MAKING SENSE OF BIOLOGY IN THE DATA AGE

Joeri Witteveen
Descartes Centre for the History and Philosophy of the Sciences and the Humanities,
Utrecht University
3512 BL Utrecht, The Netherlands
e-mail: j.witteveen@uu.nl

A review of
Data-Centric Biology: A Philosophical Study.
By Sabina Leonelli. Chicago (Illinois): University of Chicago Press. $105.00 (hardcover); $35.00 (paper). vi + 275 p.; ill.; index. ISBN: 978-0-226-41633-5 (hc); 978-0-226-41647-2 (pb); 978-0-226-41650-2 (eb). 2016.

If popular science books and magazines are to be believed, we are at the cusp of a major revolution in how science is done. For example, *Wired* magazine has notoriously claimed that the advent of computer-based mining for patterns in large datasets will ring in "the end of theory" (Anderson 2008) and mark "the death of the theorist" (Steadman 2013). It is easy (but necessary) to criticize inflated claims such as these for presenting an overly simplistic and unrealistic image of what science is and where it is heading. It is much harder, though, to provide a constructive account of how scientific epistemology in the age of data-intensive research should be understood. What are the implications of the superabundance of collected data for prevailing modes of science? Does the advent of digital, disembodied, computer-processable formats herald a fundamental change in how knowledge is acquired and circulated? And what about the role of theory in algorithmic mining and nonparametric modeling of large swaths of data?

In *Data-Centric Biology: A Philosophical Study*, Sabina Leonelli provides the first book-length study of what is novel about data-centric research in the life sciences. Her choice of the term "data-*centric*" in the title already signals that she is not a fellow traveler of those who claim that the mere increase in quantity of data ("*Big* Data") or the supposedly purely inductive method enabled by large datasets ("data-*driven* biology") is what distinguishes life sciences today from 30 years ago. Her focus on data-centricity encompasses a much wider range of practical, epistemic, institutional, technological, and socioeconomic developments that are relevant to understanding the contemporary nature of knowledge production in the life sciences. This is reflected in the volume's continuous engagement with the literature on social studies of science, the interweaving of interviews with key people in the field, and the author's insightful reflections on her participation in steering committees and policy debates.

Arguably, the subtitle of the book somewhat undersells this eclectic mix of approaches that can be found inside; the volume is much more than just a "philosophical study." However, Leonelli would probably respond that a "thick" and multifaceted perspective is exactly what we should require from a philosophical analysis of scientific inquiry. She distances herself from those approaches to philosophy of science that strip away all of the messy details until a supposedly "pure" core of epistemic and metaphysical commitments remains. Such a strategy amounts to throwing away the baby with the bathwater on the "philosophy of science in practice" approach that the author sides with. The idea is that we cannot make sense of science without paying detailed attention to the constraints and opportunities provided by the people, platforms, and policies that shape quotidian scientific practice.

It is easy to sympathize with this practice-based approach. But it also sets a high bar: a successful account of this kind calls for profound scholarly and practical knowledge of several fields of study. Fortunately, Leonelli delivers on the promise. Her book is a laudable achievement in integrating insights from a methodologically permissive and pluralistic outlook. Given the volume's ambition and breath of scope, it is unsurprising that it is more convincing in some areas than others. As I sketch some of the key themes of the book in the remainder of this review, I will make some critical observations on where the attempt

The Quarterly Review of Biology, September 2017, Vol. 92, No. 3
For permission to reuse, please contact journalpermissions@press.uchicago.edu.

at making sense of data-centric biology should be amended or could be further improved.

Making Data Travel

The volume has a clear structure that organizes the seven main chapters into three parts. The first part (Chapters 1 and 2) provides an insightful sketch—informed by relevant case studies—of how contemporary data-gathering and dissemination practices emerged, and of the curatorial, technological, and institutional scaffolds that maintain them. In Part Two (Chapters 3 to 5), Leonelli probes and embellishes this complex sociotechnological picture of data-centric biology by raising philosophical questions about the nature of data, experiment, and theory. Part Three closes the book with a discussion of the implications for science, science policy, and philosophy (Chapters 6 and 7).

At the start of the first part, the author explains her well-motivated decision to focus throughout the book on data-gathering and dissemination practices surrounding model organism research (in particular in the study of *Arabidopsis thaliana*, which has been the focus of much of Leonelli's earlier research). In the first chapter we also meet what might well be considered the hero of the volume: the database curator. Here and elsewhere the author emphasizes the pivotal importance of this underappreciated expert, tasked with the development and management of databases. She warns against the tendency among database users to think of curators as second-class scientific citizens, whose only job is to be serviceable to scientists. Leonelli drives home that curation is neither simply a service nor a one-way street. In their choices of how to "package" data, curators make scientifically informed decisions with important epistemic implications. Their choice of data format, keywords, metadata, and means of data representation directly affect how data are circulated and accessed. Close collaboration and consultation between curators and scientists is indispensable to ensure successful circulation of data among users from different epistemic cultures. Putting blind trust in serviceable curators is not a viable long-term strategy—curators ought to be respected as scientific equals. The author therefore stresses the need to develop better systems of credit for curatorial activities. This last point feeds into a broader discussion in the second chapter, about the (all-to-easily overlooked) institutional scaffolding that brings curators, regulators, and users into contact.

A particularly eye-catching feature of these first two chapters is the vocabulary it is couched in. The pages are permeated with metaphorical language about "making data travel," "data journeys," "packaging tools for data," "vehicles for data," and other associations with movement and travel. At the end of the first chapter Leonelli defends this deliberate choice of language by noting that speaking of "journeys" and "travels" highlights that the circulation of data is hardly ever unproblematic and straightforward. As with human journeys, data journeys involve complex infrastructures, breakdowns, dead-ends, unexpected material and communicative obstacles, and changes of vehicle along the way. These complications make travel talk far more appropriate than the smoothness that is implied by popular rhetoric of "data deluges" or "data flows." The latter metaphors are not innocent; they suggest that computer-based solutions will allow us to sieve information from readily available streams of data. The author's travel metaphors instead accentuate the complex interpersonal, infrastructural, and institutional scaffolding required to bring data into circulation, and the repercussions this has for how (and which) data are presented and interpreted.

Yet, travel talk has its own limitations. For example, unlike human travel, data travel is hardly ever linear and trained on a specific destination. The role of the curator is to make data available and accessible for many potential users. Also, the extent of modification and change in format that regularly occurs as part of data journeys has no obvious equivalent in human travel. Leonelli is aware that taking travel talk to seriously can lead us astray. But despite her own cautions not to take travel metaphors at face value, I will argue that they do engender a somewhat skewed discussion in the second part of the book.

The Medium and the Message

The three chapters that constitute Part Two address the questions: What Counts as Data? What Counts as Experiment? What Counts as Theory? In the first of these, the author develops what she calls a "relational framework" of data, on which data are characterized as material objects or artifacts that serve as potential evidence for claims about phenomena, and that are appropriately packaged for travel (p. 78). The last two conditions underscore that what counts as data is situation-dependent. The first condition, however, raises questions. Surely, the idea of data being material fits nicely with the metaphorical picture of "stuff" that can be made to travel, but arguably it is the metaphor that distorts the philosophical analysis here. There is no question that data can be embedded in and extracted from material, but Leonelli goes further by suggesting that we should *identify* materials with data and holds that an object *constitutes* its content. Yet, a popular catchphrase notwithstanding, the medium is typically *not* the message. Indeed, the author's repeated talk of "extracting" data from objects, of "formatting" data to expedite dissemination, and of data that "change medium" all betray an implicit commitment to distinguishing between data and their material bearers.

She recognizes a related objection when she raises the question "If data change format as they travel . . . how can we account for them being in some respect 'the same objects'?" (p. 82). In answering this question, she draws an analogy between the ontology of data and that of biological individuals. The idea is that just like organisms maintain integrity while continuously shedding and growing cells, data maintain integrity while moving from one medium and format to another. In other words, a "data lineage" is spatiotemporal sequence of "data objects" just like an organism is a spatiotemporal sequence of cells and tissues. But how does this answer the question at stake? The fact that two cells belong to the same organism at different times does not entail that they are the same type of cell. Similarly, saying that two "data objects" are part of a "data lineage" does not help to answer why we are dealing with the same dataset at the different points in time. The solution seems obvious: instead of speaking of object *as* data, we should say that objects *carry* data. This would also allow us to make sense of claims about the same type of data being distributed through different tokens.

Unfortunately, Leonelli closes off this option when she asserts that "the distinction between token and type does not help to make sense of the epistemology of data" (p. 83). She argues that in the light of the incredible diversity in media, formats, and contexts in which data are produced and consumed "the distinction between types and tokens becomes meaningless" (p. 84). This seems overstated. The complexity of data travel surely raises questions about how, in practice, curators and database users determine whether they are dealing with tokens of the same data. But this is not a good reason to abandon the basic distinction that renders such inquiries meaningful.

Another recurring theme in the author's discussion of the nature of data is the point that traditional—often theory-centric and physics-focused—philosophical discussions of data versus phenomena and models fail to consider the complex interpretative and coordinative steps involved in disseminating data. This point is well taken. But instead of concluding that this continuous processing of data blurs the difference with phenomena and models, Leonelli keeps her eye trained on finding a clean definition of "data." This focus on definitions is somewhat surprising in the light of her practice-oriented approach to the philosophy of science. More generally, one might wonder whether her philosophical interests are best served by asking the three "what is X?" questions that guide the middle chapters of the book. A different series of philosophical questions—along the lines of "how does X work (in data-centric biology)?" or "what role does X serve?"—could have put definitions more explicitly in the service of practices, while placing the latter in the philosophical limelight.

Classificatory Theory?

The discussion in Chapter 5, What Counts as Theory?, illustrates this mild tension between the author's overarching aims and the specific questions she pursues. The main goal of this chapter is to specify how data-centric biology differs from theory-driven research, while steering clear of the fantasy that it is purely inductive and theory-free. Leonelli's answer is that the dissemination of data through databases relies on classifications that constitute theories. Her prime example of "classificatory theories" in this domain are bio-ontologies: machine-processable networks of terms that denote the relations between entities and processes in a domain. Bio-ontologies such as gene ontology (GO) were already discussed in the first part of the book, where the author pointed to their pivotal role in circulating data from different areas of biology by specifying the relation between processes and phenomena. The formalization of these relations requires experimentalists and curators to formulate keywords and definitions that capture cross-disciplinary meanings. When Leonelli returns to these practices of formulating definitions in Chapter 5, she argues that they constitute a veritable form of theory-making. But it remains unclear why this is so. The act of building an ontology rather seems like a process of negotiating a set of widely accepted theoretical terms for the purposes of communication across fields. The GO works well as a classificatory tool because it rests on a condensation of well-confirmed theory knowledge, not because it sticks its neck out to make daring new predictions, as theories do. In reading this chapter, one sometimes gets the impression that the author's defense of classificatory theory hinges on a false contrast between mere classifications "as lists" (p. 128) and classificatory theory as "an achievement that goes well beyond listing" (p. 130). But is hardly controversial that classifications are more than mere listings, and that the practice of classification can be heavily theory-laden. It is not clear that ontologies are special in this regard.

The broader worry, though, is that even if we accept Leonelli's characterization of ontologies as classificatory theories, this hardly helps to illuminate what is novel and distinctive about the uses of theory in data-centric biology. In Leonelli's view, classificatory theory pervades biology, from the representation of normal stages in developmental biology to the Linnaean system of classification in 18th-century botany (p. 126). This leaves readers wondering what is so new about the role theory in the context of data-intensive biology. It seems plausible that some uses of theory in this area are novel and unorthodox, say, in the deployment of sophisticated mining techniques and the application of nonparametric tools of analysis. A further philosophical discussion of these aspects of data-centric biology would help to respond more directly and persuasively to the "theory-free" rhetoric.

Lessons for the Future

The third and final part of the volume identifies expectations and hurdles for biology (Chapter 6) and outlines some general lessons for philosophy of science (Chapter 7) in the light of the preceding chapters. I will not go into detail on either of these chapters, but note that they articulate some valuable lessons about the broader implications of data-centric science. Chapter 6 is particularly exciting. After presenting a lucid account of varieties of data integration, the author shows its broader relevance by offering a socially engaged account of the opportunities and dangers related to integrating and circulating data. From philosophy to (implicit) policy advice in one chapter: this is where Leonelli's practice-oriented philosophy of science truly shines.

REFERENCES

Anderson C. 23 June 2008. The end of theory: the data deluge makes the scientific method obsolete. https://www.wired.com/2008/06/pb-theory/.

Steadman I. 25 January 2013. Big data and the death of the theorist. http://www.wired.co.uk/article/big-data-end-of-theorist.