Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science

Abstract

I provide a philosophical perspective on the characteristics of data-centric research and the conceptualization of data that underpins it. The transformative features of contemporary data science derive not only from the availability of Big Data and powerful computing, but also from a fundamental shift in the conceptualization of data as research materials and sources of evidence. A *relational* view of data is proposed, within which the meaning assigned to data depends on the motivations and instruments used to analyze them and to defend specific interpretations. The presentation of data, the way they are identified, selected and included (or excluded) in databases and the information provided to users to re-contextualize them are fundamental to producing knowledge - and significantly influence its content. Concerns around interpreting data and assessing their quality can be tackled by cultivating governance strategies around how data are collected, managed and processed.

Keywords
data philosophy; data history; data-centric research; inference; data management; data curation; modelling.

There has been much debate around the extent to which the advent of big data and data-intensive methods is heralding the "end of theory" in scientific research, and the start of a "data-driven" approach. Yet a simplistic opposition between inductive and deductive procedures does not help to understand how data science is changing the research landscape. As philosophers have long shown, there is no such thing as 'raw data,' since data are forged and processed through instruments, formats, algorithms, and settings that embody specific theoretical perspectives on the world. Nor can there be direct, unmediated inference from data: data interpretation involves recourse to models and various other kinds of conceptual and material scaffolding (Callebaut, 2012). Moreover, big data is arguably not new. Sciences such as astronomy, meteorology, and taxonomy have long grappled with how to manage, order, and visualize large and complex datasets. These efforts spurred key developments in the skills, tools and institutions used to collect and manage data. These included: the international standardization of terminology via thesauri, registries and databases; the creation of guidelines and legislation for the governance of confidential data; and the emergence of techniques, including statistical methods and visualization tools, to integrate and sustain diverse data collected over long periods of time (Aronova et al., 2017). Rather than a break with the past, the emergence of data science and related computational approaches could be construed as the culmination of these efforts (November, 2013; Daston, 2017).

In what follows, I argue that what makes the contemporary manifestation of data science both novel and revolutionary to scientific knowledge production is not only the existence of large datasets or the application of powerful data analytics. It is also the result of a fundamental shift in the conceptualization of data as research materials and source of evidence, with which the research world and wider society are still struggling to come to terms. Ever since the creation of scientific journals such as *Philosophical Transactions of the Royal Society* in the 17th century, data have been conceptualized and managed as fundamentally private objects, which are owned by the scientists who produce them and have value only within well-delimited spaces of inquiry. In other words, only those familiar with the circumstances, methods, and goals of data generation would be expected to be able to interpret them and/or assess their validity. Within this approach, the usefulness of data lies in their function as evidence for a specified hypothesis.

This overarching perception has shifted in the last few decades, with data increasingly portrayed as powerful yet unpredictable objects, whose evidential value is not fixed and may increase the more data are shared and scrutinized across multiple contexts. Rather than a data-*driven* approach, we are witnessing the rise of a data-*centric* approach to research (Leonelli, 2016). This is characterized by three main features: (1) data are no longer by-products of administrative and research processes, but rather research outputs and valuable commodities in their own right; (2) efforts to mobilize, integrate, disseminate and visualize data are viewed as central contributions to discovery, since the more data travel, the higher the chance that they will acquire new significance and meaning; and (3) consultation of data resources, typically mediated by complex digital infrastructures and databases, constitutes the first step in any process of inquiry and plays a key heuristic role in determining future research directions. Within data-centric research, data have become public objects whose scrutiny is open to many different types of expertise – a phenomenon underscored by the increasing emphasis on making data FAIR (Findable, Accessible, Interoperable, and Reusable; Wilkinson et al., 2017).

What does this shift in status indicate about data as research components? It is tempting to think that the scientific significance of data lies in their context-independence. And yet, fifteen years of in-depth empirical studies of data practices have taught me that a spectroscopic image of animal tissue, a meteorological measurement, or a genomic sequence do not have fixed evidential value. How data are interpreted often changes depending on the skills, background knowledge, and circumstances of the researchers involved, which is why looking at the same dataset from a variety of viewpoints often yields new knowledge. Think of the famous case of the X-ray diffraction image known as "Photograph 51" and its different interpretations by Rosalind Franklin and James Watson.

Even more remarkably, studies of data re-use across contexts show that the expectations and abilities of those handling and mobilizing data determine what is regarded as 'data' in the first place (Leonelli, 2012; Borgman, 2015). Researchers make choices about which of the objects produced through their interactions with the world – whether they be experimental interventions, observation studies, or measurements – deserve the most attention as potential evidence for claims about phenomena or specific courses of action. Biologists, clinicians and plant breeders differ considerably in the data they will consider most useful towards studying gene-environment interactions (Leonelli, 2012, 2016); and there are many documented cases in archaeology, astronomy, biomedicine, and physics where objects considered as data at the start of an investigation no longer have that status by the end of it, or vice versa (Leonelli and Tempini, 2019). A set of photographs taken in a forest, for example, could constitute useful data for the study of phenomena as diverse as the morphological development of a given tree species, the symptoms of an infection, the effect of certain meteorological conditions on photosynthesis, and the presence of parasites in a specific location. Each of these interpretations is affected both by the physical features of the photos (definition, level of detail, focus of attention, color schemes) and by the manner in which whoever handles these objects accentuates their usability as data (for instance by zooming on a specific detail, adding metadata, and/or changing format to foster interoperability with other botanical data). Hence, while the features of the objects considered as data certainly shape their use and interpretation, it is often possible to obtain different information from the same objects depending on how these are managed and interpreted. A particular combination of interests, abilities and accessibility determine what is identified as data in each instance.

In light of these considerations, data analysis cannot be conceived as a purely objective process in which the meaning of data is uncovered through context-independent methods. Data can be used to represent various aspects of reality and each interpretation will depend on the specific circumstances of analysis, including the skills and technical premises that allow people and/or algorithms to organize and visualize data in a way that corroborates a certain conceptualization of reality. In other words, the interpretation of data is constantly mediated by the view point and abilities of those using it.

The conceptualization of data as objects with fixed and contextually-independent meaning is at odds with this key observation and engenders mistaken expectations relating to how data provide information on the world. One such expectation is the idea that there can be universal ways of measuring data quality and reliability. There is no underestimating the importance of methods for error detection and countering misinformation in contemporary data science, particularly in the wake of the replicability crisis and given the ease with which unreliable data sources can infiltrate the growing, complex ecosystem of online databases. Nevertheless, most existing approaches are tied to domain-specific estimations of what counts as quality and reliability—and for what purposes. These estimations cannot be easily transferred across fields, and sometimes even across specific cases of data use (Floridi and Illari, 2014). This is a big obstacle to the development of overarching checks for data quality and begs the question of whether producing such context-independent methods is the most useful way to tackle the problem.

An alternative approach is to recognize that data are not fixed representations of reality, and that it is impossible to regard only certain types of objects (such as numbers or symbols) as legitimate data. Instead, data are objects that are treated as potential or actual evidence for claims about phenomena in ways that can, at least in principle, be scrutinized and accounted for —a view I have called "relational" in my philosophical work (Leonelli, 2015). The meaning assigned to data depends on their provenance, their physical features and what these features are taken to represent, as well as on the motivations and instruments used to visualize them and to defend specific interpretations. The reliability of data thus depends first and foremost on the credibility and strictness of the processes used to produce and analyze them. This framework acknowledges that any object can be used as a datum, or stop being used as such, depending on the circumstances—a well-known consideration to anyone dealing with historical data, often held in forgotten archives and therefore reduced to meaningless objects. The presentation of data; the way they are identified, selected, and included (or excluded) in databases; and the information provided to users to re-contextualize them are fundamental to producing knowledge and significantly influence its content. The unwillingness to acknowledge the epistemic importance of data handling processes translates into an unwillingness to give these processes attention and document them so as to make them visible and open to constructive criticism. By contrast, the relational view acknowledges that objects regarded as data are often altered in their transit through different production, dissemination and reuse sites. For instance, changes in data format – as most obviously involved in digitalization, data compression or archival procedures— can have a significant impact on where, when, and who uses the data as source of knowledge. The relational view also explains how, depending on the research perspective interpreting it, the same dataset may be used to represent different aspects of the world. When considering the full cycle of scientific inquiry from the viewpoint of data production and analysis, it is at the stage of data *modelling* that a specific representational value is attributed to data (Figure 1; Leonelli, 2019a).

The relational view of data encourages care and attention to the history of data, highlighting their continual evolution and sometimes radical alteration, and the impact of this feature on the power of data to confirm or refute hypotheses. It explains the critical importance of documenting data management and transformation processes, especially with Big Data that transit far and wide over

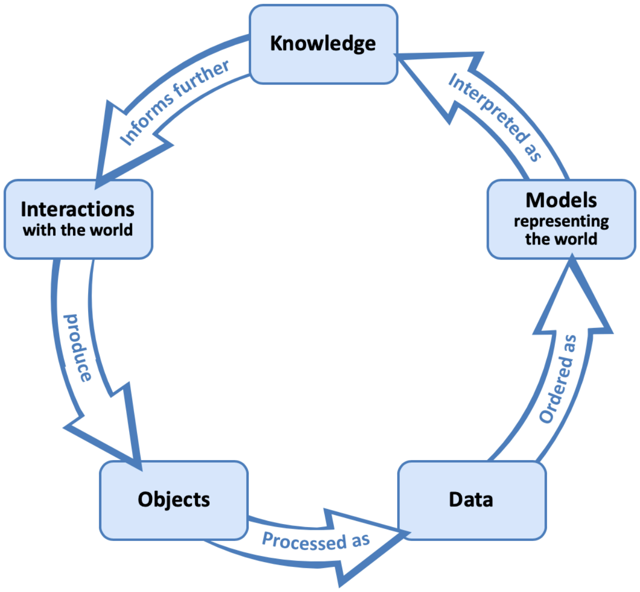

Figure 1: *Figure 1. The process of scientific inquiry according to the relational view of data (Leonelli, 2019a). The stages of inquiry include the set-up of an interaction between humans and the world, which is documented through the collection of objects; decisions about how to process such objects so that they can be credibly regarded as data; decisions about how data should be ordered and visualized, so as to function as representations for a phenomenon of interest (in other words, as "models"); and the extraction of knowledge from these representations, which in turn informs how future research is designed.*

digital channels and are grouped, analyzed, and interpreted in different ways and formats. It also explains why the rise of data-centrism involves the increasing acknowledgment of the expertise of those who produce, curate, and analyze data as indispensable to the effective use of Big Data within and beyond the sciences; and the inextricable link between social and ethical concerns around the potential impact of data sharing and scientific concerns around the quality, validity, and security of data (Boyd and Crawford, 2012; Leonelli, 2017; Tempini and Leonelli, 2019b).

Data governance, which I define as the strategies and tools employed to identify, manage, and disseminate data, thus becomes central to data interpretation. This has enormous implications for the priorities, organization and investment in data science, where data and related infrastructures are sometimes regarded as unstructured empirical fodder to the application of formal/computational analytic methods. Thus conceptualized, data-centrism encourages significant shifts in current research priorities and established hierarchies, subverting the traditional view of laboratory technicians, librarians, administrators, and database managers as marginal to knowledge production, and fostering attention and investment towards data management. It challenges existing ideas of research excellence, with funders and institutions moving away from evaluation systems focused solely on the impact of publications; and it stresses the importance of social sciences and humanities within data science training, to ensure that researchers acquire the critical skills necessary to examine the conceptual and social implications of data management strategies. Data scientists, and particularly those working with large and heterogeneous datasets, need to be able to identify and consider different ways of handling and valuing data, and explicitly articulate possible conflicts among them. Data infrastructures and analytic tools should encourage and facilitate this type of inferential reasoning. This will help to effectively balance the constraints and decisions internal to scientific reasoning with the broader landscape of opportunities, demands and limitations within which data science operates.

Acknowledgements

This research was supported by the European Research Council under award number 335925, and the Australian Research Council under award number DP160102989, and the Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Aronova, E., von Oertzen, C. and Sepkoski, D. (2017). Introduction: Historicizing Big Data. *Osiris* 32: 1-17.

Boyd, D. and Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15 (5): 662–679.

Borgman, C. (2015). *Big Data, Little Data, No Data.* Cambridge, MA: MIT Press.

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.

Daston, L. (2017). Science in the Archives. Chicago, IL: Chicago University Press.

Floridi, L. and Illari, P. (eds.). (2014). *The Philosophy of Information Quality.* Synthese Library 358. Cham, Switzerland: Springer.

Leonelli, S. (2012). When Humans Are the Exception: Cross-Species Databases at the Interface of Clinical and Biological Research. *Social Studies of Science* 42(2): 214-236.

Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science* 82: 810-821.

Leonelli, S. (2016a). *Data-Centric Biology: A Philosophical Study.* Chicago, IL: Chicago University Press.

Leonelli, S. (2016b). Locating Ethics in Data Science: Responsibility and Accountability in Global and Distributed Knowledge Production. *Philosophical Transactions of the Royal Society: Part A.* 374: 20160122.

Leonelli, S. (2019a). What Distinguishes Data from Models? *European Journal for the Philosophy of Science* 9: 22. https://doi.org/10.1007/s13194-018-0246-0

Leonelli, S. (2019b) *La Recherche Scientifique à l'Ère des Big Data: Cinq Façons Donc les Données Massive Nuisent à la Science, et Comment la Sauver.* Éditions Mimésis.

Leonelli, S. and Tempini, N. (eds). *Varieties of Data Journeys.* Manuscript in preparation.

November, J. (2015). *Biomedical Computing.* Baltimore, MD: Johns Hopkins Press.

Tempini, N. and Leonelli, S. (2018). Concealment and Discovery: The Role of Information Security in Biomedical Data Re-Use. *Social Studies of Science* 48(5): 663-690.

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18