Filozofia Nauki (*The Philosophy of Science*)
ISSN 1230-6894
2018, vol. 26(4) [104]: 5-28
DOI: 10.14394/filnau.2018.0021

BILLY WHEELER*

HOW TO FIND PRODUCTIVE CAUSES IN BIG DATA: AN INFORMATION TRANSMISSION ACCOUNT**

Abstract

It has been argued that the use of big data in scientific research obviates the need for causal knowledge in making sound predictions and interventions. Whilst few accept that this claim is true, there is an ongoing discussion about what effect, if any, big data has on scientific methodology and, in particular, the search for causes. One response has been to show that the automated analysis of big data by a computer program can be used to find causes in addition to mere correlations. However, up until now it has only been demonstrated how this can be achieved with respect to difference-making causes. Yet it is widely acknowledged that scientists need evidence of both "difference-making" and "production" in order to infer a genuine causal link. This paper fills in the gap by outlining how computer-assisted discovery in big data can find productive causes. This is achieved by developing an inference rule based on a little-known causal process theory called the information transmission account.

Keywords: causation, big data, data-intensive science, machine learning, conserved quantities, causal processes

A debate has recently emerged concerning to what extent — if any — the use of big data as a form of scientific evidence overturns traditional forms of scientific theorizing. One particular issue is whether or not the introduction of big data obviates the need for making causal inferences. Advocates of big data, such as Chris Anderson (2008), Jim Gray (2007), and Viktor Mayer-Schönberger and Kenneth Cukier (2013), argue that the sheer size of big data

* Department of Philosophy (Zhuhai), Sun Yat-Sen University, Tangjia Bay, 519082 Zhuhai, Guangdong Province, China, billy.wheeler@cantab.net.

** An earlier version of this paper was presented to the Seventh Workshop on the Philosophy of Information *Conceptual Challenges of Data in Science and Technology*, hosted by University College London, March 30-31, 2015. I am grateful to the attendees who provided useful ideas which helped shape the argument of the paper. I am also indebted to the advice of an anonymous reviewer at *Filozofia Nauki*.

is such that "correlation can supersede causation" as the primary goal of scientific investigation. As Anderson puts it, this overturns traditional thinking, which often values causation over correlation:

> Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between x and y (it could just be a coincidence). . . . But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete. . . . There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot. (2008: 2-3)

Big data's unique nature can be characterized by the so-called "three Vs": greatness in *volume*, diversity in *variety*, and quickness in the *velocity* in which it is acquired and analyzed (Laney 2001). This has been made possible by rapid advances in technology (especially the internet and data storage capacity) over the past few decades. And it is precisely this technological change in the way data is gathered and analyzed which some believe is radically affecting the way science is practiced. For example, Gray calls the increased use of big data a "new paradigm of scientific exploration" (2007: xix). On the issue of causation, Schönberger and Cukier call the traditional understanding of science as engaged in finding causal mechanisms a "self-congratulatory illusion" which is "overturned by big data" (2013: 18).

In section 1, I lay out some of the reasons why enthusiasts of big data claim it heralds a new age of science — one free from theory and causal inference. But it should be noted that most philosophers of science have tended not to agree with these extreme claims. Although they agree big data changes some practices of science, they argue this does not overturn the need, nor the desire, for causal knowledge. One type of response to the challenge from big data is to argue that — despite appearances — causal connections can be searched for in big data using the very same computer programs which are typically believed to only search for correlations. An example of this response is given by Wolfgang Pietsch (2016). He outlines a difference-making account of causation which is compatible with a number of existing big data search methods.

Difference-making accounts of causation form one half of a well-known dichotomy between two concepts of causation employed in science. Whereas difference-making accounts focus on patterns and probabilities, "productive" concepts focus on the ability of a cause to "bring about" or "produce" its effect. Productive accounts typically appeal to causal processes or mechanisms in order to do this. It has been demonstrated that when scientists make

causal inferences they look for evidence of both difference-making and productive causes (Russo, Williamson 2007, 2012, Clarke et al. 2013, 2014).

This raises the question: can computers be designed to do more than just find difference-making causes in big data? Can they be programmed to find productive causes as well? This paper will argue that they can. This is achieved by developing a recent version of the transfer theory of causation called the "information transmission account." This account, which has its origins in John Collier (1999, 2010), but has been more recently advocated by Phyllis Illari (2011, 2014), claims that causation is the *transmission of information* between two objects or events. Given that information is principally a property or characteristic of data, this potentially allows one to search for productive causes by measuring the size of data.

In section 2, I outline the main tenets of the information transmission account and how it relates to similar accounts of causation such as those of Wesley Salmon and Philip Dowe. In its current form, the information transmission account relies on a qualitative measure of information; but if it is to provide a suitable basis for the automated search of big data, then it needs to be supplemented with a quantitative measure. In section 3, I evaluate the suitability of three existing quantitative measures and argue that information understood as algorithmic complexity provides the best with respect to big data. In section 4, I outline a potential inference rule that could be used to search for productive causes. There I briefly illustrate how this rule might be utilised in practice in a well-known data-intensive scientific field: exposomics.

1. BIG DATA AND CAUSAL INFERENCE

David Hume once said "all reasonings concerning matters of fact are founded on the relation of cause and effect" (1978: 649), and whilst not all scientific investigation is about finding causal connections in nature, a significant portion of it is. Much scientific knowledge, especially in the fields of biology, medicine, and the social sciences, concerns the identification and accurate description of causally connected variables. It seems remarkable therefore to claim, as some have done, that the arrival of big data should overturn this. To help see why, consider a well-known success story of big data: Google Flu Trends.

In 2008, Google developed a model to predict influenza spread based on search queries in areas identified as endemic by the Centers for Disease Prevention and Control (CDC). Once a correlation had been found, Google was

able to predict outbreaks of influenza in the US with a reported success rate of 97% (Ginsberg et al. 2009). What was different about Google's model is that it did not depend on any prior theory or causal understanding of which terms would be searched by individuals with influenza: it only looked for correlations. The speed at which Google was able to make predictions was impressive. Unlike the CDC, which relied upon records from thousands of healthcare providers, Google could read directly in real time the search terms entered into its site. This gave it a ten-day advantage over the CDC at predicting outbreaks (Helft 2008).

For Mayer-Schönberger and Cukier (2013: 1-3), examples like this reveal the advantages of big data analysis over traditional theorizing. Without the need for a hypothesis, researchers at Google were able to predict the spread of influenza much faster than traditional methods. They effectively bypassed the scientific method as we currently know it: there was no need for causal understanding — correlation could do "just as good." Another example that Anderson cites comes from the inventor of "synthetic biology," Craig Venter. By using "shotgun" gene sequencing methods of entire eco-systems, Venter was able to identify thousands of previously unknown bacteria and other life-forms. It is the size of the data gathered and its quick analysis by a computer program that Anderson claims changes everything here. Ventor did not know much about the life-forms he had discovered (for example, their species or how they lived), yet despite this — according to Anderson — he has "advanced biology more than anyone else of his generation" (2008: 3).

Few mainstream philosophers of science believe big data really spells the end of causal inference in science. Even Mayer-Schönberger and Cukier state that the true value of big data is that it might quickly and more efficiently

> point the way for causal investigations. By telling us which two things are potentially connected, they allow us to investigate further whether a causal relationship is present, and if so, why. (2013: 66)

Among philosophers who have written on this topic, two different types of responses have emerged.

The first, which might be called the "hidden causal thesis," is that big data practices already use a broad range of causal knowledge — both in the design of the study and in subsequent experimental investigation. Examples of this interpretation can be found in the work of Sabina Leonelli (2014) and Stefano Canali (2016).

Leonelli focusses on the construction of databases for model organisms in experimental biology such as FlyBase and WormBase. She argues that the construction of these databases requires a significant amount of data cura-

tion, given the complexities surrounding the ways in which the data was originally gathered. She claims that the data in these collections undergoes three distinct phases: de-contextualization, re-contextualization, and re-use, and in each case evidence can be found of significant amounts of prior theory being employed as well as knowledge of existing causal connections (2014: 8).

Canali finds similar results in current exposomics research. More specifically, he outlines a "meet-in-the-middle approach" in the use of big data by Chadeau-Hyam et al. (2011) in their study of breast and colon cancer. According to Canali, the identification of a correlation between exposure and disease is only one part of the story: the next step in their investigation was the discovery of mechanisms or processes in the middle that connect the two. These mechanisms involve connections between internal responses, or "biomarkers" (see section 4 for more on exposomics and biomarkers). From this, he infers that there was an obvious need for causal knowledge in subsequent investigation of the data (2016: 4-5).

The work of Leonelli and Canali therefore shows that when big data is used by scientists there is often hidden causal knowledge at work, either in the curation of the data or in subsequent investigations. However, this is not the only way the claims made about causal inference by Anderson and others have been questioned.

This brings me to the second response, which might be called the "automated causal thesis." This interpretation is more resonant with the ambitions of big data enthusiasts in that it recognizes that the use of big data, especially its automated search by a computer, brings about genuine changes in scientific methodology. However, it disagrees that causal inferences can no longer be made. Representative of this approach is Pietsch (2016). He argues that search algorithms can be used to do more than just search for correlations. In fact, on certain conceptions of causation and causal inference, it is possible to find causes within big data.

Pietsch achieves this by focusing on a rule for causal inference based around eliminative induction and Mill's "method of difference":

> The best known and arguably most effective method of eliminative induction is the so-called method of difference that establishes causal relevance of a condition C_X by comparing two instances which differ only in C_X and agree in all other circumstances C. If in one instance, both C_X and A are present and in the other both C_X and A are absent, then C_X is causally relevant to A. (2016: 148)

By focusing on regularity and correlation, Pietsch's conception of causation and causal inference is easily compatible with automated search programs. In fact, he argues, even existing search programs, such as those used in machine

translation and microtargetting, already analyze data in ways that allow one to identify causes via eliminative induction (2016: 152-156).

Pietsch's conception of causation, as given by Mill's method of difference, is an example of "difference-making." It should be contrasted with another important conception of causation: "production" (Hall 2004, Godfrey-Smith 2010). Difference-making causes typically refer to whether or not the presence of one event is *statistically relevant* to the occurrence of another. Examples of difference-making accounts of causation are those that focus on regularities, probabilities or counterfactuals. However, there is another important conception of causation which focusses not on relevance but *responsibility*. Productive accounts attempt to describe what it is about one event or state of affairs that is "responsible for" or "brings about" another event. Productive theories of causation focus on mechanisms, processes, and other ways in which two objects or events can be connected together.

Recent studies show that when scientists make causal inferences, they look for evidence of *both* difference-making *and* production. Frederica Russo and John Williamson have studied the concepts at play when causal inferences are made in healthcare (2007), autopsy reports (2011), and the EnviroGenomarkers project (2012). They argue that although evidence of probabilistic connections is vital, this is rarely enough to warrant a causal connection. Causal relevancy is good for making predictions in a known sample or population, but it does not allow us to extrapolate to unknown groups. Russo and Williamson identify knowledge of production (in the form of mechanisms) as an additional component necessary for establishing general causal claims:

> The existence of a mechanism provides evidence of the stability of a causal relationship. If we can single out a plausible mechanism, then that mechanism is likely to occur in a range of individuals, making the causal relation stable over a variety of populations. If no mechanism were found, that may be because the correlation is particular to a specific sample population or a specific set of circumstances — i.e., it is a "fragile" relationship — and not sufficiently repeatable. In other words, mechanisms allow us to generalize a causal relation: while an appropriate dependence in the sample data can warrant a causal claim "C causes E in the sample population," a plausible mechanism or theoretical connection is required to warrant the more general claim "C causes E." (2012: 159)

Pietsch's interpretation of big data analysis and its search for causes therefore gives us only half the story. A full search for causes within big data by a computer program should look for both difference-making and production. But can a program be designed to search for productive causes? And if so, what concept of production is suitable for use in the automated search of big data by a computer? I now turn to examine a recently proposed productive ac-

count of causation called the "information transmission account." I argue that this has many of the right features to function as a suitable concept for big data analysis. However, in its current form, it requires some amendments in order to be fully utilized in practice.

2. THE INFORMATION TRANSMISSION ACCOUNT

The information transmission account is an attempt to provide a concept of causation that underscores our productive intuitions about causality. To highlight the difference between productive and difference-making concepts, consider the following example. Billy and Suzy are both throwing stones at a bottle in an attempt to break it. Billy throws his stone at t_1 followed shortly by Suzy's at t_2. Billy's stone strikes the bottle and breaks it. However, it is also true that had Billy's stone missed, Suzy's would also have hit, and the bottle would have broken anyway. In this example, whether or not the bottle breaks depends on both Billy's and Suzy's throws. They are both *causally relevant* factors for whether or not the bottle breaks. But there is also an important sense in which it was Billy's stone alone that was *responsible* for the breaking of the bottle. It is this "responsibility" aspect of causation that productive accounts attempt to explain.

The information transmission account is a version of a transfer theory, and attempts to improve on previous transfer theories such as those of Ron Aronson (1971), David Fair (1979), and Philip Dowe (2000). These previous theories explain responsibility as the transfer of a physical quantity such as force, momentum, or energy. In the case above, the reason why Billy's stone was responsible for the bottle breaking is that it transferred some of its physical properties — such as energy — to the bottle. This transferred energy caused structural failure resulting in the bottle breaking. No such transfer occurred between Suzy's stone and the bottle, and so her throw was not responsible for the bottle breaking.

Phyllis Illari criticizes these previous transfer theories on the basis of their general applicability:

> The only properties the conserved quantity theory could possibly pick out as relevant are conserved quantities. These are relatively few, such as charge, mass, momentum and so on. But in the vast majority of cases of causality in the special sciences, these are not the relevant properties at all. . . . Charge, mass and momentum seem incidental to such causal claims as "smoking causes cancer," since the various sciences of cancer do not concern themselves with charge, mass or momentum. (2011: 98)

Many causal claims do not involve reference to fundamental quantities from physics. If we are to have a theory of causation that is applicable in a range of sciences then we need a concept which is more universally applicable.

Illari identifies this concept as "information." She agrees with John Collier, who writes that "the basic idea is that causation is the transfer of a particular token of a quantity of information from one state of a system to another" (1999: 215). In this regard, the Illari–Collier view is similar to another transfer theory: Salmon's (1984) mark transmission theory. However, whereas Salmon defined a causal process as one which has the *ability* to transmit a mark or message, Collier and Illari argue that causal processes are those that *actually* transmit information — and transmit the *same* information.

As Collier puts it: "the connection in this case is identity, which is perhaps the strongest connection one can have, and requires information transmission across time: it is the identical token of information" (2010: 11-12). So according to the information transmission account two objects, states of affairs, or processes are causally connected provided there is the exchange or transmission of information between them. The concept of information is suitably general to cover the range of causal claims made across the sciences. It also seems a particularly appropriate concept for finding causes in big data given that data-intensive sciences trade in large amounts of information and its processing.

Despite this, the information transmission account is too abstract to be of practical use by scientists in its current form. Although it might provide a useful account of the metaphysics of causation, it is too vague to help in its epistemology. For how should we look for information transfer? How would we know information had been successfully transferred between two objects or states of affairs? When we think of information, we tend to think of *meaningful* information so that the "same information" carries the "same meaning." Yet if two objects or events are causally connected by transferring the same information in this way, it might require us to think of objects as communicating with one another — or it at least requires a semantic interpretation of physical states as carriers of data.[1]

The problem here is similar to that with the previous theories of Aronson and Fair. Like Collier and Illari, they too demanded that the transferred quantity retain its identity during the exchange. Literally, when the stone hits the bottle it exchanges the *very same energy* that was present in the stone. But as Dowe has previously argued (2000: 55-59), it is not obvious that

[1] Collier states that a causal process is not just one that transmits the same "form" but one that transmits the same "fact" between two states of affairs (1999: 9).

physical magnitudes retain their identity in this way when exchanged. Consider a case where Billy and Suzy are pushing a car uphill together. It is not clear that we can say, metaphysically, which part of the momentum of the car is provided by Suzy and which part by Billy. The most it seems we can say is that each contributes a particular amount to its momentum. We can measure this amount and give it a number, but we cannot divide it into tokens and track their metaphysical identity through time.

For Dowe, this suggests that what matters in causal interaction is not identity of transferred quantities but their equality *vis-à-vis* their numerical magnitude. We might then wonder whether the same could be said about the information transmission account. Instead of requiring, as Collier and Illari do, that the information is the same in terms of meaning, we might demand only that the information is the same in terms of quantity.[2] Two objects or events are causally connected provided the amount of information is conserved before and after the exchange. Information is widely thought to be constant within a closed system, so information would qualify as a conserved quantity in Dowe's sense.[3]

Following Dowe's original treatment of causal process and causal interaction, this modified version of the information transmission account yields the following definitions:

> CAUSAL PROCESS: A causal process is a world line of an object that conserves information.
>
> CAUSAL INTERACTION: A causal interaction is an intersection of causal processes whose sum total of information is conserved.

The case of single causal processes is included in order to explain how objects that are not interacting are nonetheless responsible for their future states. For example, a single wrench spinning in the vacuum of space is responsible for its future states even though it is not exchanging information or physical magnitudes with any other objects.

Can we now define an inference rule for finding productive causes in big data? Even though it is possible to formulate the information transmission account using a purely quantitative notion of information transfer, we are not yet in a position to do this. The reason is we have not yet said *how* the infor-

---

[2] Collier (1999, 2010) also develops versions of the information transmission account that are quantitative and based on the ideas of entropy and complexity. I discuss these interpretations in more detail in section 3.

[3] Black holes might provide an exception here. See Preskill (1992) and Hawking (2015) for more context on this debate.

mation is to be measured. The literature on information measurement contains many different accounts and not all of these are going to be practical for use in searching big data. Therefore, I now turn to evaluate three of the more well-known measures: "knowledge update," "entropy," and "algorithmic complexity." I argue that the last of these provides the best measure of information for finding causes in big data.

3. THREE CONCEPTS OF INFORMATION

3.1. INFORMATION AS KNOWLEDGE UPDATE

The first concept comes from the field of epistemic logic. This branch of logic concerns itself with modelling knowledge states and how they change when new information is received. The basic idea is that when an agent receives some new piece of information, their epistemic state changes. The epistemic state of an agent is modelled using Kripke semantics: each possible world available to the agent is a possible way the world could be.

To illustrate, suppose an agent does not know which day of the week it is. Then there are seven possibilities. She is told "it is a weekend day". This reduces the number to two. Furthermore, she is told "it is not Saturday". She updates her state, and there is only one possibility remaining: it must be Sunday. Every time the agent receives a new piece of information, her state changes and we can measure "how informative" a message is in terms of changes in her state.

What would the information transfer account look like when supplemented with the notion of information as knowledge update? Let us imagine that the world line of an object provides a knowledge update to an agent, which we may take to be some kind of measurement or observation on a particular occasion. Then we can define a causal process and causal interaction in the following way:

> CAUSAL PROCESS: A world line is a causal process if the epistemic update received by an agent at time t_1 excludes the same possibilities as an epistemic update received by an agent at time t_2, where t_1 and t_2 are different points along the world line.
>
> CAUSAL INTERACTION: There is a causal interaction between two causal processes A and B if the sum total epistemic updates received by an agent observing A and B at time t_1 excludes the

> same possibilities as the total epistemic updates received by an agent observing A and B at time t_2, where t_1 is a point prior to intersection, and t_2 is a point after intersection.

In the case of a single causal process, how much information an agent receives by observing an object remains the same, no matter when they observe it. This would be opposed to a pseudo-process, which can become more or less informative at different times. This seems highly plausible in the case of knowledge updates. If we go back to the example of the wrench, it appears that whatever the current state of the agent, they will receive the same knowledge when observing the wrench — no matter when they observe it.

It is clear that the agent in the definitions above needs to be defined relative to a specific epistemic context. This is because what an agent already knows affects the value of a given piece of information. How detrimental is the inclusion of an ideal agent into this proposal? In terms of finding a rule for causal inference the inclusion of an ideal agent is not that troubling. This is a common device in many approaches to scientific reasoning. Scientists do not reason "in a vacuum" and, provided we are explicit about what knowledge they have during the reasoning process, we can lay down rules for good and bad inferential practices. The inclusion of an ideal agent is more problematic in terms of giving a metaphysical analysis of causation "as it is in the world." Yet since our aim here is with the epistemology of causation, rather than with defending a particular metaphysical view, I think we can put this worry to one side.

In the case of causal interaction, two objects or events are causally connected provided an ideal agent receives the same total epistemic updates after interaction as they would prior to interaction. Unfortunately, a simple thought experiment demonstrates how this is not possible in practice.

Imagine two particles α and β passing through space with different values for momentum, energy, and charge. At a point along their world lines they collide and transfer some of their physical quantities to each other. Even though their sum physical magnitudes remain constant, it is clear that knowledge updates do not. Let t_1 be a time along their world lines prior to collision and t_2 a time along their world lines after collision. An agent observing the particles at t_1 knows something about the particles at that time — namely "their energy, momentum, and charge at t_1." Like all updates, this excludes a number of possibilities about the world and gives her a particular amount of information. But an ideal agent who observes the particles at t_2 cannot exclude these same possibilities — as she has no access to the properties of α and β before t_2. Given her current observations, there are more possibili-

ties available for the energy, momentum, and charge at t_1 — since a number of different configurations are compatible with her current observation.

This demonstrates that it is possible to learn something at t_1 which cannot be learned at t_2 and so for that reason information as "knowledge-update" is not always conserved during interaction. Information as "knowledge update," therefore, seems unsuitable as a basis for the information transmission account.

3.2. INFORMATION AS ENTROPY

The next concept of information has been influential in electrical telecommunication, where it has provided rigorous mathematical definitions of optimal coding, noise, and channel capacity. The basic idea is that the less likely a message is (out of all possible messages), the more informative it is and vice versa. For any given message or symbol produced by a source, there is an assumed probability distribution. The "entropy" (H) contained within a message x is given by its probability p(x) according to the following equation (Shannon, Weaver 1949):

ENTROPY: $H(x) = -\Sigma p(x) \log_2 p(x)$

How useful is this idea for thinking about productive causes and causal inference? The first thing to say is that, in its original use, the concept of information as entropy provided a nice model of causation understood as the *flow* or *transfer* of information. The original application was for copper wires transmitting messages via electrical wave or impulse (Pierce 1961). If we think of causal processes as taking place along a channel, then we can readily appreciate the relationship. Secondly, by taking the negative log of the probability, we get the quantity of information that is additive: $H(x) + H(y) = H(x + y)$. This means it can avoid the worries we had with the knowledge update view as the sum totals will be conserved during interaction.

At the moment I have not said how we measure the entropy of a process or channel. This could amount to one of two things. It could be the "entropy rate," which is the *average information* carried by a message per second, or it could be the "self-information," which is the amount of information *received by an individual* at the end of the channel or process.

It is not obvious how the first of these relates to causation. The entropy rate is an average, so by definition this will remain constant in all processes. This makes it difficult to separate genuinely causal processes from pseudo-processes on the basis of conservation. Likewise, thinking of the entropy as the property of a channel with respect to a particular message is problematic

if a channel produces messages with different likelihoods and therefore different entropies.

Fortunately, the nature of causal processes suggests we can model them as channels transmitting single messages through space-time. Providing the causal process is not interacting, it will transmit just one message, with a single amount of self-information:

> CAUSAL PROCESS: A causal process is a channel which transmits a message with constant entropy value.

This has prima-facie plausibility with respect to the solitary wrench in space. The entropy it carries remains constant and would inform the receiver equally regardless of when they intersected it.

On this basis, we can develop an account of causal interaction as well:

> CAUSAL INTERACTION: A causal interaction occurs between two channels A and B if the sum total of entropies before interaction equals the sum total of entropies after interaction.

This definition says that the chance of a message occurring remains constant through interaction and therefore the total information *vis-à-vis* entropy remains constant.

Although it characterizes causation differently, this version of the information transmission account inherits familiar problems raised against difference-making accounts — the most obvious being how to explain where the values in the probability distribution come from. How we do this depends on which interpretation of probability we take. I am keen to avoid a protracted discussion of the pros and cons of various interpretations of probability (see Gillies 2000 for an overview of the existing literature). However, I think it is worth looking at three of the most relevant interpretations in order to highlight the difficulties any advocate of entropy faces when attempting to formulate an inference rule on the basis of conserved information.

(i) The *relative frequency interpretation* gives an objective value for the probability of x occurring based on the number of actual occurrences of x out of all possible occurrences of x. This interpretation proves particularly problematic for causal inference. Firstly, there is an issue regarding how to ascertain the values. Sampling is our best option here, which is already consistent with scientific practice, but our sampling method may fall short through sampling bias or statistical irregularity. The case of big data does provide some reprieve: if our sample contains N = all, then these biases and coincidences can be minimized. But there is a second problem. The probability for

any outcome based on past occurrences is constantly changing. This means that its entropy value will also be constantly changing. Even a causal process which transmits a single message cannot be expected to have constant entropy, since presumably instances of the event-type are happening elsewhere in the universe, thus changing its distribution. Interpreting probability as relative frequency, therefore, has the undesirable consequence that entropy cannot be a conserved quantity.

(ii) The aforementioned problem might be overcome if we adopt a *physical propensity interpretation* (Popper 1959). This view claims that probability is given by the propensity of a mechanism or system to produce a certain outcome. For example, flipping a fair coin has a propensity (as a real tendency or dispositional property of the system) to produce heads 1/2 of the time. The trouble with the propensity interpretation, as has been discussed before (Gillies 2000: 825), is that values are underdetermined by the evidence. If an event occurs once, its relative frequency is 1, but its physical propensity may be different. Naturally, therefore, this raises questions about our ability to ever know the propensities of physical processes. There are also metaphysical worries: philosophers who adopt causal process theories usually do so on empirical or Humean grounds. But, to borrow an expression from John Earman (1984), propensities fail the "empiricist loyalty test" since two worlds could agree on all occurrent/observable facts but differ over the chances for physical systems.

(iii) The last option to consider equates probability with *subjective degrees of belief* (Ramsey 1990). This interpretation has the virtue of having been extensively discussed already through Bayesian confirmation theory (Howson, Urbach 1993). However, this interpretation is also problematic for thinking about causation as conservation of entropy. Like the epistemic update view, this notion would depend on an ideal agent and their background beliefs. It is quite possible that here subjective degrees of belief are not conserved in causal interaction at all — especially when the outcome of that interaction is "surprising" to the agent. Alexander Flemming's combined degrees of belief of there being penicillin mould and bacteria in his petri dish may be far higher than his belief that the petri dish would contain penicillin mould and dead bacteria. It is hard to see how conservation could be guaranteed in such cases even though causal interaction was clearly involved.

This section has shown that the main difficulty for measuring information in terms of entropy is its dependence on probability. Whilst this provides a quantitative theory, it requires some explanation of the origin of the prob-

ability distribution. Well-known accounts all seem to be problematic in this respect; the difficulties discussed above will have to be dealt with before measuring information in terms of entropy becomes a viable option.

3.3. INFORMATION AS ALGORITHMIC COMPLEXITY

The final measure of information originates from algorithmic information theory (AIT), which was developed independently by Ray Solomonoff (1964) and Andrei Kolmogorov (1965). The basic idea is that informativeness is connected to complexity: the more *complex* an object, the more *information* is required to describe it. The size of the information is measured formally as the *length* of a program running on a universal computing device that can produce a description of the object. Compare, for example, the following two strings:

(a) 01101001100101101001011001101001…

(b) 01011010111001010101111101001000…

At first glance (a) and (b) appear random, but on closer inspection (b) is revealed to contain greater structure than (a). The arrangement in (b) is known as the Thue-Morse sequence and can be produced by a simple set of rules.[4] This makes string (b) computationally less complex than (a). In order for a computer to output (a), it would need to repeat the entire message, whereas for (b) it need only execute a simple algorithm. AIT defines the amount of information in a string S as the *length of the shortest program* which, when executed, outputs S. This quantity, "K(S)," is known as the *algorithmic complexity* of S.

Algorithmic complexity looks like a suitable concept for information transfer. It is easily measurable (the size can be given simply by counting the number of bits) and it is additive: $K(S_1) + K(S_2) = K(S_1 + S_2)$.[5]

A version of the information transfer theory that adopts algorithmic complexity could look something like the following:

> CAUSAL PROCESS: A causal process is the world line of an object that conserves algorithmic complexity.

---

[4] This sequence is produced by starting with 0 and then repeating every consecutive symbol according to two rules: (i) If "0" print "01" and (ii) If "1" print "10."

[5] For finite strings the additivity rule may not be met because short strings with structure may not be compressible if the size of their algorithm is large. The difference dissipates as their size increases. So, assuming S_1 and S_2 are relatively large strings, we can assume that additivity is met. Given most data sets are large, this seems like a safe assumption.

> CAUSAL INTERACTION: There is a causal interaction between processes A and B if the sum total of algorithmic complexity of A and B before interaction equals the sum total of algorithmic complexity of A and B after interaction.

In the case of the wrench in space, the first definition looks plausible. Provided it does not interact with anything else, then regardless of which time we describe it, the total amount of resources required should remain the same. In this case we can say that the single process conserved information over time.

It also looks plausible in cases of causal interaction. Consider a very simple example. Suppose a scientist wishes to investigate whether or not changing the value of a variable A has any effect on another variable B. For simplicity, suppose these two variables only come in two values: 0 and 1. The scientist then records the values of each variable at different points in time and produces two data sets:

Set-1	A_{t1}	0	0	0	0	0	0	0	0	0
	B_{t1}	0	0	0	0	0	0	0	0	0

Set-2	A_{t2}	1	1	1	1	1	1	1	1	1
	B_{t2}	1	1	1	1	1	1	1	1	1

Table 1. Data for variables A and B at two time points

In terms of "meaning" there is clearly no information conservation between the two data sets. The values for A and B at t_1 are not the same as those at t_2. There is therefore no *semantic* conservation. But there is conservation of *structure*, and because of this, both set-1 and set-2 could require the same amount of computational resources to describe. In the case of set-1, we need only note the value of one instance of A (0), the number of instances of A (n = 9), and the rule "v(A) = v(B)" (where "v(x)" stands for the value of x). These three instructions provide a complete description of the data in set-1. But these instructions equally produce all the data we need for set-2. As the descriptions of set-1 and set-2 are the same (and therefore the same size), so there is equality of complexity. If this set of instructions is the shortest possible, then we can say there is conservation of algorithmic complexity (K) and that there is *information conservation* between the states of A and B at time t_1 and their states at t_2.

Intuitively in this case we would be willing to infer that there is a causal connection between A and B, and that changing the value of A "brought

about" or was "responsible" for a change in B. Our intuitions about productive causes in this simple example, therefore, seem to match up with information conservation understood as algorithmic complexity.

One potential worry is that K is language-dependent. As we are measuring K in terms of the number of symbols, its length depends on our choice of encoding. Imagine that I use language L_1 to describe the properties of the wrench before some particular time t. However, after t, I describe it using a different language L_2. Since K is language-dependent this means its complexity will not be conserved and that the amount of information carried by the causal process is not conserved.

There is a solution to this problem we could appeal to. We can exploit a formal result in AIT known as the "invariance theorem":

$$\text{INVARIANCE THEOREM: } (\forall S) \; | K_{L1}(S) - K_{L2}(S) | \leq c$$

This states that for all strings S the difference in their complexities equals a constant c, whose value depends only on the computational resources required to translate one coding language to another (Li, Vitanyi 1993). If the strings are themselves long relative to a translation program, then the difference becomes minimal. In the limit, as the size of S tends towards infinity, the choice of encoding becomes irrelevant.

If comparisons are made using different languages, then we need to be aware that a margin of error will be present and that all inferences are subject to an error given by the value for c. However, given that working within margins of error is standard practice within statistical analysis, the use of complexity as a measure of information does not bring with it any new problems. A better way to avoid this would be to set the requirement that all descriptions are carried out using a specific language. Provided scientists continue to use the same coding language, causal processes will conserve complexity. Given that most data-intensive research requires a period of data curation — where the data is made readable for analysis by a computer — we can be confident that most comparisons will be made relative to a fixed language.

In conclusion, I propose that algorithmic complexity — as given by the value for K — provides the best measure of information with respect to big data analysis. This does not mean that any other measure could not work in practice, only that information as algorithmic complexity seems to pose the least problems when finding causation via the transmission of information.

4. INFERRING CAUSES USING K: AN ILLUSTRATION FROM EXPOSOMICS

It is now time to make good on the work from section 3 and show how productive causes can be found by searching for algorithmic complexity conservation. I will do this by outlining a potential inference rule and illustrate how it might help in the scientific field of exposomics. Exposomics is a relatively recent field. It makes use of big data and new data gathering technologies and has explicitly recommended the use of computer-assisted discovery in the search for causal links (Manrai et al. 2017). It is an ideal field, therefore, to illustrate the potential benefit of a new inference rule for finding productive causes.

Exposomics is the study of the "exposome" and its effect on our health. According to Christopher Wild — who coined the phrase "exposomics" — "at its most complete, the exposome encompasses all life-course environmental exposures (including lifestyle factors), from the prenatal period onwards" (2005: 1847). It is known that many diseases are causally affected by a combination of genetic and environmental factors. Exposomics aims to provide a counterpart to genomics by producing detailed models of risk factors for environmental conditions (such as pollution, radiation, water contamination, etc.) and their connection to certain diseases. One of the innovative elements of exposomics is its use of "omics technologies" that take recordings of biomarkers (internal biological processes) that may be symptomatic of environmental exposure or the onset of a disease. In this way, it aims to provide a more complete description of the pathway that leads from exposure to disease:

$$\text{Environment} \rightarrow \text{Biomarkers} \rightarrow \text{Disease}$$

Data is collected at each stage in this process, and exposomics research groups are using innovative data collection methods in the hope of providing a more fine-grained picture of how environment affects health. For example, the EU-funded project "EXPOsOMICS" hosted at Imperial College London hopes to investigate the effect of pollution on individuals by utilizing smartphone technologies (such as GPS tracking and accelerometers) and have designed portable air quality monitoring equipment that can be carried by individuals (Vineis et al. 2017: 146). This provides much larger amounts of data at higher resolution than traditional standalone monitory stations.

As already seen, the inclusion of biomarkers has led Canali (2016) to believe that the use of big data in exposomics points to a conscious effort to do more than just find correlation but to search for causal connections as well. In a recent article by Illari and Russo (2016), they go one step further and

claim that the relationship between the environment, biomarkers, and disease is best conceptualized as the "flow of information":

> We want to understand the whole system, all the bits, how they interact, quantitatively, build reliable models of the dynamic evolution of whole systems under many different exposure conditions, and the concept giving the dynamic evolution is information. The flow is in the link, and the link, we suggest, is given by information. (2016: 187)

By using biomarkers, they argue exposomics faces an "evidential puzzle." The inside of a human body is a busy place, with many different processes interacting at any given time. The challenge posed by using data gathered from biomarkers is in deciphering whether or not they are part of a causal pathway that links an environmental factor to a disease.

Because Illari and Russo work with a qualitative measure of information, they believe the only way to find evidence of causal processes is through the "interpretation" and "reconstruction" of the data by a human scientist (2016: 186). But, by using a quantitative measure of information as outlined in section 3, I believe it is possible to do more than just this. By using a quantitative measure, it is possible to automatically search for evidence of stages along the causal processes.

To see how this might work, think about what it means for a causal process to conserve information between two or more variables. Let X, Y, and Z be variables that exist along a causal pathway. If the information provided by X is conserved along this pathway, then knowing the information from X allows us to know the information provided by Y and Z. But as I have already made clear, the information is not the same. What is conserved is not the content of the information but the size, and that size is measured in terms of complexity. So to say X, Y, and Z conserve information is to say that the size of the shortest description of X is the same as the size of the shortest description of Y and Z. On this basis, we can develop an inference rule for deciding whether or not a given variable is part of a causal process:

> CAUSAL PROCESS: For variables X and Y, if K(X) = K(Y), then X and Y form part of a causal process.

In the case of exposomics, the inference rule would work as follows. Given exposure (E) and disease (D), the question is whether or not one or more biomarkers (B_1-B_n) forms part of a causal pathway linking E to D. A causal search program can check for this by measuring the sizes of the algorithmic complexity of E and D and then comparing them to the algorithmic complexity of a given biomarker B. If K(B) = K(E) (or alternatively, K(B) = K(D)), then we have evidence of complexity preservation and therefore evidence that B is part of a causal pathway linking environmental exposure to disease. In the proposed

study by Vineis et al. (2017), the inference rule could help by isolating biomarkers which are related to increased exposure to pollution. Mobile monitoring methods, such as those in the study of Klompmaker et al. (2015), can provide data about individual exposure patterns to a variety of pollutants of different size (e.g., PM10, PM2.5, and NO_2) as well as ultrafine particles and soot. By comparing these data with biomarkers from the same individuals, it is possible to identify pathways of contamination for various types of pollutants. Longer term studies or studies with available biomarkers and disease can search for complexity conservation allowing for the establishment of a route between pollutant, biomarker, and the onset of diseases such as heart disease or lung cancer.

The conservation of algorithmic complexity, therefore, provides a simple test to check whether a variable forms part of a causal pathway along two previously established members. But it does not tell us about the direction of the interaction. For example, suppose we have two biomarkers B_1 and B_2 which are equal in their complexity to E and D. This provides evidence that both B_1 and B_2 are part of the body's internal response to the exposure and the onset of a disease, but it does not tell us whether or not one is causally responsible for the other. Yet this information could be very useful to medical practitioners. It could help predict the progression of a disease within an individual as well as assist in therapeutic remedies. It is likely, therefore, that we need to know more than just whether a variable is part of the causal process: we also need to know what role it plays within that process.

The direction of causal interaction can be measured using the idea of "conditional algorithmic complexity" (Budhathoki, Vreeken 2018). Put simply, the conditional algorithmic complexity between two variables $K(Y \mid X)$ is the size of the shortest program that describes Y given X as input. If there is information flow between two variables X and Y in the direction of $X \rightarrow Y$, we would expect that Y would be more easily compressed given a description of X than vice versa. This can be described more formally as follows:

> CAUSAL INTERACTION: For variables X and Y, if $K(X) + K(Y \mid X) < K(Y) + K(X \mid Y)$, then X is a cause of Y.

In this instance, we are comparing compressions of the whole set of data in the causal process given different variables as starting points. Note that this does not contradict the previous claim that there is conservation between variables in a causal process. Here, we are not just comparing K(X) with K(Y), we are comparing their relative complexities when given information about each other. If, as a matter of fact, the best description of the data provided by measuring Y is one which includes a description of X, then this will already factor into the value for K(Y).

Discovering whether one biomarker is a cause of another requires a comparison of their respective conditional complexities. If the data provided by B_2 can be compressed more easily given B_1 than the other way around, then there is evidence that B_1 is a cause of B_2. This provides the justification for positing a link and allows one to make predictions and interventions on B_1 such as medical therapies. By detecting the presence of B_1 in an individual who lives in an area with known high concentration of a given pollutant, it is possible to predict that further biomarkers will be present and help the prevention of the development of a disease. One can also infer causal direction if data is being gathered by a diachronic study that records the values of X and Y at intervals over time. If changes in X occur before Y but both form part of a causal process, then it is natural to infer that X was responsible for the change in Y. This is not always available of course, and some interactions might be instantaneous or near instantaneous. In the ideal scenario, data gathered over time would be used in conjunction with the comparison of conditional complexities.

These two inference rules show how, in principle, it is possible to find productive causes in big data using the notion of complexity preservation. But this leaves one question remaining: can these inference rules be implemented by a computer program in order to automate the search for productive causes? One challenge concerns the nature of K itself. It is widely known that K is *non-computable*: given finite resources, it is not possible to know whether or not K has been calculated (Li, Vitanyi 1993). However, this has not stopped scientists in a number of fields from using the concept of a "best compression" and comparing the lengths of compressions of various sets of data. In practice, scientists turn to a related method called the "minimum description length" (MDL) principle. The MDL for an object, such as a string of symbols, is similar to its algorithmic complexity: it is a measure of its length when best compressed. But whereas K talks about the "best *possible* compression," MDL refers only to the "best compression available" given a limited set of coding languages and methods (Grünwald 2007).

Many examples can be given to show that MDL is routinely measured by a computer program and used to make automated inferences in scientific research.[6] In fact, MDL is behind some of the most important applications of machine learning (Grünwald, Myung, Pitt 2013). If MDL has a good track record of use by computers, then it provides assurance that it could be used in the analysis of big data, with values for best compression given and compared to those of existing data sets. Perhaps the best evidence that algo-

[6] See, e.g., the applications by Iba, De Garis, and Sato (1994), Allison, Edgoose, and Dix (1998), and Tan and Dowe (2003).

rithmic compression can be used to search for causal connections comes from recent work by Kailash Budhathoki and Jilles Vreeken (2018). They have designed a computer program ORIGO, which compares the conditional complexity of various data sets. In trials using both synthesised and real-world data, ORIGO is able to correctly identify causal interactions in a significant number of cases (2018: 296-304).

CONCLUSION

The introduction of big data as a source of scientific knowledge does not mean that causes are no longer important for scientific understanding, but it does show that new methods, such as automated search, form important new tactics in uncovering causal connections. If it is true that scientists need evidence of both difference-making and productive causes in order to infer genuine causal links, then automated search programs need to do more than just find difference-making causes in big data. I have shown how it is possible for such programs to find productive causes as well through the information transfer account of causation. When information is measured as algorithmic complexity, programs can be designed to determine whether or not an event is part of a causal process and whether or not it is responsible for one or another event in that causal process. In practice, the method outlined here should be conjoined with a difference-making method: together they provide the strongest evidence of causal connections within big data.

BIBLIOGRAPHY

Allison L., Edgoose T., Dix T. I. (1998), *Compression of Strings with Approximate Repeats* [in:] *Intelligent Systems in Molecular Biology: Proceedings from the American Association for Artificial Intelligence*, J. Glasgow (ed.), Menlo Park, CA: AAAI Press, 8-16.

Anderson C. (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired Magazine*, June 23. https://www.wired.com/2008/06/pb-theory.

Aronson J. (1971), "On the Grammar of Cause," *Synthese* 22(3-4), 414-30.

Budhathoki K., Vreeken J. (2018), "ORIGO: Causal Inference by Compression," *Knowledge and Information Systems* 56(2), 285-307.

Canali S. (2016), "Big Data, Epistemology and Causality: Knowledge in and Knowledge out in EXPOsOMICS," *Big Data and Society* 3(2), 1-11.

Chadeau-Hyam M., Athersuch T. J., Keun H. C., De Iorio M., Ebbels T. M., Jenab M., Sacerdote C., Bruce S. J., Holmes E., Vineis P. (2011), "Meeting-in-the-Middle Using

Metabolic Profiling — A Strategy for the Identification of Intermediate Biomarkers in Cohort Studies," *Biomarkers* 16(1), 83-88.

Clarke B., Gillies D., Illari P., Russo F., Williamson J. (2013), "The Evidence that Evidence-Based Medicine Omits," *Preventative Medicine* 57(6), 745-747.

Clarke B., Gillies D., Illari P., Russo F., Williamson J. (2014), "Mechanisms and the Evidence Hierarchy," *Topoi* 33(2), 339-360.

Collier J. (1999), "Causation is the Transfer of Information," *Australasian Studies in History and Philosophy of Science* 14, 215-245.

Collier J. (2010), *Information, Causation and Computation* [in:] *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*, G. Crnkovic, M. Burgin (eds.), London: World Scientific, 89-106.

Dowe P. (2000), *Physical Causation*, Cambridge: Cambridge University Press.

Earman J. (1984), *Laws of Nature: The Empiricist Challenge* [in:] D. M. Armstrong, R. J. Bogdan (ed.), Dordrecht: D. Reidel Publishing Company, 191-223.

Fair D. (1979), "Causation and the Flow of Energy," *Erkenntnis* 14(3), 219-250.

Gillies D. (2000), *Philosophical Theories of Probability*, London: Routledge.

Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L., Smolinski M. S., Brilliant L. (2009), "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature* 457 (19 February), 1012-1014.

Godfrey-Smith P. (2010), *Causal Pluralism* [in:] *The Oxford Handbook of Causation*, H. Beebee, C. Hitchcock, P. Menzies (eds.), Oxford: Oxford University Press, 326-337.

Gray J. (2007), *Jim Gray on eScience: A Transformed Scientific Method* [in:] T. Hey, S. Tansley, K. Tolle (eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery*, Redworth: Microsoft, xvii-xxxi.

Grünwald P. (2007), *The Minimum Description Length Principle*, Cambridge, MA: MIT Press.

Grünwald P., Myung J., Pitt M. (2013), *Advances in Minimum Description Length: Theory and Applications*, Cambridge, MA: MIT Press.

Hall N. (2004), *Two Concepts of Information* [in:] *Causation and Counterfactuals*, J. Collins, N. Hall, L. A. Paul (eds.), Cambridge, MA: MIT Press, 198-222.

Hawking S. (2015), "Stephen Hawking Says He's Solved a Black Hole Mystery, but Physicists Await the Proof," accessed 10.04.2015. http://phys.org/news/2015-08-stephen-hawking-black-hole-mystery.html.

Helft M. (2008), "Google Uses Searches to Track Flu's Spread," access 10.04.2015. https://www.nytimes.com/2008/11/12/technology/internet/12flu.html.

Howson C., Urbach P. (1993), *Scientific Reasoning*, Chicago: Open Court.

Hume D. (1978), *A Treatise of Human Nature*, L. A. Selby-Bigge, P. H. Nidditch (eds.), Oxford: Clarendon Press.

Iba H., Garis H., Sato T. (1994), *Genetic Programming Using a Minimum Description Length Principle* [in:] *Advances in Genetic Programming*, K. Kennear (ed.), Cambridge, MA: MIT Press, 265-284.

Illari P. (2011), "Why Theories of Causality Need Production: An Information-Transmission Account," *Philosophy & Technology* 24(2), 95-114.

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

Illari P., Russo F. (2016), "Information Channels and Biomarkers of Disease," *Topoi* 35(1), 175-190.

Klompmaker J., Montagne D. R., Meliefste K., Hoek G., Brunekreef B. (2015), "Spatial Variation of Ultrafine Particles and Black Carbon in Two Cities: Results from a Short-Term Measurement Campaign," *Science of the Total Environment* 508(1), 266-275.

Kolmogorov A. (1965), "Three Approaches to the Definition of the Quantity of Information," *Problems of Information Transmission* 1(1), 1-7.

Laney D. (2001), "3D Data Management: Controlling Data Volume, Velocity, and Variety," *Application Delivery Services* 949, 1-4.

Leonelli S. (2014), "What Difference does Quantity Make? On the Epistemology of Big Data in Biology," *Big Data and Society* 1(1), 1-11.

Li M., Vintanyi P. (1993), *An Introduction to Kolmogorov Complexity and its Applications*, New York: Springer-Verlag.

Manrai A. K., Cui Y., Bushel P. R.,..., Patel C. J. (2017), "Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health," *The Annual Review of Public Health* 38, 279-94.

Mayer-Schönberger V., Cukier K. (2013), *Big Data: A Revolution that will Transform how we Live, Work and Think*, London: John Murray.

Pierce J. (1961), *An Introduction to Information Theory: Symbols, Signals, and Noise*, New York: Dover.

Pietsch W. (2016), "The Causal Nature of Modeling with Big Data," *Philosophy and Technology*, 29(2), 137-171.

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

Preskill J. (1992), *Do Black Holes Destroy Information?* [in:] *Black Holes, Membranes, Wormholes, and Superstrings*, S. Kalara, D. V. Nanopoulos (eds.), Hackensack, NJ: World Scientific, 1-18.

Ramsey F. (1990), *Philosophical Papers*, Cambridge: Cambridge University Press.

Russo F., Williamson J. (2007), "Interpreting Causality in the Health Sciences," *International Studies in the Philosophy of Science* 21(2), 157-170.

Russo F., Williamson J. (2011), "Generic versus Single-Case Causality: The Case of Autopsy Reports," *European Journal for the Philosophy of Science*, 1(1), 47-69.

Russo F., Williamson J. (2012), "EnviroGenomarkers: The Interplay Between Mechanisms and Difference Making in Establishing Causal Claims," *Medicine Studies* 3(4), 249-262.

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

Shannon C., Weaver W. (1949), *The Mathematical Theory of Communication*, Urbana: University of Illinois Press.

Solomonoff R. (1964), "A Formal Theory of Inductive Inference: Part I," *Information and Control* 7(1), 1-22.

Tan P. J., Dowe D. L. (2003), *MML Inference of Decision Graphs with Multi-Way Joins and Dynamic Attributes* [in:] *AI 2003: Advances in Artificial Intelligence: 16th Australian Conference on AI Proceedings*, T. D. Gedeon, L. C. C. Fung (eds.), Berlin: Springer, 269-281.

Vineis P., Chadeau-Hyam M., Gmuender H.,..., EXPOsOMICS Consortium (2017), "The Exposome in Practice: Design of the EXPOsOMICS Project," *International Journal of Hygiene and Environmental Health* 220(2), 142-151.

Wild C. (2005), "Complementing the Genome with an 'Exposome': The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology," *Cancer Epidemiology, Biomarkers and Prevention* 14(8), 1847-1850.