Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445 🅭
http://www.rifanalitica.it

Seventh Workshop on the Philosophy of Information

"Conceptual Challenges of Data in Science and Technology"

[London, March, 30th-31st 2015]

Stefano Canali

I would argue that, by now, it is safe to say that the philosophy of information has developed in the last few years as a proper philosophical field of research. In the pages of *RIFAJ* we followed this development in two main occasions, when we interviewed Luciano Floridi in our second issue, in 2011, and when we reviewed Floridi's first Italian publication, in 2013. The present report may be seen as an addition to the two former publications and, at the same time, as a way of trying to present a broader consideration of the philosophy of information.

In spite of its young age, the philosophy of information is a broad field of research. It is broad not only in the sense of the topics and questions covered, but also from the methodological perspective: within the literature, one can find work on the technical information theory and computer science, on philosophy of science applied to information and digital technologies, on metaphysics and the debates about realism, on the ethical and societal aspects of information and communication technologies, etc. This mix of different methodologies and areas of research makes for what I think is a very vibrant and active environment, in which, moreover, significant philosophical insights can come from research which many would not even define as 'philosophical'. As a further consequence of this, the approach I found during the workshop was open to considerations and, possibly, critiques of different kinds.

Within this broad range of topics and fields, the organisers of the workshop – Phyllis Illari and Giuseppe Primiero – decided to focus on data and its related conceptual challenges in science and technology. Data can be considered a traditional subject of research in the philosophy of science, as for instance the work of Bogen and Woodward (1988) and Hacking (1983) show, but has become a central theme in more recent research, as a consequence of the increasingly important role data plays in both science and other elements of the human society (think, for instance,

Author. Stefano Canali, stefanocanali@me.com.

of the importance big data and data more in general have in current discussions about the economy as well as policy-making). Within this framework, conceptual research and theoretical considerations of data can prove to be useful and relevant.

Speaking of the specific talks of the workshop, here, for matters of space, I had to focus on six presentations. In particular, I start off with the report of Emma Tobin's talk about the classification of proteins through data. In the talk, Tobin argued against traditional monism, suggesting that the case of proteins can be considered as a new element highlighting the flaws of monism and natural kins essentialism. In order to show this point, Tobin focused on scientists' practices of classification through online databases.

Similarly to Emma Tobin, Sabina Leonelli considered what scientists practically do with data and how they *curate* it on order to extract useful knowledge. In the talk, this kind of research was extended to include what happens when something goes wrong with the data and was used to argue against what can be considered the received view on data (i.e. data as something which is there). As a consequence of the problems of such received view, Leonelli proposed a new, relational, characterisation of what data is.

Rob Kitchin's talk was an especially useful one, as it clearly summed up the different definitions of big data and the different views on its influence on epistemology. In fact, in the talk Kitchin discussed the question regarding how big data is changing traditional ways of doing research in the different sciences, including the social sciences and the humanities, wondering whether we can really talk of big data as a paradigm-shift for science.

Causation and its philosophical importance and characterisation was one of the recurring elements of the talks. Billy Wheeler considered recent views in the philosophy of causation, according to which causation is the transfer of information, and, a part from describing the features and benefits of these, focused on a definition of what is it that is transferred, in the sense of the best way of characterising information from the perspective of causation.

Within a similar framework to the one of Wheeler's talk, Wolfgang Pietsch presented his view about the epistemological challenge of big data and a consequent kind of science hugely reliant on data. In particular, Pietsch's main goal in the talk was proposing a specific account of causation which he finds capable of explaining current data practices and debates about the use of data in science.

As for the debates about the role of data in science, Teresa Scantamburlo analysed the assumptions and philosophical underpinnings of disciplines where data is increasingly central, such as machine learning and pattern recognition. In the talk, Scantamburlo argued that these assumptions are significantly similar to a Humean kind of empiricism and, in particular, its approach towards reason and theories.

References

- James Woodward and James Bogen (1988). "Saving the Phenomena". In: *The Philosophical Review*, 97(3), pp. 303-352.

- Ian Hacking (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Contents

1 *Data in Protein Classification*

Emma Tobin (University College London)

In her talk, Emma Tobin addresses a classical issue in philosophy of science – natural kinds and classification (see Bird and Tobin, 2015) – and considers the specific case of data in protein classification. In general, in the classification literature we can find what Tobin calls a *great divide*. In fact, on the one hand, as a consequence of the so-called species problem, many philosophers hold a pluralistic view on the classification of biology: that is, since scientists have many and different ways to define and classify species, philosophers tend to conclude that there is not a single, best, natural way of dividing animals in species (see e.g. Ereshefsky, 1998). On the other hand, chemical elements have been traditionally taken to be the instances of the fact that there is only one way of dividing nature, because nature has an order we can reflect in classification; this is why many philosophers hold a monistic view on the classification of chemistry (see e.g. Hendry, 2006). Within this framework, Tobin thinks that proteins are an interesting case, because, being biochemical entities, they lie at the interface of the divide and, thus, lead to the following question: should we argue that monism can be extended to macromolecules (bottom up approach), or is there a species problem for proteins as well and we should be pluralistic (top down approach)? This is the main question of the talk.

In order to try and answer the question, Tobin starts off with a definition of proteins. Generally, proteins are defined as <<linear chains of amino acids bonded in peptide bonds>> (Tobin, 2009), that is they are essentially defined in terms of amino acid sequences. As a consequence, we may think that the structure of the amino acid sequences is the criterion of the classification of proteins, thus arguing in favour of a structural kind of monism, also knows as microstructuralism. However, Tobin highlights that the problem with this position is that proteins' structure is actually a process divided in different steps, in the sense that the amino acid sequences are the initial primary structure which then folds in upper level structures: the path from the amino acid sequence to the folding is not always the same and, for instance, can be affected by external elements; moreover, the amino acid sequence is not necessarily connected to the protein's function, as for example proteins with the same structure do different things when in different places. Hence, according to Tobin, the latter and other phenomena suggest that structures are not really a good basis for classification: by focusing on structure only, one would miss out on many other features of proteins which are fundamental for classification. In other words, on Tobin's view, microstructuralism is not a tenable position.

In order to better sustain her position against microstructuralism, Tobin argues that it has empirical grounding: with a move typical of recent philosophy of science in practice, she focuses on the way scientists practically classify proteins. And this is where data comes in: as a matter of fact, currently most of the results of the classification work on proteins is uploaded by scientists on online databases. In particular, Tobin considers the Protein Data Bank (PDB), which is the primary repository of protein structures: what happens with the PDB is that scientists determine structures of proteins with a number of techniques and then their results are given an identifier and released on the database; journals require the PDB identifier before publishing a protein discovery. As a consequence, one could argue that the PDB case supports a monistic, bottom up view on proteins, in the sense that PDB identifiers are the unique and natural way of classifying proteins. Nevertheless, Tobin thinks that scientists' practices actually suggest the opposite. In fact, the techniques scientists use in order to find out about proteins' structure are highly indirect and do not consist in the direct imaging of the structure. For example, one of these techniques – X-ray crystallography – requires proteins to be crystallised, which is not possible for every protein and uses much idealisation and approximation; after the crystallisation, the crystallised proteins are beamed by X-rays and, from the different angles and intensities of the diffracted beams scientists design 3-D electron density maps. Moreover, another element of X-ray crystallography which lets us see that it is a very indirect process is the strong presence of mathematical representation, for instance in the generation of the coordinates and 3-D maps. As a consequence, Tobin argues that what we see in the PDB is not simply the structure of the protein, as it is very idealised and dictated by contextual things (technology, funding, etc.).

Furthermore, Tobin argues that another reason why the monistic approach based on structure is flawed is that with proteins' classification we can find a situation which is similar to the one of species in biology. As a matter of fact, the PDB is the primary but not the only and unique database for protein classification: there are hundreds of other databases, which use different criteria to classify proteins. For instance, the CATH database divides the protein structures of the PDB into structural domains, which in turn are grouped in evolutionary superfamilies; similarly, the SCOP (Structural Database of Proteins) focuses on the structural and evolutionary relationships between proteins of which we know the structure. The presence of these different database is interesting because they divide proteins differently, to the extent that certain proteins are classified in different ways in the different databases: for example, *papain* is considered as a single domain by SCOP, while it is split in two domains by CATH. The presence of different criteria of classification and the fact that the same elements are classified as different kinds is very similar to the species problem of biology, to the point that, in Tobin's opinion, we could argue that there is a species problem in proteins' classification as well.

The monist, though, could reply by highlighting that, actually, secondary databases such as SCOP and CATH take the data from the PDB, and could thus suggest that, metaphysically, we can be monists about protein structure and that the different ways in which data is organised reflect a data deluge problem, which is an epistemological – not metaphysical – problem. For Tobin, the problem with this response is that there is no agreed way of dividing the PDB data. As a matter of fact, before using the secondary databases, scientists have to identify – "choose", as scientists call it – the so-called domains of the proteins, which are parts of the structure capable of independently existing and functioning; thanks to the division in domains, scientists can reduce the complexity of the structure to simpler units. Once again, the point here is that there are different (both manual and automatic) ways of doing domain partitioning, which itself is an indirect process relying on existing knowledge. More particularly, although there is a benchmarking dataset (P-Domains) measuring the consensus about the domains, scientists agree only on very simple cases: for proteins with more complex structures, domain partitioning is subjective and requires a choice. As a consequence, Tobin argues that the microstructural response is flawed. The monist, though, may have another response, saying that one day we will know which is the right database and the right way of classifying proteins, it is just that we do not know it yet. However, Tobin highlights once again how classifying proteins via structure is difficult and, crucially, the structure does not tell us enough about proteins themselves.

Therefore, Tobin concludes that structural monism about proteins is not tenable and argues in favour of a pluralist, top down approach, similar to the one many philosophers hold in the case of biological species. Scientists' practices with data regarding proteins and databases are significant, insofar as they enlighten this point.

References

- Alexander Bird and Emma Tobin (2015). *Natural Kinds*. Ed. by Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: http://plato.stanford.edu/entries/natural-kinds/.
- Marc Ereshefsky (1998). "Species Pluralism and Anti-Realism". In *Philosophy of Science* 65, pp. 103-20.
- Robin Findlay Hendry (2006). "Elements, Compounds, and Other Chemical Kinds". In: *Philosophy of Science* 73 (5), pp. 864-875.
- Emma Tobin (2009). "Microstructuralism and Macromolecules: The Case of Moonlighting Proteins". In: *Foundations of Chemistry*, 12(1), pp. 41-54.

2 *Data Journeys: Openness and Shadows*

Sabina Leonelli (University of Exeter)

Sabina Leonelli's talk can be seen as a way of reflecting on a foundational aspect of the philosophical framework she has recently established. As a matter of fact, in the last few years Leonelli has focused on a philosophical consideration of data as used in the scientific practices (biology and model-organism biology in particular), highlighting their assumptions, epistemic features and more generally philosophical relevance (see e.g. Leonelli, 2014). Her talk begins with a consideration of the usual conceptualisation of data – i.e., data as a given –, then touches on a few of the topics and concepts she has mostly focused on in her research (data journeys and data reuse) and, within this framework, reflects on the conceptual consequences of new issues relating data travels (data absence, shadows of data, dark data, etc.).

Leonelli starts off by suggesting that the discourse around big and open data seems to be very much connected with ideas about what is available and what are the best ways to exploit the values of what is there. For instance, when we speak of open data, we usually speak of the ways in which we should open up data which is already there in order to exploit its value. Similarly, big data discourse normally involves issues such as the gathering, integration and analysis of data as an already available resource. These elements are now also reflected in data policies, whose idea is opening up e.g. government public spending in order to be more transparent and accountable about what is going on and – again – what is there. Even from an etymological perspective, data means something which is given. On this view, data seems an entity which exists and, thus, can be used as evidence for statements of different kinds. While this, in a way, could be seen as the received view on data, Leonelli suggests that it might not be enough when it comes to the diverse activities which are now possible with data. As a matter of fact, data is not only something already available, because it can actually also be made and produced under very specific conditions. These elements can be found in the discourse about open and big data, insofar as data is presented as a commodity, the precious outcome of labour and investments, but the emphasis is mostly about the subsequent passage, i.e. the access, exploitation and re-use of the data when it is made available. Is this view of data as a given a good way of accounting for the epistemological value of data? Is it the only possible view?

In order to find a possible answer, Leonelli suggests that we focus on databases and data journeys. The idea, here, is that the cases of databases and data journeys are a good window for exploring data practices and the epistemological value of data. Leonelli has written extensively on these topics, especially by studying the data practices of scientists working in modal-organism biology (Leonelli and Ankeny, 2012) and what it takes for data to *travel* from the laboratory in which it is produced in the first place to new laboratories in which it can be used for possibly different goals. When it comes to databases, for example, this kind of research consists in looking at the ways in which data is produced, submitted to the database, how it is curated, visualised and made searchable so that as many scientists as possible can reuse it. Why are these practices interesting from the perspective of philosophy of science? Because the study of data practices reveals the epistemic conditions under which data can travel and be used as evidence for scientific claims; such epistemic conditions include the way in which data is donated and/or submitted to the database, the institutional support for curators, the conditions and presence of the infrastructures (databases, but also data-journals), the *packaging* competences and technologies (the procedures of cleaning, selecting, mining data and organising it through common formats, metadata, labels and visualising tools), etc. That is, the research on the data practices of scientists highlights the complexities of data-travelling and the possible problems affecting it.

Having summed up the most important elements of her research on data travels and their conditions, Leonelli turns to consider situations in which data is not there, is not given, but can nonetheless represent a useful piece of information and be used for good scientific research. What happens in these cases and how should we conceptualise data so that we can understand them? For instance, data may be: missing or incomplete; negative, i.e. data giving you evidence for something which is not there, for the absence of some phenomenon; unobtainable, e.g. because of lack of resources or costs; unreliable, e.g. produced in non-reproducible conditions; invisible or ignored, e.g. not seen as relevant data by the curators and thus not circulated; untagged and unclassified, i.e. unusable because it is not retrievable; unintelligible, e.g. data about an organism about which there is no previous knowledge; inaccessible, e.g. because it is private or confidential; immobile, i.e. it cannot be made to move because of, for instance, costs, lack of infrastructure (e.g. a very big archive which cannot be digitalised and thus has to stay in a place); loss or missed, e.g. where the labels, tags and other packaging features fail completely.

When thinking about data journeys and the latter forms of data absence, Leonelli argues that a few general considerations can be drawn: the epistemic role of data, the extent to which it is going to be useful to produce knowledge, is heavily dependent on how data has been organised, processed, disseminated and contextualised and on whether it gets missed, stuck, abandoned, etc.; that is, data journeys affect what does and does not count as data and for whom. So, which kind of conceptualisation of data can capture the previous considerations? According to Leonelli, we should completely give up on conceptualisation based on manipulation: for instance, Ian Hacking (1983) proposed to consider data as whatever comes out of the machines in the lab; the problem with this view, for Leonelli, is that often what we consider as data does not come out of usual laboratory machines (e.g. data can be the result of simulations). Equally, we should also give up a notion of data based on its intrinsic properties, i.e. data as representations of some kind that can be used independently of the context. Leonelli proposes a different way to conceptualise data: we should think of data as any product of research activities which is collected, stored and disseminated in order to be used as evidence for knowledge claims; that is, data is a relational concept, because any object may be – and shift to become – data as long as it fulfils the previous features. In this relational sense, Leonelli argues that we can better understand the epistemology of data and why data can be useful even when it is absent: data should not be considered as an immutable commodity (as, for instance, Latour (1986) does), something which is relevant only if it is there and is given; the relevance of data can change and the change depends on the journeys, the relations established with the data.

References

- Ian Hacking (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

- Sabina Leonelli (2014). "Data Interpretation in the Digital Age". In: *Perspectives on Science*, 22 (3), pp. 397-417.

- Sabina Leonelli and Rachel A. Ankeny (2012). "Re-thinking organisms: The impact of databases on model organism biology". In: *Studies in History and Philosophy of Science*, 43 (1), pp. 29-36.

- Bruno Latour (1986). "Visualization and Cognition: Thinking with Eyes and Hands". In: *Knowledge and Society*, 6, pp. 1-40.

3 *Big Data, New Epistemologies and Paradigm Shifts* Rob Kitchin (National University of Ireland Maynooth)

In his talk, Rob Kitchin addresses one of the big questions which regards big data and comes up in different forms and with different levels of depth in other talks. The question is about the way in which big data is changing epistemology: are we witnessing a paradigm shift as a consequence of big data? In other words, is big data a revolution on the epistemological level, is it challenging established epistemologies? Positive answers to these questions can often be found in the literature as well as in more general discussions about big data: for instance, Gray (see Hey, *et al.* 2010) argues that revolutions in science are usually preceded by revolutions regarding measurement and Boyd and Crawford (2012: 665) suggest that big data <<is a profound change at the levels of epistemology and ethics>>. Kitchin's talk intends to critically assess these views.

First of all, what is big data? Usually, big data is defined in terms of three dimensions (see e.g. Beyer and Laney, 2012), that is in terms of the high volume and variety of the data collected and the high velocity of the collection. However, in Kitchin's opinion this definition is not enough and it is necessary to consider other specific features of big data, which stand out in comparison with small data: big data is *exhaustive*, in the sense that it can capture entire domains and does not need samples; it has a high level of *resolution* and is indexical in *identification*; it is strong in *relationality* and capable of conjoining different sets; it is highly *flexible* and scalable (see Kitchin, 2014: 1). As for practical examples, big data is, for instance, the number of transactions collected by supermarkets, or, in the context of cities, big data is the result of collecting data in a direct and manual (e.g. CCTV), automated (e.g. phones automatically sending data to providers) or freely volunteered (e.g. wearable devices, social media) way. So, what can be done with big data? As a consequence of its features, big data is necessarily messy and unstructured data and needs to be analysed to be useful: in order to analyse the data, what is used are techniques of machine learning, capable of automatically mining the data, finding the patterns and making predictions.

So, do big data and automated analytics imply a new paradigm-shift in science? In order to reply, Kitchin begins with Kuhn, who famously introduced the notion of scientific paradigm and paradigm-shift, in the sense of the historical moment in which an accepted set of theories, notions, experimental techniques and methodologies, etc. – a paradigm – declines and is changed in favour of a new one. According to Gray (see Hey *et al.*, 2009), Kuhn's notion of paradigm-shift should be applied to the case of measurements: that is, real paradigm-shifts in science take place when the nature of data and the analysis concerning data change; in particular, Gray identifies three main paradigms in the history of science and argues that with big data we have entered a fourth paradigm. What is this fourth paradigm? According to many, it is a radical form of empiricism: the idea is that, thanks to the automated analysis of a huge amount of data, it is not necessary to actively engage with data through theory because data can speak for itself. For instance, this is what Chris Anderson (2008) thinks, when he argues that big data implies <<the end of theory>>. But, why and how is big data sufficient? The point is that the computational power of automated analytics makes it possible to apply an ensemble approach, which consists in using every type of algorithms and see which one is the best and works, while normally scientists would choose and apply only one method. As a consequence, the idea here is that the answers we get from the ensemble approach are better, because they are not subject to the biases of humans choosing one analytical method, and are objective explanations, because they are not the subjective applications of a theory. As a consequence, big data are enough because its patterns and correlations give us answers that are not subject to human biases and theories: there is no need for any *a priori* model, hypothesis or subjective choice, as the patterns of the data are always useful and true, value-free and universal, to the point you just need data-scientists or software rather than domain experts. In Kitchin's opinion, these ideas regarding big data and epistemology are powerful and fascinating, but are not free of flaws and can be criticised. As a matter of fact, first of all, the idea of big data as capturing whole domains is flawed, because, even if data is big, it still remains a sample: for example, Twitter is a very big and quite inclusive kind of sample, but it is still a sample as not everyone is on Twitter. As a consequence, big data is not free of any bias, since it is at least subject to sampling bias. Moreover, the fact that algorithms are capable of making automatic discoveries does not entail that discoveries are theory-free or that the data speaks for itself: algorithms are designed by humans, who rely on scientific theories and act in certain contexts with certain values. Linked with the previous points, it is not either true that data can speak for itself and be meaningful independently of the context in which it was generated and to which its patterns refer.

In contrast to the former forms of radical empiricism and their problems, Kitchin argues that a different view on the epistemology of big data can be found, i.e. *data-driven science*: data-driven science can be considered as a mixed approach, according to which one can start off with an initial exploration of the data only, by searching for correlations and patterns and generating hypotheses from the data rather than the theory; theory, however, guides the whole process, at the level of choosing the algorithms, the most interesting correlations and patterns, etc. The idea, then, is a sort of mix between induction, used to generate hypotheses from the data, abduction, used to guide the formulation of hypotheses, and deduction, used to assess the validity of hypotheses. As such, one could argue that data-driven science is a new scientific paradigm, because it is a new way of generating knowledge starting from the data. Presented in this way, hence, data-driven science is very different from the ideas of data speaking for itself and the end of theory and the point is that the revolutionary epistemology of big data consists in this initial exploration, which informs the generation of scientific hypotheses.

After having analysed what he thinks are the two main epistemologies related to big data – empiricism and data-driven science –, Kitchin switches to considering specific disciplines where the application of big data epistemologies does not seem so straightforward: the social sciences and humanities. In these two broad disciplines, traditionally there is not much statistical analysis and, even where quantitative methods are traditionally used, as in economics, political science, human geography, sociology, etc., more recently there has been a move towards qualitative approaches.

Hence, can big data be applied to the social sciences and the humanities? As for the social sciences, big data is seen as an opportunity by positivistic social scientists (who think that the scientific method can be used to study and explain social phenomena): in fact, thanks to big data, social scientists are able to design social models that are much finer-grained and wider-scale; all of this can be used by positivistic scholars to respond to the classical critiques and issues of their views, such as reductionism and universalism. However, big data is an opportunity for post-positivistic social scientists as well, for example because of the presence of a huge amount of new (e.g. social media) or previously inaccessible (e.g. digitalised archives) data. At the same time, though, big data poses challenges: carrying out mechanistic analyses seems too simple for many cases; social trends may not entail causes, thus not being very useful; in big data there is a lot of what, but not much how; big data is sometimes seen as a treat to certain expertise not based on data. In similar ways, big data is both an opportunity and a challenge for the humanities, in particular in the form of so-called *digital humanities*. Kitchin argues that, in the digital humanities, one can find two main approaches to the role of big data: according to some scholars, big data and related technologies bring methodological rigour and objectivity to disciplines which were previously lacking them; on the other hand, others think that big data epistemologies can improve current methodologies, of which they may become a sort of extension, but not a replacement. Considering the challenges of big data, many highlight how big data methods may make the humanities mechanist and reductionist, sacrificing depth for width. Hence, the use of big data and related analytics in the social sciences and the humanities seems more complex than it is for the other sciences. An additional and specific challenge concerning both the fields regards the role of *small data*: in this picture, what happens to small data, on which these fields have successfully been based up until now? It is difficult to think that big data methodologies will entirely replace or delete the study and use of small data, which have a proven track record of giving powerful insights. Moreover, most of big data was not originally produced to be subject of research in the social sciences or humanities: for example, Twitter data was never produced to give information about health. This means that most of big data needs to be re-purposed (this has problems to be solved, see e.g. Illari, 2014) and, in addition, that big data can provide an interesting but surface snapshot, opposite to the very specific and deep insights which are the goal of small data research. However, Kitchin thinks it will increasingly be possible to apply big data methodologies to small data as a consequence of the sharing, opening up, reusing policies which scale the infrastructure of small data.

So, concluding his talk, Kitchin draws a few general conclusions about big data and its consequences on the scientific epistemologies. In his opinion, big data and related analytics are a disruptive kind of technology, insofar as, by radically altering the nature of data, they broaden the objects of research and provide new and powerful ways to analyse phenomena. As such, thus, there is no doubt that big data is capable of influencing and radically changing the epistemologies of the sciences; at the same time, big data poses new social, political and ethical questions. As for epistemology, the big question is how precisely big data is going to change the ways we do science, and the talk has consequently focused on critically assess ideas on how this change may take place. For the sciences, the radically empiricist approach of the end of theory and data speaking for itself is quite popular in many discussions, but seems to be flawed; on the other hand and as a consequence of flaws of the empiricist approach, the data-driven one seems more promising and likely to win out in the long run as a new paradigm. As for the social sciences and humanities, the application of big data is more complex and, while big data surely offers many significant opportunities to these disciplines, it seems difficult that the current and established epistemologies, based on small data, will be replaced; probably, big data lead to more pluralistic approaches. Therefore, the question about whether big data is going to establish a new in the sciences remains an open question, but Kitchin's guess is that more pluralistic and "mixed" approaches will be the ones to stand out.

References

- Chris Anderson (2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". In *Wired*. URL: http://wrd.cm/1nS6mjC.

- Sinan Aral, Erik Brynjolfsson and Marshall W. Van Alstyne (2010). "Harnessing the Digital Lens to Measure and Manage Information Work". *SSRN*.

- Mark A. Beyer and Douglas Laney (2012). *The Importance of 'Big Data: A Definition*. Gartner.

- Danah Boyd and Kate Crawford (2012). "Critical Questions for Big Data". In: Information, Communication & Society, 15:5, pp. 662-679.

- Tony Hey, Stewart Tansley and Kristin Tolle (2009). "Jim Grey on eScience: A transformed scientific method". In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pp. xvii–xxxi. Ed. by T. Hey, S. Tansley and K. Tolle. Redmond: Microsoft Research.

- Phyllis Illari (2014). "Big Data and Information Quality". In: *The Philosophy of Information Quality*. Ed. Luciano Floridi and Phyllis Illari. Springer.

- Rob Kitchin (2014). "Big Data, New Epistemologies and Paradigm Shifts". In: Big Data & Society, 1–12.

4 *Causation and Information: What is Transferred?* Billy Wheeler

In his talk, Billy Wheeler considers the recent philosophical view on causation, according to which causation, in the world, is actually the transfer of information; in his opinion, this is a promising view on the philosophical level and the practical one, especially for the design of algorithms and analytical methods for data.

The starting point of Wheeler's talk is the so-called Conserved Quantities View (CQV). Usually, when we think of causation, we tend to think of relations between events in time; the CQV takes a different approach, focusing on causal processes rather than events and suggesting that causal processes (in contrast with pseudo processes, see Salmon, 1977) are those processes which possess a conserved quantity (e.g. charge, momentum, etc.). That is, for example, considering the charge of an object, we can speak of a causal process between t(1) and t(2) if the charge has been conserved between t(1) and t(2). Within the CQV, then, the traditional way of thinking of causation as the interaction between two things producing something is explained in terms of the exchange of the conserved quantity between two causal processes. While the view is good for a number of reasons, two big problems have been highlighted in the literature: we often invoke the absence of an object or a process as a cause of something (e.g. not watering plants causes their death), but it is difficult to see how there can be exchange of a conserved quantity with an absent object or process; secondly, the CQV has an issue of applicability to the special sciences (and, consequently, their datasets), because in the latter very few quantities are governed by a conservation law. These issues can be seen as the motivation for a new version of the CQV and, in particular, an information-based view of causation; this has firstly been proposed by Krajewski (1997) and more recently by Collier (2011) and Illari and Russo (2014). The basic idea of the view, which Wheeler calls i-CQV, is that what is conserved in causal processes is information. The advantages of the view is that, by using information as a reference, the problems affecting the CQV are potentially solved: as for the problem of absence, in information theory absence can be data and thus a piece of information (e.g. the fact that the alarm clock has not ring yet is itself a piece of information, notifying that the pie is not ready); as for applicability, information is a more general concept compared to physical quantities and can be applied to the special sciences and a wider number of cases. In addition, the fact that we deal with information makes the i-CQV a more suitable concept for, possibly, writing algorithms searching for causation in data.

Thus, i-CQV seems a very useful and interesting way of treating causation. But, if causation is really the transfer of information, what is information? What is it that we measure as a conserved quantity? Having defined the i-CQV and highlighted its potential benefits, Wheeler switches to considering these questions about the nature of information and, specifically, he analyses three notions of information: information as 'knowledge update', information as 'entropy' and information as 'computational complexity'. Wheeler states that he is not an advocate of any of these views in particular, as he has not made a decision about which is the best one, and that his consideration is not aimed at assessing these notions in themselves as views of information, but rather as for how good they are for analysing causation.

So, the first concept Wheeler considers is the idea of information as a knowledge update. This seems the notion of information which is presupposed by epistemic logic, i.e. the idea that an agent has a number of hypotheses about how the world is and, every time she learns something new and gains knowledge, her range of hypotheses goes down; this notion of information is probably the most intuitive and the closest to our ordinary use of the term 'information', the idea that, when you are informed of something, this changes the way you see the world. Moreover, it is a qualitative theory and gives a semantic notion of information. But, is this view good for analysing causation? In other words, how would information be conserved within this view? Wheeler suggests that, here, the most natural suggestion would be in the following terms: the sum total of updates received by the agent from A and B at a time t(1) is conserved insofar as it equals the sum total of updates received by the agent from A and B at time t(2). Would this work? Wheeler thinks that there are problems. Firstly, on this view knowledge can only be updated once: once the agent learns something new, the range of hypotheses goes down and, if you learn something new again, there is no more change in range; we could solve the issue through counterfactuals ("the agent gets a knowledge update at time t(2); however, if she had had exposure to that information at time t(1), her range reduction would have been equal to what it would have been at t(2)"), but this may create problems of circularity when analysing causation and the CQV itself has been criticised in the past because of its reliance on counterfactuals. The second problem is that a person's range of hypotheses may change between t(1) and t(2) and the information acquired at t(2) may be different depending on what she has learnt between t(1) and t(2) (e.g.

it might be less or more informative and the agent might exclude or include different hypotheses); a possible solution – using an Hintikka approach on time fixing – would make the view useless for conservation. Thirdly, the knowledge update conception of information also presents the issue of the influence of background beliefs; defining an ideal agent might be the solution here, but it would be very arbitrary. As a consequence of these three issues, Wheeler argues that this view is hard to defend.

The second concept of information Wheeler considers in his talk is the one based on entropy, which comes from Shannon's mathematical theory of communication (Shannon and Weaver, 1949). According to this view, the informativeness of a message is defined in terms of the uncertainty that is resolved at the end of the receiver. As such, this notion has been vastly influential and is philosophically interesting, as we can think of causal processes as Shannon's communication channels. More specifically, a possible way to think of the conservation of information from the perspective of entropy may be arguing that the sum total of uncertainty resolved at points A and B at time t(1) equals the sum total of uncertainty resolved at A and B at t(2). Would this concept work? Wheeler thinks that it is better than the previous one, but it still has problems. Firstly, it requires an intervention, in the sense that it requires the presence of a receiver intervening to receive a message in a channel; interventions are problematic because they may already presuppose a concept of causation, but Wheeler is not sure as to whether measuring entropy really counts as an intervention or not. This problem might be overcome by defining entropy as choice of a source rather than a receiving end, but that is problematic if you want to measure at each stage of the process and not just at the beginning. A second objection to this view may be that it just reduces to the familiar definition of causation in terms of probability; a possible response to that would be that this is a probabilistic account which is very different from the traditional Reichenbach-inspired views of causation as raising of chances, but, in any case, talking of probability would probably require an interpretation of probability itself.

The last notion Wheeler talks about is the computational complexity view, also known as algorithm or Kolmogorov complexity. This is the idea that the informativeness of a message is equal to the sum total of computational resources that is required to produce that message and goes back to Kolmogorov and Solomonoff's work in the 1960s. A possible way in which this could work within the i-CQV is that the sum total of computational resources required at time t(1) equals the sum total of computational resources used to describe A and B at t(2). This seems the concept of information that Collier (1999) presupposes in his work, as his idea is that what is transferred is essentially the amount of complexity. Moreover, this view has significant advantages: interventions are not necessary, as any particular point in a world line can have a fixed amount of information expressed in terms of computational resources; interpretations of probability are not necessary either; in addition, the notion is general enough to be applied to the causal process of all scientific fields and very suitable to be used as a basis for designing algorithms to search for causality in big data.

Wheeler does not think that there are major problems with this view, but there rather is an open question: if we want to measure complexity by measuring the length of computational resources, we have to measure data; then, what is the data in the causal process? We could say that it does not really matter: we could measure data in any language and the difference in length of complexity would not matter, because the invariance theorem of complexity theory shows that any structural feature demonstrated from encoding in one language is automatically going to hold in another language. The problem with this, though, is that, if we change language between t(1) and and t(2), complexity will not be conserved but this will not imply that causation does not take place. A second way of thinking about complexity is in physical terms, as Collier does when he argues that for physical systems it is energy which is conserved. The problems with this proposal is that, as we have already seen, energy does not seem to work outside the physical sciences; Collier (1999) responds to that by saying that it does not matter, since each field will have its own interpretation of substance, but the response is problematic as well because in many fields the interpretation is not obvious and, anyway, inter-field causation would be impossible. Hence, as a consequence of all these problems, Wheeler argues that we should go for a radical view, according to which the physical world is not basic, but is emergent out of a more basic reality which is computational and, thus, physical processes are actually computational processes; this is the view originally given by John Archibald Wheeler as the "it from bit" hypothesis, sometimes called *digital realism*. The best mathematical model for this view is the concept of 'cellular automata' developed by Wolfram (2002): the idea is taking causal processes as series of computations in the basic cells and, then, defining information as the length of the program in the operating language of those cellular automata; in this way, the language is fixed by the identification of a transcendent operating system. Of course, Wheeler does not suggest that this rules out significant questions about this view, which could as well be considered crazy and making metaphysical assumptions going beyond basic empiricist constraints; other problems regard how we know the basic operating language of the cellular automata and the fact that the idea of programs running the automata seems very similar to the idea of laws of nature and we would thus need a definition of causation based on laws, which might be problematic as well. Nevertheless, the it from bit hypothesis may prove to be the best way to describe what is really transferred during causal processes.

References

- John Collier (1999). "Causation is the transfer of information". In: *Causation, Natural Laws and Explanation*, 279-331. Ed. by Howard Sankey. Dordrecht: Kluwer.

- John Collier (2011). “Information, Causation and Computation”. In: *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*. Ed. by Gordana Dodig Crnkovic and Mark Burgin. Singapore: World Scientific.
- Phyllis Illari and Federica Russo (2013). “Information Channels and Biomarkers of Disease”. In: *Topoi*.
- Władysław Krajewski (1997). “Energetic, Informational, and Triggering Causes”. In: *Erkenntnis* 47 (2), pp. 193-202.
- Wesley C. Salmon (1977). “An “At-At” Theory of Causal Influence”. In: *Philosophy of Science*, 44, pp. 215–224.
- Claude E. Shannon and Warren Weaver (1949). *The Mathematical Theory of Communication*. University of Illinois Press.
- Stephen Wolfram (2002). *A New Kind of Science*. Wolfram Media.

5 *Difference-Making as a Notion of Causation for Data-Intensive Science*

Wolgang Pietsch (Technische Universität München)

The basic question of Wolfgang Pietsch’s research concerns the way the reliance on data and related technologies in science is changing the methodology of science itself. According to many, as highlighted in Kitchin’s talk, scientific methodology has changed towards a new kind of science, where data is a sufficient guide thanks to its massive availability: scientists just need to analyse the data and look for correlations, so that they do not need theories, because data can *per se* tell us everything, and do not need to find causation, because correlations are enough. A good synthesis of these positions can be found in Anderson (2008), who argues that <<the new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all>>. According to Pietsch, it is quite easy to debunk statement such as the one of Anderson and, in fact, this has largely been done in the literature (see e.g. Boyd and Crawford, 2012); however, the real challenge to philosophy of science is identifying the grain of truth rather than simply debunking, because there may be some truth under these exaggerated positions.

First of all, Pietsch reflects on the very idea of big data. What is big data in the first place? As explained in Kitchin’s talk, the usual definition of big data is the 3V definition. Similarly to Kitchin, Pietsch thinks that this definition may be considered problematic, because it uses relational concepts only and thus one would need to clarify the specific point of the volume/variety/velocity. In addition and more importantly, the definition mostly refers to the technical challenges of big data and, hence, is not really useful for analysing data-intensive science and methodological elements. A possibly more useful way of defining big data deals with the idea that there is something happening to sampling, in the sense that thanks to big data we no longer need to choose a specific sample because we the data may represent all – or at least a significant subset of – the configurations of phenomena. Another crucial aspect of big data definitions which are useful for understanding the scientific use of data regards the automation of scientific processes. For example, Jim Gray (Hey *et al.*, 2009: xvii-xix) argues that the availability of a huge amount of data and data-handling technologies enables scientists to ask questions about more generally as well as causally complex. As for the issues this kind of data-intensive science deals with and the ways it solves them, Pietsch thinks that it is mostly about predictions, many instances of observations and thus variables, nonparametric modelling. These issues are similar to the ones of statistics, for which the presence of big data poses many challenges and produces significant changes: this is why, according to Pietsch, there is currently a paradigm-shift developing in statistics.

So, what happens to causation in the light of data-driven methods? Against the naive idea of causality being superseded by correlations as a consequence of big data, Pietsch wants to propose an account of causality which is capable of dealing with data-intensive science and/or is useful to analyse the methods of data-intensive science. In order to do that, an account of causality should meet a few requirements: it should fit the variational nature of evidence; it should not require a strong notion of intervention, because data has usually an observation-based nature; it should in some way account for the intuition that data-intensive science is theory-free, or at least suggest a new role for theory in inductive rather than deductive terms (this is one of the reasons Pietsch thinks that mechanist accounts of causation, here, may have problems); it should account for the contextuality of causation. So, Pietsch begins with a basic idea of counterfactuals, firstly formulated by Hume (1739: 70), according to whom, <<if the first object had not been, the second never had existed>> and then specified by Lewis in terms of a causal chain of events of which, if one had not happened, the other would not have happened either, and in terms of the semantic framework of possible worlds, to evaluate the truth-values of the counterfactuals' conditionals (see Menzies, 2014). As a consequence, the account of causation that Pietsch presents is a difference-making account which, inspired by Mill's method of difference (see Pietsch, 2014), is based on the counterfactual idea and also includes a notion of causal irrelevance, introduces context dependence. While the notion of causal irrelevance does not play a substantial role in the philosophical discussion on causation, Pietsch thinks that it is a powerful tool; for instance, causal irrelevance is useful for the evaluation of counterfactuals and in the context of analogical inference and probabilistic independence. The account is presented as follows: "in the context B in which the conditions C and the phenomena A occur, C is causally relevant to A if and only if the following counterfactual holds, if C had not occured, A would not have occured either"; 'in the context B in which the conditions C and the phenomena A occur, C is causally irrelevant to A if and only if the following counterfactual holds, if C had not occured, A would still have occured". As for the context dependence, the context needs to be constant in the sense that only the causally irrelevant elements may change. As for the counterfactuals' evaluation, Pietsch suggest that the two main traditional evaluation methods – Goodman's one and the more popular one by Lewis, based on the similarity between possible worlds – should be ditched in favour of this different approach, inspired by the method of difference, relying on causal irrelevance: the comparisons does not take place between possible worlds, but rather between phenomena which in the world differ only in terms of the causally irrelevant circumstances. According to Pietsch, this account of causation fits quite well with what is currently happening in data-intensive methods: as a matter of fact, in data-intensive science what happens is that, as a consequence of the huge amount of data, instances are *compared* between the data and the goal is getting predictions from that. Furthermore, this account fulfils well the previous conditions of adequacy, insofar as it fits the variational nature of evidence, does not rely on a strong notion of intervention, does not use underling knowledge of mechanisms and explains the importance of contextuality. As a practical example of application of this notion to a case in data-intensive science, Pietsch mentions the usage of the algorithms of classification trees, which in some simple cases is identical to the method of difference and, in more complicated ones, is significantly similar (for instance, the condition of the stableness of context is equally required).

References

- Chris Anderson (2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". In *Wired*. URL: http://wrd.cm/1nS6mjC.
- Danah Boyd and Kate Crawford (2012). "Critical Questions for Big Data". In: Information, Communication & Society, 15:5, pp. 662-679.
- Tony Hey, Stewart Tansley and Kristin Tolle (2009). "Jim Grey on eScience: A transformed scientific method". In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pp. xvii-xxxi. Ed. by T. Hey, S. Tansley and K. Tolle. Redmond: Microsoft Research.
- David Hume (1739). *An Enquiry Concerning Human Understanding*. Edited by Stephen Buckle. Cambridge (UK): Cambridge University Press.

- Peter Menzies (2014). *Counterfactual Theories of Causation*. Ed. by Edward N. Zalta. The Stanford Encyclopedia of Philosophy.
  URL: http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/.
- Wolfgang Pietsch (2014). "The Structure of Causal Evidence Based on Eliminative Induction". In: Topoi 33, pp. 421–435.

6 *Big Data: The Empiricist Approach and Its Philosophical Underpinnings*

Teresa Scantamburlo (University of Ca' Foscari, Venice)

Teresa Scantamburlo's talk looks at the philosophical underpinnings of the current scholarship in machine learning and pattern recognition for big data analysis, suggesting that they are very much related to traditional empiricism.

First of all, Scantamburlo starts off with a definition of big data. As stands out from the other talks, defining big data is a main issue within current academic work and often scholars, before even arguing something about big data, have to specify the definition they think is the best one. In this case, Scantamburlo essentially agrees with Rob Kitchin's broad definition of big data (volume, velocity, variety, exhaustivity, resolution and indexicality, relationality, flexibility) and contrasts it with Viktor Mayer-Schönberger and Kenneth Cukier (2013)'s view, which she thinks syntheses well the main trends of current machine learning views. As a matter of fact, according to Mayer-Schönberger and Cukier, the most important and characterising features of big data are the following: the possibility of seeing phenomena from several angels and perspectives; the fact that you can get a sense of the main general directions of phenomena; the superiority of predictions based on correlations as opposed to explanations and causation. In addition to Kitchin's view, Mayer-Schönberger and Cukier's points can be considered similar to what Boyd and Crawford (2012) call *mythologies of big data*, including the idea that big data entails the end of theory because data can speak for itself and the triumph of correlations over causation. Scantamburlo highlights how, for Boyd and Crawford, most of these ideas regarding big data are, precisely, mythologies and, for instance, claims of objectivity are misleading, bigger data is not necessarily better data, big data is not always universal and loses meaning when out of context and has often limited access.

As a consequence of the latter and other critiques regarding the myths of big data, Scantamburlo believes that we are currently witnessing a sort of reconciliation, somehow trying to recombine the radical empiricist approach according to which data can speak for itself, correlations are enough, etc. with theoretical models and, more in general, the sphere of reason; this alternative approach is what has been defined as data-driven science. On Scantamburlo's view, the efforts of reconciling data and theory can be seen as a sort of solution of Hume's division between reason and matters of fact. In other words, Scantamburlo thinks that, while we are trying to find an alternative and critical way of looking at big data, this alternative way is an opposition to Hume's notion of induction; at the same time, in fact, the development of big data analysis and machine learning is the result of a Humean view of induction and distinction between different kinds of knowledge. That is, looking at the big data discourse from a Humean perspective can enlighten the roots of the discourse and let us better understand why data is increasingly trusted, while at the same time being unreasonable (see Halevy *et al.*, 2009).

So, according to Scantamburlo, certain features of the big data discourse can be better understood by analysing their philosophical underpinnings and, particularly, having Hume's anti-rationalist approach in mind. Hume introduced an idea of induction based on probable reasoning and regularity, in the sense that we know the world just by repeating experiences and it is a spontaneous process that we tend to naturally trust. This is the main way in which machine learning and pattern recognition developed the idea of inductive inference: you have some instances that you have observed, and this is useful insofar as, when a new instance occurs, you can make a prediction on it. Statistical learning theory basically repeats the same story: an algorithm takes some training examples on a particular target and then, after the training phase, each time a test instance appears, through a mapping function the algorithm can predict its outcome. Interestingly, this way of thinking about inductions has led machine learning and patter recognition researchers to think of models of data as if they were models of phenomena, to the point that data analysis models are seen as equivalen to theoretical and scientific models; the problem, though, is that data analysis models comprise a limited knowledge of phenomena, while theoretical models are more general because they directly refer to phenomena.

So, having highlighted the two Humean philosophical underpinnings of the conception of induction in machine learning and pattern recognition and taking these into account, Scantamburlo suggest two main questions which remain open and need further research. The first one regards the way in which we should consider induction itself in these two disciplines: the two main approaches – abstraction and generalisation – are correlated, but are not really the same; in the machine learning ad pattern recognition literature, though, they are often treated as if they were the same and, as a consequence, it is often difficult to distinguish them and understand their conclusions and results. Secondly, Scantamburlo argues that another question regards how we can use machine learning and pattern recognition algorithms for models of data and models of phenomena, without making confusion between the two of them.

References

- Danah Boyd and Kate Crawford (2012). "Critical Questions for Big Data". In: Information, Communication & Society, 15:5, pp. 662-679.

- Alon Halevy, Peter Norvig, and Fernando Pereira (2009). "The Unreasonable Effectiveness of Data". In: *IEEE Intelligent Systems*, 24, pp.8–12.

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