Published in Special Issue of *Studies in the History and the Philosophy of the Biological and Biomedical Sciences: Part C 43:1, 2012*

Editorial: Making sense of data-driven research in the biological and biomedical sciences

Sabina Leonelli, University of Exeter, s.leonelli@exeter.ac.uk

Debates about the emergence, significance and long-term impact of 'big data' have become ubiquitous across most scientific disciplines. Thanks to new technologies for generating and storing information, data production is said to have increased on an unprecedented scale, together with the expectation that data should be made freely accessible to global research networks as a common resource from which new knowledge can be harvested (as often emphasised by editorials in *Nature* and *Science* over the last decade). The biological and biomedical sciences are no exception, and are in fact widely seen as fields where the difficulties and potential rewards of handling big datasets are most pronounced. This is partly due to the huge diversity in the types of available data and organisms from which data are taken. The complexity of biological and biomedical phenomena is also seen as particularly challenging, especially given the rise of systemic/integrative approaches wishing to understand how entities and processes at different levels of organisation, ranging from genes and cells to organisms, populations and ecosystems, shape and construct each other. Further, the social and economic stakes in these areas are enormous, not only because of the vast investment of resources devoted to them by both the public and the private sectors, but also due to the tantalizing promises attached to biological and biomedical discovery. Biologists are expected to yield an understanding of life that helps humans to make sense of themselves and their role in their environment, while clinicians are charged with providing improved

diagnoses and effective treatments for all diseases, as well as informing the public about how best to preserve their health and well-being. Big data have been widely publicised as an important tool towards reaching these goals, as illustrated by the public portrayal of personalised medicine and whole genome sequencing as revolutionary ways to identify and prevent potential pathologies. Last but not least, what makes biology and biomedicine exemplary of big data science is the sophistication of the technologies developed over the last two decades to produce, store and disseminate particularly genomic data. Thanks to high-throughput technologies such as sequencing and micro-array analysis, the activity of data gathering in molecular biology has become increasingly automated and technology-driven, resulting in the production of billions of data-points in need of a biological and/or biomedical interpretation. Evidence-based medicine has also fostered an attention shift to large-scale data collection, by placing data obtained through clinical trials at the top of the hierarchy of evidence. A consequence of this focus on data collection has been a commitment to data sharing, with massive efforts devoted to building infrastructures (databases, repositories, biobanks) for the dissemination of data across research locations, disciplines and specific projects.

This special issue aims to examine critically epistemic and historical claims on big data biology, such as the ones reported above. In particular, we probe the underlying vision of big data as fostering a new mode of scientific research, which some commentators refer to as 'data-driven'. The sense that data-driven science constitutes a novel approach, or even a new paradigm, has been widely discussed in the scientific literature;[1] and at the same time, questions have been raised about the extent to which current research is driven by data, and

[1] See for instance Hey, T., et al. (Eds.). (2009) *The fourth paradigm: data-intensive scientific discovery.* Redmond, WA: Microsoft Research.

whether this is desirable and/or productive.[2] A general characterisation of data-driven methods is hard to achieve, given the wide range of activities and epistemic goals currently subsumed under this heading. Still, two key features are often highlighted as pillars of this approach: one is the intuition that induction from existing data is being vindicated as a crucial form of scientific inference, which can guide and inform experimental research; and the other is the central role of machines, and thus of automated reasoning, in extracting meaningful patterns from data. Notably, both champions and critics of data-driven research recognise that even these two features are highly controversial and difficult to apply to specific research contexts. For a start, it is not clear what is meant by induction. Using data for the purposes of discovery can happen in a variety of ways, and involves a complex ensemble of skills and methodological components. Inferential reasoning from data is tightly interrelated with specific theoretical commitments about the nature of the biological phenomena under investigation, as well as with experimental practices through which data are produced, tested and modelled. For instance, extracting biologically meaningful inferences from high-throughput genomic data may involve reliance on theories about gene expression and regulation, models of the biological processes being regulated and familiarity with the instruments and organisms from which data were obtained. In this context, 'inductive' clearly does not mean 'hypothesis-free'; nor can automated reasoning be seen as a substitute to human judgement based on specific expertise and laboratory experience. This methodological and epistemic complexity plays a key role in data-driven research, and needs to be taken into

[2] E.g. Allen, J. F. (2001) Hypothesis, induction and background knowledge. Data do not speak for themselves. *BioEssays*, 23, 861—862; Valencia, A. (2002) Search and retrieve. Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? *EMBO Reports*, 3, 396—400; Kell, D.B., and Oliver, S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era, BioEssays, 26(1), 99-105; Weinberg, R. A. (2010) Point: hypotheses first. *Nature*, 464, 678—678. More references can be found in the papers in this issue.

account when attempting to pinpoint its epistemic characteristics as an emerging mode of inquiry.

Philosophers and historians of science are well-positioned to investigate the history and current epistemic role of inductive and automated reasoning, and more generally to assess whether data-driven research constitutes something special and new, and in which respects, if at all, it differs from other modes of inquiry. This intuition is what drove the participants to this special issue to meet at the University of Exeter in April 2010. The discussion at Exeter revolved around the following three questions, which in turn drive the arguments expressed in the resulting papers:

1. How do practices and technologies for data collection, dissemination and use affect the production of scientific knowledge?

2. What is the role of theories and hypotheses within research practices that are currently referred to as data-driven, and what are the relationships more generally between data-driven and hypothesis-driven research?

3. If data-driven research constitutes a distinctive mode of knowledge production, how can it be characterised, and how innovative is it with respect to existing or past scientific practices?

The first paper in the collection, by Staffan Müller-Wille and Isabelle Charmantier, questions the novelty of data-driven approaches by highlighting the crucial role that information processing techniques played already in the 18th century in shaping the production of scientific knowledge about the natural world. Müller-Wille and Charmantier document the strategies used by Carl Linnaeus to compile, archive, and revise data on new species that were pouring into his studio from all corners of the planet. Their careful analysis of these

strategies, which included adequate nomenclature, lists, index cards, book annotations and filing cabinets, illuminates the parallels between Linneaus' situation and the current information overload in biology; most notably, the importance of processing techniques and classification tools in directing future research, by establishing ontological commitments and generating a scaffold for new question and problems to be formulated and understood. A revealing example is the very concept of genus, which provides a key ontological component of what biologists and naturalists call the 'natural system', yet was initially invented by Linneus as a 'place-holder', to be slowly filled with relevant information (p.X). Jumping forward in time, Miguel Garcia-Sancho analyses the consequences of introducing new computing technologies – most notably, the minicomputer – for the research goals and practices of biologists working on the nematode *C. elegans* throughout the 1980s. He argues that data production techniques shaped the formulation of hypotheses about the genetic programme of *C. elegans*, thus expanding the epistemic role of data from empirical evidence to conceptual as well as organisational platform for research on the nematode. Sabina Leonelli and Rachel Ankeny extend this argument to the present day by considering the epistemic impact of computational tools, and particularly of community databases for online data dissemination, within contemporary model organism biology. They explore the ways in which databases affect the use of data in experimental biology and the conceptualisation of organisms that underlies those practices. This analysis explicitly stresses the social context – in this case, the collaborative ethos – in which data-driven methods have developed and flourished. It also emphasises the relation between the digital and the material aspects of research: the physical circumstances in which data are collected deeply affect the ways in which they can be interpreted and thus constitute crucial information to determine their evidential value.

Peter Keating and Alberto Cambrosio explore yet another crucial dynamic within data-driven research: that is, its controversial and context-specific intertwining with hypothesis-driven research, when the latter is taken to encompass statistical analysis. Their paper focuses on one specific type of data, microarrays, and their clinical use within oncology. In this domain, 'hybridisation between statistical hypothesis testing and algorithm-driven data analysis is underway' (p. X); and the scale and automation of microarray data collection and analysis pose serious challenges to basic tenets of statistical analysis, such as the choice of comparative methods, class predictors (such as survival rates) and sample size. Through their analysis of the interaction between two styles of practice (bioinformatics and biostatistics) in making clinical sense of microarray studies, Keating and Cambrosio draw attention to the increasingly critical role of data-driven methods within translational research, and particularly the fledging field of personalised medicine; and the complex array of expertises necessary to build bioinformatics tools for data handling. These themes also animate the paper by Maureen O'Malley and Orkun Soyer, which explores the practices used in system biology to make sense of large-scale datasets. While Keating and Cambrosio use the term 'hybridisation', O'Malley and Soyer prefer the broader notion of 'integration', which in their analysis comes in three main forms: methodological, explanatory and data integration. As suggested by the fact that data feature prominently in only one of these types of integration, O'Malley and Soyer see the dichotomy between data- and hypothesis-driven methods as a restrictive and inaccurate representation of biological practices. They argue that the notion of integrative systems, which involve combining 'a range of approaches for the purposes of exploring a general question about a biological system' (p.X), is a better way to capture what is new in contemporary biology, where integration 'has become something that happens not just occasionally or accidentally, but an activity that must be aimed at systematically and achieved consistently' (p. X). This optimistic reading of data-intensive science as one piece

of a much broader puzzle is counterbalanced by Ulrich Krohs' more critical assessment of the extent to which contemporary biology is driven by increasingly entrenched technologies and standards for data production, which can be integrated with existing practices only at the cost of restricting the scope of exploratory research. Krohs coins the term 'convenience experimentation' to describe the ways in which the increasing automation of data collection and the strategic pursuit of 'low-risk' experimentation, especially in the so-called 'omic' disciplines, has resulted in a re-focusing of modelling and experimental efforts on 'lab-mediated description of the molecular makeup of biological systems' (p.X).

By addressing the twin aspects of integration and experimentation in contemporary biology, O'Malley, Soyer and Krohs provide some of the philosophical foundations from which to ponder the role of theory in data-driven research. Werner Callebaut targets this issue in a direct and provocative manner, by wondering whether data-driven biology lacks a guiding vision or 'big picture' - and therefore the capacity to explicitly formulate and evaluate its long-term impact on our understanding of the natural world. Callebaut rejects the idea that there is no theory in data-driven biology, even if such theory is presently hard to explicitly formulate; importantly, he also rejects the notion that theory needs to take the form of one unifying vision on which all biologists agree, and which provides a unique way to understand the world. Callebaut proposes scientific perspectivism as a remedy against such expectations, and argues that data-driven biology in its broadest characterisation, including data-intensive fields such as system biology, multiscale modelling and biocomplexity, does (and needs to continue to) generate multiple theoretical perspectives, each of which capture some aspects of the complex, multiscale nature of biological phenomena.

As already evident from this brief overview, the take-home lessons offered in this special issue are diverse and varied, and yet several threads emerge as common results of all of these lines of investigation. The collection as a whole vividly illustrates the several meanings,

epistemic roles and material forms that can be attached to the very notion of 'data'; and the difficulties in evaluating the implications of data-driven methods for biology and biomedicine at a time when the potential of these methods is far from being fully realised. Bruno J. Strasser observes in his commentary how several of the papers, as well as his own work, trace historical continuities between natural history and contemporary biology, especially concerning the role of standards, material samples and hypotheses in shaping scientific knowledge. Wild claims about the novel and revolutionary nature of data-driven science are thus being questioned, while the links to and intersections with existing lines and methods of inquiry are emphasised by all of the papers. Yet, contributors also agree that several features of data-driven research are innovative and have the potential to greatly affect science in the future. In his overview, Strasser singles out three elements that he believes to be unique to data processing in contemporary data-driven science: 'the analysis of data is carried out by researchers with different disciplinary backgrounds than those who produce it, the analysis is heavily dependent on statistical tools, and the analysed data come from the laboratory, not the field' (p.X). In her discussion of the more philosophically focused contributions, Jane Calvert also highlights as novel the role played by interdisciplinary tools and statistical and mathematical modelling in data-driven biology; and reflects on system and synthetic biology as fields that feature prominently in this collection, notably because they exemplify the extent to which 'technological developments can result in conceptual ones' (p.X). As evident in Calvert's commentary, contributors to this issue do not see the large scale of data collection and dissemination efforts as the main locus of innovation in data-driven biology, and are not too impressed by the staggering numbers associated to the data deluge. Rather, the scale of data-driven science becomes epistemologically interesting insofar as it affects and challenges the ways in which science is organised and practiced, and especially the forms of collaboration, division of labour and integrative strategies (of models, data, theories,

software) set up to deal with such an accumulation of resources. The increasing quantity of data, and the parallel development of technologies to produce and manage them, is engendering a shift in the quality and conceptual content of research by affecting how scientists evaluate, compare, interpret and re-use available datasets. In other words, data-intensive methods are changing what counts as good science in ways that are cannot be captured by simply attributing primacy to data over theory (and models, technological instruments, methods, values and goals) as motors of research.

The above set of insights is of course only a starting point for further research. This special issue constitutes the first concerted attempt to make sense of data-driven science within the fields of history and philosophy of science, and I hope that it will spark further interest in this fascinating area, resulting in an ever improving understanding of the practices, methods and epistemology characterising 21st century biology and biomedicine. In closing, I wish to acknowledge funding from the British Academy and the Economic and Social Research Council, which made it possible to hold the conference whose results are collected in this special issue; and to warmly thank all the authors in this collection, the other participants to the Exeter conference, and my colleagues in Egenis and elsewhere (especially John Dupré, Annamaria Carusi, James Griesemer, Giovanni Boniolo and Mary Morgan) for contributing intellectual energy and wisdom to the overall project of understanding data-driven science.

Sabina Leonelli
ESRC Centre for Genomics in Society (Egenis)
University of Exeter
Byrne House
St Germans Road, EX4 4PJ Exeter
s.leonelli@exeter.ac.uk

15 July 2011