Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2016: 1–11

DOI: 10.1177/2053951716669530
bds.sagepub.com
SAGE

Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS

Stefano Canali

Abstract

Recently, it has been argued that the use of Big Data transforms the sciences, making data-driven research possible and studying causality redundant. In this paper, I focus on the claim on causal knowledge by examining the Big Data project EXPOsOMICS, whose research is funded by the European Commission and considered capable of improving our understanding of the relation between exposure and disease. While EXPOsOMICS may seem the perfect exemplification of the data-driven view, I show how causal knowledge is necessary for the project, both as a source for handling complexity and as an output for meeting the project's goals. Consequently, I argue that data-driven claims about causality are fundamentally flawed and causal knowledge should be considered a necessary aspect of Big Data science. In addition, I present the consequences of this result on other data-driven claims, concerning the role of theoretical considerations. I argue that the importance of causal knowledge and other kinds of theoretical engagement in EXPOsOMICS undermine theory-free accounts and suggest alternative ways of framing science based on Big Data.

Keywords

Big Data epistemology, data-intensive science, EXPOsOMICS, causality, complexity, biomarkers

Introduction

Big Data is increasingly used in different areas of human activity, with an arguably significant impact on our society. If one looks at discussions regarding Big Data, its value is often presented in terms of the abundance of correlations and the consequent possibility of data-driven research, without strong theoretical considerations and causal knowledge (Mayer-Schönberger and Cukier, 2013; Anderson, 2008). Recently, however, a number of scholars have raised questions about these claims, by highlighting the difficult and complicated tasks needed for the use of Big Data in the sciences (Leonelli, 2014a), the technical impossibility of many data-driven claims (Kitchin, 2014b) and the role of hypotheses in research based on Big Data (Ratti, 2015).

In this article, I assess data-driven claims on causal knowledge by focusing on the research of EXPOsOMICS. EXPOsOMICS is a current biomedical project, where Big Data is used to study the associations between exposure and disease (Wild, 2012). The project has many features which make it a very interesting case study, including its novel, multidisciplinary and methodologically aware nature (Illari and Russo, 2013: 175). EXPOsOMICS is a particularly relevant case to look at from the perspective of the data-driven view, as the use of Big Data, the focus on associations and the absence of a strong theory of disease causation (Illari and Russo, 2013: 177) may make it look like the perfect exemplification of the view. In the article, nevertheless, I argue that causal knowledge plays a crucial role in EXPOsOMICS, as one of the key sources and outputs of the project.

Methodologically, this paper is the result of case study research based on different sources. In order to concretise and clarify the argument, I use the studies published by

Leibniz University Hannover, Germany

Corresponding author:
Stefano Canali, Leibniz University Hannover, Institute of Philosophy, Am Klagesmarkt 14–17, 30159 Hannover, Germany.
Email: stefano.canali@philos.uni-hannover.de

scientists as part of their research and concerning specific exposure and disease associations. Moreover, the attention to methodology and the significant number of articles researchers dedicate to the discussion of how they work with Big Data, correlations and causality have allowed me to study closely the methodological choices of the project. Another significant source for the argument of the paper is the interview I carried out with the Principal Investigator of EXPOsOMICS, Professor Paolo Vineis (see Supplementary material), who gave me key insights about Big Data research in epidemiology, the curation activities of the project and its approach towards the issues of correlations and causality. Finally, the work on these sources on EXPOsOMICS was carried out on the basis of the current debate in critical data studies, developed in different disciplines including philosophy, sociology, computer science, etc.

The article is structured as follows. In the first section I present EXPOsOMICS, highlighting the features which make it novel and challenging biomedical research and suggesting that it should be seen as a Big Data project according to a broad and multi-featured definition of Big Data. Next, I directly engage with data-driven claims about causal knowledge. In the second section, I describe how researchers have developed a specific methodology for working with Big Data and I argue that this is a way of studying disease causation. I show that researchers' engagement with causality is a consequence of two distinct kinds of complexity, one affecting large datasets and the other the target systems examined in the project. In the third section, I argue that the use of Big Data in EXPOsOMICS is substantially based on various sources of knowledge and cannot be reduced to a mechanical 'extraction' of causality from the data. In the fourth section, I show how causal knowledge is not only important as an input for Big Data research: it is equally crucial as an output to try to meet the goals of the project. Hence, I conclude that EXPOsOMICS shows how causal knowledge is a necessary element of Big Data research and data-driven claims regarding causality are flawed. Finally, in light of the results on causal knowledge, in the fifth section of the paper I consider the theory-free claim of the data-driven view; I argue that this is in contrast with the role of theoretical considerations in crucial stages of EXPOsOMICS. In the conclusion, I present new questions coming up from the results of the article, especially concerning their validity in other Big Data projects.

EXPOsOMICS and the Big Data literature

EXPOSOMICS is a current biomedical collaborative project, structured as a consortium of 13 research centres in Europe and the US and coordinated by the Department of Epidemiology and Biostatistics of Imperial College London. At the end of 2012, the European Commission funded the project with a budget of €8.7 million. The group of researchers is highly interdisciplinary (Illari and Russo, 2013: 175), comprising experts in epidemiology, biology, bioinformatics, statistics, information and communication technologies, public health and risk assessment. EXPOSOMICS studies the associations between exposure and disease in order to assess the levels of disease risk connected with elements and features of the environment. The project has many innovative and interesting features. The first one I would like to highlight regards the new concept of exposure researchers have developed. Usually, we think of exposure as exposure to external factors; this is reflected in traditional epidemiology, which looks at exposure to environmental factors in order to account for disease. On the other hand, more recently genomics has developed a way of explaining disease relying on an internalist perspective, by focusing on the genetic features of the organism and using 'omics' technologies (which collect data about genes, mRNA, proteins and metabolites), EXPOsOMICS aims at filling in the gaps between these two approaches: since many diseases 'develop predominantly from a combination of environmental exposures played out on a particular genetic background' (Wild, 2012: 24), their explanation cannot be reduced to either environmental factors or internal genetic features. Consequently, researchers in EXPOsOMICS have developed a new concept, the 'exposome' (Wild, 2005), which is the total sum of exposures affecting individuals during their lifetime and comprises what is called 'internal' and 'external' exposures. That is, when for instance trying to explain how chlorination by-products in water may lead to bladder cancer (Vineis and Chadeau-Hyam, 2011: 101), researchers studied both contamination due to the by-products (external exposure) and consequent responses in the body (internal exposure). Another example is breast and colon cancer, for which Chadeau-Hyam et al. (2011) investigated both the diet and lifestyle of patients and the metabolic responses to them. Through this 'global' approach to exposure, scientists intend to develop more accurate evaluations of environmental factors and identify crucial stages of disease evolution (Illari and Russo, 2013: 175). Practically, the study of the exposome is carried out through the search for what scientists call 'biomarkers'. These are biological markers, i.e. elements or features of the environment and the organism which can be measured and indicate biological processes (think of e.g. proteins or metabolites). In particular, scientists try to find associations between biomarkers of exposure and biomarkers of disease, thus following the path of disease

right from exposure and tracking its initial steps. These conceptual novelties make EXPOsOMICS a case of 'frontier research' (Illari and Russo, 2013: 175) in the realm of epidemiology. Yet, these are not the only innovations of the project. Another novel aspect regards the fact that scientists retrieve or collect Big Data as sources for their study of biomarkers. In particular, among the main sources of the project are large cohort studies, which collect data about thousands of patients for a number of years. Researchers are creating new cohorts, but they also rely on existing databases (e.g. the European Prospective Investigation into Cancer and Nutrition). In addition, scientists often use databases that focus on particular chemicals or biological samples (e.g. the Humane Metabolome Database). The data used in the project is usually produced by high-throughput technologies: for internal exposure, omics clearly stands out as the most used kind of technology (Vineis et al., 2009); for external exposure, researchers rely on a variety of technologies, including sensors, geo-referencing, satellites, smartphones, etc. (Wild, 2012: 26). The use of Big Data brings about both exciting possibilities and challenges, which concern how to process, analyse and handle different datasets and consequently require experts in statistics, bioinformatics and information and communication technologies. In addition to these, the interdisciplinary group of researchers in EXPOsOMICS also comprises experts in public health, policy and risk assessment. This allows me to introduce another key feature of this innovative project. We have seen that, through the research on the exposome, scientists intend to follow the development of diseases at different stages. This is where the role of the aforementioned policy experts comes in, since predictions and knowledge about disease evolution are used to inform policy. For instance, as a consequence of the study of chlorination by-products in water and their effects on bladder cancer, researchers may be able to determine acceptable and risk levels of chlorine in water. That is, the innovative theoretical and methodological elements of EXPOsOMICS research are considered capable of having a significant impact on issues of primary importance, like risk assessment and public health.

I have highlighted that the use of Big Data is one of the innovative aspects of EXPOsOMICS research. But what is Big Data and why can EXPOsOMICS be defined a Big Data project? On the face of it, defining Big Data looks trivial: Big Data seems to simply refer to the fact that we now have technologies capable of collecting, storing and processing 'big' datasets; on this view, EXPOsOMICS would be a Big Data project insofar as its researchers work with large amounts of data. However, as highlighted in the discussions and literature on Big Data, the definition is less straightforward. In official documents like those of the Intellectual Property Office (2014) and the Parliamentary Office of Science and Technology (2014), three elements are used to define Big Data: volume, velocity and variety, which are known as the 'three Vs'. According to the three Vs definition, Big Data surely is the result of the great volume of data collected, but also of the high velocity with which data is collected (in terms of accumulation and event rates) and the large variety of the data itself (in terms of the variety of sources and data structure). The idea that Big Data is simply *Big* Data seems even more problematic if we consider that the three Vs definition has been criticised in the literature. The philosopher Floridi (2012: 435–436) highlights the relational nature of the three categories, arguing that relationality requires more specificity and hence other categories. In recent years, several other categories and consequently broader definitions of Big Data have been proposed in the social sciences literature (see Kitchin and MacArdle, 2016: 1–2). On the basis of a review of these definitions, Kitchin (2013: 262–263, 2014a: 27–28) focuses on the categories making Big Data qualitatively different from small data. In addition to the three Vs, features which are unique to Big Data include the possibility of conjoining different types of data, the strong detail and resolution allowing for unique identifications, the ability of focusing on entire populations and the high scalability of data production. Such a broader definition of Big Data, taking into account many different aspects, is the most suited to describing EXPOsOMICS. Indeed, while researchers in the project certainly benefit from the volume, velocity and variety of the datasets they work with, they need the high level of resolution allowing them to study exposure at both the external and internal levels. Moreover, we have also seen that researchers work with a variety of technologies and sources and, from this perspective, the possibility of conjoining different types of datasets is another key feature. Therefore, I would argue that EXPOsOMICS is a Big Data project in the sense suggested by Kitchin's broad definition of Big Data, which is the one I shall stick to throughout the paper.

In the literature on Big Data, discussions do not simply concern its definition but also what we can do with it; this is where the epistemological questions come in. In general, Big Data is often considered capable to offer 'unprecedented opportunities for data-driven discovery and decision-making in virtually every area of human endeavor' (US National Science Foundation, 2015: 4). In the literature, the main focus of the debate is on the nature of these opportunities and the way they can be achieved. According to Mayer-Schönberger and Cukier (2013), these opportunities are due to the incredible amount and variety of data,

which allows us to find an equally incredible amount of meaningful correlations and, on the basis of them, make sound predictions. For Mayer-Schönberger and Cukier this implies that, while normally we would have had to design a hypothesis first and then test it against a dataset, data is now so big and comprehensive that we can work with correlations only and use them to make predictions; theoretical elements like hypotheses or theories are not necessary. Moreover, the value and abundance of correlations in Big Data entail that there is no need to study the causal relations between the correlated variables; causal knowledge is not necessary, neither as a guidance for nor as an output of research. Mayer-Schönberger and Cukier's data-driven view can be synthesised in two elements, which I will now label for use throughout the article: (i) Big Data allows for predictions and discoveries free of theoretical elements; (ii) correlations found in the data are enough and causal knowledge is not necessary for the exploitation of Big Data. However powerful and fascinating, the idea that Big Data can make science data-driven has been quite criticised in the literature. Kitchin (2014b: 3–5) highlights how crucial elements in Big Data research, such as methodological choices are neglected by the data-driven view. Kitchin argues that large datasets, although aimed at capturing entire domains, are still samples requiring methodological considerations to be correctly picked. Methodology, according to Kitchin, is also important after the stage of data collection, as statements based on Big Data are not universal and independent of their context. From a similar perspective, boyd and Crawford (2012) argue that data analysis always requires methodological choices and is 'most effective when researchers take account of the complex methodological processes that underlie' it (boyd and Crawford, 2012: 668). Critiques can also be found in the philosophical literature. Floridi (2014) argues that exploiting the value of Big Data is not as simple as many suggest: the value of Big Data lies in the possibility of finding 'small patterns', but these require significant theoretical engagement and the design of the 'right questions' about what we want to collect, look for, infer, etc. A practical instance of this theoretical engagement can be seen in the work of the philosopher of science Leonelli on scientists' practices of data curation. Leonelli shows how curation is necessary to exploit the scientific value of Big Data and consists in a significant number of activities requiring explicit and tacit knowledge (Leonelli, 2014b: 404–411). Data-driven claims on causality have been criticised as well, especially in the social sciences literature. For instance, Titiunik (2015) focuses on causal inference in political science, arguing that correlations in Big Data are no substitute to causal knowledge, which is always required in order to make valid causal inferences (for a review of recent discussions in the field, see Clark and Golder, 2015). This is a good starting point, but I think that more work needs to be done concerning how exactly Big Data can be used for causal inference and whether this is generally possible in the sciences. As for biomedical sciences in particular, while philosophers have vastly argued that causal and mechanist knowledge plays a crucial epistemological role (see e.g. Bechtel and Abrahamsen, 2005), it is not clear whether this is still the case in Big Data science. At the same time, whilst authors such as Leonelli (2014a: 2–3, 8) and Kitchin (2014a: 135) mention causal knowledge as one of the major issues of Big Data, they do not discuss it in detail and do not analyse case studies directly related to it. Therefore, I think that there are important questions to answer regarding causal knowledge in Big Data science and this is what I will focus on in this paper, investigating EXPOsOMICS from the perspective of what I have defined as the second element of the data-driven view.

The meet-in-the-middle approach, complexity and causality

EXPOsOMICS is a Big Data project where scientists look for associated biomarkers, capable of tracing exposure and disease. The proponent of the data-driven view may say that this project is the perfect example showing how Big Data research consists in gathering large amounts of data, analysing it, looking for correlations between biomarkers of exposure and biomarkers of disease and making predictions. This would show how correlations are enough and there is no need for causal knowledge. In order to assess these claims, I will now look at the methodology of the project, as used in specific studies and explained in a number of articles. As a way to make the argument more precise, I shall mainly refer to the aforementioned study of breast and colon cancer carried out by Chadeau-Hyam et al. (2011), which clearly shows the main features of researchers' methodology. In this case the first step consisted in looking for associations in the data to produce lists of putative associated markers of exposure and disease (Chadeau-Hyam et al., 2011: 85). According to the data-driven view, researchers should have stopped at this analysis of data and associations; however, associations were used only as a starting point and – I argue – as a way of studying disease causation. That is, in EXPOsOMICS, when a correlation between biomarkers is identified as statistically significant, scientists apply what they call the 'meet-in-the-middle approach' in order to look for intermediate biomarkers between the ones of exposure and disease, i.e. what lies 'in the middle' of correlations. In the breast and colon cancer case, the application of the meet-in-the-middle

approach consisted in the comparison between two lists of associated makers and, then, the search for a third biomarker at the intersection between the two (Chadeau-Hyam et al., 2011: 85–86). As made clear by researchers' words, this search for intermediate biomarkers is considered a way of investigating disease causation. Chadeau-Hyam et al. define intermediate biomarkers as '*causal links* between exposures and disease' (2011: 84; emphasis added). Assi et al. define the meet-in-the-middle approach as way 'to unravel utmost important steps in the *aetiology* of disease' (2015: 752; emphasis added). When introducing the approach in 2007, Vineis and Perera characterised the use of the term intermediate biomarker 'in a very broad sense to encompass all measurable markers (in body fluids or in cells) that lie within the putative *causal pathway* linking an exposure to the onset of disease' (2007: 1959; emphasis added). Vineis also highlights that the inspiration for the approach was Salmon's (1984) characterisation of causal processes in terms of mark transmission.[1]

Hence, in contrast with the data-driven view, researchers working with Big Data in EXPOsOMICS do not think that correlations are enough and have designed a specific methodological approach for the investigation of causal links. But why do they think that studying causality is necessary? I will argue that this is precisely because correlations are not enough, since they are affected by a number of issues and need to be 'validated'. That is, as Vineis and Perera explain, it is necessary to understand whether the relation between associated elements is a causal one or is only a consequence of e.g. side effects or confounding (2007: 1961). In other words, scientists study disease causation as a way to handle issues affecting correlations in big datasets, which in turn are due to two different kinds of complexity. A first kind of complexity affects the data itself: the volume of data is so large that it comprises thousands of variables; consequently, when scientists analyse the dataset looking for correlations, the problem is not finding correlations but rather finding too many of them. Additionally, a second kind of complexity affects the target systems studied by scientists. As Vineis explains, this is quite evident in the case of cancer, for which we do not know a necessary cause and, probably, single instances of cancer are linked to a variety of (both strong and weak) exposures. Therefore, scientists cannot focus on just one variable in their datasets or a correlation between variables, but need to study a variety of them. At the same time, though, the complexity of the target systems also implies that the variables in the datasets interact in non-simple ways: for one, variables are not only correlated with their causes, but also with their effects, the other effects of its causes, etc. For example, scientists may find correlations between biomarkers tracing, say, high levels of glucose and biomarkers signalling the development of breast and colon cancer. The problem is that, at this stage of research, scientists may not be able to tell if the data about levels of glucose increases their knowledge of cancer development or can be used to suggest policy, because it might be an effect of cancer or both the high level of glucose and cancer might be caused by something else. Consequently, in this situation scientists need to be sure that none of these issues is in place, i.e. they need to validate the association between glucose and cancer; this, we have seen, is done through the meet-in-the-middle approach and the search for an intermediate causal link. Practically, in the breast and colon cancer case, Chadeau-Hyam et al. ran different statistical tests on the lists of associated makers and ditched possibly positive results as a consequence of considerations regarding statistics, previous experiments, disease mechanisms and causation. For instance, the ranking of significance was found consistent between limited and larger samples, but this was seen as possibly leading to 'uncontrolled confounding for the matching variables' (Chadeau-Hyam et al., 2011: 86). Moreover, researchers relied on existing causal knowledge on disease mechanism, through for instance the Human Metabolome Database, which gathers qualitative descriptions, labels, visualisations, etc. on metabolomic mechanisms (Wishart et al., 2007). Thanks to this combination of data, theoretical and methodological considerations and existing causal knowledge, they studied the causal links between exposure and disease, identifying the dietary intake of fibres as a probable intermediate biomarker (Chadeau-Hyam et al., 2011: 86).

Formal causal models and the crucial role of *knowledge in*

In the previous section, I have focused on the methodological novelty designed in EXPOsOMICS, the meet-in-the-middle approach, arguing that researchers use it to study disease causation and solve correlations' issues. The approach shows how, in contrast with data-driven claims, causality is necessary for the Big Data research of EXPOsOMICS. Yet, in spite of these points, the proponent of the data-driven view may point out that certain analytical tools, like formal causal models, make it possible to 'extract' causality from datasets, without strong theoretical considerations and in a significant data-driven way.[2] Formal causal models are quantitative tools, which apply a statistical interpretation to the data in order to study the causal relations of a target system. They are composed of two elements: a number of variables, which describe the causal structure of the system and may be displayed in terms of tables, directed graphs

or structural equations; a probability distribution attached to each variable (as synthesised by Hitchcock, 2009: 300–301). When studying a system, two are the main scopes of formal causal models (as summarised by Illari and Russo, 2014: 62–64). They can be used either to model the causal relations in order to make predictions, if something is known about the causal relations of the system, or to model phenomena in order to get information and make discoveries about causal relations, if these are unknown. Considering these features and their use in big datasets (Illari and Russo, 2014: 60–61), formal causal models may be considered the way to go in EXPOsOMICS. For instance, in the case of breast and colon cancer, researchers may describe the causal relations by listing a number of variables (features of diet and lifestyle, such as smoking, levels of sugar, cholesterol, etc. and elements of the metabolic system, like levels of glucose, creatinine, lipids, etc.) and, on the basis of data analysis, assigning probabilities to the variables.[3] In this way, scientists may be able to make predictions. If, say, a certain level of cholesterol is observed, scientists may substitute this value for the variable within the system, resolve the equations and obtain a predicted level of, for instance, glucose. In principle, this would avoid one of the kinds of complexity we have seen in the previous section: the simplification of datasets through equations would reduce their dimension and allow researchers to focus on just a few variables, thus avoiding complexity due to the volume of data. In turn, proponents of data-driven claims may say that the use of formal causal models warrants a weaker version of their view: while scientists may find causality necessary for their research, the idea would be that they could study it on the basis of data only and without relying on existing causal knowledge or strong theoretical considerations.

However, while researchers in EXPOsOMICS do assign probability distributions to their variables and use various statistical models, they do not use formal causal models. Again, the problem here is complexity, specifically the kind of complexity affecting target systems, which leads to huge gaps between the systems and formal causal models. That is, while the representation of a system might be approximated so that formal causal models can be used, approximations would leave out too many of the system's important aspects. For example, consider the causal Markov condition, one of the conditions formal causal models need to meet. It states that, for each variable V of the system, the direct causes of V screen off V from anything but the direct effects of V: that is, the probability of V only depends on its direct causes and is independent of anything else except its effects (see Hitchcock, 2009: 306–308; Illari and Russo, 2014: 68–69). The condition is quite problematic for the target systems studied in EXPOsOMICS, because important causes may not be direct or it may be difficult to understand which causes are direct. That is, often scientists do not know enough about the target system to apply the causal Markov condition and approximating the representation of the system so that the condition can be applied to the model may make it radically different and spurious.[4]

Because of complexity, thus, the minimal theoretical engagement and the arguably data-driven nature of formal causal models are not enough. Researchers need to use other statistical tools, which I will argue are based on an extensive use of different theoretical elements, including causal knowledge. The importance of these methodological elements and the role played by causal knowledge are highlighted by scientists in a significant number of articles dedicated to introducing the different statistical models used in the project. For instance, in the case of breast and colon cancer, researchers assigned a probability distribution to the variables of the system, but then used a 'multivariate' statistical approach called O-PLS (Chadeau-Hyam et al., 2011: 85). Multivariate approaches work with data obtained through multiple measurements and sources by finding a reduced number of principal variables – called Principal Components, PCs – in the data. That is, O-PLS and other multivariate approaches do not work on the whole structure of the datasets, but aim at finding the PCs which are capable of 'reducing' the large datasets of EXPOsOMICS into structures. Significantly, PCs are identified thanks to considerations about statistical performance and disease mechanisms (see Trygg and Wold, 2002). In addition to multivariate models, the complexity of both the target systems and big datasets of EXPOsOMICS require the use of other models. There are cases where the reduction to few PCs is not useful, because these may not reflect the whole data in its diversity or the studied phenomenon may not work as a consequence of a few independent variables (Chadeau-Hyam et al., 2013: 548–549). Consequently, researchers also use variable selection models, which allow them to directly select a specific subset of variables and predictors, so that their performance as estimators is improved (the selection works as a penalised regression). Again, variables are selected relying on existing knowledge about the causal nature and the statistical features of the estimator (see a technical overview of these selection methods in Yuan and Lin, 2006). Another type of model used in EXPOsOMICS is the univariate one (Chadeau-Hyam et al., 2013: 544–546). This works separately on data, by associating a predictor (e.g. omics measurement) with outcomes of interest (e.g. presence/absence of disease) and tends to use linear models, as a result of the continuous nature of most omics measurement

(having the form of e.g. levels of cholesterol, percentages of metabolomic developments, etc.). Univariate models are very useful because they are computationally efficient, flexible in accommodating different types of data and available in most statistical software. Considerations regarding the data and use of causal knowledge are crucial in this case too, since, especially in the more complex situations, it is necessary to make hypotheses about the nature of the causal relation between variables, predictor and outcome.

Hence, we have seen that the study of causality by scientists in EXPOsOMICS cannot consist in extracting causality from the data with a minimal engagement with theoretical and methodological considerations. In contrast with the data-driven view, scientists consider causality a crucial element of their Big Data research and carry out this research relying on existing causal knowledge. I would define this use of causal knowledge in terms of *knowledge in*, meaning what researchers have to 'put in' their routine in order to handle complexity issues, study the causal links between exposure and disease and exploit the value of their big datasets. This use of existing causal knowledge is in line with significant work in the literature on causal discovery in the sciences, especially with the idea of 'no causes in, no causes out' developed by the philosopher Cartwright. According to her, when trying to scientifically investigate the causality of phenomena, 'old causal knowledge must be supplied for new causal knowledge to be had' (Cartwright, 1989: 39). Having seen how old causal knowledge is supplied in EXPOsOMICS, I will now turn to what it means to have new causal knowledge as a goal of the project.

EXPOsOMICS goals and *knowledge out*

In the previous sections we have seen how researchers working with Big Data in EXPOsOMICS think that correlations are not enough and they need to study causality, on the basis of theoretical considerations and existing causal knowledge. Now, I will show that studying causality in EXPOsOMICS is necessary not only to validate the associated biomarkers of big datasets, but also to try and meet the project's goals: if scientists only worked with correlations and did not study causality, they would unlikely meet these aims. Let me go back to the previous example of glucose, for which I imagined that scientists found associations between biomarkers tracing levels of glucose and biomarkers signalling the development of breast and colon cancer. In such a situation, the level of glucose may be a good predictor of cancer even if it is not causally linked with the disease, as for instance it may have a causal relation with another element which is, in turn and separately, linked with cancer. Therefore, one may say that correlations are enough. However, we should remember that one of the main goals EXPOsOMICS is informing policy interventions. For this sort of goals, correlations are not enough: it is crucial to understand whether the correlated items are also causally linked or not; if they are not, policies and interventions may be useless. Consider again the glucose example: on the basis of correlations only, scientists would unlikely suggest policies regarding, say, the presence of glucose in diets. The problem here is that correlations alone will not tell scientists whether e.g. levels of glucose are linked with effects or causes of cancer, which is fundamental for interventions. That is, if certain levels of glucose are linked with effects of cancer, intervening on them will not yield positive results; on the other hand, if they are linked with causes of cancer, implementing policies on glucose may change the presence or evolution of the disease. For these goals, therefore, studying the causal path between exposure and disease is necessary and the search for intermediate biomarkers is precisely aimed at that, since intermediate biomarkers are either indicators of the causal path or causal links themselves.

Hence, in EXPOsOMICS scientists do not simply extract correlations or causality from the data, but, thanks to theoretical, statistical and methodological considerations, explicitly intend to add new causal knowledge to the existing literature in order to meet the goals of the project. In other words, researchers aim at producing new casual knowledge on the basis of Big Data research. Such a production of causal knowledge can be seen as an instance of *knowledge out*, i.e. the output of scientists' activities, potentially capable of having an impact on external situations like public health and policy-making. Therefore, for the EXPOsOMICS case causal knowledge plays a crucial and necessary role, both 'in', as one of the sources of research, and 'out', as one of the outcomes of the project. This fundamentally undermines what I have labelled as one of the main claims of the data-driven view (ii), according to which correlations found through Big Data analysis are enough and causal knowledge is redundant. In contrast with these claims, evidence from EXPOsOMICS suggests that causal knowledge should be considered a legitimate and necessary element of Big Data research.

The role of theoretical considerations in EXPOsOMICS

In the previous sections, as the main focus of the article, I have argued that causal knowledge is necessary for the Big Data research of EXPOsOMICS. Before focusing on the methodology of the project, I have distinguished two main claims composing the data-driven view,

i.e. the ideas of research free of theoretical elements (i) and causal knowledge (ii). As I have already said, the first element of the data-driven view has been widely examined in the literature. Now, I would like to connect the case study research of the article to this literature by looking at (i) from the perspective of EXPOsOMICS. For this purpose, I will emphasise again the role of theoretical, methodological and statistical considerations, which – as we have seen – play a fundamental role in scientists' research, one without which research would not be possible. For instance, the colon and breast cancer case was shaped by researchers' methodological considerations, like the decision to use the multivariate statistical model. This echoes boyd and Crawford (2012) and Kitchin's (2014b) critiques of (i), which emphasise the importance of hypotheses, sampling and methodology. In particular, the following statement by boyd and Crawford may be seen as a quite appropriate characterisation of EXPOsOMICS: 'data analysis is most effective when researchers take account of the complex methodological processes that underlie the analysis of that data' (boyd and Crawford, 2012: 668). Another stage of research where the literature has significantly highlighted the role of theoretical elements is data curation, which is the set of activities surrounding the management, organisation, storage, sharing, etc. of datasets and has become a fundamental aspect of research in different parts of the life sciences (see e.g. Leonelli and Ankeny, 2012). The idea is that, if data is not stored in a continuously maintained accessible database, if it is not easily searchable, if it is not properly labelled (Boem, 2016), i.e. if it is not curated, not much can be done with it; since curation is also based on theoretical elements, Big Data research cannot be considered theory-free (Leonelli, 2013). This can also be seen in the practices of EXPOsOMICS. When researchers collect data, for instance analysing blood samples through mass spectrometry, they carry out what Vineis calls 'pre-processing practices', mainly consisting in removing impossible results and nuisances (data cleaning) or selecting the elements which are most relevant to the study (data selection). Vineis highlights how these are grounded in scientists' knowledge: results are cleaned as a consequence of considerations about biological plausibility and are selected depending on the sort of study researchers are carrying out (for instance, researchers may discard water-related associations because the study focuses on air-related ones).

Therefore, EXPOsOMICS is significantly based on knowledge and theoretical elements. As a consequence, I would argue that these features of EXPOsOMICS are in line with critiques of (i) expressed in the literature. In addition to this, I would also say that the EXPOsOMICS case can give interesting suggestions on alternative ways of framing the role of theoretical elements in Big Data research: indeed, while these significantly shape research, it is not really correct to say that they 'come first' and drive research. For example, in the breast and colon cancer case, research started with data from the European Prospective Investigation into Cancer and Nutrition,[5] which had not been personally collected and curated within EXPOsOMICS. Theoretical considerations were carried out in subsequent steps, like when scientists chose the multivariate statistical model, applied the meet-in-the-middle approach, used mechanist knowledge about the disease, etc. Hence, sometimes it may be that scientists begin with the data and do not apply strong a priori hypotheses, theories or models; yet, theoretical elements need to be used subsequently. In other cases, however, theoretical elements may need to be used a priori, for instance when scientists rely on *knowledge in*, choose to focus on the collection of data regarding specific elements of the target system and accordingly carry out data cleaning. Hence, while theoretical elements do not always drive research, saying that data always drives the project would not be accurate either. Consequently, the EXPOsOMICS case suggests that we should abandon accounts of Big Data science according to which there is a strict relationship between theoretical considerations and data, so that either theoretical considerations or data drive research. In contrast with both theory- and data-driven views, in EXPOsOMICS data and theoretical aspects are mutually influencing elements: as in a loop, research may start with and come back to data and theoretical elements more than once. For example, in the breast and colon cancer case researchers did start with the data, but then at the more theoretical level used the meet-in-the-middle approach, again there was an analysis of the data for the intermediate biomarkers, then results were compared with the literature and existing causal knowledge, etc. The value of these suggestions drawn from EXPOsOMICS can be seen in the comparison with other accounts of Big Data science proposed in the literature. Kitchin (2014b: 5–7) argues that Big Data research starts off with an initial exploration, used to find and generate new hypotheses 'born from the data'; such an exploration is guided by theoretical elements like existing knowledge. This view correctly highlights the crucial role played by theoretical considerations, but it may still seem that data always drives research; on the other hand, we have seen that in EXPOsOMICS scientists often begin with hypotheses which are not born from the data. Ratti (2015) distinguishes between Big Data projects based on a hybrid data- and hypothesis-driven approach and 'data mining studies' like EXPOsOMICS, where 'researchers look for robust regularities in metadata associations' (Ratti, 2015:

210). Ratti argues that in data mining studies the elimination of putative results depends only on the robustness of regularities found in the data and the overall goal is obtaining predictions or generalisations, not discovering mechanisms. I would say that EXPOsOMICS (if considered a data mining study) shows that things are different. In the breast and colon cancer case, scientists removed the results of the putative lists of biomarkers on the basis of statistical as well as theoretical considerations; moreover, data and correlations were used to understand disease causation and the mechanistic link between biomarkers (as also argued by Illari and Russo, 2013: 187–188).

Conclusion

In this paper, I have shown the flaws of the data-driven account of Big Data science. Focusing on EXPOsOMICS, I have highlighted how researchers study causality as a consequence of correlations' issues and different kinds of complexity. I have shown that causal knowledge is a necessary element of the project, both as a source to shape research (*knowledge in*) and as an output to meet its goals (*knowledge out*). I have consequently concluded that data-driven claims about the redundancy of causal knowledge are flawed and causal knowledge should be considered a necessary element of Big Data science. Besides, I have argued that the EXPOsOMICS case is also relevant to the discussion on theoretical elements, insofar as in the project it is not that either data or theoretical elements strictly drive research, but rather both of them play a crucial role and mutually influence each other.

Considering the results of the paper, one may question their generality and applicability, arguing that EXPOsOMICS is a special case with radically different features from most of Big Data science. While more research needs to be done on these points, I think that the aspects of EXPOsOMICS which undermine the data-driven view – the significant presence of complexity and goals of suggesting policy – exist in many other Big Data projects. As for the level and kinds of complexity, these – especially the complexity of target systems – of course vary in each case, but the kind of complexity and the consequent correlations' issues due to the volume of Big Data are typical of most Big Data research (as also argued by Illari and Russo, 2013: 183). Moreover, we should not think that complexity regards scientific projects only, as it also affects the economic and social environment where Big Data is considered to potentially have a great impact.[6] As for policy suggestions, the use of Big Data for this is not exclusive to EXPOsOMICS: actually, one of the contexts where Big Data is often considered capable of making a difference is precisely policy-making.[7] Therefore, while other Big Data projects may be substantially different from EXPOsOMICS when it comes to research areas, methodology, novelty, etc. the reasons why correlations are not enough and causal knowledge is necessary for Big Data research may still be present. These results about causal knowledge are not to be seen only as critiques of the data-driven view: rather, on the positive side, they suggest that other projects may benefit from a stronger consideration of causal knowledge. In addition, new questions on the role of causal knowledge arise. For instance, it remains unclear how Big Data research may change as a consequence of a stronger consideration of causal knowledge and whether the production of causal knowledge is always possible or limited by the features of specific projects. Moreover, since we have seen that causal knowledge is necessary for policy interventions, another important issue regards the consequences on privacy and ethics, i.e. questions concerning whether causal knowledge requires that we know more and at a more specific level. Thus, the results of this article indicate the need for further research and raise new questions about the appropriate framing of Big Data science. Indeed, studying the way we frame Big Data research is decisive, as improving our use of Big Data also depends on what we think it can do and what kind of data-driven, methodologically engaged, causality-free, etc. ideas we associate with it.

Acknowledgement

I wish to thank the *Deutsche Forschungsgemeinschaft* (DFG) for supporting this work as part of the research training group GRK 2073 "Integrating Ethics and Epistemology of Scientific Research" (https://grk2073.org). I am deeply indebted to Phyllis Illari, Billy Wheeler and Thomas Reydon for useful comments and suggestions on previous drafts. This paper was crucially improved due to detailed comments from the anonymous referees and the editors of this special issue, Federica Russo and Andrew Iliadis. I am also grateful to Paolo Vineis for discussions on these issues. Remaining errors are, of course, my own.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: the research for this paper was funded by the *Deutsche Forschungsgemeinschaft* (DFG) as part of the research training group GRK 2073 "Integrating Ethics and Epistemology of Scientific Research".

Notes

1. By reporting Vineis' words, I want to highlight that researchers in EXPOsOMICS explicitly aim at investigating causality as a crucial part of their research and I do not intend to endorse Salmon's view as an account of causality in EXPOsOMICS. My aim in this paper is not arguing that EXPOsOMICS suggests a specific account of causality, but rather that it shows how causal knowledge is necessary for Big Data science. However, one may say that I need to specify what I mean by causality and the account I commit to. The focus on links, mechanisms and the reference to Salmon suggest that a production approach to causality may be the right account of causality in biomarkers research. Indeed, Illari and Russo (2013) draw on this tradition in order to develop their informational account of the causal links scientists look for in EXPOsOMICS. As the references I use throughout the article suggest, I find this account quite promising and capable of correctly depicting the approach of the project (see also Russo and Vineis, forthcoming). Yet, I also think that defending Illari and Russo's account is beyond the scope of this paper and investigating its validity in Big Data science more generally needs further research, as for instance Pietsch (2016: 147–148) has argued that causal modelling in Big Data is based on the difference-making account of causation.
2. For formal causal models I refer to Pearl (2000), whose use of structural equations generalises causal Bayesian Networks. For an introduction to the literature on formal causal models, see Illari and Russo (2014: 60–85).
3. Here, I apply Hitchcock's (2009: 302) characterisation of formal causal models to a disease example.
4. The issue of not having enough knowledge is also highlighted by Hausman and Woodward (1999: 580).
5. For an introduction to the project and data collection practices, see Riboli et al. (2002).
6. See Mitchell (2009) for a summary of complexity in different contexts, including economy and policy-making.
7. See e.g. the goals and vision of the Alan Turing Institute for Data Science, launched in November 2015: 'inform scientific and technological discoveries, create new business opportunities, accelerate solutions to global challenges, inform policy-making, and improve the environment, health and infrastructure' (2015).

References

Alan Turing Institute (2015) The vision. Available at: https://turing.ac.uk/ (accessed 26 March 2016).

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 23 June. Available at http://www.wired.com/2008/06/pb-theory/ (accessed 26 March 2015).

Assi N, Fages A, Vineis P, et al. (2015) A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study. *Mutagenesis* 30(6): 743–753.

Bechtel W and Abrahamsen A (2005) Mechanistic explanation and the nature-nurture controversy. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Boem F (2016) Orienteering tools: biomedical research with ontologies. *Humana.Mente Journal of Philosophical Studies* 30: 37–65.

boyd D and Crawford K (2012) Critical questions for Big Data. *Information, Communication and Society* 15(5): 662–679.

Cartwright N (1989) *Nature's Capacities and Their Measurement*. Oxford, UK: Clarendon Press.

Chadeau-Hyam M, Athersuch TJ, Keun HC, et al. (2011) Meeting-in-the-middle using metabolic pro ling – A strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers* 16(1): 83–88.

Chadeau-Hyam M, Campanella G, Jombart T, et al. (2013) Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environmental and Molecular Mutagenesis* 54: 542–557.

Clark WB and Golder M (2015) Big Data, causal inference, and formal theory: Contradictory trends in political science? *PS: Political Science & Politics* 48(1): 65–70.

Floridi L (2012) Big Data and their epistemological challenge. *Philosophy & Technology* 25: 435–437.

Floridi L (2014) Big Data and information quality. In: Floridi L and Illari P (eds) *The Philosophy of Information Quality*. Cham (CZ), Switzerland: Springer International Publishing.

Hausman DM and Woodward J (1999) Independence, invariance and the causal Markov condition. *British Journal for the Philosophy of Science* 50: 521–583.

Hitchcock C (2009) Causal modelling. In: Beebee H, Hitchcock C and Menzies P (eds) *The Oxford Handbook of Causation*. Oxford, UK: Oxford University Press, pp. 299–314.

Illari P and Russo F (2013) Information channels and biomarkers of disease. *Topoi* 35: 175–190.

Illari P and Russo F (2014) *Causality: Philosophical Theory meets Scientific Practice*. Oxford, UK: Oxford University Press.

Intellectual Property Office (2014) Eight great technologies: Big Data, a patent overview. *Report of the Intellectual Property Office*, UK, June.

Kitchin R (2013) Big Data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography* 3(3): 262–267.

Kitchin R (2014a) *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles, CA: Sage.

Kitchin R (2014b) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12.

Kitchin R and McArdle G (2016) What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3(1): 1–10.

Leonelli S (2013) Classificatory theory in biology. *Biological Theory* 7(4): 338–345.

Leonelli S (2014a) What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society* 1(1): 1–11.

Leonelli S (2014b) Data interpretation in the digital age. *Perspectives on Science* 22(3): 397–417.

Leonelli S and Ankeny RA (2012) Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 29–36.

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that Will Transform How We Live, Work and Think*. London, UK: John Murray.

Mitchell SD (2009) *Unsimple Truths: Science, Complexity, and Policy*. Chicago, IL: The University of Chicago Press.

Parliamentary Office of Science and Technology (2014) Big Data: An overview. *Report of the Parliamentary Office of Science and Technology*, UK, July.

Pearl J (2000) *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Pietsch W (2016) The causal nature of modeling with Big Data. *Philosophy & Technology* 29: 137–171.

Ratti E (2015) Big Data biology: Between eliminative inferences and explanatory experiments. *Philosophy of Science* 82(2): 198–218.

Riboli E, Hunt K, Slimani N, et al. (2007) European Prospective Investigation into Cancer and Nutrition (EPIC): Study populations and data collection. *Public Health Nutrition* 5(6b): 1113–1124.

Russo F and Vineis P (forthcoming) Opportunities and challenges of molecular epidemiology. In: Boniolo G and Nathan MJ (eds) *Philosophy of Molecular Medicine: Foundational Issues in Research and Practice*. London, UK: Routledge.

Salmon WC (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton, MA: Princeton University Press.

Titiunik R (2015) Can Big Data solve the fundamental problem of causal inference? *PS: Political Science & Politics* 48(1): 75–79.

Trygg J and Wold S (2002) Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* 16: 119–128.

US National Science Foundation (2015) Critical techniques, technologies and methodologies for advancing foundations and applications of Big Data sciences and engineering. Available at: http://www.nsf.gov/pubs/2016/nsf16512/nsf16512.htm (accessed 26 March 2016).

Vineis P and Chadeau-Hyam M (2011) Integrating biomarkers into molecular epidemiological studies. *Current Opinion in Oncology* 23(1): 100–105.

Vineis P and Perera F (2007) Molecular epidemiology and biomarkers in etiologic cancer research: The new in light of the old. *Cancer Epidemiology Biomarkers Prevention* 16(10): 1954–1965.

Vineis P, Khan AE, Vlaanderen J, et al. (2009) The impact of new research technologies on our understanding of environmental causes of disease: The concept of clinical vulnerability. *Environmental Health* 8(54).

Wild CP (2012) The exposome: From concept to utility. *International Journal of Epidemiology* 41: 24–32.

Wild CP (2005) Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers Prevention* 14(8): 1847–1850.

Wishart DS, Tzur D, Knox C, et al. (2007) HMDB: The human metabolome database. *Nucleic Acids Research* 35: D521–D526.

Yuan M and Lin Y (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society* 68: 49–67.

This article is a part of Special theme on Critical Data Studies. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/critical-data-studies.