Penultimate draft. To be published in Illary, P, and Russo, F. (eds) *The Routledge Handbook of Causation and Causal Methods*

Between quantity and quality: competing views on the role of Big Data for causal inference

Stefano Canali, Politecnico di Milano

Emanuele Ratti, University of Bristol

Abstract

When does more data help and when does it not in the sciences? In the past decade, this question has become central because of the phenomenon of Big Data. While these discussions started as a result of somewhat naive ideas that have been closely analyzed and mostly rejected in the philosophy of data, the question about the epistemic difference that more or less data make still matters, especially in light of the impressive performance of data science tools, which seem to improve their performance the more data are trained on. In several areas of the sciences, having more data is connected to methodological and epistemic benefits and something that research should strive towards. More data is often equated to better science: this elicits crucial questions about the epistemic value of the quantity of data. In this chapter, we discuss this problem in light of current discussions in the life and health sciences and the philosophy of data.

1. Introduction

When does more data help and when does it not? We interpret this as a question about the evidential basis of scientific and causal inference, in the context of ongoing discussions on the epistemological and methodological values and consequences of data-intensive methods in the sciences. While these discussions started as a result of somewhat provocative ideas on the epistemology of Big Data (Anderson, 2008) that have been closely analyzed and significantly criticized rejected in the philosophy of data (boyd and Crawford, 2012; Iliadis and Russo, 2016; Leonelli, 2014; Ratti, 2015), the question about the epistemic difference of more or less data still matters. In several areas of the sciences, having more data is connected to methodological and epistemic benefits and something that research should strive towards. More data is often equated to better science and more specifically the possibility of higher-quality causal inference and the identification of more precise causal relations. At the same time, it should also be noted that an appeal to naive inductivism (which presupposes the positive value of more data) is also rejected.

This creates a difficult situation, which elicits crucial questions about the epistemic value of the quantity of data for science in general, and for causal inference in particular.

In this chapter, we discuss this problem in light of current discussions in the life and health sciences and the philosophy of data. In the first part of the chapter, we present and review claims to the effect that more data entails better science. While the original provocative claims about the epistemic potential of big data for causal inference are often dismissed, similar arguments nonetheless inform the background of research programmes and approaches in the life and health sciences. We provide evidence for this claim with reference to attempts to give a philosophical basis to the inferential structure of Big Data (Pietsch, 2021), for instance in the literature on induction and discussions in areas including genomics and postgenomics, personalized and precision medicine, digital epidemiology.

In the second part of the paper, we present and critically dissect the literature against the idea that more data should be equated to better science and higher-quality causal inference. There are two sources for this literature. First, there is a literature in statistics showing that more data does not necessarily decrease the total error, but in some cases it can even increase it by magnifying the systematic error (Msaouel, 2022). Second, there is a literature that identifies the difficulties in data quality, interpretation, handling, curation, etc. in the sciences (Cai and Zhu, 2015), for instance in the context of epidemiology and medicine (Fleming et al., 2017; Vogt et al., 2019). In philosophy of science, this position has been presented in connection to arguments about the epistemic costs of dealing with more and different data (Leonelli, 2016).

2. Big data as a way of doing better causal inference

As we have seen in the introduction, there is a tendency to consider data as a valuable commodity, and the use of expressions such as ‘data is the new oil’ is not rare. The idea that more data is generally a good thing is especially connected to inductive frameworks. Intuitively, having more instances of something can be seen as a way to strengthen a particular model or law. At the same time, there has been a sharp opposition to naive inductive methods by the scientific community. Pietsch (2021) emphasizes this contrast. He describes this as something akin to a paradox. He says that in the context of Big Data, scientists and engineers are generally excited by the idea of abundance of data, while at the same time they reject the reasoning framework that most accommodates the use of large quantities of data.

According to Pietsch, the paradox can be solved when we are more precise in the kind of inductive framework we have in mind. When scientists and engineers reject induction, they usually have in mind enumerative induction, which is notoriously fraught with difficulties. However, Pietsch says, (most) Big Data methods do not implement enumerative induction; rather, they exploit another type of induction, which is called variational induction. Variational induction is immune to typical critiques made to enumerative induction. In other words, variational induction accommodates large quantities of data with induction without creating any friction. Because of his thorough analysis of various types of induction, we take Pietsch's work to be the most systematic and rigorous argument in favor of the claim that more data is generally better and can be used as a basis for high-quality causal inference. Pietsch's account of variational induction, as we will see, is presented as a tool for causal inference, similarly to other variational epistemologies (Russo, 2009). In order to grasp his argument, we will briefly introduce both enumerative induction and variational induction, detail how the claim for 'more data' changes depending on the type of induction one has in mind and discuss the presence of these philosophical positions in the biomedical literature.

2.1 Enumerative Induction, Variational Induction, and Big Data

Let's start with induction. In his books (2021, 2022) Pietsch distinguishes between three types of induction: enumerative, eliminative, and variational. We will not consider eliminative induction here because it is less relevant for the topic of this chapter, and there are also doubts that it constitutes an inductive inference in the first place (Ratti, 2015).

Enumerative induction is the most common type of induction. It is a type of inference that is taken to establish a general dependency between two types of events, let's say A and B. By having a sufficient number of instances where these two types of events co-occur and no instances in which they do not, we are justified in believing that there is a general dependency. Enumerative induction is a kind of ampliative inference. If we take all the co-occurring instances as premises, then enumerative induction is ampliative because the conclusion (the general dependency) is not contained in the premises. As it is well-known, there are many common limitations to enumerative induction. For instance, it is difficult to establish how many instances are necessary to justify the ampliative inference. A second limitation is the absence of criteria that can be used to distinguish relevant and non-relevant conjunctions of events. Despite these (and other) limitations, enumerative induction is considered a crucial type of induction, both in science and in everyday reasoning. Seeing more data as an epistemic good seems often connected to enumerative induction, given the emphasis placed on having as many instances as possible that can justify the ampliative inference.

Variational induction is Pietsch's alternative approach that looks at big data as a good basis for causal inference, without the limitations of enumerative indication. The term is rather new, and Pietsch borrows it from Russo's work (Russo, 2009; Russo and Williamson, 2007). In fact, it is known by other names: according to Pietsch, a representative case of variational induction is Mill's canons of induction, in particular the method of difference and the method of agreement, which are methods for causal inference (Mill 1886). Both methods, according to Pietsch, infer a causal relationship (or at least causal relevance) between a phenomenon and its circumstances "by relying on variational evidence, i.e., on evidence that tracks changes in a phenomenon resulting from systematic variations of circumstances" (Pietsch, 2021, p. 29). He describes this idea by means of a simple example. Imagine how one can learn that a light switch is causally related to a light. One way is to compare a situation where the switch is turned off and the light is off, with another situation where the switch is turned on and light is on. The comparison must be done carefully in order to ensure that nothing else that might be relevant has changed. The focus here is on changes in circumstances. There are a number of problems with variational induction (e.g. unrealistic conditions for the method of difference; issues in determining causal factors and alternative causes) that Pietsch tries to overcome with his own version of variational induction. While there is no space here to recall the specificities of Pietsch's variational induction, it is important to emphasize that, according to him, variational induction is what happens in most Big Data analytics and causal inference based on Big Data. There are indeed examples of Big Data analytics that are more akin to enumerative induction (e.g. association rules), but the most successful algorithms such as decision trees or neural networks have a variational inductive structure.

But how does exactly variational induction imply the idea that more data is generally better? At first, one may be inclined to interpret Pietsch's work in another direction, given the tensions between enumerative and variational induction. Differently from enumerative induction, confirmation in variational induction increases with the variety of evidence rather than with the number of co-occurrences. One can then say that the intuitive idea of 'more data' is not preserved. However, increasing the variety of evidence may well mean increasing the quantity of data as well – an increase in specific types of data. Rather than just accumulating data of the same type, we may want to increase the number of features considered, in the sense of "observing as many different situations in terms of changing circumstances as possible" (Pietsch, 2021, p. 30). But Pietsch provides also other useful indications. In responding to the objection that more data will increase the risk of spurious correlations, he adds that we can certainly be satisfied with a sufficient amount of data correctly prepared, but also that "any arbitrary amount of further data will not prevent the analysis in terms of variational induction" (Pietsch, 2021, p. 61). It is possible to conclude that, in the absence of precise indication of when enough data is enough, more data is always wise for induction as a basis of causal inference.

2.2 Big Data and Causal Inference in the Life and Health Sciences

Pietsch's work is based and developed on a mostly theoretical stance. This has the merit of elucidating the theoretical background of positive answers to the question of whether quantity of data is a positive gain for scientific epistemology and causal inference in particular. In addition, it has the advantage of providing a theoretically-grounded answer to this question, in a context where several similar claims have associated the quantity and volume of scientific data with revolutionary changes for the epistemology of scientific research, but often with little theoretical and epistemological depth. Famously, around the start of the increasing presence of Big Data as a buzzword in public and academic discussions, commentators and scholars such as Chris Anderson

(2008), Viktor Mayer-Schönberger and Kenneth Cukier (2014), and others argued that the increasing availability of larger and larger volumes of data was the dawn of a new epistemological revolution, with specific and critical changes towards a diminishing role of theories and conceptual assumptions, a new approach based on powerful correlations at the expense of traditional causal inference, and overall new revolutions and paradigms for scientific epistemology. These claims were often based on unsound and approximate views of scientific practice and epistemology, and have largely been rebutted by philosophical analyses of the last decade, as philosophers have turned their attention to the epistemic role of data in the sciences and presented accounts of the ways in which scientific Big Data are used and treated. For instance, large quantities of data are often laden and entangled with theoretical and conceptual assumptions, and they are still very much used to build causal and mechanistic arguments, while not being necessarily or automatically instances of revolutions and paradigms for scientific research. As we have seen in this section, however, Pietsch's work goes in a different direction and presents a theoretically-grounded positive answer to the question we discuss in this chapter. At the same time, we want to illustrate claims in various areas of the scientific literature that go in a direction that is similar to Pietsch's account.

In the life sciences and genomic research in particular, the availability of more omics and sequencing data is traditionally associated with the possibility of new discoveries and clinical applications, for instance in cancer research the discovery of new cancer genes may be used by drug discovery teams to design a new drug; conversely, gaps in knowledge of relevant disease causal processes and treatment are often associated to a general lack of data (Garraway and Lander, 2013; Martínez-Jiménez et al., 2020). Many of these claims are gaining more significance and specificity with the use of machine learning and deep learning models in this context, where the reliability and effectiveness of these methods is predicated upon the availability of larger quantities of data for learning (Jiao, et al., 2020). Similar perspectives seem to be expressed in other areas of the life and health sciences, including epidemiology (Fleming et al., 2017) and personalized and precision medicine (Topol, 2014), where the availability of large datasets is seen as a way of grounding stronger causal inference on population and public health and individual determinants of health and disease (see also the chapter by Bas De Boer in this volume).

While these positions from the scientific literature do not necessarily present clear references to epistemological or methodological choices on the use of larger volumes of data, we can interpret some of them under the light of our previous discussion of Pietsch's work. In particular, elements of Pietsch's position on variational approaches to inductive and causal inference can be seen in the context of this literature. Some claims on the role of larger volumes of data in genomics seem to argue for an increase of types of data in the sense of an increase of the variety of evidence, and thus the possibility of observing more and more diverse situations and circumstances and producing better causal inference. This can for instance be seen in the context of the development of new -omics features and related data and technical tools, which is often presented as the basis of new discoveries and clinical applications as these allow to diversify and expand the evidential basis of genomics. For instance, in a landmark review, Vogelstein et al. (2013) praise the advancement in genome sequencing that has allowed the construction of detailed models of cancer and causal knowledge (e.g. how certain proteins disrupt certain cellular processes) in this field. However, they also highlight the limitations of drug discovery approaches that focus, for practical reasons, mostly on oncogenic alterations, while neglecting tumor suppressor gene mutations. But this is because the knowledge of biological pathways that can be exploited to at least indirectly drug tumor suppressor genes is limited. Therefore, they plead for more data, not just in quantity, but also in variety (e.g. genomics, transcriptomics, proteomics, etc), which is a hallmark of variational induction. This is an interesting take, because the limitations of an approach (i.e. the focus on cancer genes and the limits of their 'druggability') are not addressed by modifying the approach; rather they are addressed by having more of the same. In a bioinformatics analysis of whole genomes (Martínez-Jiménez et al., 2020, p. 82), members of ICG and TCGA motivate the creation of such big consortia by the need to systematically document "somatic mutations that drive common tumor types". In order to do this, more data is seen as the solution. In fact, they endorse the 'more data' mantra at the end, by saying that the insights of the analysis "can be obtained only from an integrated analysis of all classes of somatic mutations on a whole-genome scale" (Martínez-Jiménez et al., 2020, p. 92). This speaks to the idea of variational induction, given the emphasis on 'all classes of somatic mutations'.

Personalized and precision medicine approaches are also often pitched as a new way of closing existing data gaps and more specifically providing more variations in terms of the groups of patients that usually participate in biomedical research. In this way, claims for expanding the evidential basis of biomedical research seem to associate the possibility of new discoveries and types of causal inferences on the basis of expansion in the volume and the variety of research participants and thus data. For example, according to the bioinformatics analysis that we have just mentioned, a barrier for precision medicine is constituted by the lack of "knowledge banks comprising tens of thousands of patients with comprehensive clinical characterization" (Martínez-Jiménez et al., 2020, p. 92), where 'comprehensive' refers to ideal data sets covering all dimensions of patients' physiology and biology. In similar terms, the need for more monitoring, surveillance, and datafication of new aspects of our lives in quantifiable biomedical terms seems is usually justified on the basis of variational ideas, whereby expanding the volume and variety of data types is seen as a way of closing research gaps and developing better and more specific interventions (Celi, 2022).

3. Big Data as an issue for better causal inference

As we have mentioned in the introduction, several of the bolder claims about the value of large datasets in the sciences and their epistemological consequences have been significantly criticized in the philosophical literature. This line of work has generally led to more specific and situated accounts of the use of large datasets in the sciences, which show substantial variety over different research contexts, methodologies and approaches to data, and paint more detailed and subtle pictures of the epistemology of Big Data in science. As a result, the tendency to consider data as a valuable commodity has been counteracted with works looking at the significant epistemic work (and difficulties) that are part and parcel of the employment of using large volumes of scientific data, including the need to interpret the data, label and curate them, assess and measure their quality, and beyond.

In this section, we present recent work in philosophy of science that can be seen as a negative answer to the main question we discuss in this chapter. In particular, we start by presenting recent work on statistical issues and considerations on the use of large datasets, and then discuss various work in philosophy of science that identifies the difficulties in data quality, interpretation, handling, curation, with a specific focus on biomedical research.

3.1 The Big Data Paradox

The expression “Big Data paradox” designates the idea that “the more the data, the surer we fool ourselves” (Meng, 2018, p. 2). Msaouel (2022) draws a general lesson from this principle, arguing that the widespread enthusiasm for large quantities of data should be toned down, because having more data does not translate into better statistical inferences.

In a nutshell, this works as follows. Confidence intervals are ranges of estimates of the true value of a parameter. For instance, if you are saying that the mean height of adult European males is 175 cm ± 1 cm, you are saying that the mean height falls within that range. When we have a wider confidence interval, then the probability that a new observation will contain the true value increases, just because the range is wider and contains more values. Msaouel argues that when confidence intervals become narrower, the probability that a new observation will include the true value decreases. In other words, when you have a narrow confidence interval, then there is a smaller chance of drawing an observation that lies within that interval. In a way, a “wider interval implies less precision, and a narrower interval implies more precision” (Rothman, 2012, p. 169). We know that confidence intervals are a function of different factors, including sample size, and we know that when sample size increases, confidence intervals get narrower. This is, in a nutshell, the “Big Data paradox” as Msaouel conceives it: “as the size of a study increases, the probability that its confidence intervals will include the truth decreases” (Msaouel, 2022, p. 1). One defining feature of Big Data is the effort to get bigger and bigger data sets, and hence bigger sample sizes; therefore, having more data can lead to lower probabilities that the true value of the parameter lies in the confidence interval that one calculates, just because the range of values is narrower.

Two things should be noted about this formulation. First, the notion of paradox is a bit of a misnomer, because the situation is paradoxical only on the view that more data is always better, which is both an empirical and theoretical thesis that is uncritically assumed. In other words, there is really no reason why we should expect that more data is always better, so why are we puzzled once we discover that this claim is unwarranted? Another thing to point out is that the idea that narrower confidence intervals will have less probability of containing the truth (which is Msaouel's formulation) is a bit misleading. It seems to us that confidence intervals and confidence levels are conflated here. A confidence interval is, as we have discussed, the range of values where you think the true value of the parameter lies, while the confidence level is how much you would expect to get those values when drawing an observation from a sample outside your initial sample. But this does not mean that by narrowing the confidence interval the confidence level will decrease. Therefore, it is not clear exactly what is meant with this initial formulation of the paradox.

There is also another way to explain the 'paradox', which is less confusing. In fact, it does not make any confusing reference to a 'paradox' at all. Rothman formulates this view in the following way:

"if an error in a study would decrease if the study were larger, then that error is a random error, whereas an error that would not decrease if the study were larger is a systematic error" (Rothman, 2012, p. 143)

The alternative explanation hence is built by referring to the different types of errors conceptualized in statistics. As widely known, statistics is fraught with difficulties and uncertainties. One important goal is being able to estimate what is called the 'total error', which is the error of when one tries to estimate the true value of a parameter. The total error is commonly conceptualized as the sum of the reducible error and the irreducible error. The latter refers to errors in measurement that are due to unknown factors not represented in the dataset, and it is assumed that it cannot be eliminated. The reducible error is the error that, as the name suggests, can be reduced, and it is usually understood as the sum of the systematic error and the standard error. The standard error is simply how spread are measurements from the established mean. The distance between the real function and the model itself is called 'bias' or 'systematic error', and it is systematic because it depends on constraints and assumptions behind the statistical model.

Now, when the sample size of an experiment like a Randomized Controlled Trial in medicine increases, we should expect the standard error to shrink. Confidence intervals will shrink as well, given the connections between confidence intervals and the standard error. The slogan "more data is always better" can thus be rephrased as the idea that diminishing the standard error by increasing the sample size will necessarily lead to a better statistical inference, and possibly, better causal inference. However, by diminishing the standard error, we are only considering one component of the reducible error (i.e. the standard error) while ignoring other components (i.e. the systematic error). The problem is that with this procedure we assume that the standard error is more costly than the systematic error, but we know that systematic errors can lead to false inferences and "unreliable predictions even when the standard error is low" (Msaouel, 2022, p. 7). Therefore, one can increase the sample size in order to reduce one type of error, but by doing this the totality of the reducible error is not necessarily reduced, as more data will also not address other types of errors, as well as gaps in quality assessment that we discuss next.

3.2 Difficulties in Data Quality, Interpretation, Curation

Beyond the aforementioned theoretical motivations, there are several, more concrete issues that make the appeal to the volume of data problematic. These are more ‘in-practice’ issues, as they show that with bigger data sets come also more practical issues to make them reliable sources. These issues can be addressed with unlimited time and resources or with a more pluralist approach to methodology, which combines Big Data for hypothesis generation with other approaches. However, resources are limited and Big Data often tends to take the lead on other approaches (Leonelli, 2020; see also the chapter by Dingmar van Eck and Kristian Gonzalez Barman in this volume) – this situation can make Big Data more of a burden rather than a blessing.

The first issue is data quality (see also Zahle in this volume). The bigger the data set, the higher is the probability that the data set will contain gaps and mistakes in it. As one can expect, with poor quality data, the reliability of statistical and causal inference decreases. We can conceptualize this possibility as a random error. This is also discussed by Msaouel, albeit not in terms of random error. One may of course place equal emphasis on both data quality and data quantity, but the more data quantity, the less are the resources that we can dedicate to issues of data quality. Even though for different (and more theoretical) reasons, this is called the ‘data quality-quantity’ tradeoff by Meng (2018). For example, this is a significant issue in areas of biomedical research that make increasing use of commercial and digital technologies, such as wearable devices: the variability of types of devices and sensors in this context, added to the large volumes of data wearables can collect, make it difficult to assess and test the quality of data (Canali et al., 2022). In this context, data quality is sometimes discarded as an issue that can be prevented by the sheer size of large volumes of data – there is often the implicit assumption that, as more and more data are available, quantity can trump quality. And yet, on top of the considerations of Msaouel and the data quality-quantity tradeoff identified by Meng, several analyses of the concrete uses of large volumes of data in the sciences have questioned this assumption.

In spite of ideas of Big Data being the basis of complete descriptions of the phenomena under study for causal inference, questions of representativity and sampling cannot be easily dismissed when the use of large datasets is at the center of research (boyd and Crawford, 2012; Leonelli, 2020). As the evidential basis of research expands and more and more data points can be collected, this does not automatically translate into the possibility of collecting *all* the necessary data points and studying phenomena completely. On the contrary, questions of sampling need to be increasingly at the center of Big Data research. Larger volumes of data often mean that some data points are more represented than others and more data are disproportionately available. For example, data collection practices involving the use of digital technologies in epidemiology – often presented in terms of digital epidemiology (Salathé, 2018) – often suffer from the fact that some socio-economic and demographic groups are disproportionately more represented than others (Fiske et al., 2022; Zinzuwadia and Singh, 2022). In this sense, Big Data is often problematic because the sheer volume of data runs the risk of suggesting that sampling considerations are less or not at all necessary for causal inference. Similar considerations have been developed in the context of discussions on issues of bias and quality of the data used for machine learning in the sciences, for instance in medicine (Grote and Berens, 2023; Ratti and Graves, 2022; see also chapters by Giorgia Pozzi and Juan M. Durán and Giuseppe Primiero and Alberto Termine in this volume).

Issues of representativity and data quality as issues related to the quantity of large datasets have also been discussed in terms of gaps in the causal representations provided by Big Data, especially when used as a basis for causal inference. For example, in the context of personalized medicine there is a general tendency to expand screening campaigns and the collection of more individual and thus larger volumes of data (Vogt et al., 2016). As a result of this increase in volume of data from screening, Green and Vogt have argued that personalized medicine runs significant risks of over-representing the presence of some pathologies in the general population and thus leading to overdiagnosis, which is very harmful for individual patients as well as healthcare systems (Green and Vogt, 2016; Vogt et al., 2019). In this case, one of the issues seems to be that, as the volume of data increases, datasets might end up representing different types of phenomena – for instance, different causal stages and types of disease, which might not be equally harmful and therefore should not necessarily require medical interventions. This is a critical issue for causal inference in medicine that is increasingly based on large datasets. In turn, it also brings up additional and different questions on the epistemic validity and overall quality of Big Data, further showing that increases in the quantity of scientific data does not simply equal to epistemic gains in the sciences – rather, in many cases, they change the types of research questions and methods that can and should be asked (Leonelli, 2016).

On top of these discussions on the quality and evidential content of large datasets in the sciences and issues therein, an emerging body of literature has also looked at the epistemic and practical costs of dealing with increasing volumes and levels of variety and diversity involved in Big Data. This literature presents an additional issue for ideas of Big Data as a positive development for scientific research – epistemic benefits need to be balanced against the fact that in some cases the use of large datasets can be an obstacle to scientific research. Here, empirically-grounded work has documented the substantial work that is involved in using Big Data in scientific research, with a specific focus on the life sciences. For instance, in biological research on model organisms, substantial work is dedicated to packaging data to travel, including labeling, cleaning, curating data so that they can be used in the first place (Leonelli, 2016, 2009; Leonelli and Tempini, 2020). This work is crucial to enable data to "travel" from the initial point of creation and collection to new areas of research and uses, thus potentially fulfilling one of the more general promises of Big Data as an expansion of the evidential basis of research. At the same time, this epistemic role of data curation, preparation, and handling can expand the epistemic value of a dataset, by making not only data accessible and reusable for other researchers, but also associable to a broader range of phenomena – thus also countering ideas of Big Data in the sciences as a case of too many data or data deluge (Müller-Wille and Charmantier, 2012; Strasser, 2019) substantial work is dedicated to packaging data to travel, including labeling, cleaning, curating data so that they can be used in the first place.

And yet, the epistemic costs of dealing with more and different data are also connected with the aforementioned issues of reliability and quality of Big Data for causal inference. The sheer volume of data make it increasingly difficult to test and validate the quality of the data, particularly when the same dataset is employed for uses that are significantly distant from the initial type of research they were collected for – several approaches, including the test of data quality at the collection stage or the use of automated and algorithmic approaches have had little success so far (Leonelli, 2017). This is increasingly happening in research contexts where several standards for high-quality research and data are introduced – consider discussions on reproducibility, fairness, explainability – and still it remains largely unclear which tools and principles of quality should be used (Leonelli, 2018). As a result, these obstacles connected to the use of Big Data can lead to the preference towards only some types of data and research methods that are deemed high-quality. This has been discussed in terms of "convenience experimentation" by Krohs (2012), as data-driven research practices have become entrenched in the life sciences and are perceived as lower risk than others. For instance, omics data and related technologies and experimental techniques remain central modes of inquiry and are often privileged and preferred tools for causal inference in the context of postgenomics (Canali, 2020).

In addition to these issues, the coupling of Big Data sets and Machine Learning (ML) methods has created another problem. It has been shown that ML systems are especially vulnerable to external tamperings. A famous example is the one of adversarial attacks (Watson 2019). In (2015), Goodfellow et al show that just introducing a small perturbation within pixels of a picture can fool a classifier like GoogLe Net on ImageNet in mislabelling a panda as a gibbon. This vulnerability to small perturbations is not just an innocuous curiosity of the ML community; it is a profound challenge. For instance, in medicine (Fynlainson et al 2019) it has been proved that even highly accurate medical classifiers can be tampered with quite easily; a classifier labeling as 'benign' an image of a mole with a confidence of 99%, has shifted its classification to 'malignant' with 100% confidence after a small perturbation. As this simple example easily shows, the consequences for fields like epidemiology or precision medicine can be daunting.

What does this problem have to do with Big Data? It should be noted that the issue is not necessarily epistemic; rather it is about security, namely how one can tamper with algorithms and data sets for malicious purposes. But the security issue is exacerbated by our inability to explore and get a grasp at bigger and bigger datasets – and hence more data become less desirable in such a context. To understand this point, consider the opportunities that one possibly has in perturbing ML systems. Papernot et al (2018) identify such opportunities at the stage of 'training' and at the stage of ‘inference’ (that is, when training is completed and “the model is deployed to infer predictions on inputs unseen during training” p 401). In the training phase, one can perturb the way the model is constructed by the algorithm by accessing training data. Because small perturbations may suffice, one needs to have access only to small portions of training and testing data. For those who have to defend the system against this type of vulnerability by monitoring datasets in real-time, it is literally searching for a needle in the haystack, where the ‘needle’ is the small perturbation. In the inference phase, one might use information about the model architecture, parameters or, again, training data to identify where the model is vulnerable, and altering a new input accordingly. Alternatively, one can just modulate inputs and see when the model performance just drops. In both cases, big amounts of data make ‘surveillance’ more daunting. First, (similarly to the problems during the training phase) because big data sets are more difficult to explore and understand. Second, because with bigger datasets might come bigger architectures, which will be increasingly difficult to survey given the well-known problem of opacity (Creel 2020). This is to say that having more data can potentially lead us to having more vulnerable ML systems, which is something that we should avoid, at least in principle.

4. Conclusions

In this chapter, we have critically engaged with the arguments in favor and against the idea that more data has significant epistemic benefits for causal inference. We have started with the positive answers. We have identified in Pietsch’s work the most systematic way of defending the epistemic benefits of Big Data, and we have also identified ways in which the scientific discourse (especially in the life sciences) seems to embed the same attitude that Pietsch displays towards the benefits of inductive processes. Finally, we have identified two classes of arguments that run against the intuition that more data leads to substantial epistemic benefits for scientific inference. The first argument is more theoretical, and it is based on the so-called ‘Big Data paradox’, while the second is an ‘in-practice’ argument, and it emphasizes issues of data quality and the costs of data curation.

References

Anderson, C., 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete | WIRED. Wired.

boyd, danah, Crawford, K., 2012. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society 15, 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cai, L., Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. CODATA 14, 2. https://doi.org/10.5334/dsj-2015-002

Canali, S., 2020. Making evidential claims in epidemiology: Three strategies for the study of the exposome. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 82, 101248. https://doi.org/10.1016/j.shpsc.2019.101248

Canali, S., Schiaffonati, V., Aliverti, A., 2022. Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness. PLOS Digital Health 1, e0000104. https://doi.org/10.1371/journal.pdig.0000104

Celi, L.A., 2022. PLOS Digital Health, a new journal driving transformation in the delivery of equitable and unbiased healthcare. PLOS Digit Health 1, e0000009. https://doi.org/10.1371/journal.pdig.0000009

Fiske, A., Degelsegger-Márquez, A., Marsteurer, B., Prainsack, B., 2022. Value-creation in the health data domain: a typology of what health data help us do. BioSocieties. https://doi.org/10.1057/s41292-022-00276-6

Fleming, L., Tempini, N., Gordon-Brown, H., Nichols, G.L., Sarran, C., Vineis, P., Leonardi, G., Golding, B., Haines, A., Kessel, A., Murray, V., Depledge, M., Leonelli, S., 2017. Big Data in Environment and Human Health, in: Oxford Research Encyclopedia of Environmental Science. Oxford University Press. https://doi.org/10.1093/acrefore/9780199389414.013.541

Garraway, L.A., Lander, E.S., 2013. Lessons from the Cancer Genome. Cell 153, 17–37. https://doi.org/10.1016/j.cell.2013.03.002

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. http://arxiv.org/abs/1412.6572

Green, S., Vogt, H., 2016. Personalizing Medicine: Disease Prevention in silico and in socio. HUMANA.MENTE Journal of Philosophical Studies 9, 42.

Grote, T., Berens, P., 2023. Uncertainty, Evidence, and the Integration of Machine Learning into Medical Practice. The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine 48, 84–97. https://doi.org/10.1093/jmp/jhac034

Iliadis, A., Russo, F., 2016. Critical data studies: An introduction. Big Data & Society 3, 205395171667423. https://doi.org/10.1177/2053951716674238

Krohs, U., 2012. Convenience experimentation. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43, 52–57. https://doi.org/10.1016/j.shpsc.2011.10.005

Leonelli, S., 2020. Scientific Research and Big Data. Stanford Encyclopedia of Philosophy.

Leonelli, S., 2018. Rethinking Reproducibility as a Criterion for Research Quality, in: Fiorito, L., Scheall, S., Suprinyak, C.E. (Eds.), Research in the History of Economic Thought and Methodology. Emerald Publishing Limited, pp. 129–146. https://doi.org/10.1108/S0743-41542018000036B009

Leonelli, S., 2017. Global Data Quality Assessment and the Situated Nature of “Best” Research Practices in Biology. Data Science Journal 16, 32. https://doi.org/10.5334/dsj-2017-032

Leonelli, S., 2016. Data-centric biology: a philosophical study. The University of Chicago Press, Chicago London.

Leonelli, S., 2014. What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society 1, 205395171453439. https://doi.org/10.1177/2053951714534395

Leonelli, S., 2009. On the Locality of Data and Claims about Phenomena. Philosophy of Science 76, 737–749. https://doi.org/10.1086/605804

Leonelli, S., Tempini, N. (Eds.), 2020. Data Journeys in the Sciences. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-37177-7

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., Gonzalez-Perez, A., Lopez-Bigas, N., 2020. A compendium of mutational cancer driver genes. Nat Rev Cancer 20, 555–572. https://doi.org/10.1038/s41568-020-0290-x

Mayer-Schönberger, V., Cukier, K., 2014. Big data: a revolution that will transform how we live, work, and think, First Mariner Books edition. ed, An Eamon Dolan book. Mariner Books, Houghton Mifflin Harcourt, Boston New York.

Meng, X.-L., 2018. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. Ann. Appl. Stat. 12. https://doi.org/10.1214/18-AOAS1161SF

Mill, J.S., 1886. System of Logic. Longmans, Green & Co, London.

Msaouel, P., 2022. The Big Data Paradox in Clinical Practice. Cancer Investigation 40, 567–576. https://doi.org/10.1080/07357907.2022.2084621

Müller-Wille, S., Charmantier, I., 2012. Natural history and information overload: The case of Linnaeus. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43, 4–15. https://doi.org/10.1016/j.shpsc.2011.10.021

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and Privacy in Machine Learning. Proceedings - 3rd IEEE European Symposium on Security and Privacy, EURO S and P 2018, 399–414. https://doi.org/10.1109/EuroSP.2018.00035

Pietsch, W., 2022. On the Epistemology of Data Science: Conceptual Tools for a New Inductivism, Philosophical Studies Series. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-86442-2

Pietsch, W., 2021. Big Data, 1st ed. Cambridge University Press. https://doi.org/10.1017/9781108588676

Ratti, E., 2015. Big Data Biology: Between Eliminative Inferences and Exploratory Experiments. Philosophy of Science 82, 198–218. https://doi.org/10.1086/680332

Ratti, E., Graves, M., 2022. Explainable machine learning practices: opening another black box for reliable medical AI. AI Ethics. https://doi.org/10.1007/s43681-022-00141-z

Rothman, K.J., 2012. Epidemiology: an introduction, 2. ed. ed. Oxford Univ. Press, New York, NY.

Russo, F., 2009. Variational Causal Claims in Epidemiology. Perspectives in Biology and Medicine 52, 540–554. https://doi.org/10.1353/pbm.0.0118

Russo, F., Williamson, J., 2007. Interpreting Causality in the Health Sciences. International Studies in the Philosophy of Science 21, 157–170.

https://doi.org/10.1080/02698590701498084

Salathé, M., 2018. Digital epidemiology: what is it, and where is it going? Life Sci Soc Policy 14, 1. https://doi.org/10.1186/s40504-017-0065-7

Strasser, B.J., 2019. Collecting experiments: making big data biology. The University of Chicago Press, Chicago.

Topol, E.J., 2014. Individualized Medicine from Prewomb to Tomb. Cell 157, 241–253. https://doi.org/10.1016/j.cell.2014.02.012

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W., 2013. Cancer Genome Landscapes. Science 339, 1546–1558. https://doi.org/10.1126/science.1235122

Vogt, H., Green, S., Ekstrøm, C.T., Brodersen, J., 2019. How precision medicine and screening with big data could increase overdiagnosis. BMJ l5270. https://doi.org/10.1136/bmj.l5270

Vogt, H., Hofmann, B., Getz, L., 2016. The new holism: P4 systems medicine and the medicalization of health and life itself. Med Health Care and Philos 19, 307–323. https://doi.org/10.1007/s11019-016-9683-8

Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. Minds and Machines, 29(3), 417–440. https://doi.org/10.1007/s11023-019-09506-6

Zinzuwadia, A., Singh, J.P., 2022. Wearable devices—addressing bias and inequity. The Lancet Digital Health S2589750022001947. https://doi.org/10.1016/S2589-7500(22)00194-7