuals to whom the data refers had acted differently than they did?" A solution could be to employ the data-driven method in conjunction with the traditional theory-driven method, using the former approach for an initial exploration of the data in order to give direction to the more formal analysis or, conversely, using the theory-driven approach to get an indication of cause-and-effect in order to make the choice of data, on which a machine learning algorithm can then be used (Breiman, 2001).

The hypothesis that with the available computing power applied to Big Data it is possible to discover all the relevant correlations that are so decisive in producing predictive models as to be able to dispense with causation must, therefore, be taken with caution (Batini, 2020).

Digital data have become so pervasive that they are the main resource used in scientific research, the economy and society. With reference to scientific research, a data-centric approach is emerging in all scientific disciplines, addressing problems that in the past were considered difficult or impossible to tackle. Beyond the advantages and enormous opportunities provided by Big Data, we must be aware of the limits and consequent risks. In fact, the opacity of an algorithm, i.e. the ignorance of the assumptions on which it is based and the failure to adequately consider the limits of its applicability, leads to its incorrect use or to interpreting the predictions provided by the algorithm in terms that are not always scientifically grounded, and therefore leads to making decisions without the necessary justifications and explanations. In the situation where, digital data have become so pervasive as to be the main resource used in scientific research the social researcher should exploit the information potential of Big Data without negotiating his key role in the process.

The emergence of data-centrism highlights several challenges concerning not only the collection, classification, mathematical and computational tools developed to analyze Big Data, but also legal and ethical implications.

Data ethics can be considered as a new branch of ethics that studies and evaluates the moral issues related to data in their life cycle (from acquisition to use by the end user), to the algorithms that make use of data, to the related practices in order to formulate morally sound solutions and behaviors (Floridi, 2016). In particular, in order to clarify the ethical implications of the three phases of the data lifecycle, Floridi argues that ethical issues begin in the collection and analysis of large datasets and in issues related to the use of big data in biomedical and social science research. In this field, the most relevant issues concern the identification of individuals by means of techniques for extracting knowledge from data, the linking and integration of datasets, as well as the risk of violating not only individual privacy but also the "privacy of groups or communities", leading to risks of discrimination of groups on the basis of, for example, ethnicity or sexual orientation. In the ethics of algorithms, crucial challenges arise from the increasing complexity and autonomy of logarithms, i.e. the moral responsibility and accountability of both algorithm designers with respect to unforeseen and undesirable consequences (e.g. discrimination or promotion of anti-social content). The central themes of the ethics of practices (including professional ethics and deontology) are personal privacy, consent, and the uses of data after the primary use (secondary use) with the aim of defining an ethical framework that favors both the progress of data science and the protection of the rights of individuals and groups. Floridi's approach is interesting because it extends ethical reflection to the moral dimensions of all types of data, even to so-called raw data by shifting the focus

of ethical questions from what Floridi calls information-centric to data-centric ones (Floridi, 2016).

This highlights the need for ethical analyses to focus on the content and nature of computational operations on data – the interactions between hardware, software and data – rather than the variety of digital technologies that enable them; and emphasizes the complexity of the ethical challenges posed by data science. Social preferability should therefore be the guiding criterion within the data-centric approach in order to reconcile individual rights with collective rights.

References

Anderson, C. (2008), The end of theory: The data deluge makes the scientific method obsolete, *Wired magazine,* 16.7.

Arvidsson, A. (2016), Facebook and finance: on the social logic of the derivative, *Theory Culture & Society*, 33(6), 3-23.

Batini, C. (ed.) (2020), *La scienza dei dati,* http://hdl.handle.net/10281/295980

Breiman, L. (2001), Statistical modeling: the two cultures, *Statistical Science* ,16.3, 199-231.

Butler, D. (2013), When google got flu wrong*, Nature News,* 494.7436.

Callebaut, W. (2012), Scientific Perspectivism: A Philosopher of Science's Response to the Challenge of Big Data Biology, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 69-80.

Couldry, N., Mejias, A.U. (2019), Data colonialism: rethinking Big Data's relation to the contemporary subject, *Television & New Media*, 20, 336-49.

Elliott, K. C. (2012), Epistemic and methodological iteration in scientific research, *Studies in History and Philosophy of Science,* 43, 376-382.

Elliott, K.C., Cheruvelil, K.S., Montgomery, G.M., and Soranno, P.A. (2016), Conceptions of Good Science in Our Data-Rich World, *BioScience*, 66(10), 880-889.

Floridi, L. (2016), Information ethics: On the philosophical foundation of computer ethics, *Ethics and Information Technology*, 1, 37-56.

Giere, R. (2006), *Scientific Perspectivism*, Chicago, University of Chicago Press.

Gooding, D.C. (1990), *Experiment and the Making of Meaning*, Dordrecht & Boston, Kluwer.

Kitchin, R. (2014), Big Data, new epistemologies and paradigm shifts, *Big Data and Society,* 1(1) April-June.

Lazer, D. (2014), The parable of google flu: traps in big data analysis, *Science,* 343.6176, 1203-1205.

Leonelli, S. (2016), *Data centric biology: a philosophical study*, Chicago, Chicago University Press.
Leonelli, S. (2018), *La ricerca scientifica nell'era dei Big Data*, Meltemi Editore.
Massimi, M. (2012), Scientific perspectivism and its foes, *Philosophica,* 84, 25-52.
O'Malley, M., Elliott, K.C., Haufe, C., Burian, R. (2009), *Philosophies of funding*, Cell, 138, 611-615.
O'Neill, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Crown.
Pearl, J., Mackenzie, D. (2019), *The Book of why*, Penguin.
Pietsch, W. (2015), Aspects of Theory-Ladenness in Data-Intensive Science, *Philosophy of Science*, 82(5), 905–916.
Popper, K. (1934), *La logica della scoperta scientifica*, Torino, Einaudi.
Radder, H. (2006), The Philosophy of Scientific Experimentation: A Review, *Automated Experimentation,* 1(1), 2.
Ricaurte, P. (2019), Data epistemologies, The coloniality of Power, and resistance, *Television & New Media*, 20 (4), 350-65.
Waters, C. K. (2007), The Nature and Context of Exploratory Experimentation: An Introduction to Three Case Studies of Exploratory Research, *History and Philosophy of the Life Sciences,* 29(3), 275-284.
Williamson, J. (2004), A dynamic interaction between machine learning and the philosophy of science, *Minds and Machines*, 14(4), 539-554.