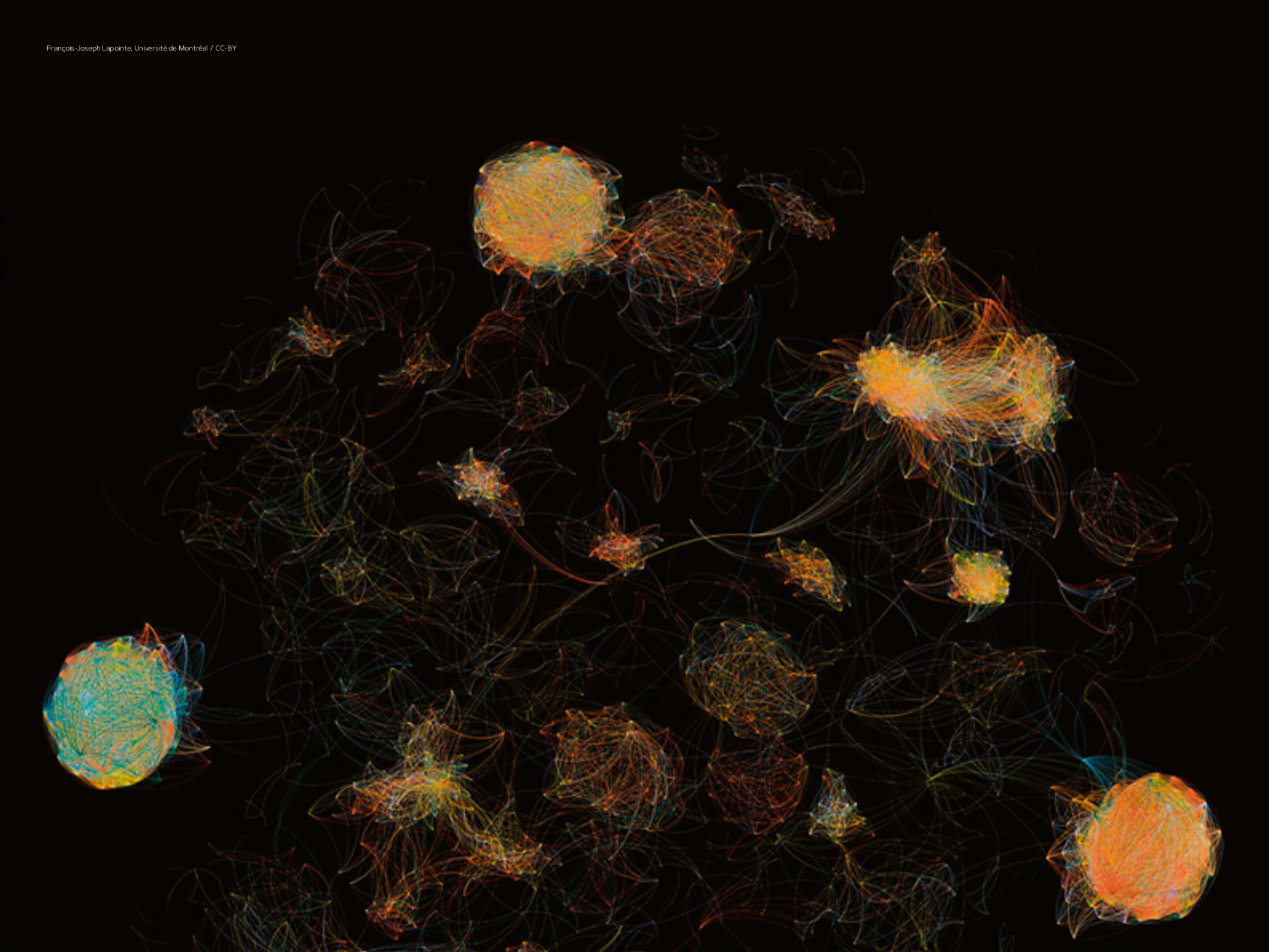


DOCUMENT
Mètode Science Studies Journal, 11 (2021): 6–13. University of Valencia.
https://doi.org/10.7203/metode.11.15258
ISSN: 2174-3487. eISSN: 2174-9221.
Submitted: 23/05/2019. Approved: 27/09/2019.

ON BIG DATA

How should we make sense of them?

Fulvio Mazzocchi

The topic of Big Data is today extensively discussed, not only on the technical ground. This also depends on the fact that Big Data are frequently presented as allowing an epistemological paradigm shift in scientific research, which would be able to supersede the traditional hypothesis-driven method. In this piece, I critically scrutinize two key claims that are usually associated with this approach, namely, the fact that data speak for themselves, deflating the role of theories and models, and the primacy of correlation over causation. My intention is both to acknowledge the value of Big Data analytics as innovative heuristics and to provide a balanced account of what could be expected and what not from it.

Keywords: Big Data, data-driven science, epistemology, end of theory, causality, opacity of algorithm.

The pictures illustrating this text are part of the series «Microbiome selfies», from the performance *1,000 Handshakes* by the artist and biologist François-Joseph Lapointe. The artist shook hands with over 1,000 people, gradually changing the invisible microbial community in the palm of his hand. Every 50 handshakes, the microbes on his palm were sampled and analysed in the lab to reveal how our contact with others shapes the microbes between us. This ongoing project has been performed in different cities around the world (including Copenhagen, Montreal, San Francisco, Perth, Berlin, and Baltimore) as a way to map our collective microbiome using scientific data. Production of the «Microbiome selfies» involved many different steps. Following the collection of the microbiome samples, bacterial DNA was extracted, amplified, and sequenced to create the bioinformatic data shown in this series. The nodes of the network represent bacterial DNA sequences, and two nodes are connected by a line when the bacterial DNA sequences have more than 95 % similarity. The different colours correspond to distinct samples collected every fiftieth handshake, from 0 to 1,001.

On these pages, *Microbiome selfie* by François-Joseph Lapointe, after 550 handshakes during the performance *1,000 Handshakes*.

«THE END OF THEORY» AND OTHER BIG DATA INNOVATION CLAIMS

According to some (e.g., Anderson, 2008), there will be no future for the hypothesis-driven scientific method. The «end of theory» was proclaimed, since we are at the beginning of a new stage in scientific research, driven by petabytes of information and supercomputing. The future belongs to a new form of empiricism, which is technologically driven, and to its powerful tools, which include highly refined algorithms and statistical techniques. Such tools are capable of digging through huge amounts of data, gathering information that may be transformed into knowledge.

Big Data's supporters describe this approach as revolutionary at many levels, pointing above all at two key innovations. First, it is possible to derive meaningful patterns from data analysis. These patterns originate directly from data. As a result, an atheoretical turn is postulated, according to which there would be no need for a priori hypotheses, theories, or models. Second, in the realm of Big Data, «correlation is enough» (Anderson, 2008), and there is no need to investigate the causal links between the associated variables. Hence, correlation supersedes causation.

Actually, the advent of Big Data brings about genuine novelties on the technological ground. Not only are they characterized by volume, velocity and variety, but also by being exhaustive in scope, fine-grained in resolution, highly relational, as well as flexible and scalable in production (Kitchin, 2014). Machine learning techniques are able to automatically mine data and detect regularities, with the belief that «much of what is generated has no specific question in mind or is a by-product of another activity» (Kitchin, 2014, p. 2). By using an ensemble approach, multiple algorithms can be applied to datasets for optimizing predictive performance. What is claimed is that «an entirely new epistemological approach for making sense of the world» is arising. In fact, «rather than testing a theory by analysing relevant data, new data analytics seek to gain insights "born from the data"» (Kitchin, 2014, p. 2).

There is no doubt that the Big Data approach is contributing to change the current epistemic landscape. Besides, data mining techniques are also opening new opportunities in scientific research. For example, there is the possibility to compare hundreds of cancer genomes, and through DNA sequencing, to establish the frequency of many potentially meaningful mutations across different cancer types, together with their functional consequences: this may even contribute to the development of new therapies (Golub, 2010). More generally speaking, by means of these techniques it is possible to discover potentially meaningful patterns within large amounts of data, some of which were previously unnoticed or unknown, because of their complexity.

However, supposing that Big Data represent a genuine epistemological paradigm shift (at least in the sense specified above) is quite another story. In fact, there is no reason to believe that they allow the creation of a new mode of knowledge production in which theoretical assumptions and hypotheses do not play any role and the idea of causation can be disregarded.

There have been strong reactions against both Big Data claims. For example, if we look at both data generation and analysis we find that a «hypothesis-neutral» way of creating knowledge – one depending only on induction and statistical manipulation – is hardly possible.

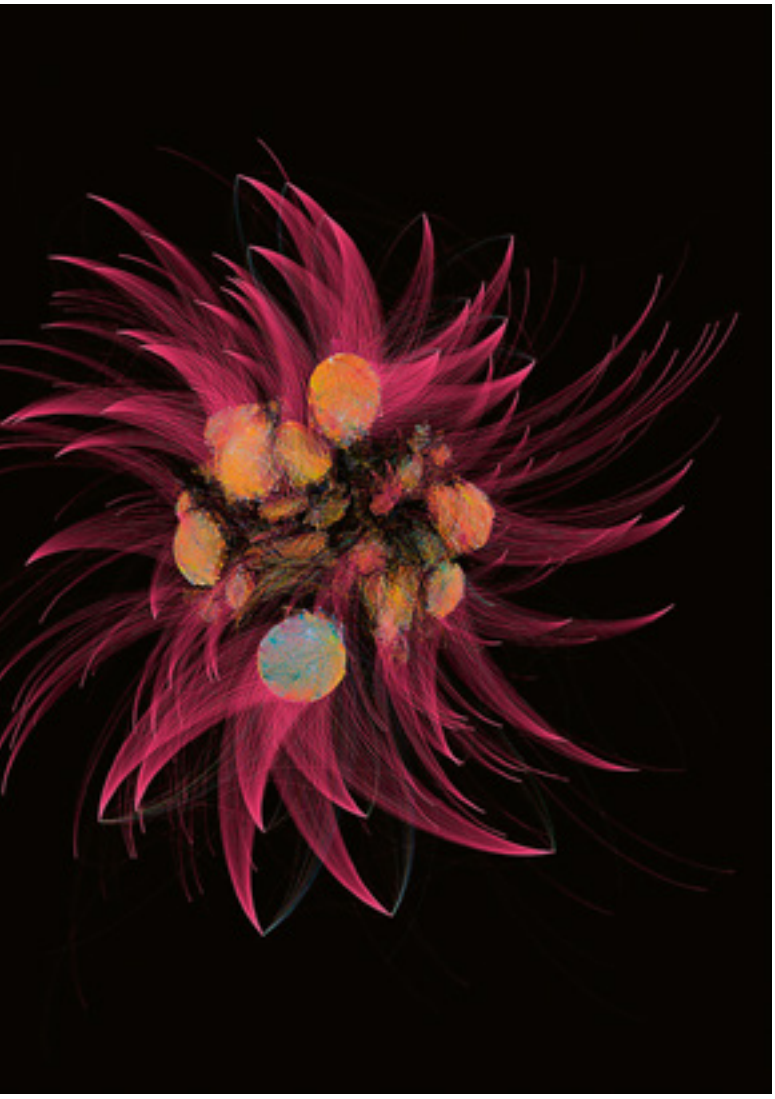
«The "end of theory" was proclaimed, since we are at the beginning of a new stage in scientific research, driven by petabytes of information and supercomputing»

First, data do not arise from nowhere. In twenty-first century's philosophy of science, much has been said about the role of preconceived notions, beginning with Karl Popper (1959, for instance). In his view, hypotheses play an essential role in scientific research, as they settle what to seek and which data to gather. Another well-known argument is the «theory-ladenness» of data and observation, that is, their being «contaminated» by theoretical presuppositions.

Actually, nature is not investigated at random. What comes to be inspected and measured is influenced by the background knowledge, interests and research strategies of the investigator. Even designing experiments depends on specific theoretical, methodological and technical constraints. Therefore, data always result from the interaction between the researcher (who is part of a given school of thought) and the world, provided that suitable material conditions are met (Leonelli, 2015; Mazzocchi, 2015).

Second, data or numbers do not speak for themselves. Potentially meaningful regularities can be computationally found, but what counts most is to find an explanation for them. This presupposes a «framework of analysis», such as a theoretical lens on

François-Joseph Lapointe, Université de Montréal / CC-BY



Microbiome selfie by François-Joseph Lapointe, after 650 handshakes during the performance *1,000 Handshakes*.

«Data or numbers do not speak for themselves. What counts most is to find an explanation for them»

which the way data came to be interpreted depends: it is here that domain-specific knowledge plays a crucial role. Boyd and Crawford (2012, p. 667) point out that «[a]ll researchers are interpreters of data (…) A model may be mathematically sound, an experiment may seem valid, but as soon as a researcher seeks to understand what it means, the process of interpretation has begun».

Several data scientists, and, in the biological field, many bioinformaticians, believe that understanding statistics can be enough to make sense of data. Patterns are supposed to be inherently meaningful, i.e., their meaning transcends context or domain, and there is no need to go outside data. According to their view, theoretical knowledge «depend[s] on reductionist generalisations that abstract from reality in problematic ways» (Chandler, 2015, p. 847). Quite the opposite, the computational approach would allow us to access interconnected data sets and gain a more holistic – beyond disciplinary barriers – understanding of complex phenomena. It is, however, a bit paradoxical to expect that data, which have been produced within a particular context (e.g., biology), could be easily interpreted out of any context. Let me stress this concept again: domain-specific knowledge matters.

Furthermore, even machine learning algorithms are imbued with particular assumptions; for instance, assumptions about what should be considered as a regularity pattern: each algorithm has its own way to develop strategies for finding relationships between data sets, and it is not unlikely that different algorithms will find different types of patterns (Hales, 2013). This is something that is recognized even by several Big Data specialists.

The second claim, i.e., the idea that «correlation is enough», exaggerates the value of prediction from correlations. Perhaps there are particular circumstances, like advertising, in which such an idea could make sense. Yet, surely this is not the case for scientific research.

Correlations may suggest potentially interesting connections. They may even be helpful for generating or assessing new hypotheses, albeit this task will always be guided by some underlying theoretical assumptions and existing knowledge (Kitchin, 2014). However, correlations do not inform us about the underlying cause of relationships.

Establishing causal connections is crucial in science, even to know how to effectively intervene in high-priority situations, e.g., to cure a disease. Therefore, scientific research does not stop at correlations. There is the need for further analysis

and testing: correlations have to be someway «validated». Reliable knowledge is gained only at the end of this process. This depends also on the fact that, especially in very large databases, most correlations are false positives (Calude & Longo, 2017). Owing to the huge volume of data, the problem is how to cope with too much correlation, and to distinguish the meaningful associations from the confounding (i.e., spurious) ones.

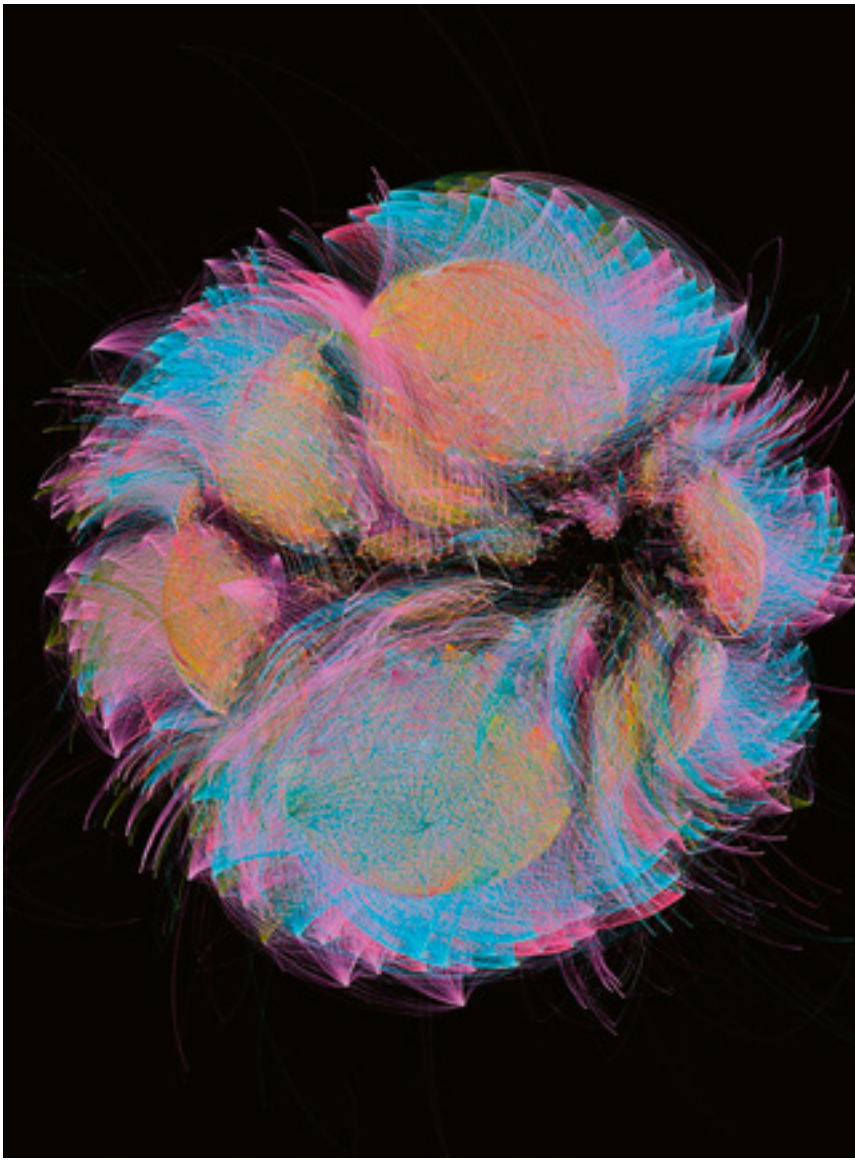
THE EXPOSOMICS CASE HISTORY

Let us now analyze a case history from Big Data biomedicine, i.e., the EXPOsOMICS project. Such a project investigates the associations between exposure and disease, referring to the novel notion of «exposome», i.e., the total amount of exposure that affects individuals during their lifetime. «Internal» and «external» exposures are included in this notion: for instance, Chadeau-Hyam et al. (2010), in their study of breast and colon cancer, analyzed both the diet and lifestyle of patients (external exposure) and the metabolic responses to them (internal exposure).

A key role in this type of research is played by biomarkers, i.e., measurable items of the environment and the organism that show biological processes: associations between biomarkers of exposure and biomarkers of disease are, in fact, searched. Significantly, the investigation on biomarkers is carried out relying on Big Data, which are usually produced using high-throughput technologies as sources: omics in the case of internal exposure and sensors, satellites, and other devices for external exposure. As pointed out by Canali (2016, p. 4):

> EXPOsOMICS is a Big Data project where scientists look for associated biomarkers, capable of tracing exposure and disease. The proponent of the datadriven view may say that this project is the perfect example showing how Big Data research consists in gathering large amounts of data, analysing it, looking for correlations between biomarkers of exposure and biomarkers of disease and making predictions. This would show how correlations are enough and there is no need for causal knowledge.

Actually, this is not the case. In fact, in the abovementioned study of breast and colon cancer, the search for associations in data to identify lists of



François-Joseph Lapointe, Université de Montréal / CC-BY

Microbiome selfie by François-Joseph Lapointe, after 850 handshakes during the performance *1,000 Handshakes*.

«Nature is not investigated at random. What comes to be inspected and measured is influenced by the background knowledge, interests and research strategies of the investigator»

putative linked biomarkers of exposure and disease is only the starting point. A correlation between biomarkers can also be judged as statistically significant, but what is then searched for is a causal link between exposure and disease (Canali, 2016).

For such a purpose, there is the need to look for «intermediate» biomarkers, which are believed to be involved in disease causation. They lie at the intersection between the biomarkers of exposure and of disease. In the case of colon cancer the dietary fiber intake is identified as a likely intermediate biomarker. All this process is guided by a combination of data, statistical tests, theoretical tenets, previous experiments and existing causal knowledge on disease mechanisms, for instance making use of the Human Metabolome Database that contains information on metabolomic mechanisms (Chadeau-Hyam et al., 2010).

Actually, this account of the EXPOsOMICS project, and of many others like ENCODE (e.g., Mazzocchi, 2015), shows how Big Data's claims about the end of theory and the primacy of correlation over causation are flawed. Even if sometimes scientific research begins with data, thus without the involvement of strong a priori hypotheses or models, theoretical and experimental knowledge is still needed immediately after. Besides, methodological considerations, like the choice to use a specific kind of statistical model, play an essential role in shaping research, and in ensuring that data analysis is really effective.

BEYOND THE MYTH OF RAW DATA AND OBJECTIVITY

We can look at the claim that «data speak for themselves» yet from another angle. If we consider etymology, we find that the term *data* is the plural of Latin *datum*, i.e., «something given», corresponding to «that which is given prior to argument», and then needs no questioning. Data are conceptualized as having a «pre-analytical», unbiased nature, counting as a direct reflection or «naked» representation of a particular aspect of nature, as if they were photographs (Gitelman, 2013). This conception is encapsulated in the term «raw data». Big Data exacerbate this situation because the objectivity of data (as particular items) is coupled with the objectivity or neutrality of patterns born directly from them.

The epistemological nature of data should, however, be understood in a more sophisticated way. As already noted, data are not given and never naked; rather they are somehow «made». As reminded by Leonelli (2015, p. 820), «what counts as data is always relative to a given inquiry where evidence is sought to answer, or even formulate, a question». Therefore, data should be seen as sociocultural artefacts. Besides, in order to be usable and function as evidence, they often have to be manipulated and organized in data structure, and yet even this process is driven by theoretical considerations, thus far from being neutral (e.g., Gitelman, 2013).

The process of data generation and management indeed involves several judgements and choices, each somehow biased, for instance about what is meaningful or reliable and what is not. Such considerations may be compared with the notion of «ontic occlusion» (Knobel, 2010), i.e., a mechanism according to which any representation of something occludes other possible representations, with the consequence that the occluded items are not taken into account, and do not play any role in «shaping the narrative». In the light of this view, the process of admitting data, for example into an archive, is a process of occluding other possibilities. Since there is no possibility to overcome the finitude of the archive, several aspects of reality will not be considered or represented (Bowker, 2014).

«Several data scientists believe that understanding statistics can be enough to make sense of data»

Also, human perception and cognition, which function by projecting boundaries on reality, are means of discovery and occlusion at the same time. Different orders can be imposed on the world because of different ways of projecting boundaries. Yet the common underlying mechanism is that in creating a particular order, or in «viewing» something, something else must be excluded. In other words, our perception and cognition are intrinsically «perspectival».

Contemporary philosophers of science like Ronald Giere (2006) have also highlighted the perspectival character of science, i.e., the fact that even scientific observation and theorizing are only able to describe the natural world in the light of a given perspective.

In this respect, the Big Data approach, which postulates a model of «aperspectival» objectivity, is a step back. Claims of objectivity, according to which the algorithmic analysis of data would be a guarantee

of truth and neutrality, reflect, *de facto*, the philosophical immaturity of the field. As stated by Bollier (2010, p. 13):

> As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an «objective truth» or is any interpretation necessarily biased by some subjective filter or the way that data is «cleaned»?

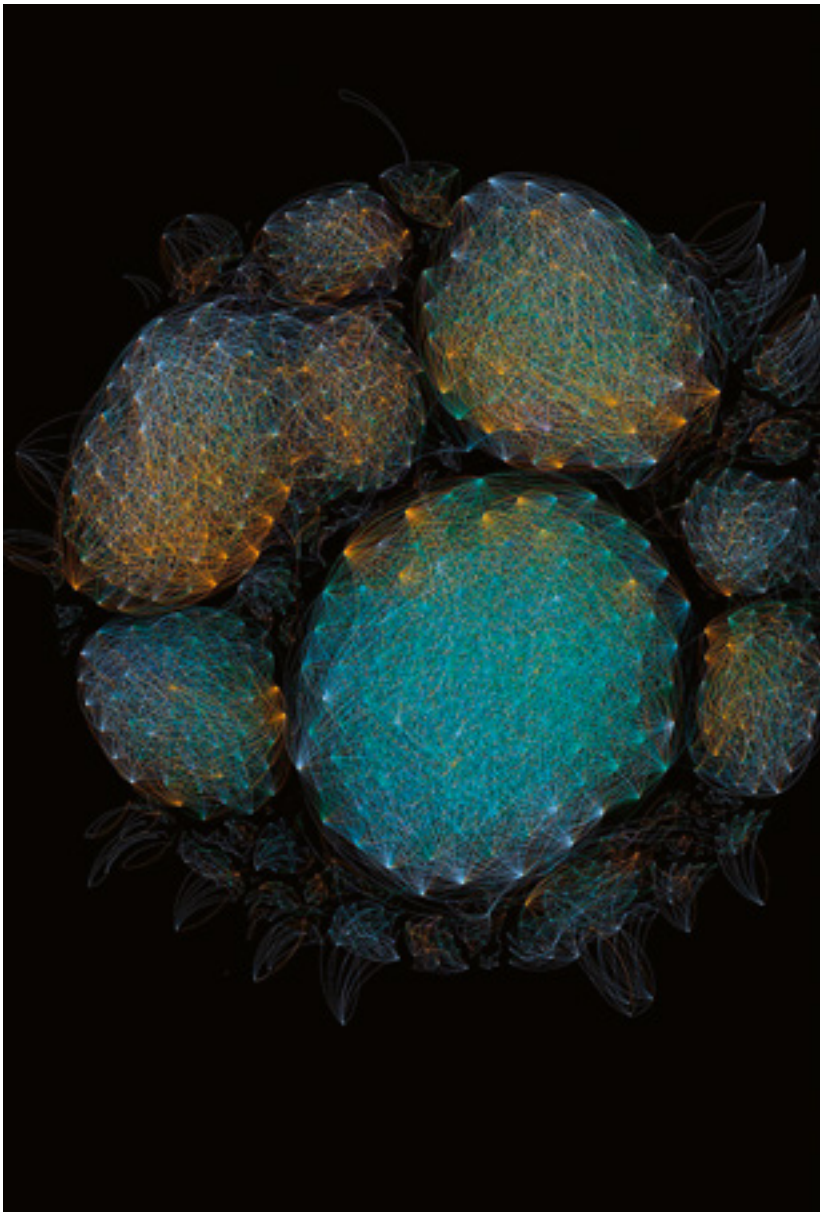
Therefore, even the nature of data should be seen as perspectival. Data and numbers will never speak for themselves, but speak only for the assumptions they incorporate. Furthermore, presuming the neutrality of data is a non-neutral position per se.

BIG DATA'S VALUE AS HEURISTICS AND THE OPACITY OF ALGORITHMS

We can make good use of the novelties brought by Big Data analytics. However, there is no need to expect that principles and procedures that have been employed and refined across many centuries of scientific research will be superseded. Today's science still grounds on theory and experiment and, very likely, will continue to do so. The value of Big Data is, instead, mostly as highly powerful and innovative heuristics.

Big Data and the computational approach contribute to strengthen the researchers' toolbox. The keyword here is *pluralism*, because by augmenting the heuristic tools, it is possible to develop multiple research strategies, which may complement each other. For instance, there is the possibility to cross-compare and establish synergies between the hypothesis and data-driven approach. Perhaps in the future we will even explore new ways of developing theories. Anyhow, a number of Big Data projects like EXPOsOMICS show that data and theoretical elements are «mutually influencing», and that both of them are repeatedly involved in the cycle of scientific research (Canali, 2016, p. 8).

In conclusion, a few words should be said about the need to not uncritically accept the algorithmic culture underlying Big Data. Even very helpful tools could contribute to create an unwanted reality. In fact, the most refined algorithms are not only tools for extracting information. They are increasingly affecting the very fabric of public and individual lives, heavily contributing to shaping them:



François-Joseph Lapointe, Université de Montréal / CC-BY

Microbiome selfie by François-Joseph Lapointe, after 350 handshakes during the performance *1,000 Handshakes*.

«The value of Big Data is mostly as highly powerful and innovative heuristics»

> We are now living in a world where algorithms, and the data that feed them, adjudicate a large array of decisions in our lives: not just search engines and personalized online news systems, but educational evaluations, the operation of markets and political campaigns, the design of urban public spaces, and even how social services like welfare and public safety are managed. But algorithms can arguably make mistakes and operate with biases. The opacity of technically complex algorithms operating at scale makes them difficult to scrutinize, leading to a lack of clarity for the public in terms of how they exercise their power and influence (Diakopoulos, 2015, p. 398).

Algorithms, especially learning algorithms, are highly performative and influencing. Yet, it is hard to grasp their functioning and implications. Even field specialists are not able to fully explain what really happens when the machine processes bulk data to gain novel information, or the reason why in particular circumstances it chooses a way to proceed, instead of another (e.g., Burrell, 2016). For such a reason, they are described as «black boxes».

This opacity to human understanding, something that may even reinforce the «power» of algorithms, is due to technical reasons and the complexity of their functioning. One way to express it is in terms of «epistemic opacity», that is, the fact that we cannot understand all the epistemically relevant factors involved in the operations (Humphreys, 2009).

«Even the nature of data should be seen as perspectival. Data and numbers will never speak for themselves»

At any rate, the growing opacity of algorithms is something that should be carefully pondered. Today, the performativity of Big Data tools is highly celebrated, even with triumphalism. The epistemic power and presumed neutrality of algorithms, which are able to do something unreachable for the human mind, are opposed to the fallibility of human interpretation and decision making. However, performativity should not be used as a reason for simply ceding authority and control to machines.

Instead of only praising the Big Data approach and its algorithms, we should ask ourselves a number of questions. For instance, «what sort of situation is that in which we use tools that are able to perfom particular complex tasks, and yet we are unable to explain how they make such tasks possible?». Nobody would doubt that technological devices like these heavily influence our representation of the world. Therefore, another question could be: «what sort of situation is that in which there are tools capable of shaping our experience of reality, and yet we are unable to fully access their underlying logic and models of representation?».

REFERENCES

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired.* https://www.wired.com/2008/06/pb-theory/

Bollier, D. (2010). *The promise and peril of big data*. The Aspen Institute.

Bowker, G. (2014). The theory/data thing. Commentary. *International Journal of Communication, 8*(2043), 1795–1799.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. http://doi.org/10.1177/2053951715622512

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science*, *22*(3), 595–612. http://doi.org/10.1007/s10699-016-9489-4

Canali, S. (2016). Big data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data & Society*, *3*(2), 1–11. http://doi.org/10.1177/2053951716669530

Chadeau-Hyam, M., Athersuch, T. J., Keun, H. C., De Iorio, M., Ebbels, T. M. D., Jenab, M., Sacerdote, C., Bruce, S. J., Holmes, E., & Vineis, P. (2010). Meeting-in-the-middle using metabolic profiling – A strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers, 16*(1), 83–88. https://doi.org/10.3109/1354750X.2010.533285

Chandler, D. (2015). A world without causation: Big data and the coming age of posthumanism. *Millennium: Journal of International Studies*, *43*(3), 833–851. http://doi.org/10.1177/0305829815576817

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism, 3*(3), 398–415. http://doi.org/10.1080/21670811.2014.976411

Giere, R. (2006). *Scientific perspectivism.* University of Chicago Press.

Gitelman, L. (Ed.). (2013). *'Raw data' is an oxymoron.* The MIT Press.

Golub, T. (2010). Counterpoint: Data first. *Nature, 464*(7289), 679. http://doi.org/10.1038/464679a

Hales, D. (2013, February 1). Lies, damned lies and big data. https://aidontheedge.wordpress.com/2013/02/01/lies-damned-lies-and-big-data/

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626. http://doi.org/10.1007/s11229-008-9435-2

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. http://doi.org/10.1177/2053951714528481

Knobel, C. (2010). *Ontic occlusion and exposure in sociotechnical systems* (Doctoral dissertation). University of Michigan, USA.

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821. http://doi.org/10.1086/684083

Mazzocchi, F. (2015). Could big data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Reports*, *16*(10), 1250–1255. http://doi.org/10.15252/embr.201541001

Popper, K. R. (1959). *The logic of scientific discovery.* Hutchinson.

FULVIO MAZZOCCHI. Biologist and philosopher. Researcher at the Institute of Heritage Science of the CNR (Rome, Italy). His research activity focuses on epistemology (e.g., epistemic pluralism, perspectivism), philosophical issues of scientific research (e.g., the reductionism-holism debate in biology, climate model validation, epistemological issues of big data), and knowledge organization. ✉ fulvio.mazzocchi@cnr.it