Adrian Currie

Book review: Sabina Leonelli // data-centric biology: a philosophical study

Article (Published version) (Non-refereed)

Sabina Leonelli // Data-centric Biology: A Philosophical Study

Reviewed by Adrian Currie

Data-centric Biology: A Philosophical Study
Sabina Leonelli
Chicago: University of Chicago Press, 2016, £73.50/£24.50 (hardback/paperback)
ISBN 9780226416335/9780226416472

Experimental biology has witnessed an industrial revolution. Increases in computational power and investment from private industry, as well as the development of sequencing techniques, has enabled the emergence of high-throughput biology on a factory scale. 'Big data' has arrived. And undoubtedly it has changed the face of biology. But what's interesting about big data philosophically, and how is a philosophical perspective illuminative of it?

Here's a tempting thought: big data matters because it pursues data *for its own sake*. Investigation doesn't aim to formulate and test hypotheses. Instead, studies are data-driven—Baconian—we generate masses of data, then hunt for meaningful patterns. Big-data science is data-driven science: theory takes a back seat, and the experimental generation of data takes on a 'life of its own'.

This thought, says Sabina Leonelli in her arresting new book, is a red-herring. She argues that what matters is the social organization, institutions, and technologies that are required to facilitate the large-scale communication, integration, and dissemination of different kinds of data, generated in different ways, towards a variety of purposes. On Leonelli's view, the action lies in how groups of scientists and their allies facilitate 'data journeys': travels from lab, to database, to being utilized in often surprising, creative ways.

Leonelli focuses on the development of data-bases in model organism biology—plant systems in particular—but it's clear that her ideas have wide scope indeed, begging to be adapted to, and incorporated into, analyses of other areas of science:

> **[…] data centrism brings new salience to aspects of scientific practice that have always been vital to successful empirical research and yet have often been overlooked policy makers, funders, publishers, philosophers of science, and even scientists themselves, who have largely evaluated science in terms of its products […] rather tha in terms of the processes through which such results are eventually achieved. (p. 188)**

Leonelli analyses science's most seductive buzzwords—'open science' and 'big data'—with a sober eye, balanced between the promise of such ideas and their potential dangers.

I'll sketch Leonelli's view, before highlighting what I take to be the central points of departure between how she and more traditional philosophers think about evidence and data.

A central character in Leonelli's story is the *data-curator*. The data-curator's job is the creation of databases. Again there is a naïve view, handily dismissed. Perhaps the creation of a database supports experimental practice by allowing the straightforward dissemination of more-or-less raw information. I run an experiment, it generates data, which is subsequently published on a database, ready to be utilized. Another scientist accesses the data and applies it to her research, black-boxing the behind-the-scenes processes that brought the data to her. On such a view, databases play a passive role—'raw' data is communicated via the database. Leonelli paints a significantly more active role for databases, and data-curators:

> **Data do not easily flow along the channels devised for their dissemination and reuse but rather undertake unpredictable journeys full of obstacles, interruptions, and setbacks, which are addressed in creative and labour-intensive ways by the many researchers involved at each stage of travel. (pp. 169-170)**

The data-curator has two tasks. First, data must be decontextualized—abstracted from its place of origin; data

must be *packaged* to facilitate travel. This involves the construction and utilization of bio-ontologies. Local idiosyncrasies are stripped, resulting in standardized data, free to interact with results from other studies. However, the story doesn't end here, because if the data is to act as evidence—if it is to inform new studies—it must be *re*-contextualized. Here, the data-curator must develop practices to promote and capture 'metadata': information about the context in which the data was produced, and its subsequent travel. This underwrites understanding of the data's evidential potential.

The data-curator, and the design and maintenance of databases, then, is a necessary condition for large-throughput data to be used in a variety of contexts. Moreover, Leonelli argues that data curation itself generates theory: bio-ontologies constitute a rich body of theoretical, biological knowledge.

How is data recontextualized? Here, Leonelli appeals to the know-how that comes from scientific training and experience—labouring at the bench endows different scientists with embodied knowledge about the nature of data. Well-packaged data with the right metadata provides the materials enabling its appropriate repurposing to different ends. Because different scientists possess different know-how, a single set of data potentially generates a plurality of evidence. In light of this perspective, Leonelli further argues that scientific understanding is distributed. Its locus is at the population level; scientific understanding is not the possession of a single individual. As different scientists have different embodied knowledge—and thus different understandings—of what data *says*, this means that no one individual has complete understanding of some datum. Although Leonelli doesn't say this explicitly, it is tempting to read her as claiming that scientific evidence is *reconstructed* rather than *communicated*. That is, we shouldn't envision a scientist taking data from a database as a simple transfer of information from one context to another. Rather, the scientist, using her embodied knowledge and guided by metadata, constructs the relevant evidence in that context.

What *is* data, then? Leonelli argues that data is 'potential evidence', organized for the purposes of communication —for being the ingredients of new studies in new contexts. Data is pluripotent, constrained by metadata and the know-how of scientists. The necessity of embodied know-how in critically assessing data suggests that there is no algorithm for scientific integration: very human kinds of reasoning play a central role in the scientific project. Data, then, evolves and changes while travelling across contexts, filtered through data curation practices, and put to use towards different aims via different embodied understandings.

It is this dynamism that underwrites what I think is most interesting about Leonelli's view.

For me, Leonelli's central insight is the recognition that data is essentially historical. Although it is true that according to some analyses of evidence, timing matters (Popperian appeals to 'novelty', for instance), evidence in philosophy typically stands in a static, linear relationship to theories and hypotheses on the one hand, and the world on the other. For many of us, then, data and evidence are frozen in time, preserved in formalism. Leonelli's data, by contrast, is fully historicized: it is generated in a particular time and place, and subsequently travels into new contexts. Dynamic data has a history, knowledge of which is necessary for its purpose of being re-contextualized. That history, data's provenance, always matters. Data, again, is pluripotent: it is reusable, re-contextualizable. This feature explains why data curation is so central—it is only in virtue of our understanding of where data came from, and how it travelled, that we are able to successfully re-employ it in new contexts. It further challenges, as Leonelli argues, the very idea that one might distinguish between the 'content' and 'context' of evidence and data:

> **[…] invoking the term 'context' becomes a way to sweep worries about the complex and ever-changing nature of research environments under the carpet, thus separati the analysis of scientific claims and methods from concerns about who produces and handles them at any point in space and time, for which reasons, and subject to which constraints. (p. 180)**

Although I've been using the term 'context', perhaps science is better thought of as *situated* in Dewey's sense, as Leonelli urges. That is, instead of thinking of data as counting as evidence in particular contexts, we should think of it, evidence, and other aspects of science as embedded within, and inextricable from, a set of social, political, and technological practices.

I want to highlight two upshots from this historicity. First, recognizing that data travels, and that it is pluripotent, leads us to think about why it ends up where it does, and how we'd *like* it to end up. Historicity's partner is contingency: there is no inevitability about what big data science looks like, what it is used for, and how it is

practiced. The promise of such science, its capacity to democratize via the open exchange of ideas, to provide wide-ranging resources, to integrate and encourage different kinds of research, and so on can be undercut by the shear expertise required to utilize databases, various kinds of gate-keeping behaviours by scientists and data-curators, and limitations in funding, time, and will.

[…] data practices may well be used to reinforce, rather than challenge, current power relations in science: but whether this turns out to be the case depends on how da journeys are managed and financed and who is involved in their development […] data dissemination strategies risk acting as a magnifying glass for existing inequaliti and disparities in research, rather than as means through which differences and disagreements can be voiced and scientific pluralism can be harnessed to expand the evidential value of data. (pp. 164–5).

Leonelli's dynamic perspective provides a framework for highlighting these sociological and political issues within science. It gives us—scientists, philosophers, policy-makers—a way to pause and consider whether the emerging features of this science are as they should be. Leonelli emphasizes the need to fund and support good data curation practices, and thus to shift incentive structures for working scientists: many biologists are discouraged from participating in data curation work as it distracts them from better rewarded activities. If databases are, as Leonelli argues, an active part of scientific epistemology—one that produces unique theoretical knowledge—then this should be reflected in how scientists are incentivized.

Second, the historicity of data, it strikes me, provides a new perspective on some traditional philosophical problems. Consider *projectability*—which predicates can underwrite general claims? Or consider the related question of *external validity*—how do we shift from localized, experimental contexts to the wild world itself? Such questions often assume that predicates, and the content of evidence, are fixed. We look for solutions by, for instance, considering whether there are natural kinds that could license proper scientific predicates or show how, under the right conditions, experimental evidence attaches to entities, processes, or structures in the world.

But Leonelli's historicized, pluripotent data entails that evidential value and content are *not* fixed. This suggests a procedural account of the projectability of a predicate: we can export knowledge between contexts, and underwrite general claims, insofar as there is a justified process of contextualization and de-contextualization, with the relevant know-how in place. Instead of worrying about 'natural kinds', we examine the categorizing practices of scientists themselves. In a sense, we switch from an abstract epistemological problem to one of implementation. Leonelli's highlighting of bio-ontologies, and argument that these themselves capture theoretical knowledge—and the role of scientific know-how in re-contextualizing that knowledge—could be read as a convincing example of the sorts of processes that produce successful projections, that is, successful journeys.

Leonelli does not have much to say about *how* experimental evidence attaches to the world—and her analysis is too subtle to be read in either naïve realist or free-floating constructivist terms. And indeed, a more traditional philosopher might complain about this: how, in all of this detail, do we determine the relationship between the image science presents and the actual world? But I think the richness of her analysis does suggest an answer to this question: science tells us about the world *with difficulty*. You might complain that this isn't much of an answer, to which the retort is: yes, but the question was ill-put in the first place. Explaining why science works and how it generates empirical and theoretical knowledge requires a local, flexible, practice-oriented analysis. The dynamic, procedural themes running throughout Leonelli's book usefully shift our philosophical focus from the *abstracta* of understanding how science works from a frozen perspective to a more useful analysis of how scientists in fact get the job done (when they do) and—more critically—how they might do it better.

Adrian Currie
Centre for the Study of Existential Risk
University of Cambridge
ac2075@cam.ac.uk

Acknowledgements

The research in this review was supported by the European Commission, and the Templeton World Charity Foundation (the opinions expressed in this publication are those of the authors and do not necessarily reflect the views of TWCF).