Studies in History and Philosophy of Biological and Biomedical Sciences 43 (2012) 85–87

Contents lists available at SciVerse ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Discussion

Data-driven sciences: From wonder cabinets to electronic databases

Bruno J. Strasser

Yale University, Program in the History of Science and Medicine, Section of the History of Medicine, P.O. Box 208015, New Haven, CT 06520-8015, USA

When citing this paper, please use the full journal title *Studies in History and Philosophy of Biological and Biomedical Sciences*

Even by the journal's own standards, this was a wild claim. In July 2008, *Wired* magazine announced on its cover nothing less than "The End of Science". It explained that "The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data".[1] Such claims about the emergence of a new "data-driven" science in response to a "data deluge" have now become common, from the pages of *The Economist* to those of *Nature*.[3] Proponents of "data-driven" and "hypothesis-driven" science argue over the best methods to turn massive amounts of data into knowledge. Instead of jumping into the fray, I would like to historicize some of the questions and problems raised by data-driven science, taking as a point of departure the three rich papers by Isabelle Charmantier and Staffan Müller-Wille on Linnaeus' information processing strategies, Sabina Leonelli and Rachel Ankeny on model organisms databases, and Peter Keating and Alberto Cambrosio on microarray data in clinical research. That a historical approach is warranted is made clear by the remark of the great book historian Robert Darnton that "every age was an age of information, each in its own way" (Darnton, 2000, p. 1). In particular, perceptions of an "information overload" (or a "data deluge") have emerged repeatedly from the Renaissance though the early modern and modern periods and each time specific technologies were invented to deal with the perceived overload (Ogilvie, 2003; Rosenberg, 2003). This commentary will explore the similarities and differences between past and present data-driven life sciences, from early modern natural history to current post-genomics.

Renaissance naturalists were no less inundated with new information than our contemporaries. The expansion of travel, epitomized by the discovery of the New World, exposed European naturalists to new facts that did not fit into the systems of knowledge inherited from the Greeks and Romans. This prompted those interested in understanding the natural world to devise new methods for managing this data, such as note-taking strategies, and new systems of classification (Blair, 2010; Ogilvie, 2006). Ironically, as Charmantier and Müller-Wille point out, these methods and systems, which were meant to tame the information overload, made it possible to accumulate even more data. But accumulation was usually only a mean to an end. These early naturalists established collections, which included specimens, drawings, and texts, so that they could compare these items systematically and draw from the comparisons conclusions about the natural world. In general, they were not testing specific hypotheses, but trying to bring order to the bewildering diversity of natural forms by examining large amounts of collected "data". This tradition continues to be central in natural history to the present day. As George Gaylord Simpson, the leading American paleontologist of the twentieth century, made clear in 1961, natural history, and taxonomy in particular, was the "science that is most explicitly and exclusively devoted to the ordering of complex data" (Simpson, 1961, p. 5). What is striking about Simpson's definition is not only that he chose the "ordering of complex data" as the most essential element of natural history, but also how similar his definition is to current characterizations of the supposedly unprecedented data-driven sciences. This should come as no surprise since, for several centuries, the natural historical sciences have fundamentally been data-driven sciences.

But was natural history driven by data alone? Most likely not, because natural history has never been free of ontological assumptions. For example most naturalists assume the existence of natural groups. As Charmantier and Müller-Wille show, Linnaeus who struggled with a data deluge of his own creation and devised numerous note-taking methods to deal with it, could only do so because he began with a hypothesis about the genus categories he used to organize his data. In other words, Linnaeus may have been driven by his data, but his approach was not exclusively data-driven. This conclusion, however, is insufficient to distinguish early modern approaches to data with contemporary ones. Indeed, as Keating and Cambrosio show in their paper, modern day biostatisticians analyzing cancer microarray data were equally driven by various hypotheses. For example, the determination of the sample size needed to produce statistically significant results required researchers to make an hypothesis about the number of classes that the data might reveal. In other words, they too were guided by

E-mail address: bruno.strasser@unige.ch

[1] *Wired*, July 16, 2008. cover.

[2] A cover and special report of the *Economist* was entitled: "The data deluge and how to handle it" *The Economist*, 394(8671), 2010; *Nature* had its cover on "Big Data", *Nature*, 455(7209), 2008.

1369-8486/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsc.2011.10.009

ontological assumptions about the natural world. Similarly, as Leonelli and Ankeny show, there are no ''raw'' data about model organisms in the databases, only data ''forced'' into ontological categories defined by convention (and not necessarily universally agreed upon), for example by the Gene Ontology Consortium. To be sure, past and present data-driven sciences have never been purely data-driven, but have always rested on a combination of hypothesis-driven and data-driven methods.

If there is any novelty to be found in today's data-driven science, perhaps it might derive from the amount and nature of the data. There is no early modern equivalent to the petabytes of digital data stored in computerized databases. Or is there? Comparing amounts of information across worlds that had different technologies to manage it is pointless. The comparison of data quantities across the analog—digital divide is particularly meaningless. Indeed, there is no common metric to compare today's petabytes of scientific data with yesterday's analog images, for example, of scientific objects. A single drawing by Ernst Haeckel of an embryo (or by a pathologist of a cancerous cell), digitalized at atomic resolution, would contain far more data than all of today's digital scientific data. Given that the amounts of analog and digital data are incommensurable, it is more useful to compare the *dynamics* of data accumulation within research fields which relied on analog and digital data, respectively. In each field, the amount of data, whether ''analog'' animal specimens or ''digital'' genome sequences, eventually surpassed the capacity to store and analyze these data. For example, much, if not most, of Louis Agassiz's collections at the Museum of Comparative Zoology, have remained in closed boxes, unanalyzed, like much of today's digital data. Similarly, his complaints about limited museum space and curatorial personnel sound remarkably like present day complaints about limited computer storage size and processing capacity (Winsor, 1991).

Even if one cannot simply juxtapose *amounts* of analog and digital data, one can attempt to compare and contrast the *nature* of the data used by the different data-driven sciences. Naturalists have often relied on data in the form of collected specimens stored in museums or herbaria. They collected, compared, and computed (without computers) their data just as contemporary researchers do. But one difference is immediately apparent. Earlier data was actually embodied in things, but today's data is entirely virtual. Or so we think. Again, Charmantier and Müller-Wille's paper allow us to revisit this simple assumption for the early modern period. Sure, Linnaeus collected plants, and parts of plants, by himself and through his extended network of students and correspondents. But he also collected massive amounts of descriptions of plants, as drawings or as texts. His data thus spanned the entire ontological range from material things to abstract representations (Latour, 1999). Similarly, Georges Cuvier not only worked on the specimens of the *Museum d'Histoire Naturelle*, but also on the numerous drawings of his ''paper museum'' (Rudwick, 2000). The early modern naturalist's data were not restricted to organisms, for they included a wide range of material and abstract entities, all of which were used as ''data''. In this respect, contemporary data-driven research is no different. As Leonelli and Ankeny's paper reminds us, the model organism databases used for data-driven research contain not only a wealth of experimental data, but also links to mutant organisms held in genetic stock centers, cell lines, and DNA clones. These physical objects too are part of today's data, which is no less diverse than the data of naturalist collections.

An overriding concern among data-driven sciences, past and present, has been the production and enforcement of standards. Because comparative approaches are so crucial to data-driven sciences, the uniformity of the data has been essential. For data to serve as a basis for grouping or dividing samples into different classes, they need to refer to the same property in all samples. Taxonomic systems for example, a key product of data-driven sciences, have rested on narrow definitions of standard taxonomic characters, as well as of collection and preservation techniques. In order to work within Linnaeus's taxonomic system, one had to adopt his definition of sexual characters, or the data produced by the observation of specimens would not be comparable to those of other observers. The same problem, on an even larger scale, has presented itself to the curators of modern model organism databases. This standardizing task was all the more daunting in that the data had been produced in a variety of experimental settings by a number of different laboratories. In order to accomplish their task, database curators addressed the most basic requirement for standardization, the creation of a common ontology, as Leonelli and Ankeny point out. Shared conventions regarding what kind of things exist in the world, and how to name them, were a prerequisite for the production of standards about how to produce data about these things and how to describe them. Database curators also attempted to standardize the experimental practices used for producing the data they would eventually receive. Yet even when technical reproducibility of the data was made possible, through the enforcement of strict experimental and nomenclature standards, there remained the problem of biological reproducibility. This point is made particularly clear by Keating and Cambrosio in the case of using microarray data in oncology. Even when technical reproducibility was achieved, a feat in and of itself, the bewildering biological variability of tumor cells severely limited the comparability of microarray data from different samples. In the nineteenth century, naturalists overcame a similar problem by agreeing that a ''type specimen'', the first specimen used to describe a species, would define the species, whether it happened to be typical or not (Daston, 2004). Whether researchers using microarrays will adopt a similar solution is too early to say.

So is there anything new after all in contemporary data-driven science or is it just a reinvention of natural history? Three features still stand out as potential differences in today's research: the analysis of the 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 analyzed data come from the laboratory, not the field. On the first point, Hooker, Darwin, Cuvier and all other modern and early modern naturalists who theorized about the data they collected had an intimate knowledge about its origins. In the plains of Lapland, in the Galapagos Islands, or in the quarries around Paris, they had personally turned plants, animals, and fossils into data (Browne, 1995; Endersby, 2008; Rudwick, 2005). Even armchair collectors, working in museums, typically had been field collectors earlier in their careers and thus had personal experience of how the data they received had been produced. They knew how imperfectly the data available in collections represented the diversity of nature or how unrepresentative the collected sample might be. But today, as Keating and Cambrosio make so clear, data are turned into knowledge by bioinformaticians and biostatisticians, most of whom have no first hand experience of producing the experimental data they are analyzing. This has contributed to an exaggerated trust in the quality and comparability of the data and to many irreproducible results. For this very reason, model organism databases are curated by experimentalists, not bioinformaticians, as Leonelli and Ankeny point out, although this only partially solves the problem. The point that bears emphasizing here is that current discussions about data-driven science focus more on the amount of data and methods of analysis than on the quality of the data. Repeated claims that more data will produce more complete knowledge of the world ignore the fact that any data-driven conclusion is only as good as the data it began with. Early modern naturalists seem to have been more careful in selecting the specimens they brought into their collections than modern day experimentalists. Contemporary databases managers, well aware of this problem, have increasingly sought to hire data analysts with some experience in the production of the experimental data, or to assemble teams of experimentalists and bioinformaticians, promoting the formation of a hybrid culture of experimentalists and data analysts.

A second important novelty of contemporary data-driven science is the omnipresence of statistical methods. All data-driven sciences, from early modern natural history to contemporary post-genomics, have essentially been comparative, identifying similarities, differences, patterns, and clusters. But only in the late twentieth century did data become grist for statistical mills. Although often associated with the rise of experimentalism, quantification was also embraced in natural history, at least in the twentieth century. George Gaylord Simpson and Anne Roe's (1939) *Quantitative Zoology*, for example, presented an introduction to the use of statistical methods in comparing data from skeletons and fossils (Simpson et al., 1939; Hagen, 2003). However, naturalists from Linnaeus to Simpson and beyond have continuously claimed the right to use subjective judgments in their analyses of data. Their avowed reliance on this epistemic value set them apart from experimentalists who cherished some form of objectivity (Strasser, 2010). Today's data-driven sciences have rejected these naturalists' inclination toward subjective judgment and aimed to replace it by a statistical kind of objectivity. This epistemic move has been reinforced by the changing relationships between collections of data and collections of physical objects. Although data, anything from numbers to images, have generally been thought to refer to physical objects, recently they have increasingly come to stand for the physical objects themselves. For example, a growing number of studies on the evolutionary history of life have been based solely on sequencing a minuscule portion of a species' genome, a practice that makes those evolutionists attached to the knowledge of whole organisms cringe. The stakes are as much epistemic as professional: they bear on *who* is the most legitimate producer of knowledge, the museum collector (the clinician, or the molecular biologist) or the statistician analyzing the data.

A third difference between post-genomics and all earlier natural history lies in the production of the data for these data-driven biomedical sciences by experimentalists in the laboratory not naturalists in the field. This matters for a number of reasons, but primarily because the moral economy of data exchange has been very different in natural history than in the experimental sciences, with some exceptions (Strasser, 2011). All data-driven sciences have relied on large collections of data provided by numerous researchers. In the natural history tradition, it has been customary for professional and amateur naturalists to donate their specimens to a museum, where they could serve as the basis for comparative work. Individual naturalists could become authors of papers or monographs even though the data on which their work was based had been provided by others. In the experimental sciences, on the other hand, data have been considered private, and only interpretations and limited sets of supporting data were made public. The idea that one person could claim authorship for the analysis of another person's experimental data has met much resistance among experimentalists until very recently. In some tightly-knit model organism communities, more communal forms of exchanges were prevalent in the twentieth century, especially when the communities were still small and led by a few charismatic researchers. Thus, in the late twentieth century, bringing experimentalists to participate in the collective production of knowledge by sharing their data, as naturalists had done for so long, proved to be a challenge. Although a change in the moral economy of data sharing is currently underway in the experimental sciences, for the time being data sharing is still achieved by scientific journals enforcing mandatory data deposition in public databases as a requirement for publication.

To conclude, it is mainly because the experimental sciences took the upper hand over natural history in the late nineteenth century and have since come to dominate the public perception of science that data-driven research is now perceived as a novel feature of twenty-first century science. Natural history had been "data-driven" for many centuries before the proponents of post-genomics approaches and systems biology began to claim the radical novelty of their methods. As I have argued here, many of what are claimed as novel features of contemporary data-driven science have parallels among earlier natural history practices. However, as this commentary has tried to make clear, there are nonetheless important differences between past and present data-driven sciences. Most significantly, much of contemporary biomedical research represents a new hybrid of naturalist and experimentalist approaches. Today's databases are as important to the experimentalists as museums were (and are) to the naturalists. Combining the data-driven and the hypothesis-driven, the comparative and the exemplary, the experimental and natural historical, current life sciences seem indeed headed in a new direction. Yet it is not one that should simply be described as "data-driven". It is far more illuminating to unpack this notion and explore the historical similarities and differences between past and present data-driven sciences, than to assume it constitutes yet another revolution in the history of science.

Acknowledgments

I would like to thank Alberto Cambrosio, Helen Curry, Staffan Müller-Wille, Rachel Ankeny and Sabina Leonelli for critical comments.

References

Blair, A. (2010). *Too much to know: Managing scholarly information before the modern age*. New Haven: Yale University Press.

Browne, E. J. (1995). *Charles Darwin: A biography—Voyaging*. New York: Knopf.

Darnton, R. (2000). An early information society: News and the media in eighteenth-century Paris. *American Historical Review, 105*, 1–30.

Daston, L. (2004). Type specimens and scientific memory. *Critical Inquiry, 31*, 153–182.

Endersby, J. (2008). *Imperial nature: Joseph Hooker and the practices of Victorian science*. Chicago: University of Chicago Press.

Hagen, J. B. (2003). The statistical frame of mind in systematic biology from quantitative zoology to biometry. *Journal of the History of Biology, 36*, 353–384.

Latour, B. (1999). *Pandora's Hope—Essays on the reality of science studies*. Cambridge: Harvard University Press.

Ogilvie, B. W. (2003). The many books of nature: Renaissance naturalists and information overload. *Journal of the History of Ideas, 64*, 29–40.

Ogilvie, B. W. (2006). *The science of describing: Natural history in renaissance Europe*. Chicago: University of Chicago Press.

Rosenberg, D. (2003). Early modern information overload. *Journal of the History of Ideas, 64*, 1–9.

Rudwick, M. J. S. (2000). George Cuvier's paper museum of fossil bones. *Archives of Natural History, 27*, 51–68.

Rudwick, M. J. S. (2005). *Bursting the limits of time: The reconstruction of geohistory in the age of revolution*. Chicago: University of Chicago Press.

Simpson, G. G. (1961). *Principles of animal taxonomy*. New York: Columbia University Press.

Simpson, G. G., & Roe, A. (1939). *Quantitative zoology; numerical concepts and methods in the study of recent and fossil animals*. New York: McGraw-Hill Book Company.

Strasser, B. J. (2010). Laboratories, museums, and the comparative perspective: Alan A. Boyden's serological taxonomy, 1925–1962. *Historical Studies in the Natural Sciences, 40*, 149–182.

Strasser, B. J. (2011). The experimenter's museum: Genbank, natural history, and the moral economies of biomedicine. *Isis, 102*, 60–96.

Winsor, M. P. (1991). *Reading the shape of nature: Comparative zoology at the Agassiz museum: Science and its conceptual foundations*. Chicago: University of Chicago Press.