Data

Aidan Lyon

November 23, 2015

Abstract

In this essay, I review some of the philosophical issues that arise in connection with the concept of data. I first ask what data are, and I evaluate a number of different answers to this question that have been given. I then examine the concept of so-called *big data* and the corresponding concept of *big science*. It has been claimed that the advent of big science calls for a fundamental change to science and scientific methodology. I argue that such claims are too strong. Finally, I review the distinction between data and phenomena, due to Bogen and Woodward 1988, and I discuss some of its connections to big data and big science.

Keywords: data, information, interpretation, phenomena, big data, big science

1 Introduction

Data are clearly very important and play a central role in our daily lives. However, this is a fairly recent development. It was not all that long ago that data were mostly confined to lab books and primarily the concern of scientists. But data are now everywhere and they are increasingly important to everyone, from teenagers who want faster data plans for their cellphones, to photographers who need larger data storage devices, to CEOs who want to know how to use their business data to improve their sales.

Although data have long had a home in science, the relationship between data and science has been changing in recent times. Terms such as "big science" and "data-driven science" refer to sciences that seem to proceed in an increasingly automated way due to their increased access to gigantic data sets—so-called *big data*. Data are now collected, filtered, stored, and even analysed automatically by electronic devices. The better humans get at building devices that can gather, store and process data, the less of a role humans seem to have in science (*cf.* Humphreys 2006.)

Since data are so important to our lives and scientific practice and since big data are apparently changing the way science is done, it is worth putting the concept of data under the philosophical microscope. In this essay, I will review three philosophical issues in relation to data. Section 2 asks what data are and reviews four analyses of the concept that have been discussed in the literature. Section 3 focuses on the big data revolution, its impact on scientific practice, and what issues in the philosophy of science it raises.

Finally, section 4 discusses the data-phenomena distinction (due to Bogen and Woodward 1988) that has been put to work in the philosophy of science.

2 What are Data?

The concept of data is clearly an important one, it appears to be central to the practice of science and the foundation of much of our modern life. Indeed, it seems that more and more things get "data-fied" each year—e.g., music, movies, books, and currencies (such as bitcoin). Since data is so important to us, it is worth asking what we mean by "data". When scientists speak of data, do they talk about the same kind of thing as when tech executives speak of our data being stored in the cloud? It is quite possible that the answer is "no" and that we have multiple concepts of the data that should be distinguished before we attempt to analyse them. However, to begin, I will assume that we have one unified concept of data and see if it can be given an interesting and coherent analysis.

Little work has been done on analysing the concept of data, and the work that has been done shows that it is a concept that is a little trickier to analyse that one might have supposed. In philosophy, the main source of work on this topic has been done by Luciano Floridi, and in particular, Floridi 2008 has provided a review of some analyses of data that seem initially plausible. In this section, I will discuss these analyses and build on Floridi's work. Floridi identifies four analyses of the concept of data: (1) the epistemic interpretation, (2) the informational interpretation, (3) the computational interpretation, and (4) the diaphoric interpretation, the last of which is his own preferred interpretation.

The epistemic interpretation of data says that data are facts. This simple analyses has some *prima facie* plausibility, especially when considered in the light of how data can be used in scientific practice. When William Herschel recorded and analysed his observations of bodies in the night sky, he was recording and analysing data, which were facts about the positions of the astronomical bodies, and the analysis allowed to him to discover new facts, such as the existence of Uranus (Holmes 2008). Similarly, when credit card companies collect data on our purchases, they seem to be collecting facts about our purchases—e.g., the fact that I just bought a coffee with my credit card at the cafe that I'm currently sitting in is now recorded in some far away data base as a datum.

However, as Floridi points out, the epistemic interpretation has at least two problems. The first problem is that it appears to be too restrictive, in that it doesn't explain processes that are common to data, such as data compression and data cryptography. Unfortunately, Floridi doesn't elaborate on this problem, and we way wonder why the epistemic interpretation doesn't explain these processes and whether it is a reasonable constraint that any interpretation should explain them (and what kind of explanation is required). One way to understand the objection is if we understand the concept of fact so that facts are not the sorts of things that can be compressed, in which case, the epistemic interpretation would entail that data cannot be compressed. Compression, at least in this context, appears to be a process that applies to strings of symbols—e.g., it is common to compress the string "for example" to the string "e.g.". So, however we understand what facts are,

if they are not understood as (or involving) strings of symbols, then it seems they cannot be compressed in the appropriate way. The most common understandings of facts understand them in (apparently) non-symbolic ways—e.g., as propositions, or as events, etc. If any such analysis of the concept fact is right, then it will turn out the data cannot be facts.

Floridi's second objection to the epistemic interpretation is that it isn't very informative, it simply trades one concept that is difficult to analyse for another. This objection seems correct, in so far as we want an analysis in terms of simpler concepts, but it can be helpful to know that we only have one concept that is difficult to analyse, rather than two. Moreover, since work has already been done in analysing what facts are (e.g., Carnap 1946, Popper 1959, Lakatos 1970), the identification of data with facts inherits the informativeness of that work.

Nevertheless, it seems that the epistemic interpretation is wrong. In addition to the problem it seems to entail that data cannot be compressed, it also seems to entail that data cannot be false. The usual understanding of facts is such that if something is a fact, then it is *true*. For example, if it is a fact that 2014 is the hottest year on record, then it is (in fact!) true that 2014 is the hottest year on record. Facts, then, cannot be false. However, it seems that data can be false. Indeed, sometimes scientists are accused of falsifying their data. To take a famous example, Ronald Fisher complained that Mendel's data on trait inheritance was probably falsified:

> "[..I]t remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected. This possibility is supported by independent evidence that the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendels expectations." Fisher 1936, p. 132.

Since facts cannot be falsified and data can, data cannot be facts.

This problem with the epistemic interpretation suggests a very natural way for us to modify the epistemic interpretation so that it avoids the problem, and, incidentally, so that it is more deserving of its name: the *epistemic* interpretation. Instead of saying that data are facts, we might try to say that data are *purported* facts. That is, data are the things that *we think* are facts. A false datum, then, is one that we think is a fact but which actually isn't. This modification of the epistemic interpretation also has the advantage that it fits quite comfortably with the original meaning of "datum" in Latin: that which is taken as given or for granted.

However, there are still problems. Whatever data are, they are things that can be stored, compressed, destroyed, encrypted, and hidden. And it seems that purported facts are not the sorts of things that can have all of these things done to them. For example, suppose that we say that a purported fact is a proposition that someone thinks is true, or one that enough people are think is true (and let's not worry about how many would be "enough"). The way we normally think about memory sticks and propositions is such that memory sticks don't store propositions, and yet they do store data. Propositions are also not compressed or destroyed or encrypted or hidden.

Another natural modification suggests itself: we might say that data are *representations* of purported facts (or representations of facts, where the representations need not be veridical). This version of the epistemic interpretation renders data as certain kinds

of symbols: symbols that are to represent purported facts. Since symbols can be stored, compressed, destroyed, encrypted, and hidden, it seems we have a substantial improvement to the original epistemic interpretation. Since this improvement also says that data are not any old symbols—i.e., since they are symbols that represent purported facts—it is easy to see how they can be false and how they have an epistemic component to them.

Although, it seems we are moving in the right direction, there are two problems with this most recent refinement of the epistemic interpretation. The first problem suggests further refinements that could be made, but the second appears to be more problematic. The first problem is that it seems that there can be data that might represent facts but which are not purported by anyone (or intended to exist). Imagine a malfunctioning microphone that is no longer able to record human voice but is still sensitive to record human coughs. Suppose the microphone has been forgotten about and accidentally left on in some office, with the recordings being stored on some far away server. It seems true to say that the server now contains data on the coughing in the office. Indeed, if someone were to ever realise this, they could analyse the coughing and use it for some purpose—e.g., to measure cold and flu activity in the office. However, until someone does realise this, the data on the server do not appear to be purported facts, since no one purports them as facts (and no one intended them to represent coughing facts). This problem suggests that concept of data may live somewhere between being an intentional and representational concept and being a non-intentional informational concept. We, therefore, might be able to further improve the epistemic interpretation by using the resources from the literature on information and representation (see e.g., Dretske 1981, Fodor 1987). For example, instead of saying that data are symbols that represent facts that are purported, we might say that they *would* represent the relevant facts if certain conditions obtained, or that the relevant facts would *would* be purported if certain conditions obtained. I shall leave these considerations here, because the second problem that I alluded to earlier appears to be more damning.

The second problem with saying that data are representations of purported facts is that it doesn't seem to do justice to a large amount of the data that we use on a daily basis. For example, these days it is common for music to be stored in the form of data on computers and various portable devices. This music is stored in a way that is fundamentally the same as the way in which, say, the outcomes of a scientific experiment are stored: as a string of '1's and '0's in a file, which is encoded in some physical format, such as point variations in magnetism in the magnetic tape of a spinning hard drive. There are many ways in which data are stored physically—e.g., solid state hard drives, compact disks, floppy disks, etc.—and these details need not concern here. So, for simplicity, I shall speak as though '1's and '0's are stored directly in some physical medium (e.g., one can think of tiny '1's and '0's written on the magnetic tape). It doesn't seem that any string of '1's and '0's in a music data file represents any purported facts. For example, the first word of the lyrics of *Hey Jude* is 'Hey' and was sung by Paul McCartney. I have that song on my computer, and when I play it, a causal process unfolds that involves a particular string of '1's and '0's on my hard drive that leads to me hearing the word 'Hey'. However, it doesn't seem correct to say that that string of '1's and '0s' represents the purported fact that, say, McCartney

sung the word 'Hey' (at some particular time and location). Although it is true that we can *infer* from this string of '1's and '0's, along with some other facts, that McCartney did in fact sing the word 'Hey', it is nevertheless false that that string represents that fact. Nor does it seem to represent *any* fact. Rather, the string represents something more like a *command* or *input* that is to be passed to, and processed by, my music player.

Music stored as data is not the only example of this kind. For example, computer programs are also stored in the same way and they appear to fail to be factive in the same way that music fails to be factive. For example, the first line of one of the programs on my computer is "import numpy", which is a command to import the numpy package, a Python scientific computing package commonly used for data analysis.[1] This command is stored as a string of 1s and 0s somewhere on my computer, but that string doesn't represent any fact about the world. So it seems that not all data are sets facts. At this point, we might try to say that we have two concepts of data here, one that is associated with the storage of music (and other media) and one that our refinements of the epistemic interpretation were starting to latch onto. This is entirely possible and the idea cannot be rejected immediately. However, we should try to resist introducing such a distinction at least to begin with. This is in part simply because of parsimony, but also because such a distinction appears *ad hoc*, at least at a first glance. Along with the 'Hey Jude' file on my computer, I have other files, some of which represent purported facts (e.g., outcomes of some experiments), and all these files are stored in fundamentally the same way: as strings of '1's and '0's.[2] So it seems that our first attempt in analysing the concept of data should be to try to find an analysis that can handle both examples of data.

This brings us to second interpretation of data that Floridi discusses: the informational interpretation. The informational interpretation of data says that data are information (or bits of information). One advantage to this interpretation is that makes sense of the locution *data mining*. Data mining typically involves running algorithms and queries on a database to obtain some fact that was not stored in an obvious way in the database. One classic example is of Target who mined their database of customer purchases to find that certain spending habits are indicative of future ones. In particular, Target found that when a woman changes from purchasing scented hand-lotion to unscented hand-lotion, that is evidence that suggests that the woman has become pregnant. In a famous incident, Target was able to work out that a teenager was pregnant before her parents were able to do so. Data mining, then, is not like coal mining; the goal of data mining is not to get more data, unlike how the goal of coal mining is to get more coal. Rather, the goal of data mining is to obtain *information*. Data mining might be better named as information mining, which suggests that the informational interpretation of data might be onto something.

However, there are serious problems with the informational interpretation. Floridi identifies two problems. The first is that the informational interpretation of data results in a circularity when combined with one of the most common interpretations of information: that information is a certain kind of true and meaningful data (*cf.* Floridi 2005). The

[1] http://www.numpy.org

[2] Of course, there are different file formats and encoding systems, but at the end of the day, all of the data are stored in the form of '1's and '0's.

second problem is that sometimes we have data without information. For example, a music CD can contain megabytes of data without containing information about anything (*ibid*). This second objection is probably stated a little too strongly, for it might be argued that the music CD at least contains the information that some music was played at some point. At any rate, this second problem is a problem with understanding information as being *necessary* for data. A third problem arises from understanding information as being *sufficient* for data. The low-charge state of a laptop battery contains information—*viz.*, that the battery has low charge—but the low-charge state does not appear to be data (or a datum). In contrast, the line in the battery-history file that records the current charge-state of the battery is both data and (contains) information. The informational interpretation of data therefore doesn't seem very promising.

The third interpretation of data, the computational interpretation, says that data are collections of binary elements—e.g., the etchings in the aluminum foil in compact discs—that are processed electronically. As Floridi states, an advantage to this interpretation is that it "explains why pictures, music files, or videos are also constituted by data" (Floridi 2008, p. 235). So the computational interpretation handles the difficulty concerning media files that our refined version of the epistemic interpretation ran into. However, the computational interpretation is surely too limited. Vast amounts of data existed long before the invention of computers, or even electricity—recall Herschel's astronomical data. And even today, plenty of data are stored non-digitally and in a non-binary way—e.g., some DJs still use vinyl records.

This brings us, finally, to the fourth interpretation, the diaphoric interpretation, which has been developed and defended by Floridi (*ibid*). The diaphoric interpretation says that a datum is a lack of uniformity in some domain. Put more formally, the interpretation says that a "datum = x being distinct from y, where x and y are two uninterpreted variables and the domain is left open to further interpretation" (*ibid*, p. 235.). Floridi claims that the diaphoric interpretation captures what is at the core of the concept data and that it is the most fundamental and satisfactory of the four interpretations.

As stated, the diaphoric interpretation of data lets too many things count as data. Indeed, *any thing* x that is distinct from some other thing y will count as a datum. This is a problem because data often play an epistemic or inferential role that helps us as truth-seeking agents. As I mentioned earlier, the original Latin meaning of 'datum' is something that is given or taken for granted, but there also appears to be something of an objective or externalist component to the concept. One of our main uses of data is to acquire knowledge or useful information (recall the Target example from before), and, as such, there are certain inferences from data that are warranted and some that are not. I won't attempt to articulate what these inferences are here; but it suffices to note that, at the very least, we can infer from that d is a datum that represents some fact, that that fact is likely—or at least that we have some evidence for that fact. However, the diaphoric interpretation doesn't respect this inferential constraint. For example, let x be the proposition that I will be a billionaire, which is distinct from $y = \neg x$, in that I wish x to be true and I don't wish y to be true. By the diaphoric interpretation, x is a datum. However, it would be crazy for me to infer that x is likely or that I have good evidence

for x from the grounds that x is a datum, since x is only a datum (by the lights of the diaphoric interpretation) because I wish x to be true.

This is just one way in which the diaphoric interpretation of data over-generates, and there will be many others, since x and y are left uninterpreted. The natural remedy would be to give x and y an interpretation, or to least put some constraints on their interpretation, and/or constrain what differences between the are relevant for datum-hood. I'm not sure what interpretations or constraints would be appropriate, but given that the content of the diaphoric interpretation is that, roughly speaking, for x to be a datum, x must differ in some way from something in a domain, it seems that most of the work will be done by specifying these details.

To close this section, I'd like to return the issue of whether we have multiple concepts of data. With the demise of the refined epistemic interpretation, I said that we should continue the search for an interpretation that treats data as a unified concept. I then reviewed three more interpretations, and they didn't fare any better—indeed, if anything, they fared worse. So it is worth considering the possibility that we have multiple data concepts. I mentioned as a reason against this idea that music files and files that report things like experiment outcomes—let's call them *fact files*—are stored in, fundamentally, the same way on our devices. However, that could be a red-herring, since we could have easily have decided store, say, music files, using strings of '0's, '1's, and '2's. And a reason *for* introducing the distinction is that we use music files and fact files in very different ways. The roles they play in our lives, and on our computers, are rather different, and they seem to differ in a way that is similar to way that works of fiction differ from works of non-fiction. So it may be fruitful to distinguish two concepts of data before we try to give any analyses. For lack of better names, I'll call the two potential concepts: *fact data* and *fictional data*. Examples of fact data include: scientific data (including falsified data), charge-state data, customer purchasing data, communications meta-data, and so on. Examples of fictional data include: some music data, some video data, some image data, some book data, some code data, and so on. We may then be able to make progress by taking some of the work done in the philosophy of fiction and using it for an analysis of fictional data. For example, Walton 1990 has analysed fictional truth as "being a prescription or mandate in some context to imagine something" (p. 39), which is analogous to the idea that music data are instructions or inputs for a music player. We then might be able to give structurally similar analyses of the two concepts. For example factual data might be (roughly speaking) symbols that represent purported facts, and fictional data might be symbols that represent intended commands.

3 Big Data

According to IBM, humans create 2.5 quintillion bytes of data every day.[3] The volume of data generated around the world each year is increasing exponentially, and so has its

[3] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

importance to businesses.[4] So-called *big data* is booming, and governments and businesses are scrambling to find and/or train enough data scientists so that they can take advantage of it.

Big data aren't limited to government and business, however. The volume and complexity of data that are now available to various scientific fields is so tremendous that some have argued that this has changed the nature of these scientific fields, so much so that they have become known as *big sciences* (*cf.* Borgman 2010). In particular, biology has seen a huge increase available data in fields such bioinformatics, computational biology, systems biology, and synthetic biology (Callebaut 2012). Physics has also seen an explosion of data, with the digitisation of particle colliders, which generate far more data than can be currently stored. Astronomy has seen its own data explosion, with the digitisations and improvements of telescopes and cameras—indeed, there is now so much data that some data processing tasks are crowd-sourced out to regular citizens (Nielsen 2011, Ch. 6). Cognitive science and psychology are "getting big" too with large synthesis of fMRI data such as Neurosynth[5] and the use of crowd-sourcing platforms such as Amazon's Mechanical Turk[6] to conduct experiments (Buhrmester *et al.* 2011, Mason and Suri 2012). These are just a few examples, and the trend is expected to continue: the sciences are becoming extremely "data rich". In many cases, the data are collected and analysed automatically, with very little—if any—manual input by the researchers (Humphreys 2006).

Explicit discussions of issues that big data may raise for epistemology and the philosophy of science are quite rare (exceptions include Humphreys 2006, Callebaut 2012, and Floridi 2012a, 2012b). At first glance, this lack of attention by philosophers work might be because "big data" is just a buzz word that doesn't mean much (cf. Floridi 2012a), or because big data is just a fad, and so no novel philosophical issues are raised—besides ethical ones (see e.g., David and Patterson 2012 and Ioannidis 2013). Although fields such as data science and analytics are growing rapidly and the sciences are become data richer, it might seem that there is nothing new here of philosophical interest, and that there are just interesting technical challenges that need to be solved. It might seem that all that is going on is that we need to develop statistical and processing methods so that we can take advantage of the large amounts of data that are available to us, but at bottom, these methods are just methods of inductive inferences. And inductive inferences are still inductive inferences: we just now have much larger total-evidence sets.

However, Floridi 2012a, 2012b has argued that big data do in fact pose a real epistemological problem, which he calls *small patterns*:

> "The real, epistemological problem with big data is small patterns. Precisely because so many data can now be generated and processed so quickly, so cheaply, and on virtually anything, the pressure both on the data *nouveau riche*, such as Facebook or Walmart, Amazon or Google, and on the data *old money*, such as genetics or medicine, experimental physics or neuroscience, is to be able to spot where the new patterns with real added value lie in their immense databases and how they can best be exploited for the creation of wealth and the advancement of knowledge.

[4] http://www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf

[5] http://neurosynth.org

[6] https://www.mturk.com/mturk/welcome

> [...W]hat we need is a better understanding of which data are worth preserving. And this is a matter of grasping which questions are or will be interesting. Which is just another way of saying that, because the problem with big data is small patterns, ultimately, the game will be won by those who know how to ask and answer questions (Plato, *Cratylus*, 390c) and therefore know which data may be useful and relevant, and hence worth collecting and curating, in order to exploit their valuable patterns. We need more and better techniques and technologies to see the small data patterns, but we need more and better epistemology to sift the valuable ones." Floridi 2012a, pp. 436–7.

So, according to Floridi, the epistemological challenge posed by big data is not to work out how we can process large data sets and draw reliable inferences from them, but rather to work out how we can *best focus* our inferential resources on our large data sets. In big data contexts, there are so many variables between which there may be useful correlations, that we cannot simply run our algorithms over all permutations of them when looking for correlations. Instead, we are forced to choose subsets of those permutations for examination, and so we need some way of making such choices as intelligently as possible. A better epistemology could help us guess that variables such as lotion purchases and pregnancies might be correlated, and help us not waste time investigating whether, say, light-bulb purchases and pregnancies might be correlated (*cf* section 1).[7]

However, Floridi's epistemological problem doesn't appear to be a new philosophical problem or a new need to make epistemology better. The problem is essentially the problem of scientific discovery, albeit in a new guise. Traditionally, the problem of scientific discovery has been broken down into two components: (i) that of hypothesis/model/theory construction, and (ii) that of hypothesis/model/theory testing, with the second component receiving the most philosophical attention (e.g., Popper 1959). However, the first component is nevertheless clearly important and it is what focuses our energies regarding the second component. As Charles Darwin once said:

> "About thirty years ago there was much talk that geologists ought only to observe and not theorize; and I well remember someone saying that at this rate a man might as well go into a gravel-pit and count the pebbles and describe the colours. How odd it is that any-one should not see that all observation must be for or against some view if it is to be of any service!" Darwin 1861

Although the problem may not be a new one, big data could still be of philosophical interest for philosophers interested in the problem of scientific discovery, who may have a lot to learn from big data case studies.

Is that all there is of philosophical interest to big data? Some of the claims that have been made by various authors seem to suggest otherwise. For example, it has been claimed that with big data we have a new paradigm of science:

> "The growth of data in the big sciences such as astronomy, physics, and biology has led not only to new models of science—collectively known as the Fourth Paradigm—but also to the emergence of new fields of study such as astroinformatics and computational biology." Borgman 2010, p. 2.

And in an article entitled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Chris Anderson, editor in chief of *Wired*, argued that when armed with big data, scientists needn't bother with developing theories and building models anymore, and they can just listen to what the data say:

[7] For all I know, these may in fact be correlated!

> "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves." Anderson 2008.

If there is something to such claims, then there may, after all, be something of philosophical interest to big data.[8] Such a dramatic change in scientific methodology seems to be of philosophical interest in of itself and perhaps it even ought to affect the work of philosophers—e.g., maybe we shouldn't worry so much about scientific explanation.

The quote by Anderson above stands in stark contrast to Floridi's epistemological problem posed by big data. Anderson appears to think that we needn't work out what questions to ask, we can just ask all of them, in effect. Anderson envisages a new scientific methodology based on Google's methodology for analysing web content:

> "Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required. That's why Google can translate languages without actually "knowing" them (given equal corpus data, Google can translate Klingon into Farsi as easily as it can translate French into German). And why it can match ads to content without any knowledge or assumptions about the ads or the content." Anderson 2008.

This methodology is the polar opposite to what philosophers of science have been studying, especially in recent years with explosion of literature on mechanistic models and mechanistic explanations (e.g., Glennan 1996, Machamer *et al.* 2000, Bechtel and Abrahamsen 2005). Indeed, Anderson states this quite explicitly:

> "The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.
>
> There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?" Anderson 2008.

Other share similar views, e.g.:

> "The methods used in data analysis are suggesting the possibility of forecasting and analyzing without understanding (or at least without a structured and general understanding). More specifically, we will argue in Section 2 that understanding occurs when we know how to relate several descriptions of a phenomenon. Instead, these connections are disregarded in many data analysis methodologies, and this is one of the key features of modern data analysis approaches to scientific problems. The microarray paradigm is at work exactly when a large number of measured variables concerning a phenomenon are algorithmically organized to achieve narrow, specific answers to problems, and the connections among different levels of descriptions are not explored. This is the perspective we propose to call *agnostic science* and at its heart is the methodological principle that we called microarray paradigm." Panza *et al.* 2011, p. 6.

So perhaps the novel philosophical issue raised by big data is not Floridi's epistemological problem, but rather the unusual absence of it. Throughout most of the history of

[8] The debate over such claims has, so far, mostly been happening outside of philosophy—e.g., see Graham 2012 for a response to Anderson.

science, theories, models, and hypotheses were developed—to steer ourselves away from the gravel-pit—and data were collected to test these products of theorising. The process of discovery in the context of big data appears to break from this tradition, and thus may allow science to proceed without models, theories, and explanations. This is what is sometimes called *data driven science*.

Anderson's claims appear to be a little stronger than they should be. Take, for example, his allusion to Google's PageRank algorithm, which analyses the statistics of incoming links. The success of the algorithm depends on the fact that there is a correlation between people's interest in a given page and the number of links going into that page (roughly speaking). The discovery of the PageRank algorithm wasn't an exercise in big data analysis conducted in a background theory/model vacuum. Indeed, Page *et al.* 1999 explicitly gave an intuitive justification of the algorithm:

> "It is obvious to try to apply standard citation analysis techniques to the web's hypertextual citation structure. One can simply think of every link as being like an academic citation. So, a major page like `http://www.yahoo.com` will have tens of thousands of backlinks (or citations) pointing to it." Page *et al.* 1999, p. 2.

The algorithm is certainly ingenious, but we can see how even commonsense would guide someone to examine whether user interests correlated with something like the numbers of incoming links. So, although Google need not know what the content of a page is to be able to intelligently rank it in response to a search query, it certainly doesn't do this without the use of any theory.

Other examples, however, seem to better support Anderson's claims. For example, Panza *et al.* describe the use of big data to train neural networks for hurricane forecasting:

> "It is remarkable that (cf. Baik and Paek 2000) the use of neural networks gives better forecasts of the intensity of winds than the best available simulations of atmospheric dynamics. In this context, the input variables X_i are sets of measured meteorological data relative to a developing hurricane, and the output variable Y is the intensity of winds of the hurricane after 48 hours. The crucial point of this method is that the structure of neural networks does not express any understanding of the hurricane dynamics. It does not mirror in any understandable way the structure of the atmosphere: the specific problem is solved, but with no new knowledge of the phenomenon. Note moreover that only the ability to access a large quantity of measurements for the hurricane during its development allows this technique to work, in line with the general tenants of the microarray paradigm." Panza *et al.* 2011, p. 21.

They describe other examples as well, such as EEG data-driven controls of seizures (p. 22) and signal-boosting techniques (p. 24). Although the examples come from different areas of science, what they have in common is that in each case, "we can see very clearly that no understanding [of] the phenomenon is gained while solving the problem" (p. 19).

Traditional approaches to the philosophy of science have focused a great deal on understanding scientific explanation. These approaches take the sciences as trying to understand and/or explain their target phenomena. In contrast, the data-driven view of science says that we can shed explanations, understandings, theories, and so on, since they are pointless: who cares about understanding or explaining some phenomenon if we can predict it with extreme accuracy? One reply might be that by developing an developing models and theories that explain a given phenomenon, we help protect ourselves against

making erroneous predictions when something about the target system changes—for example when we move from predicting what WEIRD (Western, Educated, Industrialized, Rich, and Democratic) people will do to predicting what non-WEIRD people will do (*cf.* Henrich *et al.* 2010).[9] However, the proponent of data-driven science could reply that the big data skeptic is not thinking *big enough*. This is just the beginning of big data, and eventually our data sets will include points on all of the variables that might ever be relevant and our analytic tools will be powerful enough so that we will always be shielded against such errant predictions. When something behind the phenomena changes, our predictions will change accordingly, because we'll have data on the change and algorithms that can analyse the change. We won't know what's going on, but we will nevertheless enjoy the benefits of our accurate and adaptively-predictive algorithms.

Will our data sets ever be that large and will our analytical tools ever be that powerful? It's difficult to say, but if I had to bet, I would bet there will always be some surprises and that we will always benefit from some theory. However, although we may never reach the point of a theory-less science, we will probably get a lot closer to that point than we have ever before—in that more and more science will use less and less theory. Perhaps this will indeed cause a general change in scientific methodology. Science, to-date, has been a human activity conducted from a human perspective, despite our best efforts to be "objective observers" (*cf.* Callebaut 2012). As more science is done automatically by our electronic devices, science becomes less of a human activity, and, so far, that seems to coincide with less theory and explanation (*cf.* Humphreys 2006). It may be that the only reason why we have been obsessed with causation, explanation, theories, models, and understanding for so long is simply because for so long we have only ever had access to tiny fragments of data and a limited ability to process that data. The notion of causation could simply be a useful fiction created by an evolutionary process for data-limited beings, and so perhaps all there really is in the world, after all, is correlation (*cf.* Hume 1738).

4 Data and Phenomena

In an influential paper titled "Saving the Phenomena", Bogen and Woodward 1988 introduced a distinction between what they called *data* and *phenomena*, and they used this distinction to argue against a widely held view of science, namely, that scientific theories predict and explain facts about what we observe. Bogen and Woodward argued that scientific theories are in the business of explaining phenomena, which, according to them, are mostly unobservable, and that they are not in the business of explaining data, which are mostly observable (*ibid*, pp. 305-6).

Bogen and Woodward went on to further refine the view in a series of papers (e.g., Woodward 1989, 2000, 2011a, 2011b, and Bogen and Woodward 1992, 2003) and others built upon the view or used it for other purposes (e.g., Kaiser 1991, Basu 2003, Psillos 2004, Suarez 2005, Massimi 2007, and Haig 2013), while others criticised it (e.g., McAllister 1997, Glymour 2000, Schindler 2007, and Votsis 2010).

[9] This is the gist of Graham's 2012 response to Anderson.

To introduce the distinction between data and phenomena and their respective observabilities, Bogen and Woodward use the example of how we (as scientists) would determine the melting point of lead. They first point out that to determine the melting point of lead, one doesn't take a single thermometer reading of a sample of lead that has just melted. Rather, what one does—or should do—is take *many* readings of the thermometer. This is because the readings will tend to vary due to errors in the measurement process. Bogen and Woodward say that these readings constitute *data* (*ibid*, p. 308). They then note that if certain assumptions about the source of the variations of the data points can be made, we can use the data to construct an *estimate* of the melting point of lead, which is a *phenomenon* (*ibid*). For example, if the data appear to be normally distributed and we have no reason to think they have systematic errors, then we would estimate the melting point of lead by taking the mean of the data points. In other words, the temperature measurements are the data and, if all has gone well, the data can be used as evidence for a hypothesis about the melting point of lead, which is the phenomenon in this example.

Bogen and Woodward note that in this process of determining the melting point of lead, we do not actually *observe* the melting point of lead. And whereas we don't observe this phenomenon, we do observe the data—i.e., we do observe the thermometer readings. So we see how, at least in this example, how data are observed and phenomena are not: we infer unobserved phenomena from observed data. Bogen and Woodward support this generalisation with various examples from the history of science. For example, bubble chamber photographs (data) were used to detect weak neutral currents (phenomenon), and test scores for behavioural tasks along with X-ray photographs of human skulls (data) were used to test whether a damaged frontal lobe results in a dysfunction of sensory processing (phenomenon) (*ibid*, p. 315-6).

So far, this is relatively straightforward and uncontroversial. However, Bogen and Woodward put the distinction to some heavy work. They argue that scientists tend not to develop theories with the goal of explaining the particular data points that they have acquired. Rather, what scientists tend to do is construct theories that explain the phenomena that they have inferred from their data (*ibid*, pp. 309-10). In the example of the melting point of lead, the phenomenon "is explained by the character of the electron bonds and the presence of so-called 'delocalized electrons' in samples of this element (Woodward 2011b, p. 166) and this theory doesn't explain the particular data points that were observed. An example that makes the point even more clearly is the theory of General Relativity (GR), the phenomena of gravity, and the data of Eddington:

> "[...] GR is a theory of gravitational phenomena, not a theory that purports to explain or provide derivations concerning the behavior of cameras and optical telescopes or Eddingtons decisions about experimental design." Woodward 2011b, p. 166.

Since the data points are often the result of a myriad of causal factors, many of which are peculiar to the local circumstances of the measurements (e.g., the purity of the lead sample, when Eddington decided to let sunlight fall on his photographic plates), any explanation of them would have to take into account those factors. If the theory that is constructed to explain the melting point of lead doesn't account for those factors, it won't be able to explain the data points. Bogen and Woodward claim that this is typical of

scientific practice. One consideration that gives this claim plausibility is that phenomena are understood to be robust and repeatable—"features of the world that in principle could recur under different contexts or conditions" (*ibid*)—whereas data are specific and local—e.g., Eddington's photographic plates are not going to be created again.[10]

Before moving on it is worth considering how Bogen and Woodward's distinction between data and phenomena matches the analyses of data that we considered in section 2. The most promising analysis was the refinement of the epistemic interpretation that said that data are representations of purported facts. This seems to sit fairly well with Bogen and Woodward's distinction. In the 1988 paper, Bogen and Woodward mostly give examples of data (e.g., thermometer readings) and some of the conceptual roles of data (e.g., being evidence for hypotheses about phenomena). In a later paper, Woodward gives a more explicit characterization of data:

> "Data are public records (bubble chamber photographs in the case of neutral currents, photographs of stellar positions in the case of Eddington's expedition) produced by measurement and experiment, that serve as evidence for the existence of phenomena or for their possession of certain features." Woodward 2000, p. 163. (See also Woodward 2011b, p. 166.)

This also fits quite well with the refined epistemic interpretation, since, presumably, the public records produced by measurements and experiments are meant to be representative of the outcomes of those measurements and experiments. It also makes it clear that the data are not the states of the measuring devices; instead they are our recordings of the measuring devices. Data, then, are observable in that public records are observable—e.g., you can open up the lab book and observe the numerical symbols.

If Bogen and Woodward are correct that scientific theories typically explain unobservable phenomena, then this would appear to be big—and bad—news for empiricist accounts of science. For example, according to van Fraassen's constructive empiricism, the ultimate criterion by which we judge the quality of a scientific theory is its *empirical adequacy*, which is how well it fits with what has been observed. If a theory fits the observations well, then we need only believe that it is empirically adequate, and there is no need to believe any unobservable entities that the theory apparently posits. However, if Bogen and Woodward are right, then this is the wrong construal of empirical adequacy. Instead, we should say that for a theory to be empirically adequate, it must fit the phenomena. But since phenomena are typically not observable, it is not clear how van Fraassen can maintain that we need not believe in unobservable entities (Bogen and Woodward 1988, p. 351).

We might worry, then, about these claims of unobservability. Bogen and Woodward 1988 claimed that phenomena "in most cases are not observable in any interesting sense of that term" (p. 306). However, this strong claim isn't crucial to their view and Woodward 2011b replaces it with a weaker one:

> "[...I]t is enough to show that reasoning from data to phenomena can (and not infrequently does) successfully proceed without reliance on theoretical explanation of data. It was an unnecessary diversion to claim that this was always or even usually the case. A similar point holds in connection with the claim in our original paper that phenomena, in contrast to data, are typically

[10] Of course, *other* photographic plates might be created, but those will be *other* data.

> not observable (in senses of observable closely linked to human perception). As Paul Teller has argued there are plausible examples of observable phenomena, at least if we do not restrict this notion in question-begging ways: for example, various phenomena associated with the operation of the visual system such as color constancy in changing light and pop-out effects in visual attention. Similarly, for some optical phenomena such as the so-called Poison spot which appears at the center of the shadow cast by a circular object illuminated by a point source. What we should have said (I now think) is that phenomena need not be observable." Woodward 2011b, p. 171.

Woodward doesn't comment on the other strong claim of their original paper: that data "for the most part can be straightforwardly observed" (Bogen and Woodward 1988, p. 305). However, the opposite of this seems to be case, especially given how much data is collected and stored electronically. For example, the Large Hadron Collider (LHC) generates about 200,000 DVDs worth of data per second and most of that data—about 99.999%—is automatically discarded using triggers.[11] That's a lot of data that cannot be straightforwardly observed. Even the data that is stored does not appear to be straightforwardly observable, since it is electronically on a server. Bogen and Woodward might respond that we observe the stored data when it is used to create outputs on our computer screens. But their criticism of the view that phenomena can be observed (contrary to their view) appears to preclude that response. For example, they write:

> "'But to recognize that X played a part in causing Y or to recognize that Y was produced by causes which include X is not to see X. That is why you cannot see our grandfathers, or the midwives who delivered them, by looking at us. No matter how much you know about where we came from, they are not distal stimuli and they are not what you look at when you look at us." Bogen and Woodward 1988, p. 346.

Similarly, although the LHC data plays a part in causing various outputs on our computer screens, outputs which we do observe, that doesn't mean we observe, or can observe, the LHC data.

At any rate, whether data can for the most part be straightforwardly observed is not crucial to Bogen and Woodward's view. They can retract their claims that phenomena are often unobservable, that data are often observable, and that theories often do not explain data. Even if this weaker, and more plausible, view is correct, then constructive empiricism appears to be in trouble. Trying to argue that we do observe phenomena that are causally proximate to data—which, say, the melting point of lead might be an example of—doesn't seem like a promising strategy, since actual scientific practice sometimes saves phenomena that are extremely far-removed from the data. Massimi 2007 argues that the discovery of the J/ψ particle (a flavor-neutral meson made up of a charm quark and a charm anti-quark) is such an example and uses it to present a similar criticism of van Fraassen's constructive empiricism.

As I mentioned earlier, various criticisms of Bogen and Woodward's view have been given, but one of particular interest here, given the discussion of big data in section 3, is that of Schindler 2007. Schindler argues that Bogen and Woodward have something of an Anderson-style "big data" view of scientific methodology (although, he doesn't put it in those words):

[11] See e.g., http://www.lhc-closer.es/1/3/13/0

> "Bogen and Woodward 1988 [...] have argued [...] for a bottom-up construction of scientific phenomena from data. For them, the construction of phenomena is 'theory-free' and the exclusive matter of statistical inferences, controlling confounding factors and error sources, and the reduction of data." Schindler 2007, p. 161.

If Schindler is correct, then we can think of Anderson's view of science as one that adopts Bogen and Woodward's data-phenomena distinction and rejects the need to construct theories to explain the phenomena. On such a picture, theory is not needed at all: it is not needed to explain phenomena since there is no need to explain phenomena, and it is not needed to infer phenomena from the data—if the data is "big".

However, this doesn't seem to be the right way to interpret Bogen and Woodward. Although they argue that theory is not in the business of explaining data, this doesn't entail that theory is not involved in our inferences from data to phenomena. For example, Bogen and Woodward say that in order to infer phenomena from data we need to know such as things as whether confounding factors have been controlled for, whether there were systematic sources of error, and statistical arguments of various kinds, such as whether the conditions of Central Limit Theorem (CLT) apply (Bogen and Woodward 1988, p. 334.). To know such things, we often employ theories of various kinds—e.g., a theory will tell us whether a factor is a confounding one or not, and a lot of theory can bear upon whether the CLT applies (*cf.* Lyon 2014 and Woodward 2011a, p. 797). As Psillos has nicely put it:

> "...[T]he establishment of the epistemic authority of what is normally called the observable phenomenon (e.g., the melting point of lead) is a rather complicated process which essentially relies on background theory. If all these background theories somehow fail to be adequate (or well-confirmed, I should say), the observed phenomenon is called into question. Now, this does not imply that before we establish, say, the melting point of lead, we need detailed theories of *why* lead melts at this point. These theories will typically be the product of further theoretical investigation. But it does imply that establishing the reliability (and hence the epistemic authority) of the data as a means to get to stable phenomena relies indispensably on some prior theories. So, observation is not epistemically privileged *per se*. Its epistemic privilege is, in a certain sense, parasitic on the epistemic privilege of *some* theories." Psillos 2004, p. 406. (Emphasis in original)

Moreover, in response to Schindler, Woodward clarifies his view:

> "Data to phenomena reasoning, like inductive reasoning generally, is ampliative in the sense that the conclusion reached (a claim about phenomena) goes beyond or has additional content besides the evidence on which it is based (data). I believe it is characteristic of such reasoning that it always requires additional substantive empirical assumptions that go beyond the evidence." Woodward 2011b, p. 172.

So, Bogen and Woodward's data-phenomena distinction and their understanding of the distinction in scientific practice doesn't entail a "bottom-up" approach to phenomena detection.

This last point shows again what is wrong with the big data picture of science that appears to be gaining in popularity. Even (what may seem to be) purely statistical inferences are guided by empirical assumptions that are made on the basis of prior understanding and/or theory of the phenomena being studied. Woodward gives a nice example of what is involved in making an inference from fMRI data:

> "[...C]onsider the common use of spatial smoothing procedures in the analysis of fMRI data. Each individual voxel measurement is noisy; it is common to attempt to improve the signal to noise ratio by averaging each voxel with its neighbors, weighted by some function that falls with distance. This procedure depends (among other considerations) on the empirical assumption that the activity of each voxel is more closely correlated with nearby spatial neighbors." Woodward 2011b, p. 173.

If we didn't have the assumption that the voxel activities of nearby spatial neighbors are correlated, we wouldn't know that we can apply the spatial smoothing techniques to the data.

5 Conclusion

In this essay, I have reviewed some of the philosophical issues that arise in connection with the concept of data. In section 2, I reviewed four analyses of the concept that have appeared in the literature, and suggested how one in particular—the epistemic interpretation—can be improved. Since none of the analyses seemed all that promising when considered against all of the uses and instances of data, I suggested that it may be useful to first distinguish two concepts of data and give them separate analyses, rather than trying to treat data as a unified concept. In section 3, I discussed the role of big data in science and its impact on scientific methodology. Grand claims have been made about the advent of theory-less science in the wake of big data in science. However, these claims appear to be too strong, and less-grand ones should replace them. In particular, it appears that big data may be allowing for more science to be done with less theory, but so far it seems that theory is still involved. Finally, in section 4, I discussed an important distinction between data and phenomena in the philosophy of science and related the distinction to the epistemic interpretation of section 2 and the role of big data in science discussed in section 3.

References

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired 16*(7).

Baik, J.-J. and J.-S. Paek (2000). A neural network model for predicting typhoon intensity. *Journal of the Meteorological Society of Japan 78*(6), 857–869.

Basu, P. K. (2003, June). Theory-ladenness of evidence: a case study from history of chemistry. *Studies In History and Philosophy of Science Part A 34*(2), 351–368.

Bechtel, W. and A. Abrahamsen (2005, June). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 36*(2), 421–441.

Bogen, J. and J. Woodward (1988, July). Saving the Phenomena. *The Philosophical Review 97*(3), 303.

Bogen, J. and J. Woodward (1992). Observations, theories and the evolution of the human spirit. *Philosophy of Science 59*(4), 590–611.

Bogen, J. and J. Woodward (2003). Evading the IRS. In *Pznań Studies in the Philosophy of the Science and the Humanities*, pp. 223–256.

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? In *China-North America Library Conference, Beijing*.

Buhrmester, M., T. Kwang, and S. D. Gosling (2011, January). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science 6*(1), 3–5.

Callebaut, W. (2012, March). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43*(1), 69–80.

Carnap, R. (1946, December). Theory and Prediction in Science. *Science 104*(2710), 520–521.

Darwin, C. (1861). Darwin, C. R. to Fawcett, Henry. *Darwin Correspondence Database 3257*.

Davis, K. (2012). ***Ethics of Big Data: Balancing Risk and Innovation***. O'Reilly Media: Sebastopol, CA 95472.

Dretske, F. I. (1981). *Knowledge and the Flow of Information.* MIT Press: Cambridge, MA.

Fisher, R. A. (1936, April). Has Mendel's work been rediscovered? *Annals of Science 1*(2), 115–137.

Floridi, L. (2005, March). Is Semantic Information Meaningful Data? *Philosophy and Phenomenological Research 70*(2), 351–370.

Floridi, L. (2008). Data. In W. A. Darity (Ed.), *International Encyclopedia of the Social Sciences*.

Floridi, L. (2012a). Big data and their epistemological challenge. *Philosophy & Technology 25*(4), 435–437.

Floridi, L. (2012b). The search for small patterns in big data. *The Philosophers' Magazine 59*(4), 17–18.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind.* MIT Press: Cambridge, MA.

Glennan, S. S. (1996, January). Mechanisms and the nature of causation. *Erkenntnis 44*(1), 49–71.

Glymour, B. (2000, January). Data And Phenomena: A Distinction Reconsidered. *Erkenntnis 52*(1), 29–37.

Haig, B. D. (2013, May). Detecting Psychological Phenomena: Taking Bottom-Up Research Seriously. *The American journal of psychology 126*(2), 135–153.

Henrich, J., S. J. Heine, and A. Norenzayan (2010, June). The weirdest people in the world? *Behavioral and Brain Sciences 33*(2-3), 61–83.

Holmes, R. (2008). *The Age of Wonder*. HarperCollins: London, UK.

Hume, D. (1738). *A Treatise of Human Nature* (L. A. Selby–Bigge, 2nd ed.). Oxford: Clarendon Press.

Humphreys, P. (2006). Epistemologia del Siglo XXI. In *Revista anthropos*, pp. 65–70.

Ioannidis, J. P. A. (2013, April). Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research. *The American Journal of Bioethics 13*(4), 40–42.

Kaiser, M. (1991, October). From rocks to graphs — the shaping of phenomena. *Synthese 89*(1), 111–133.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos and A. Musgrave (Eds.), *Criticism and the Growth of Knowledge*, pp. 91–195.

Lyon, A. (2014). Why are Normal Distributions Normal? *The British journal for the philosophy of science 65*(3), 621–649.

Machamer, P., L. Darden, and C. F. Craver (2000). Thinking about mechanisms. *Philosophy of Science 67*(1), 1–25.

Mason, W. and S. Suri (2012, March). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods 44*(1), 1–23.

Massimi, M. (2007, June). Saving Unobservable Phenomena. *The British journal for the philosophy of science 58*(2), 235–262.

McAllister, J. W. (1997, September). Phenomena and Patterns in Data Sets. *Erkenntnis 47*(2), 217–228.

Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton University Press: Princeton, NJ.

Page, L., S. Brin, R. Motwani, and T. Winograd (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab.

Panza, M., D. Napoletani, and D. Struppa (2011). Agnostic Science. Towards a Philosophy of Data Analysis . *Foundations of Science 16*(1), 1–20.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Basic Books: New York.

Psillos, S. (2004). Tracking the Real: Through Thick and Thin. *The British journal for the philosophy of science 55*(3), 393–409.

Schindler, S. (2007). Rehabilitating theory: refusal of the 'bottom-up' construction of scientific phenomena. *Studies In History and Philosophy of Science Part A 38*(1), 160–184.

Shelton, T., M. Zook, and M. Graham (2012). The Technology of Religion: Mapping Religious Cyberscapes. *The Professional Geographer 64*(4), 602–617.

Suárez, M. (2005). The Semantic View, Empirical Adequacy, and Application (Concepción semántica, adecuación empírica y aplicación). *Crítica: revista Hispanoamericana de filosofía 37*(109), 29–63.

Votsis, I. (2010). Making Contact with Observations. In *Making Contact with Observations*, pp. 267–277. Dordrecht: Springer Netherlands.

Walton, K. L. (1990). *Mimesis as Make-believe*. Harvard University Press: Cambridge, MA.

Woodward, J. (1989). Data and phenomena. *Synthese 79*(3), 393–472.

Woodward, J. (2000). Data, phenomena, and reliability. *Philosophy of Science 67*(Sep), S163–S179.

Woodward, J. (2011a, January). Data, Phenomena, Signal, and Noise. *Philosophy of Science 77*(5), 792–803.

Woodward, J. F. (2011b). Data and phenomena: a restatement and defense. *Synthese 182*(1), 165–179.