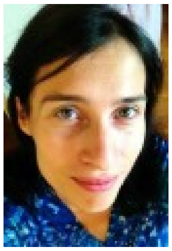
Sabina Leonelli: “What constitutes trustworthy data changes across time and space”

blogs.lse.ac.uk/impactofsocialsciences/2015/01/19/philosophy-of-data-science-series-sabina-leonelli/

1/19/2015

The next installment of the Philosophy of Data Science series is with **Sabina Leonelli**, *Principal Investigator of the ERC project, The Epistemology of Data-Intensive Science. Last year she completed a monograph titled “Life in the Digital Age: A Philosophical Study of Data-Centric Biology”, currently under review with University of Chicago Press. Here she discusses with* **Mark Carrigan** *the history of data-centric science and research practice and data’s relation to pre-existing and emerging social structures. Data types are produced by many stakeholders, from citizens to industry and governmental agencies, which means that what constitutes data, for whom and for which purposes is constantly at stake.*



Previous interviews in the Philosophy of Data Science series: Rob Kitchin, Evelyn Ruppert, Deborah Lupton, Susan Halford, Noortje Marres.

What is “data-intensive science”? How new is it?

I take data-intensive science to be any research enterprise where major efforts are devoted to the generation, dissemination, analysis and/or interpretation of data. Indeed, my preferred term to refer to this scientific approach is ‘data-centric’ rather than data-intensive, as a distinctive feature of such research is the high degree of attention and care devoted to data handling practices (which is not necessarily to the exclusion of theories, models, instruments, software and materials, since data practices are almost invariably intertwined with concerns about other components of research). Thus defined, data-centric science is definitely not new. This is clearly illustrated by the major data collection and curation efforts characterising 17th century astronomy and metereology and 18th century natural history – cases which, together with many others, are documented within the ‘Historicising Big Data’ working group which I visited last year at the Max Planck Institute for the History of Science in Berlin.

At the same time, the current manifestations of data-centric science have distinctive features that relate to the technologies, institutions and governance structures of the contemporary scientific world. For instance, this approach is typically associated to the emergence of large-scale, multi-national networks of scientists; to a strong emphasis on the importance of sharing data and regarding them as valuable research outputs in and of themselves, regardless of whether or not they have yet been used as evidence for a given discovery; the institutionalization of procedures and norms for data dissemination through the Open Science and Open Data movements, and policies such as those recently adopted by RCUK and key research funders such as the European Research Council, the Wellcome Trust and the Gates Foundation; and the development of instruments, building on digital technologies and web services, that facilitate the production and dissemination of data with a speed and geographical reach as yet unseen in the history of science. In my work, I stress how this peculiar conjuncture of institutional, socio-political, economic and technological developments have made data-centric science into a prominent research approach, which has considerably increased international debate and active reflection over processes of data production, dissemination and interpretation within science and beyond. This level of reflexivity over data practices is what I regard as the most novel and interesting aspect of contemporary data-centrism.

What are the epistemological issues raised by data-intensive science?

Some obvious issues, raised both within the sciences and the humanities, concern the notion of data itself and the patterns of reasoning and methods associated with them. What are data, and how are they transformed into meaningful information? What is the status of so-called raw data with respect to other sources of evidence? What constitutes good, reliable data? What role do theory and materials play in data-intensive research? What patterns of reasoning characterize this scientific approach? What difference do the scale (itself a multifaceted notion), technological sophistication and institutional sanctioning of widespread data dissemination make to discovery and innovation? These are issues investigated by my current ERC project 'The Epistemology of Data-Intensive Science', which analyses data handling across a range of disciplines including plant biology, biomedicine and oceanography. Philosophical analysis can help to address these questions in ways that inform both current data practices and the ways in which have been conceptualized within the social science and humanities, as well as by policy bodies and other institutions.

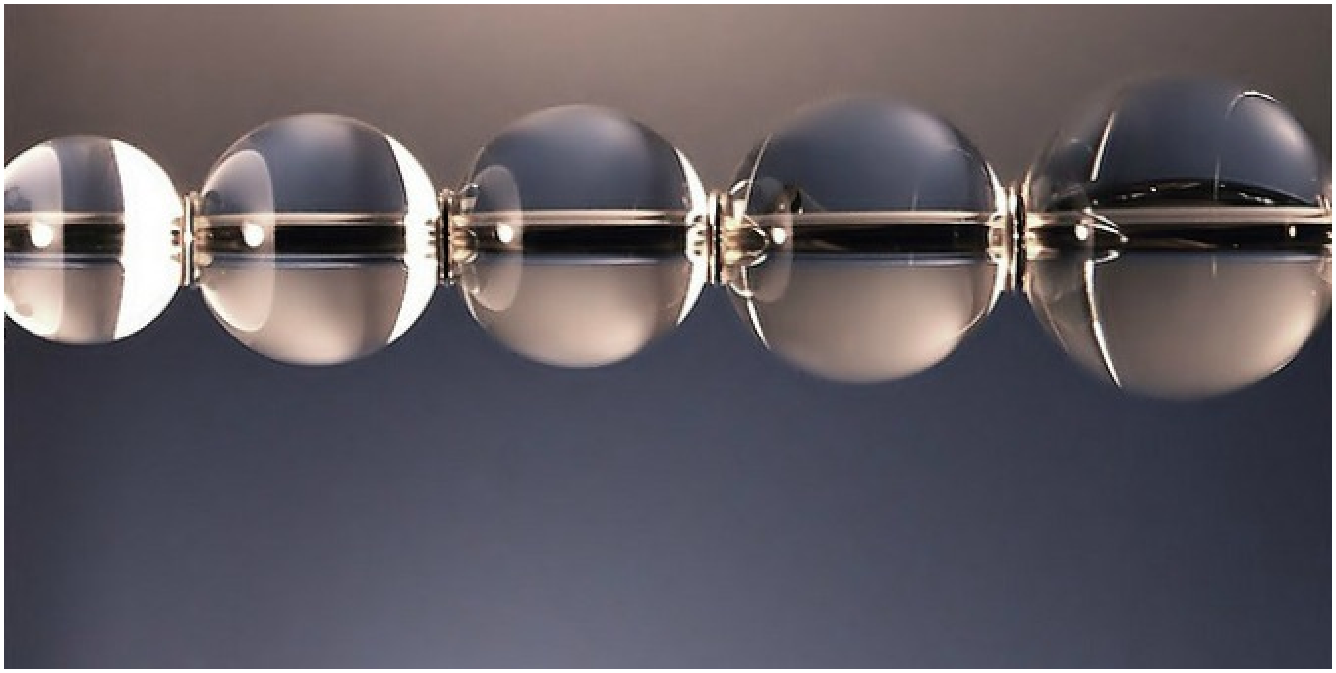


Image credit: David (Flickr, CC BY-SA)

The epistemological aspect that interests me most, however, is even more fundamental. Given the central role of data in making scientific research into a distinctive, legitimate and non-dogmatic source of knowledge, I view the study of data-intensive science as offering the opportunity to raise foundational questions about the nature of knowledge and knowledge-making activities and interventions. Scientific research is often presented as the most systematic set of efforts in the contemporary world aimed to critically explore and debate what constitutes acceptable and sufficient evidence for any given belief about reality. The very term 'data' comes from the Latin 'givens', and indeed data are meant to document as faithfully and objectively as possible whatever entities or processes are being investigated. And yet, data collection is always steeped in a specific way of understanding the world and constrained by given material and social conditions, and the resulting data are therefore marked by the historical circumstances through which they were generated: what constitutes trustworthy or sufficient data changes across time and space, making it impossible to ever assemble a complete and intrinsically reliable dataset. Furthermore, data are valued and used for a variety of reasons within research, including as sources of evidence, tokens of exchange and personal identity, signifiers of status and markers of intellectual property; and myriads of data types are produced by as many stakeholders, from citizens to industry and governmental agencies, which means that what constitutes data, for whom and for which purposes is constantly at stake.

This landscape makes the study of data into an excellent entry point to reflect on the activities and claims associated to the idea of scientific knowledge, and the implications of existing conceptualisations of various forms of knowledge production and use. This is nicely exemplified by an ongoing Leverhulme Trust Research Grant on the digital divide in data handling practices across developed and developing countries, particularly sub-Saharan Africa, which we are currently developing at Exeter – what constitutes knowledge, and a ‘scientific contribution’, varies enormously depending not only on access to data, but also on what is regarded as relevant data in the first place, and what capabilities any research group has to develop, structure and disseminate their ideas.

What are the implications of data-intensive science for the social sciences?

At a practical level, it constitutes an opportunity for social scientists to invest more time and energy in understanding the functioning of technologies geared towards data production, dissemination and analysis (such as complex data infrastructures, digital databases and software), their relation to pre-existing and emerging social structures and practices, and the ways in which they can be fruitfully and critically appropriated as research tools. It is also an occasion to revisit the importance of intertwining quantitative and qualitative data, which is particularly important at a time where regrettably few analysts work with both types of data. Spotting correlations through the analysis of so-called ‘big data’ is an exciting endeavor and excellent opportunity to devise new research directions. At the same time, the significance of such findings can only be assessed in relation to in-depth understandings of social dynamics and their history, which is typically garnered through qualitative methods such as interviews and ethnography. Just like in the natural sciences, where multi-disciplinary networks are increasingly valued, social scientists need to cooperate with each other in order to combine the qualitative and quantitative skills needed to work with big data; with computer scientists and statisticians, so as to deepen their understanding of the analytic tools and technologies available to handle and interpret data; and with the humanities, particularly history and philosophy, to ensure help with contextualizing and reflecting upon the conditions under which data are obtained, disseminated, processed and used.

This interview is part of an ongoing series on the Philosophy of Data Science. Previous interviews in the series: Rob Kitchin, Evelyn Ruppert, Deborah Lupton, Susan Halford, Noortje Marres.

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Sabina Leonelli *is Associate Director of the Exeter Centre for the Study of the Life Sciences (Egenis), Associate Editor of the journal History and Philosophy of the Life Sciences, and Principal Investigator of the ERC Starting Grant, [DATA_SCIENCE]: the Epistemology of Data-Intensive Science. Last year she completed a monograph titled “Life in the Digital Age: A Philosophical Study of Data-Centric Biology” in which her general perspective is articulated (under review with Chicago University Press). She can be found on Twitter @sabinaleonelli and the research group @DataScienceFeed.*

Mark Carrigan *is a sociologist based in the Centre for Social Ontology at the University of Warwick. He edits the Sociological Imagination and is an assistant editor for Big Data & Society. His research interests include asexuality studies, sociological theory and digital sociology. He’s a regular blogger and podcaster.*