Paper accepted for publication in the Inaugural Issue of *Big Data & Society*, 2014.

What Difference Does Quantity Make? On the Epistemology of Big Data in Biology

Sabina Leonelli, Department of Sociology, Philosophy and Anthropology & Exeter Centre for the Study of the Life Sciences (Egenis), University of Exeter, UK

s.leonelli@exeter.ac.uk

Abstract

Is big data science a whole new way of doing research? And what difference does data quantity make to knowledge production strategies and their outputs? I argue that the novelty of big data science does not lie in the sheer quantity of data involved, but rather in (1) the prominence and status acquired by data as commodity and recognised output, both within and outside of the scientific community; and (2) the methods, infrastructures, technologies, skills and knowledge developed to handle data. These developments generate the impression that data-intensive research is a new mode of doing science, with its own epistemology and norms. To assess this claim, one needs to consider the ways in which data are actually disseminated and used to generate knowledge. Accordingly, this paper reviews the development of sophisticated ways to disseminate, integrate and re-use data acquired on model organisms over the last three decades of work in experimental biology. I focus on online databases as prominent infrastructures set up to organise and interpret such data; and examine the wealth and diversity of expertise, resources and conceptual scaffolding that such databases draw upon. This illuminates some of the conditions under which big data need to be curated to support processes of discovery across biological subfields, which in turn highlights the difficulties caused by the lack of adequate curation for the vast majority of data in the life sciences. In closing, I reflect on the difference that data quantity is making to contemporary biology, the methodological and epistemic challenges of identifying and analyzing data given these developments, and the opportunities and worries associated to big data discourse and methods.

Keywords: big data epistemology; data-intensive science; biology; databases; data infrastructures; data curation; model organisms.

1. Introduction

Big data have become a central aspect of contemporary science and policy, due to a variety of reasons that include both techno-scientific factors and the political and economic roles played by this terminology. The idea that big data are ushering in a whole new way of thinking, particularly within the sciences, is rampant – as exemplified by the emergence of dedicated funding, policies, and publication venues (such as this journal). This is at once fascinating and perplexing to scholars interested in the history, philosophy and social studies of science. On the one hand, there seems to be something interesting and novel happening as a consequence of big data techniques and communication strategies, which is however hard to capture with traditional notions such as 'induction' and 'data-driven' science (partly because, as philosophers of science have long shown, there is no such thing as direct inference from data, and data interpretation typically involves the use of modelling techniques and various other kinds of conceptual and material scaffolding).[1] On the other hand, many sciences have a long history of dealing with large quantities of data, whose size and scale vastly outstrip available strategies and technologies for data collection, dissemination and analysis (Gitelman 2013). This is particularly evident in the life sciences, where data gathering practices in subfields such as natural history and taxonomy have been at the heart of inquiry since the early modern era, and have generated problems ever since (e.g. Johnson 2012, Müller-Wille and Charmantier 2012).

So what is actually new here? How does big data science differ from other forms of inquiry, what can and cannot be learnt from big data, and what difference does quantity make? In this paper, I discuss some of the central characteristics typically associated to big data, as conveniently summarised within the recent book *Big Data* by Viktor Mayer-Schönberger and Kenneth Cukier (2013), and I scrutinize their plausibility in the case of biological research. I then argue that the novelty of big data science does not lie in the sheer quantity of data involved, though this certainly makes a difference to research methods and results. Rather, the novelty of big data science lies in (1) the prominence and status acquired by data as scientific commodity and recognised output both within and beyond the sciences; and (2) the methods, infrastructures, technologies and skills developed to handle (format, disseminate, retrieve, model and interpret) data. These developments generate the impression that data-intensive research is a whole new mode of doing science, with its own epistemology and norms. I here defend the idea that in order to understand and critically evaluate this claim, one needs to analyze the ways in which data are actually disseminated and used to generate knowledge, which I refer to as 'data journeys'; and consider the extent to the current handling of big data fosters and validates their use as evidence towards new discoveries.[2]

Accordingly, the bulk of this paper reviews the development of sophisticated ways to disseminate, integrate and re-use data acquired on model organisms

[1] For a review of this literature, which includes seminal contributions such as Hacking (1992) and Rheinberger (2011), see Bogen (2010).

[2] This idea, though articulated in a variety of different ways, broadly underscores also the work of Sharon Traweek (1998), Geoffrey C. Bowker (2001), Christine Borgman (2007), Karen Baker and Francois Millerand (2010) and Paul Edwards (2011),

such as the small plant *Arabidopsis thaliana*, the nematode *Caenorhabditis elegans* and the fruit-fly *Drosophila melanogaste*r (including data on their ecology, metabolism, morphology and relations to other species) over the last three decades of work in experimental biology. I focus on online databases as a key example of infrastructures set up to organise and interpret such data; and on the wealth and diversity of expertise, resources and conceptual scaffolding that such databases draw upon in order to function well. This analysis of data journeys through model organism databases illuminates some of the conditions under which the evidential value of data posted online can be assessed and interpreted by researchers wishing to use those data to foster discovery. At the same time, model organism biology has been one of the best funded scientific areas over the last three decades, and the curation of data produced therein benefitted from much more attention and dedicated investments than data generated in the rest of the life sciences and biomedicine. Considering the challenges encountered in disseminating this type of data thus also highlights the potential problems involved in assembling data that have not received comparable levels of care (i.e. the vast majority of biological data).

In my conclusions, I use these findings to inform a critique of the supposed revolutionary power of big data science. In its stead, I propose a less sensational, but arguably more realistic, reflection on the difference that data quantity is making to contemporary biological research, which stresses both continuities with and dissimilarities from previous attempts to handle large datasets. I also suggest that the natural sciences may well be the area that is least affected by big data, whose emergence is much more likely to affect the political and economic realms – though not necessarily for the better.

2. The Novelty of Big Data

I will start by considering three ideas that, according to Mayer-Schönberger and Cukier (2013) among others, constitute core innovations brought in by the advent of big data in all realms of human activity, including science. The first idea is what I shall label *comprehensiveness*. This is the claim that the accumulation of large datasets enables scientists to ground their analysis on several different aspects of the same phenomenon, documented by different people at different times. According to Mayer-Schönberger and Cukier, data can become so big as to encompass *all* the available data on a phenomenon of interest. As a consequence, big data can provide a comprehensive perspective on the characteristics of that phenomenon, without needing to focus on specific details.

The second idea is that of *messiness*. Big data, it is argued, push researchers to embrace the complex and multifaceted nature of the real world, rather than pursuing exactitude and accuracy in measurement obtained under controlled conditions. Indeed, it is impossible to assemble big data in ways that are guaranteed to be accurate and homogeneous. Rather, we should resign to the fact that "big data is messy, varies in quality, and is distributed across countless servers around the world" (ibid., 13) and welcome the advantages of this lack of exactitude: "With big data, we'll often be satisfied with a sense of general

direction rather than knowing a phenomenon down to the inch, the penny, the atom" (ibid.).[3]

The idea of messiness relates closely to the third key innovation brought about by big data, which Mayer-Schönberger and Cukier call the 'triumph of *correlations*'. Correlations, defined as the statistical relationship between two data values, are notoriously useful as heuristic devices within the sciences. Spotting that fact that when one of the data values changes, the other is likely to change too, is the starting point for many a discovery. However, scientists have typically mistrusted correlations as a source of reliable knowledge in and of themselves, chiefly because they may be spurious – either because they result from serendipity rather than specific mechanisms, or because they are due to external factors. Big data can override those worries. Mayer-Schönberger and Cukier give the example of Amazon.com, whose astonishing expansion over the last few years is at least partly due to their clever use of statistical correlations among the myriad of data provided by their consumer base in order to spot users' preferences and successfully suggest new items for consumption (ibid., 52). In cases such as this, correlations do indeed provide powerful knowledge that was not available before. Hence, big data encourage a growing respect for correlation, which comes to be appreciated as more a informative and plausible form of knowledge than the more definite, but also more elusive, causal explanation. In Mayer-Schoenberger and Cukier's words: "the correlations may not tell us precisely *why* something is happening, but they alert us *that* it is happening. And in many situations this is good enough" (ibid., 14).

These three ideas have two important corollaries, which shall constitute the main target of my analysis in this paper. The first corollary is that big data makes reliance on small sampling, and even debates over sampling, unnecessary. This again seems to make sense *prima facie*: if we have all the data about a given phenomenon, what is the point of pondering which types of data might best document it? Rather, one can now skip that step and focus instead on assembling and analysing as much data as possible about the phenomenon of interest, so as to generate reliable knowledge about it: "big data gives us an especially clear view of the granular; subcategories and submarkets that samples can't assess" (ibid., 13). The second corollary is that big data is viewed, through its mere existence, as countering the risk of bias in data collection and interpretation. This is because having access to large datasets makes it more likely that bias and error will be automatically eliminated from the system, for instance via what sociologists and philosophers call 'triangulation': the tendency of reliable data to cluster together, so that the more data one has, the easier it becomes to cross-check them with each other and eliminate the data that look like outliers (Wylie 2002; Denzin 2006).

Over the next few sections, I show how an empirical study of how big data biology operates puts both of these corollaries into question, which in turn compromises the plausibility of the three claims that Mayer-Schönberger and

[3] Incidentally, the idea of comprehensiveness may be interpreted as clashing with the idea of messiness when formulated in this way. If we can have all the data on a specific phenomenon, then surely we can focus on understanding it to a high level of precision, if we so wish? I shall return to this point below.

Cukier make about the power of big data – at least when they are applied to the realm of scientific inquiry. Let me immediately state that I do not intend this analysis to deny the widespread attraction that these three ideas are generating in many spheres of contemporary society (most obviously, big government) and which is undoubtedly mirrored in the ways in which biological research is being re-organised since at least the early 2000s (which is when technologies for the high-throughput production of genomic data, such as sequencing machines, started to become widely used). Rather, I wish to shed some clarity on the gulf that separates the hyperbolic claims made about the novelty of big data science from the challenges, problems and achievements characterising data handling practices in the everyday working life of biologists – and particularly the ways in which new computational and communication technologies such as online databases are being developed so as to transform these ideas into reality.

3. Big Data Journeys in Biology

For scientists to be able to analyse big data, those data have to be collected and assembled in ways that make it suitable to consider them as a single body of information (O'Malley and Soyer 2012). This is a particularly difficult task in the case of biological data, given the highly fragmented and pluralist history of the field. For a start, there are myriads of epistemic communities within the life sciences, each of which uses a different combination of methods, locations, materials, background knowledge and interest to produce data. Furthermore, there are vast differences in the types of data that can be produced and the phenomena that can be targeted. And last but not least, the organisms and ecosystems on which data are being produced are both highly variable and highly unstable, given their constant exposure to both developmental and evolutionary change. Given this situation, a crucial question within big data science concerns how one can bring such different data types, coming from a variety of sources, under the same umbrella.

To address this question, my research over the last eight years has focused on documenting and analysing the ways in which biological data – and particularly 'omics' data, the quintessential form of 'big data' in the life sciences – travel across research contexts, and the significant conceptual and material scaffolding used by researchers to achieve this. For the purposes of this paper, I shall now focus on one case of big data handling in biology, which is arguably among the most sophisticated and successful attempts made to integrate vast quantities of data of different types within this field for the purposes of advancing future knowledge production. This is the development of model organism databases between 2000 and 2010.[4] These databases were built with the immediate goal of storing and disseminating genomic data in a formalized manner, and the longer-term vision of (1) incorporating and integrating any data available on the biology of the organism in question within a single resource, including data on physiology, metabolism and even morphology; (2) allowing and promoting cooperation with other community databases so that the available datasets

[4] Investigations of how other types of databases function in the biological and biomedical sciences, which also point to the extensive labor required to get these infrastructures to work as scientific tools, have been carried out by Hilgartner (1995), Hine (2006), Bauer (2008), Strasser (2008), Stevens (2013) and Mackenzie and McNally (2013).

would eventually be comparable across species; and (3) gathering information about laboratories working on each organism and the associated experimental protocols, materials and instruments, thus providing a platform for community building. Particularly useful and rich examples include FlyBase, dedicated to *Drosophila melanogaster*; WormBase, focused on *Caenorhabditis elegans*; and The Arabidopsis Information Resource, gathering data on *Arabidopsis thaliana*. At the turn of the 21st century, these were arguably among most sophisticated community databases within biology. They have played a particularly significant role in the development of online data infrastructures in this area and continue to serve as reference points for the construction of other databases to this day (Leonelli and Ankeny 2012). They therefore represent a good instance of infrastructure explicitly set up to support and promote big data research in experimental biology.

In order to analyse how these databases enable data journeys, I will distinguish between three stages of data travel, and briefly describe the extent to which database curators are involved in their realisation.

Stage 1: De-contextualisation

One of the main tasks of database curators is to de-contextualise the data that are included in their resources, so that they can travel outside of their original production context and become available for integration with other datasets (thus forming a big data collection). The process of de-contextualisation involves making sure that data are formatted in ways that make them compatible with datasets coming from other sources, so that they are easy to analyse by researchers who see them for the first time. Given the above-mentioned fragmentation and diversity of data production processes to be found within biology, there tends to be no agreement on formatting standards for even the most common of data types (such as metabolomics data, for instance; Leonelli et al 2013). As a result, database curators often need to assess how to deal with specific datasets on a one-to-one basis. Despite constant advances, it is still impossible to automate the de-contextualisation of most types of biological data.

Formatting data to ensure that they can all be analysed as a unique body of evidence is thus exceedingly labour-intensive, and requires the development of databases with long-term funding and enough personnel to make sure that data submission and formatting is carried out adequately. Setting up such resources is an expensive business. Indeed, debate keeps raging among funding agencies about who is responsible for maintaining these infrastructures. Many model organism databases have struggled to attract enough funding to support their de-contextualisation activities. Hence, they have resorted to include only data that had been already published in a scientific journal – thus vastly restricting the amount of data hosted by the database – or that were donated by data producers in a format compatible to the ones supported by the database (Bastow and Leonelli 2010). Despite the increasing pressure to disseminate data in the public domain, as recently recommended by the Royal Society (2012) and several funding bodies in the UK (Levin et al, in preparation), the latter category comprises a very small amount of researchers. Again, this is largely due to the labour-intensive nature of de-contextualisation processes. Researchers who

wish to submit their data to a database need to make sure that the format that they use, and the meta-data that they provide, fit existing standards – which in turn means acquiring updated knowledge on what the standards are and how they can be implemented, if at all; and taking time out of experiments and grant-writing. There are presently very few incentives for researchers to sacrifice research time in this way, as data donation is not acknowledged as a contribution to scientific research (Ankeny and Leonelli 2015).

Stage 2: Re-Contextualisation

Once data have been de-contextualised and added to a database, the next stage of their journey is to be re-contextualised - in other words, to be adopted by a new research context, in which they can be integrated with other data and possibly contribute to spotting new correlations. Within biology, re-contextualisation can only happen if database users have access not only to the data themselves, but also to information about their provenance – typically including the specific strain of organisms on which they were collected, the instruments and procedures used for data collection, and the composition of the research team who originated them in the first place. This sort of information, typically referred to as 'meta-data' (Leonelli 2010, Edwards et al 2011), is indispensible to researchers wishing to evaluate the reliability and quality of data. Even more importantly, it makes it possible to interpret the scientific significance of data, thus enabling researchers to extract meaning from their scrutiny of databases.

Given the challenges already linked to the de-contextualisation of data, it will come as no surprise that re-contextualising them is proving even harder in biological practice. The selection and annotation of meta-data is more labour-intensive than the formatting of data themselves, and involves the establishment of several types of standards, each of which is managed by its own network of funding and institutions. For a start, it presupposes reliable reference to material specimens of the model organisms in question. In other words, it is important to standardise the materials on which data are produced as much as possible, so that researchers working on those data in different locations can order those materials and reasonably assume that they are indeed the same materials as those from which data were originally extracted. Within model organism biology, the standardisation, coordination and dissemination of specimens is in the hands of appositely built stock centres, which collect as many strains of organisms as possible, pair them up with datasets stored in databases, and make them available for order to researchers interested in the data. In the best cases, this happens through the mediation of databases themselves; for instance, The Arabidopsis Research Database has long incorporated the option to order materials associated with data stored therein at the same time as one is viewing the data (Rosenthal and Ashburner 2002). However, such a well-organised coordination between databases and stock centres is rare, particularly in cases where the specimens to be collected and ordered are not easily transportable items such as seeds and worms, but organisms that are difficult and expensive to keep and disseminate, such as viruses and mice. Most organisms used for experimental research do not even have a centralised stock centre collecting exemplars for further dissemination. As a result, the data generated from these

organisms are hard to incorporate into databases, as providing them with adequate metadata proves impossible (Leonelli 2012a).

Another serious challenge to the development of metadata consists of capturing experimental protocols and procedures, which in biology are notoriously idiosyncratic and difficult to capture through any kind of textual description (let alone standard categories). The difficulties are exemplified by the recent emergence of a Journal of Visualised Experiments, whose editors claim that actually showing a video of how a specific experiment is performed is the only way to credibly communicate information about research methods and protocols. Indeed, despite the attempted implementation of standard descriptions such as the Minimal Information about Biological and Biomedical Investigation, standards in this area are very under-developed and rarely used by biologists (Leonelli 2012a). This makes the job of curators even more difficult, as they are then left with the task of selecting which meta-data to insert in their database, and which format to use in order to provide such information. Additionally, curators are often asked to provide a preliminary assessment of the quality of data, which can act as a guideline for researchers interested in large datasets. Curators achieve this through so-called 'evidence codes' and 'confidence rankings', which however tend to be based on controversial assumptions (for instance, the idea that data obtained through physical interaction with organisms are more trustworthy than simulation results) which may not fit all scenarios in which data may be adopted.

Stage 3: Re-Use

The final stage of data journeys that I wish to examine is that of re-use. One of the central themes in big data research is the opportunity to re-use the same datasets to uncover a large number of different correlations. After having been de-contextualised and re-contextualised, data are therefore supposed to fulfil their epistemic role by leading to a variety of new discoveries. From my observations above, it will already be clear that very few of the data produced within experimental biology make it to this stage of their journeys, due to the lack of standardisation in their format and production techniques, as well as the absence of stable reference materials to which data can be meaningfully associated for re-contextualisation. Data that cannot be de-contextualised and re-contextualised are not generally included into model organism databases, and thus do not become part of a body of big data from which biologically significant inferences can be made. Remarkably, the data that are most successfully assembled into big collections are genomic data, such as genome sequences and microarrays, which are produced through highly standardised technologies and are therefore easier to format for travel. This is bad news for biological research focused on understanding higher-level processes such as organismal development, behaviour and susceptibility to environmental factors: data that document these aspects are typically the least standardised in both their format and the materials and instruments through which they are produced, which makes their integration into large collection into a serious challenge.

This signals a problem with the idea that big data involves unproblematic access to all data about a given phenomenon – or even to at least some data about

several aspects of a phenomenon, such as multiple data sources concerning different levels of organisation of an organism. When considering the stage of data re-use, however, an even more significant challenge emerges: that of data classification. Whenever data and metadata are added to a database, curators need to tag them with keywords that will make them retrievable to biologists interested in related phenomena. This is an extremely hard task, given that curators want to leave the interpretation of the potential evidential value of data as open as possible to database users. Ideally, curators should label data according to the interests and terminology used by their prospective users, so that a biologist is able to search for any data connected to her phenomenon of interest (e.g. 'metabolism') and find what she the evidence that she is looking for. What makes such labelling process into a complex and contentious endeavour is the recognition that this classification partly determines the ways in which data may be used in the future – which, paradoxically, is exactly what databases are not supposed to do. In other publications, I have described at length the functioning of the most popular system currently used to classify data in model organism databases, the so-called 'bio-ontologies' (Leonelli 2012b). Bio-ontologies are standard vocabularies intended to be intelligible and usable across all the model organism communities, sub-disciplines and cultural locations to which data should travel in order to be re-used. Given the above-mentioned fragmentation of biology into myriads of epistemic communities with their own terminologies, interests and beliefs, this is a tall order. Consequently, and despite the widespread recognition that model organism databases are among the best sources of big data within biology, many biologists are suspicious of them, principally as a result of their mistrust of the categories under which data are classified and distributed. This puts into question not only the idea that databases can successfully collect big data on all aspects of given organisms, but also the idea that they succeed in making such data retrievable to researchers in ways that foster their re-use towards making new discoveries.

4. What Does It Take to Assemble Big Data? Implications for Big Data Claims

The above analysis, however brief, clearly points to the huge amount of manual labour involved in developing databases for the purpose of assembling big data and making it possible to integrate and analyse them; and to the many unresolved challenges and failures plaguing that process.

I have shown how curators have a strong influence on all three stages of data journeys via model organism databases. They are tasked with selecting, formatting and classifying data so as to mediate among the multiple standards and needs of the disparate epistemic communities involved in biological research. They also play a key role in devising and adding meta-data, including information about experimental protocols and relevant materials, without which it would be impossible for database users to gauge the reliability and significance of the data therein. All these activities require high amounts of funding for manual curation, which is mostly unavailable even in areas as successful as model organism biology. They also require the support and co-operation of the broader biological community, which is however also rare due to the pressures and credit systems to which experimental biologists are subject. Activities such

as data donation and participation in data curation are not currently rewarded within the academic system. Therefore, many scientists who run large laboratories and are responsible for their scientific success perceive these activities as an inexcusable waste of time, despite being aware of their scientific importance in fostering big data science.

We thus confronted with a situation in which (1) there is still a large gap between the opportunities offered by cutting-edge technologies for data dissemination and the realities of biological data production and re-use; (2) adequate funding to support and develop online databases is lacking, which greatly limits curators' ability to make data travel; and (3) data donation and incorporation into databases is very limited, which means that only a very small part of the data produced within biology actually get to be assembled into big data collections. Hence, big data collections in biology could be viewed as very small indeed, compared to the quantity and variety of data actually produced within this area of research. Even more problematically, such data collections tend to extremely partial in the data that they include and make visible. Despite curators' best efforts, model organism databases mostly display the outputs of rich, English speaking labs within visible and highly reputed research traditions, which deal with 'tractable' data formats. The incorporation of data produced by poor or unfashionable labs, whether in developed or developing countries, is very low – also because scientists working in those conditions have an even lesser chance than scientists working in prestigious locations to be able to contribute to the development of databases in the first place (the digital divide is alive and well in big data science, though taking on a new form).

A possible moral to be drawn from this situation is that what counts as data in the first place should be defined by the nature of their journeys. According to this view, data are whatever can be fitted into highly visible databases; and results that are hard to disseminate in this way do not count as data at all, since they are not widely accessible. I regard this view as empirically unwarranted, as it is clear from my research that there are many more results produced within the life sciences which biologists are happy to call and use as data; and that what biologists consider to be data does depend on its availability for scrutiny (it has to be possible to circulate them to at least some peers who can assess their usefulness as evidence), but not necessarily on the extent to which they are publicly available – in other words, data disseminated through paper or by email can have as much weight as data disseminated through online databases. Despite these obvious problems, however, the increasing prominence of databases as supposedly comprehensive sources of information may well lead some scientists to use them as benchmarks for what counts as data in a specific area of investigation. This tendency is reinforced by wider political and economic forces, such as governments, corporations and funding bodies, for whom the prospect of assembling centralised repositories for all available evidence on any given topics constitutes a powerful draw (Leonelli 2013).

How do these findings compare to the claims made by Mayer-Schönberger and Cukier? For a start, I think that they cause problems to both of the corollaries to their views that I listed above. Consider first the question of sampling. Rather than disappearing as a scientific concern, looking at the ways in which data

travel in biology highlights the ever-growing significance of sampling methods. Big data that are made available through databases for future analysis turn out to represent highly selected phenomena, materials and contributions, to the exclusion of the majority of biological work. What is worse, this selection is not the result of scientific choices, which can therefore be taken into account when analysing the data. Rather, it is the serendipitous result of social, political, economic and technical factors, which determine which data get to travel in ways that are non-transparent and hard to reconstruct by biologists at the receiving end. A full account of factors involved here far transcends the scope of this paper.[5] Still, even my brief analysis of data journeys illustrates how they depend on issues as diverse as national data donation policies (including privacy laws, in the case of biomedical data); the good-will and resources of specific data producers, as well as the ethos and visibility of the scientific traditions and environments in which they work (for instance, biologists working for private industries may not be allowed to publicly disclose their data); and the availability of well-curated databases, which in turn depends on the visibility and value placed upon them (and the data types therein) by government or relevant public/private funders. Assuming that big data does away with the need to consider sampling is highly problematic in such a situation. Unless the scientific system finds a way to improve the inclusivity of biological databases, they will continue to incorporate partial datasets that nevertheless play a significant role in shaping future research, thus encouraging an inherently conservative and irrational system.

This partiality also speaks to the issue of bias in research, which Mayer-Schönberger and Cukier also insist can potentially be superseded in the case of big data science. The ways in which big data are assembled for further analysis clearly introduce numerous biases related to methods for data collection, storage, dissemination and visualisation. This feature is recognised by Mayer-Schönberger and Cukier, who indeed point to the fact that the scale of such data collection takes focus away from the singularity of data points: the ways in which datasets are arranged, selected, visualised and analyzed becomes crucial to which trends and patterns emerge. However, they assume that the diversity and variability of data thus collected will be enough to enable counter the bias incorporated in each of these sources. In other words, big data are self-correcting by virtue of their very unevenness, which makes it probable that incorrect or inaccurate data are rooted out of the system because of their incongruence with other data sources. I think that my arguments about the inherent imbalances in the types and sources of data assembled within big biology casts some doubt as to whether such data collections, no matter how large, are diverse enough to counter bias in their sources. If all data sources share more or less the same biases (for instance, they all rely on microarrays produced with the same machines), there is also the chance that bias will be amplified, rather than reduced, through such big data.

[5] While a full investigation has yet to appear in print, STS scholars have explored several of the non-scientific aspects affecting data circulation (e.g. Martin 2001, Bowker 2006, Harvey and McMeekin 2007, Hilgartner 2013).

These considerations do not make Mayer-Schönberger and Cukier's claims about the power of big data completely implausible, but they certainly dent the idea that big data is revolutionising biological research. The availability of large datasets does of course make a difference, as advertised for instance in the Fourth Paradigm volume issued by Microsoft to advertise the power of data-intensive strategies (Hey et al 2009). And yet, as I stressed above, having a lot of data is not the same as having all of them; and cultivating such illusion of completeness is a very risky and potentially misleading strategy within biology – as most researchers whom I have interviewed over the last few years pointed out to me. The idea that the advent of big data lessens the value of accurate measurements also does not seem to fit these findings. Most sciences work at a level of sophistication in which one small error can have very serious consequences (the blatant example being engineering). The constant worry about the accuracy and reliability of data is reflected in the care put by database curators in enabling database users to assess such properties; and in the importance given by users themselves to evaluating the quality of data found on the internet. Indeed, databases are often valued because they provide means to triangulate findings coming from different sources, so as to improve the accuracy of measurement and determine which data are most reliable. Although they may often fail to do so, as I just discussed, the very fact that this is a valued feature of databases makes the claim that 'messiness' triumphs over accuracy look rather shaky. Finally, considering data journeys prompts second thoughts about the supposed primacy of correlations over causal explanations. Big data certainly do enable scientist to spot patterns and trends in new ways, which in turn constitutes an enormous boost to research. At the same time, biologists are rarely happy with such correlations, and rather use them as heuristics that shape the direction of research, without necessarily constituting a discovery in itself. Being able to predict how an organism or ecosystem may behave is of huge importance, particularly within fields such as biomedicine or environmental science; and yet, within experimental biology the ability to explain why a certain behaviour obtains is still very highly valued - arguably over and above the ability to relate two traits to each other.[6]

5. Conclusion: An Alternative Approach to Big Data Science

In closing my discussion, I want to consider its specificity with respect to other parts of big data science, but also the general lessons that may be drawn from such a case study. Biology, and particularly the study of model organisms, represents a field where data have been produced long before the advent of computing and many data types are still generated in ways that are not digital, but rather rely on physical and localised interactions between one or more investigators and a given organic sample. Accordingly, biological data on model organisms are heterogeneous both in their content and in their format; are curated and re-purposed to address the needs of highly disparate and fragmented epistemic communities; and present curators with specific challenges to do with the wish to faithfully capture and represent complex,

[6] The value of causal explanations in the life sciences is a key concern for many philosophers, particularly those interested in mechanistic explanations as a form of biological understanding (e.g. Bechtel 2006; Craver and Darden 2013).

diverse and evolving organismal structures and behaviours. Readers with a experience in other forms of big data may well be dealing with cases where both data and their prospective users are much more homogeneous, which means that their travel is less contested and tends to be curated and institutionalised in completely different ways. I view the fact that my study bears no obvious similarities to other areas of big data use as a strength of my approach, which indeed constitutes an invitation to disaggregate the notion of big data science as a homogenous whole, and instead pay attention to its specific manifestations across different contexts. At the same time, I maintain that a close examination of specialized areas can still yield general lessons, at the very least by drawing attention to aspects that need to be critically scrutinized in all instances of big data handling. These include, for instance, the extent to which data are – and need to be – curated before being assembled into common repositories; the decisions and investments involved in selecting data for travel, and their implications for which data get to be circulated in the first place; and the representativeness of data assembled under the heading of 'big data' with respect to other (and/or pre-existing) data collection activities within the same field.

At the most general level, my analysis can be used to argue that characterisations of big data science as comprehensive and intrinsically unbiased can be misleading rather than helpful in shaping scientific as well as public perceptions of the features, opportunities and dangers associated with data-intensive research. If one admits the plausibility of this position, then how can one better understand current developments? I here want to defend the idea that big data science has specific epistemological and methodological characteristics, and yet that it does not constitute a new epistemology for biology. Its strength lies in the combination of concerns that have long featured in biological research with opportunities opened up by novel communication technologies, as well as the political and economic climate in which scientific research is currently embedded. Big data brings new salience to aspects of scientific practice which have always been vital to successful empirical research, and yet have often been overlooked by policy-makers, funders, publishers, philosophers of science and even scientists themselves, who in the past have tended to evaluate what counts as 'good science' in terms of its products (e.g. new claims about phenomena or technologies for intervention in the world) rather than in terms of the processes through which such results are eventually achieved. These aspects include the processes involved in valuing data as a key scientific resource; situating data in a context within which they can be interpreted reliably; and structuring scientific institutions and credit mechanisms so that data dissemination is supported *and* regulated in ways that are conducive to the advancement of both science and society.

More specifically, I want to argue that the novelty of big data science can be located in two key shifts characterising scientific practices over the last two decades. First is the new **prominence** attributed to data as commodities with high scientific, economic, political and social value (Leonelli 2013). This has resulted in the acknowledgment of data as key scientific components, outputs in their own right that need to be widely disseminated (for instance, through so-called 'data journals' or repositories such as Figshare or more specialised

databases) – which in turn is engendering significant shifts in the ways in which research is organised and assessed both within and beyond scientific institutions. Second is the emergence of a new set of methods, infrastructures and skills to **handle** (format, disseminate, retrieve, model and interpret) data. Stephen Hilgartner has talked about the introduction of computing and internet technologies in biology as a change of communication regime (Hilgartner 1995). Indeed, my analysis has emphasised how the introduction of tools such as databases, and the related opportunity to make data instantly available over the internet, is challenging the ways in which data are produced and disseminated, as well as the types of expertise relevant to analysing such data (which now needs to include computing and curatorial skills, in addition to more traditional statistical and modelling abilities).

When seen it through this lens, data quantity can indeed be said to make a difference to biology, but in ways that are not as revolutionary as many big data advocates would advocate. There is strong continuity with practices of large data collection and assemblage conducted since the early modern period; and the core methods and epistemic problems of biological research, including exploratory experimentation, sampling and the search for causal mechanisms, remain crucial parts of inquiry in this area of science - particularly given the challenges encountered in developing and applying curatorial standards for data other than the high-throughput results of "omics" approaches. Nevertheless, the novel recognition of the relevance of data as a research output, and the use of technologies that greatly facilitate their dissemination and re-use, provide an opportunity for all areas in biology to reinvent the exchange of scientific results and create new forms of inference and collaboration.

I end this paper by suggesting a provocative explanation for what I argued is a non-revolutionary role of big data in biology. It seems to me that my scepticism arises because of my choice of domain, which is much narrower than Mayer-Schönberger and Cukier's commentary on the impacts of big data on society as a whole. Indeed, biological research may be the domain of human activity that is least affected by the emergence of big data and related technologies today. This is precisely because, like many other natural sciences such as astronomy, climatology and geology, biology has a long history of engaging with large datasets; and because deepening our current understanding of the world continues to be one of the key goals of inquiry in all areas of scientific investigation. While often striving to take advantage of any available tool for the investigation of the world and produce findings of use to society, biologists are not typically content with establishing correlations. The quest for causal explanations, often involving detailed descriptions of the mechanisms and laws at play in any given situation, is not likely to lose its appeal any time soon. Whether or not it is plausible in its implementation, the big data epistemology outlined by Mayer-Schönberger and Cukier is thus unlikely to prove attractive to biologists, for whom correlations are typically but a starting point to a scientific investigation; and the same argument may well apply to other areas of the natural sciences.[7] The real revolution seems more likely to centre on other areas

[7] The validity of this claim needs of course to be established through further empirical and comparative research. Also, I should note one undisputed way in which big data rhetoric is

of social life, particularly economics and politics, where the widespread use of patterns extracted from large datasets as evidence for decision-making is a relatively recent phenomenon. It is no coincidence that most of the examples given by Mayer-Schönberger and Cukier come from the industrial world, and particularly globalised sales strategies as in the case of Amazon.com. Big data provides new opportunities for managing goods and resources, which may be exploited to reflect and engage individual preferences and desires. By the same token, big data also provide as yet unexplored opportunities for manipulating and controlling individuals and communities on a large scale – a process that Rita Raley (2013) characterised as "dataveillence". As demonstrated by the history of quantification techniques as surveillance and monitoring tools (Porter 1995), data have long functioned as a way to quantify one's actions and monitor others. 'Bigness' in data production, availability and use thus needs to be contextualised and questioned as a political economic phenomenon as much as a technical one (Davies, Frow and Leonelli 2013).

Acknowledgments

I am grateful to the "Sciences of the Archive" Project at the Max Planck Institute for the History of Science in Berlin, whose generous hospitality and lively intellectual atmosphere in 2014 enabled me to complete this manuscript; and to Brian Rappert for insightful comments on the manuscript. This research was funded by the UK Economic and Social Research Council (ESRC) through the ESRC Centre for Genomics and Society and grant number ES/F028180/1; the Leverhulme Trust through grant award RPG-2013-153; and the European Research Council under the European Union's Seventh Framework Programme (*FP7/2007-2013)* / ERC *grant agreement* n° 335925.

Bibliography

Ankeny, Rachel and Sabina Leonelli. 2015, in press. Valuing Data in Postgenomic Biology: How Data Donation and Curation Practices Challenge the Scientific Publication System. In: Hallam Stevens and Sarah Richardson (eds) *PostGenomics*. Duke University Press.

Baker, Karen S., and Florence Millerand. 2010. "Infrastructuring Ecology: Challenges in Achieving Data Sharing." In *Collaboration in the New Life Sciences*, edited by John N. Parker, Niki Vermeulen and Bart Penders, 111-138. Farnham, UK: Ashgate.

Bastow, Ruth, and Sabina Leonelli. 2010. "Sustainable digital infrastructure." *EMBO Reports* 11 (10): 730-735.

Bauer, Susanne. 2008. "Mining Data, Gathering Variables, and Recombining Information: The Flexible Architecture of Epidemiological Studies." *Studies in History and Philosophy of Biological and Biomedical Sciences* 39: 415-426.

Bechtel, William. 2006. *Discovering Cell Mechanisms. The Creation of Modern Cell*

affecting biological research: the allocation of funding to increasingly large data consortia, to the detriment of more specialised and less data-centric area of investigation).

Biology. Cambridge University Press.

Bogen, James. 2013. "Theory and Observation in Science." In *The Stanford Encyclopedia of Philosophy (Spring 2013 Edition)*, edited by Edward N. Zalta. Last accessed February 20 2014. http://plato.stanford.edu/archives/spr2013/entries/science-theory-observation/

Borgman, Christine, L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: The MIT Press.

Bowker, Geoffrey C. 2001. "Biodiversity Datadiversity." *Social Studies of Science* 30 (5): 643-684.

Bowker, Geoffrey C. 2006. *Memory Practices in the Sciences*. Cambridge, MA: The MIT Press.

Craver, Carl F. and Lindley Darden. 2013. *In Search of Biological Mechanisms: Discoveries across the Life Sciences*. Chicago, IL: University of Chicago Press.

Davies, Gail, Emma Frow, and Sabina Leonelli. 2013. "Bigger, Faster, Better? Rhetorics and Practices of Large-Scale Research in Contemporary Bioscience." *BioSocieties* 8 (4): 386-396.

Dezin, N. 2006. *Sociological Methods: A Sourcebook*. Aldine Transaction.

Dupré, John. 2012. *Processes of Life*. Oxford, UK: Oxford University Press.

Edwards, Paul N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: The MIT Press.

Edwards, Paul N., Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. 2011. "Science Friction: Data, Metadata, and Collaboration." *Social Studies of Science* 41 (5): 667-690.

Gitelman, Lisa (ed). 2013. "Raw Data" is an Oxymoron. Cambridge: MIT Press.

Hacking, Ian. 1992. "The Self-Vindication of the Laboratory Sciences." In *Science as Practice and Culture*, edited by Andrew Pickering, 29-64. Chicago, IL: The University of Chicago Press.

Harvey, Mark, and Andrew McMeekin. 2007. *Public or Private Economics of Knowledge? Turbulence in the Biological Sciences*. Cheltenham, UK: Edward Elgar Publishing.

Hey, Tony, Stewart Tansley, and Kristine Tolle, editors. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Hilgartner, Stephen. 1995. "Biomolecular Databases: New Communication Regimes for Biology?" *Science Communication* 17: 240-263.

Hilgartner, Stephen. 2013. "Constituting large-scale biology: Building a regime of governance in the early years of the Human Genome Project." *BioSocieties* 8: 397-416.

Hine, Christine. 2006. "Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work." *Social Studies of Science* 36 (2): 269–298.

Johnson, Kristin. 2012. *Ordering Life: Karl Jordan and the Naturalist Tradition*. Baltimore, MD: Johns Hopkins University Press.

Kelty, Christopher M. 2012. “This is not an article: Model organism newsletters and the question of ‘open science’.” *BioSocieties* 7 (2): 140-168.

Leonelli, Sabina and Rachel A. Ankeny. 2012. “Re-thinking organisms: The impact of databases on model organism biology.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 29-36.

Leonelli, Sabina. 2010. “Packaging Small Facts for Re-Use: Databases in Model Organism Biology.” In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, edited by Peter Howlett and Mary S. Morgan, 325-348. Cambridge, UK: Cambridge University Press.

Leonelli, Sabina. 2012a. “When humans are the exception: Cross-species databases at the interface of clinical and biological research.” *Social Studies of Science* 42 (2): 214-236.

Leonelli, Sabina. 2012b. “Classificatory Theory in Data-Intensive Science: The Case of Open Biomedical Ontologies*.” International Studies in the Philosophy of Science* 26(1): 47-65.

Leonelli, Sabina. 2013. “Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production and the Political Economy of Contemporary Biology.” *Bulletin of Science, Technology and Society* 33 (1/2): 6-11.

Levin, Nadine, Dagmara Weckoswka, David Castle, John Dupré and Sabina Leonelli. Manuscript in Preparation. “How Do Scientists Understand Openness? Assessing the Impact of UK Open Science Policies on Biological Research.”

Mackenzie, Adrian, and Ruth McNally. 2013. “Living Multiples: How Large-Scale Scientific Data-Mining Pursues Identity and Differences Theory." *Culture & Society* 30: 72-91.

Martin, Paul. 2001. "Genetic governance: the risks, oversight and regulation of genetic databases in the UK." *New Genetics and Society* 20 (2): 157-183.

Mayer-Schoenberger, Viktor, and Cuckier, Kenneth (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray Publisher.

Müller-Wille, Staffan, and Isabelle Charmantier. 2012. “Natural history and information overload: The case of Linnaeus.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 4-15.

O’Malley, Maureen and Soyer, Orkun S. 2012. “The Roles of Integration in Molecular Systems Biology.” *Studies in the History and the Philosophy of Biological and Biomedical Sciences* 43(1): 58-68.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Raley, Rita. 2013. “Dataveillance and Countervailance.” In Lisa Gitelman (ed.) *“Raw Data” Is An Oxymoron.* MIT Press, pp. 121-146.

Rheinberger, Hans-Jörg. 2011. "Infra-experimentality: from traces to data, from data to patterning facts." *History of Science* 49 (3): 337-348.

Rosenthal, Nadia, and Michael Ashburner. 2002. “Taking stock of our models: the function and future of stock centres.” *Nature Reviews Genetics* 3: 711–717.

Royal Society. 2012. “Science as an Open Enterprise.” Accessed 14 January 2014.

http://royalsociety.org/policy/projects/science-public-enterprise/report/.

Stein, Lincoln D. 2008. “Towards a Cyberinfrastructure for the Biological Sciences: Progress, Visions and Challenges.” *Nature Reviews Genetics* 9 (9): 678-688.

Stevens, Hallam. 2013. *Life Out of Sequence: Bioinformatics and the Introduction of Computers into Biology*. Chicago: University of Chicago Press.

Strasser, Bruno J. 2008. "GenBank—Natural History in the 21st Century?" *Science* 322 (5901): 537-538.

Traweek, Sharon. 1998. “Iconic Devices: Towards and Ethnography of Physical Images”. In Gary Downey and Joseph Dumit (eds.) *Cyborgs and Cytadels*. University of Washington Press.

Wylie, Alison. 2002. *Thinking From Things: Essays in the Philosophy of Archeology*. Berkley, CA: University of California Press.