Is there anything special about the ignorance involved in Big Data practices?

María del Rosario Martínez-Ordaz

Universidade Federal do Rio de Janeiro
martinezordazm@gmail.com

Abstract

It is a fact that the larger the amount of defective (vague, partial, conflicting, inconsistent) information the more challenges scientists face when working with it. Here, I address the question of whether such challenges are of the same kind when working with ordinary-sized datasets than when working with Big Data. In order to respond to such a question, I focus on one particular epistemic challenge that comes naturally when dealing with large amounts of information, namely, ignorance. I submit that the ignorance that emerges when using Big Data in the empirical sciences is quite unique: it is a limit case of *ignorance of theoretical structure with reliable consequences*. I illustrate this with a case study from cosmology.

Keywords— Epistemology of Big Data, ignorance, ignorance of theoretical structure, epistemic opacity, defective data.

1 Introduction

Cosmology is the branch of astronomy which concerns the studies of the origin and evolution of the universe; some of its objects of enquiry include galaxies, dark matter and dark energy, among others. For long time, cosmology had been regarded to be very different from other empirical disciplines, its basis in particular, despite its successful predictions and observational discoveries, it was in general perceived as too speculative −having a status even closer to philosophy than to other areas of physics (Cf. Massimi and Peacock 2015). Nonetheless, this has changed in the last decades, mostly, thanks to the development of new technological and formal resources that allow scientists to receive, order and integrate enormously large amounts of data. This data is later used in surveys, like Kepler, Gaia and DES, SDSS, DESI, LSST, Euclid and WFIRST, which increase the scope of the cosmologists' predictions, makes more accurate their models and allows them to discover new phenomena.[1]

[1]Thanks to this, much progress is being made in the study of the nature of dark matter and the formation and evolution of galaxies due to the possibility of ordering, integrating and even

But cosmology is not the only scientific discipline that has been benefited from the emergence of Big Data and data science; as a matter of fact, the same happened for other sciences such as geology, climatology and biology which have a long history of working with large datasets. Yet, despite the positive outcome, the incorporation of these new resources to any scientific discipline has come with some problems. In particular, with the increase in the amount of data comes an increase in the degree of ignorance. And it is a fact that the larger the ignorance the more challenges scientists face when working with it, making more difficult for them to achieve any type of scientific success in defective contexts.

The combination of the above gives the impression that, when scientists move from dealing with defects in minimal sets of data to dealing with them in immensely large datasets, the challenges that they face would, at least, multiply. The ways in which these challenges might increase −either in severity or in frequency− will be crucial for determining how to preserve the rationality of cosmologists and any other scientist that participate in Big Data practices.

While philosophy of science has already started discussing the different epistemic and ethical difficulties that come with the use of Big Data in the scientific endeavor, very little attention has been paid to the individual agents and the ways in which they overcome ignorance and acquire both knowledge and understanding when depending on Big Data. The novelty of this paper lies in paying attention to the problems that individual epistemic agents face when using Big Data in the empirical sciences.

Here I aim at contributing to the epistemology of Big Data and its applications in the empirical sciences. In what follows, I address the question of whether the epistemic challenges that scientists face are of the same type when working with ordinary-sized datasets than when working with Big Data. In order to respond to such a question, I focus on one particular epistemic challenge: ignorance. I argue that the ignorance that scientists suffer from is different when working with small datasets than when working with big Data; and therefore, the requirements for overcoming ignorance will be different in both scenarios. I illustrate this with a case study from cosmology.

The plan for the paper goes as follows.The plan for the paper goes as follows. In Sec. 2, I discuss the epistemological worries about the use of Big Data in the empirical sciences. Later on, in Sec. 3, I scrutinize the relation between these epistemological worries and ignorance, and in Sec. 4 I argue that the ignorance that underlies Big Data practices in the empirical sciences is *ignorance of theoretical structure with reliable consequences* and I present a case study from cosmology to illustrate this. Sec. 5 is devoted to drawing some conclusions on the connections between ignorance and Big Data practices in the empirical sciences.

visualizing the data that different telescopes report. A great example of this are the famous images of the *bullet cluster* which integrate optical data, X-ray data, and a reconstructed mass map, and that work as evidence in favor of the existence of dark matter.

2 Epistemological worries

Incorporating Big Data into scientific practices has changed much of the traditional epistemology of science; in particular, it has had an important impact on how scientific knowledge is pursued, achieved, assimilated and shared −both collectively as well as individually. These changes in epistemology can be classified into two main categories: the methodological and the understanding-directed ones. Here, I aim at discussing the most distinctive epistemological worries associated to such changes. To do so, the section is divided in three main parts: Sec. 2.1. introduces some preliminary concepts, Sec. 2.2., summarizes the main methodological worries and Sec. 2.3. presents the main concerns related to understanding.

2.1 Preliminaries

Let *data* be anything that can be soundly recorded in a relational database respecting semantic and pragmatical requirements. "The semantics require that the recordings be understood as true or false statements. The pragmatics suggest that we favor recording what seem to be concrete facts (i.e., singular and relatively weak statements) and that interpreted recordings be true statements" (Frické 2014: 652). *Data science* is the interdisciplinary field that concerns the analysis of data, as well as the extraction information from databases, and the modeling and prediction the behaviour of such information.

Machine learning "is a marriage of statistics and computer science that began in artificial intelligence (...) It is concerned with the design of algorithms that run on a machine to solve some or another problem of our choosing, and it further addresses the question of which problems are tractable enough to admit a computational solution and which are not" (Wheeler 2019: 324). Machine learning has been of the interest of philosophers of information, logicians and philosophers of probability. *Data mining* "is a set of techniques for analyzing and describing structured data, for example, finding patterns in large data sets" (Ivezić *et al.* 2019: 6).

Big Data is the field that concerns ways to work with datasets whose size is beyond the ability of typical database software tools to capture, analyze, store, and manage (Cf. Manyika et al., 2011). Note that the name *Big Data* does not only indicate the amount of data that is managed but, more importantly, the range of computational methods used to work with such data (Cf. Arbesman 2013, Boyd and Crawford 2012). Big Data practices are grounded in data science and, due to the human agents' cognitive limitations, make constant use of machine learning algorithms to process, retrieve, analyze and extract information from immensely large and complex datasets. There are five main characteristics of these datasets: *volume*(the amount of data that is being managed, measurable in terabytes, petabytes, and even exabytes), *velocity* (the data generation rate and the processing time requirement), *variety* (the data-type, which can be structured, semi-structured, unstructured, and mixed), *veracity* (how accurate or truthful a dataset or a data source may be) and *value* (the possibility of turn

data into something useful).

Epistemology of science is a branch of philosophy which concerns how are scientific products obtained and the different ways in which "we seek to support with sufficient evidence that they are worthy of belief" (Bird 2010: 5). Due to the fact that science is often seen as *the business of generating knowledge*, epistemology of science is the study of how such an knowledge is pursued and achieved from very different perspectives −going from considering individual cognitive predispositions to collective practices and methodologies. The *epistemology of Big Data* is an area of philosophical research that analyzes the methodologies behind Big Data practices and their consequences for the acquisition of knowledge; it deals with problems associated to new inferential mechanisms that scientists and computer programs can use when faced with immensely large amounts of data, as well as the paths that they follow to accept, trust and understand this data.[2]

From the outset I want to be clear about the main purpose of the paper. From now on, I only focus on the epistemic issues that individual agents deal with when working with Big Data in the sciences. I leave aside philosophical discussions regarding the logical grounds of information and machine learning algorithms, the philosophical approaches to computability, the connections between Artificial Intelligence and the human mind, among others. My aim in the rest of this section is to show that Big Data practices have introduced important challenges to the scientific activity.

2.2 The methodological worries

Big Data has changed the ways in which sciences are practiced, it has affected information gathering processes and simulations making them more comprehensive and faster than ever before. Nonetheless, it has been argued that the increase of data that Big Data brings to scientific practice must not be conceived as essentially problematic. "Yes, there is an obvious exponential growth of data on an ever-larger number of topics, but complaining about such overabundance would be like complaining about a banquet that offers more than we can ever eat (...) We are becoming data-richer by the day; this cannot be the fundamental problem" (Floridi 2012: 436). As a matter of fact, the novelty of Big Data, at least for epistemology of science, should not lie in the sheer quantity of data involved, but rather in

> (1) the prominence and status acquired by data as commodity and recognized output, both within and outside of the scientific community and (2) the methods, infrastructures, technologies, skills and knowledge developed to handle data. (Leonelli 2014: 2)

With respect to (1), the first epistemological worry is to determine the purposes and uses of such data. The identification of which questions are interesting or

[2]If interested in a comprehensive philosophical analysis of information, logic and computability see [Floridi 2011] and [Floridi 2019], and for introductory discussions regarding the epistemology of Big Data see [Floridi 2012] and [Leonelli 2014].

even essential to answer at a certain moment as well as the production and selection of the relevant answers constitute the key epistemological problem for the use of Big Data in the sciences (Cf. Floridi 2012).

Regarding (2), the most notorious change when moving to Big Data driven scientific practices consists of moving from mistrusting *correlations* to ground scientific activity in the search for them.[3] Correlations are now seen as a form of knowledge −even if compared to explanatory knowledge−, "the correlations may not tell us precisely why something is happening, but they alert us *that* it is happening. And in many situations this is good enough" (Mayer-Schönberger and Cukier 2013: 14). This suggests that the traditional way in which scientific knowledge was conceived is not enough for capturing and describing the statistical epistemic practices that nowadays ground many scientific disciplines.

With the acceptance of correlations as keystones for scientific development came the second worry: how can scientists justify their beliefs. In particular, under which circumstances can scientists rationally trust their epistemic products. This in light of the fact that, most of the time, they are not in a position in which they can explain neither why something is the case nor how exactly they arrived at this conclusion −specially considering the high software dependency that Big Data practices require.

While this problem has been largely tackled by studying its implications for policy, law and ethics, the methodological worry of trusting Big Data can be formulated in terms of *epistemic opacity*.[4] Many of our epistemic processes are, in different degrees, opaque to us, but what is distinctive of the ones involved in Big Data is that many of them would be *essentially* opaque to human agents. It is a fact that scientists are sometimes active parts and witnesses of an important range of the computer operations that ground Big Data practices, yet, the high dependency of machine implementation as well as the complexity of computation (which most of the time exceeds the human capacities) and the intractability of some inferential products make their active role extremely limited.

This worry concerns two main cases of epistemic opacity:

- **Opacity regarding the status of the products:** Right now, it is not clear whether the models that are created by computer-based methods are substitutes for empirical experiments in empirically inaccessible contexts or they are closer to theoretical abstractions (Cf. Barberousse and Vorms 2014, Morrison 2015). This opacity has an impact in the way in which these models are and should be endorsed by the scientists and the doxastic commitments that they might have towards them.

- **Opacity regarding the procedures:** Big Data practices are extremely

[3] Let *correlations* be defined as "the statistical relationship between two data values, are notoriously useful as heuristic devices within the sciences" (Leonelli 2014: 3). For long time, correlations in the scientific enterprise were considered to be confusing and even misleading; as they do not suffice for explanation, it seemed unclear how much could correlations get scientists closer either to truth or to knowledge.

[4] A process is *epistemically opaque* "to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process" (Humphreys 2009:).

> collaborative journeys, meaning that, the processes for recollecting, storing, classifying and analyzing data require joint efforts between different types of experts within the same team (programmers, curators, analysts, among others), as well as between different research communities. To track the explanations as well as to transmit the (procedural) knowledge behind the production of certain results, is at its best hard, and at its worst, impossible. This opacity has an impact in the way in which scientists acquire justification for their beliefs as well as in how they can explain the reliability of the procedures through which the products are obtained.

Very often scientific communities rely fully on the quality of the datasets that are being shared making these practices to be based on a new type of epistemic trust, what is being transmitted and acquired is not only what traditional epistemology would considered knowledge by (expert) testimony. In Big Data practices, the source of knowledge is not an individual expert that can provide better explanations to back up her claims if asked to do so. The source of knowledge is a combination of methodologies (from mathematics, informatics, datascience, and many more scientific disciplines) plus machine implementation, observational reports from very diverse sources, and many more epistemic echelons that are not necessarily susceptible of back-tracking.

While the methodology of Big Data has been recurrently explored during the last decade very little has been said about the repercussions of the methodological changes into the achievement of knowledge and understanding in the sciences. In the rest of the section, I discuss very briefly the understanding-directed worries of Big Data.

2.3 The understanding-directed worries

Understanding "consist of knowledge about relations of dependence. When one understands something, one can make all kinds of correct inferences about it" (Ylikoski 2013: 100). Scientific understanding is a fundamental component of any successful scientific enterprise; understanding a theory allows scientists to find new domains of application for it, and understanding an empirical domain makes it possible to build new theoretical approaches to that domain.

There is a common agreement on the fact that the increase of data that the sciences receive, storage and manage nowadays should lead scientists to an ever greater understanding of the world. Unfortunately, according to the traditional literature, the more scientists rely in correlations and statistics, while losing grasp of causal explanations, the further away from understanding they are (Leonelli 2014).[5] As a matter of fact, for achieving understanding, "the ability

[5] A group of epistemologists of science characterize understanding as an epistemic achievement that comes only after having obtained explanatory knowledge; this type of understanding has received the name of *explanatory understanding* (Cf. Kvanvig 2003; Grimm 2006, 2014; Morris 2012; Strevens 2013, 2017; Kelp 2014; Sliwa 2015; Lawler 2016, 2018). If understanding is essentially explanatory, it would be available only if (i) scientists can provide (causal) explanations for what is being understood, and (ii) the content of their beliefs is true.

to explain why certain behaviour obtains is still very highly valued – arguably over and above the ability to relate two traits to each other" (Leonelli 2014: 6).

The first worry that comes when pursuing understanding in Big Data practices: understanding requires explanatory knowledge, correlations do not suffice for explanation, and the salient product of Big Data methodology is the recognition of new correlations. Therefore, understanding and Big Data methodology might just be going in opposite directions −and more frequent than desirable, one might have to choose between gaining understanding and identifying new patterns.[6]

A second worry is that due to the involvement of epistemic opacity, agents would not be able to identify the relations of dependence between their beliefs. For instance, as long as they ignore important parts of the procedures that originated certain chunks of information, the dependence relations between these chunks will remain opaque for them, preventing understanding to take place.

A third worry concerns the quality of the data that inform the agents' beliefs. Scientific data is, and has been, often *defective*; this is, vague, partial, conflicting or even inconsistent.[7] This defective character of information is not only ubiquitous, but inevitable; for this reason, an important part of the scientific activity consists of tolerating the defects of the scientific data while aiming at acquire some scientific success −such as increase of either predictive or explanatory power, accuracy, empirical adequacy, among others. So, it should not come as a surprise that the data that scientists get when working with immense datasets is defective. However, there is a consensus on the factive character of understanding; this is, the content of what will be understood should be true. In the case of defective information, the satisfaction of this factive condition does not look so straight forward, and therefore, understanding seems still too distant.[8]

To summarize: I take this section to have shown that, when science incorporates Big Data to its epistemic practices there are, at least, five important worries to consider. On the one hand, from a methodological point of view, the first worry is to determine the purposes and uses of the data that is being storage, analyzed and used. The second worry is to justify the scientists'

[6]There is an alternative account for scientific understanding which does nor require the previous acquisition of explanatory knowledge; however, it still requires that the content of the beliefs that will be related and understood is known to be true, which will also conflict with the third understanding-directed worry. If the reader is interesting in this view, see [Pettit 2002]; [Elgin 2004, 2007, 2017]; [De Regt and Dieks 2005]; [De Regt 2009, 2015]; [Khalifa 2013]; [De Regt and Gijsbers 2017].

[7]The fact that scientists work most of the time with defective information has an important impact in the way they construct and apply their theories and models. Some historical examples of defective scientific theories include: Aristotle's theory of motion which, allegedly, contained mutually conflicting assumptions (Cf. Priest and Routley, 1983), Bohr's theory of the atom, which combined importantly partial information (Cf. Bueno and French, 2011), and Classical Electrodynamics which, allegedly, is internally inconsistent (Cf. Frisch, 2004), among others.

[8]When having two mutually conflicting or inconsistent sets of data, the shared intuition is that at least one of them contains a falsehood; for that reason, it would not be easy to decide which of both chunks is reliable and even worst, which contains only true statements.

beliefs while overcoming two types of epistemic opacity −one about products, the other about procedures. On the other hand, from an understanding-directed point of view, there are three main worries to take into account: first, to achieve understanding even if trusting correlations pulls in the opposite direction from explanatory knowledge. The second worry is to overcome the two kinds of epistemic opacity in order to achieve scientific understanding. Finally, the third worry consists in identifying a way to make compatible the presence defective data with the chase for understanding. The next section is devoted to explain how these five worries have a common ground: ignorance.

3 Types of ignorance

The connection between understanding, knowledge and ignorance is undeniable. This considered, the study of ignorance and the different ways in which we deal with it is expected to be revealing about rationality and the acquisition of both knowledge and understanding.

Here, I take that the study of ignorance can shed light on important peculiarities of Big Data practices in the empirical sciences; this section is devoted to characterize the different types of ignorance that have been recognized in traditional epistemology. The section is divided in two main parts: Sec. 3.1. explains very briefly how the worries introduced in the previous section indicate different types of ignorance. Sec. 3.2., provides an overview of the different types of ignorance that epistemologists have recognized and they might put forward against the scientific activity.

3.1 A common ground

In the previous section, I argued that there are five worries associated to Big Data and its epistemic practices. While these worries might seem very different from each other, they have a common ground: ignorance.[9]

For instance, needing to determine the purposes of data reveals that, because Big Data methodology consists in accumulating as much data as possible without a privileged purpose, when possessing access to immense datasets, scientists often *ignore* the specifics of the domains of application for such data as well as the problems that it can help to solve within the discipline. In addition, the two types of epistemic opacity that I discussed before are clear instances of ignorance, scientists ignore the nature of the products of simulations as well as the procedures through which they were obtained −and most of the time, they cannot perform the inferential procedures that originated such products.

In addition, the trust of correlations when also aspiring to (causal) explanation reveals the previous acknowledgment of ignorance of a causal link. As I had already mentioned, as epistemic opacity indicates ignorance, when it conflicts with understanding, it can be said that because there is still a blank that should

[9]Here just assume the intuitive characterization of ignorance as *lack of knowledge about x*. In Sec. 3.2., I present a more technical approach to ignorance.

be filled −such a blank could be about the status of the models and simulations, or about the mechanisms that generated such models − understanding remains out of reach for the scientists.

For the case of the emergence of defective data the role that ignorance plays should not be marginalized. There is the trivial sense in which having incomplete, partial or vague information is only a direct consequence of ignoring important bits of such information −how it connects, how it behaves, how does it relate to other datasets, among other aspects. Yet, there is also a more substantial interpretation of the ignorance behind the use of defective data, which is, even if we know that two mutually conflicting or even inconsistent chunks of information cannot be true at the same time, what scientists ignore is how to determine the truth values of the propositions contained in each chunk, and that uncertainty is what prevents the achievement of understanding

In sum, as the epistemological worries coincide in being the result of apparently different types of ignorance, the role that the latter might play in Big Data practices becomes a subject worth of philosophical analysis when trying to understand the epistemology of Big Data. In what follows I present four different types of ignorance, this to be able to address, in Sec. 4., the kind of ignorance that underlies Big Data practices.

3.2 Ignorance about what?

First of all, some preliminaries.

Traditionally, ignorance has been understood as *lack of knowledge*. In this sense, one can be ignorant via the non-satisfaction of any of the basic conditions for knowledge. This is, by failing at fulfilling a doxastic condition (S believes that p), an alethic condition (p is true), a justificatory condition (S believes that p with justification) or a Gettier-proofing condition (S's justification for believing that p must withstand Gettier-type counterexamples) (Cf. Le Morvan and Peels, 2016: 18).

Following such characterization, ignorance is often classified in, at least, the following types: (i) *absence of factual knowledge*, (ii) *absence of objectual knowledge*, (iii) *absence of procedural knowledge*, and very recently, another type of ignorance has been added to the list: (iv) *absence of knowledge of theoretical structure* [see Martínez-Ordaz, 2020]. Orthogonally, one can also recognize (v) *absence of answers to questions*. However, because at this point, there is no clarity regarding its status compared to the other types or whether this ignorance reduces to any of the others, in what follows, I do not focus on this particular type expecting that the characterization of the other four is broad enough to capture the large majority of cases of lack of answers to questions.[10]

Considering the main purpose of the paper, I address these types of ignorance by paying special attention to corresponding challenges that they (might) impose to the scientific activity in general and to scientific reasoning in particular.

[10]This type of ignorance has also been known as "erotetc ignorance". For more detailed analyses on erotetic ignorance see: [Rescher, 2009], [Nottelmann, 2016] and [Peliš, 2017].

(i) **Factual ignorance (or absence of factual knowledge):** this ignorance consists in lacking knowledge of either facts or the truth of specific propositions. For instance, let p be ⌜cosmic microwave background is electromagnetic radiation which fills all space and it is considered to be a remnant from an early stage of the universe⌝. When an agent S is factually ignorant of p the agent fails at determining the (correct) truth value for the proposition in question. This could happen due to: S holding a false belief, S struggles at assigning an alethic value to p[11] or S' cognitive limitations prevent her from knowing a particular fact.

This type of ignorance conflicts with scientific reasoning in a very peculiar way: if one grants Modus Ponens (rule of detachment) to play a privileged role in human reasoning, when faced with a conditional of the form $p \rightarrow q$, if scientists ignore the truth of p, detachment seems impossible. In particular, if factual ignorance is of the kind caused by S's failure to assign an alethic value to p, every conditional in which p is the antecedent, will remain in its conditional form −which for the scientific practice might mean that there will be a chunk of propositions that scientists would not be able to assert as they directly depend on something that is ignored, p. To restrict the use of one of the most basic inferences should be seen as loss, a loss that is directly caused from ignoring the truth of just one proposition.

(ii) **Objectual ignorance:** this ignorance requires absence of knowledge of a particular object. This *object*-category might include, but not be restricted to, people, appearances, sensations (tastes, smells, looks, etc.), places, among others.

Objectual ignorance is often expressed in terms that resemble factual ignorance, if S ignores the taste of apples that could be expressed by saying that S ignores that ⌜the taste of apples is sweet⌝, this is, S ignores p. Yet, the main characteristic of this type of ignorance is not that one ignores the truth of one specific proposition, but the fact that one ignores a whole set of properties that an object possesses and that are regarded to be indicative of such an object. For this reason, objectual ignorance prevents agents to connect lists of properties to a particular object.

This type of ignorance conflicts with scientific reasoning in the sense that, even if knowing that there is an x which has the properties p_1 and p_2, and knowing that there is a y that has the properties p_1, p_2 and p_3, one cannot determine whether there is any relation between x and y until we come to know them. Therefore, the main problem that comes with objectual ignorance is the impossibility of relating lists of properties to

[11] When believing p, S cannot determine whether p is the case or if she is mistaken. This type of ignorance is often not caused by the phenomenon itself but by a temporary lack of resources to test the truth of p" (Martínez-Ordaz 2020: 7). For instance, the cosmic microwave background was discovered in 1965, and before that, while its existence was predicted already, the truth of p was still undetermined by lack of empirical evidence; therefore anyone believing .

a common object, which prevents scientists from doing both identifying (new) entities or phenomena and naming them.

(iii) **Procedural ignorance (or absence of procedural knowledge):** this type of ignorance requires agents to not know how to perform a certain task, such as riding a bike, baking a cake, operating a computer, and so on. According to some epistemologists, this type of ignorance also resembles factual knowledge (Cf. Williamson 2001, Snowdon 2004). Sometimes, to be able to follow procedures can be translated into knowing lists of causal relations, this is, knowing what has to be done to obtain certain outcome. According to some, to be procedurally ignorant would be indicated by lack of answers to questions of the type *what are the sufficient steps that one has to follow in order to do x?* [12]

This type of ignorance conflicts with scientific practice in, especially, experimental contexts. If S is ignorant of how to operate a telescope or a computer, or if she is ignorant of how to run a program in a certain way, this limits significantly the results that S could achieve by her own. Yet, nowadays, scientific practice is done mostly collectively. With this in mind, consider a scenario in which all members of a particular scientific community are ignorant of how to reproduce an experiment in order to validate other team's reports; this absence of procedural knowledge becomes an impediment for the other team's results.

(iv) **Ignorance of theoretical structure:** this type of ignorance consists in lacking knowledge of

> the (relevant) inference patterns that scientific theories allow for. When ignoring (the relevant parts of) the theoretical structure of a theory, scientists are not capable of grasping abstract causal connections between the propositions of their theory, they can neither identify the logical consequences of the propositions that they are working with nor can explain under which conditions the truth value of such propositions will be false. (Martínez-Ordaz 2020: 12)

This ignorance is common in the early stages of scientific theories. Common sense would dictate that if the theory is very young the inferential patterns that connect parts of the theory might still remain unknown. Nonetheless, even if the theory is already well accepted by the community, the connections between it and other theories of the same domain can still remain unknown. This gives the impression that ignorance of

[12] However, the fact that in many interesting cases of procedural knowledge, such as riding a bike or calming a crying baby, agents often cannot provide causal explanations of how to perform certain task −despite being able to perform it successfully− suggests that to answer questions like the above might not be necessary for ascribing procedural knowledge to someone. And thus, lack of answers to those questions might not be enough to ascribe procedural ignorance, unless the agent actually can neither explain nor perform the task in question. This type of ignorance comes with epistemic opacity regarding procedures.

theoretical structure is very likely to never be fully overcome, it will be always a part of the inference patterns of a theory that scientists keep ignoring.

When, for long time, despite the development of new instruments and experimental resources, scientists still struggle assigning an alethic value to specific propositions, this is explained by the lack of access to the segment of the structure that governs such propositions within the theory. Having access to a relevant part of the structural conditions of the theory allows scientists to infer the value of certain proposition. Nonetheless, the scope of this will be constrained by the theory; this is, the proposition will have a specific value in a world like the one described by the theory, but this does not necessarily extend to the actual world.

There are two main consequences of the analysis that has been carried out in this section. First, there are three important challenges that scientists face when suffering any of these types of ignorance: reduced inference power (specially with regard to Modus Ponens), incapability for grouping and integrating properties with respect to a particular object, and incapability for following and explaining experimental procedures. Second, in order to overcome these difficulties, it is necessary to respond to a more general challenge: ignorance.

In addition, after analyzing these types of ignorance, it becomes clear that the five epistemological worries that I discussed in the Sec. 2 are significantly related to different types of ignorance. But, how is this presence of ignorance *significantly* problematic? In particular, it is a fact that human agents are epistemically limited in the sense that they are constantly ignorant of different things at different moments. So, if ignorance is not only common but essential to human agents, why should we worry about it when using Big Data in the sciences? These questions lead nicely to the study of the ignorance involved in Big Data practices when applied to the empirical sciences; so, in the next section I focus on discussing the particular kind of ignorance that surrounds these practices.

4 Big Data, Big Ignorance?

In a nutshell: I contend that the ignorance that underlies Big Data practices is, most of the time, ignorance of theoretical structure and that it can be overcome in order to gain scientific understanding of a specific phenomenon. I take two complementary paths. First, I explain that, in the majority of cases, ignorance of theoretical structure underlies any of the other type of ignorance when using Big Data. Second, I address which is the type of understanding available to the scientists when overcoming ignorance of theoretical structure and illustrate this with a case study from cosmology.

The section is divided in four parts: Sec. 4.1. I briefly acknowledge how important is to identify the ignorance that is involved in Big Data practices and the ways to over come it. Sec. 4.2. addresses the type of ignorance that

underlies the Big Data practices, Sec. 4.3. sketches the type of understanding that is achievable through these practices and Sec. 4.4. provides a case study from cosmology to illustrate how this ignorance has been overcome.

4.1 Why should we care?

Big Data-methodology consists of the recollection of very different types of data (images, redshifts, time series data, and simulation data, among others) that relates to different aspects and facets of the studied phenomena −this is, in the large majority of cases, scientists receive partial information about their object of study. This recollection involves integrating data from various sources and formats which initially might not be fully compatible.

Also the data is produced, transmitted and analyzed at an extremely high velocity, which prevents individual agents to keep a detailed track of how the data changes and relates. In addition, it is well known that the use of defective (partial, incomplete, conflicting, inconsistent) data comes with the price of different degrees of ignorance (see Wimsatt 2007, Norton 2008). But scientific rationality is only met either when the degree of ignorance can be maintained or reduced, or when scientists do not hold any doxastic commitments towards the information that they are working with −in particular, if they do not trust neither the information that they are working with nor their results.[13]

The combination of the above poses the following dilemma against scientific rationality: unless scientists find a efficient way to low the level of ignorance, they are irrational for trusting data that at its best is defective and at its worst might be false; or they are irrational for reasoning under high degrees of ignorance –regardless their doxastic commitments towards the products of using Big Data. So either we explain how scientists can reliably lower their degrees of ignorance when working with Big Data or we accept the fact that they are irrational.

4.2 The ignorance behind Big Data

Ignorance of theoretical structure is commonly the cause of persistent instances of any of the other types of ignorance. For instance,

(i) if an agent cannot assign the value of a specific proposition, this could occur because she ignores the relevant part of the theoretical structure that determine whether the proposition is true within the theory in question (Cf. Martínez-Ordaz 2020).

[13]While there is not a uniform view on what *scientific rationally* is exactly, there is a common agreement on the fact that reliable indicators of it include: the achievement of knowledge and understanding, instances of scientific success (accurate predictions, the provision of explanations, manipulability via experimentaiton, among others), reliable mechanisms for constructing, testing, revising, and selecting theories, among others. For the purposes of the paper, I focus only on the relation between scientific rationality, knowledge and understanding.

(ii) if an agent cannot identify whether distinct sequences of properties refer to the same object, this exhibits her lack of knowledge about the theoretical structure of these properties with regard to a specific object.

(iii) if an agent ignores how follow a specific procedure to solve a problem x, this could happen because she ignores the inference patterns that are allowed when addressing x in a theoretical framework y.

However, even if this is correct for scientific practices when working with minimal datasets, one should still argue that the same will occur in Big Data practices. In order to do this, let's pay attention to two differences between more traditional ways of doing science and science data driven science: (a) in Big Data contexts, the human agents' cognitive limitations prevent them from computing, classifying and analyzing by themselves the data at the speed it is received, as well as (b) the massively cooperative nature of Big Data practices makes the transmission and acquisition of knowledge very opaque processes.

With respect to (a), the phenomena that are nowadays studied with the help of Big Data and data science are phenomena that were initially thought to be inaccessible to human agents for different reasons: their complexity, their scale, the physical or temporal distance between the object of study and the scientists, among others.[14] The tools that nowadays grant scientific access to these objects include instruments that collect high dimensional data from such objects, and formal apparatuses for extracting information from the data in question.

The latter category includes machine learning algorithms that help scientists to detect, classify, integrate, visualize and clean data regardless its complexity or the remoteness of the sources. The combination of both physical instruments and formal tools has helped to automate much of the scientists' processes (like pattern recognition and classification) as well it has facilitated big tasks (like processing vast amounts of data in hours instead of the months or years it would take for a group of human agents by themselves), producing extremely comprehensive descriptions of the newly recognized objects. This together has allowed scientists to acquire knowledge regarding the objects that were initially inaccessible; this is, to attain objectual knowledge. However, as the selection of data depends heavily on the algorithms and the instruments, scientists end up losing sight of most of the particularities of the data that is received and neglected by the tools they are using. The price to pay is factual ignorance.

Regarding (b), to incorporate Big Data into a scientific discipline is a highly cooperative activity, meaning that it requires scientists from extremely different areas of knowledge to trust results that they cannot fully explain −programmers might ignore what the curators know, and hardware engineers might ignore what the scientists who select the data would know. Making the process of analyzing the data the result of a eclectic combination of methodologies (from mathematics, informatics, datascience, and many more scientific disciplines)

[14] This clearly resembles a case of objectual ignorance: while scientific theories could predict or require the existence of a specific object, due to their limited access to them agents ignore these objects.

plus machine implementation, observational reports from very diverse sources, and many more epistemic echelons that are not necessarily susceptible of backtracking.

The inclusion of different scientific communities with diverse backgrounds causes that scientists not only ignore the procedures that are followed in different stages of the data analysis. This is, scientists might also be lacking procedural knowledge about the ways in which some results are obtained. In the long run, this has the effect of scientists being unable to provide explanations about procedures; in addition, this prevents them from providing causal explanations about the novel phenomena and the paths that lead to their discovery.

So far, we can say that there are two types of ignorance that seem proper of Big Data practices: factual and procedural ignorance. However, I consider that these instances of ignorance are only symptoms of a more profound lack of knowledge: ignorance of theoretical structure.

The fact that scientists are able to overcome their objectual ignorance and provide accurate descriptions of specific objects should not be mistaken by a partial overcoming of ignorance of the relevant theoretical structure.[15] In the same way, the factual and procedural ignorances that are exhibited must not be disregarded as the mere result of lack of more sophisticated instruments, they should be seen as symptoms of a more profound illness.

Given the multiplicity of approaches undertaken by scientists to extract useful information from big data it could be said that even in those cases where this information points out to the existence of an object and some of is properties, it does not do so in a way that suffices for full blown knowledge of theoretical structure. Crucially, the disjointness in methods, types of information and models used to arrive at the object leaves gaps in the scientific understanding of theoretical structure, both in terms of inferences and properties and in terms of experimental and operational procedures for its use, and this is why neither the alethic blanks can be filled nor the procedures can be explained.

When working with Big Data, scientists are trading knowledge of some parts of theoretical structure in exchange for access to inaccessible objects. While in other contexts they would aim at gaining as most control as possible of their their theories, their instruments and the mathematics that underlie the theories' applications, the incorporation of Big Data to the empirical sciences has brought a new preference: "answers are found through a process of automatic fitting of the data to models that do not carry any structural understanding beyond the actual solution of the problem itself" ((Napoletani, Panza, and Struppa 2014: 486. in [Leonelli 2020]).

But, considering the loss of knowledge of theoretical structure and its negative impact on achieving knowledge and understanding, one should start worrying about how reliable are both the data and the results of data-analysis. "In terms of logistics, having a lot of data is not the same as having all of them,

[15] *Partial overcoming of ignorance of theoretical structure* means that, when tolerating a contradiction, scientists need not to identify the *ultimate* or the total structure of their theory, but that they can provide a set of inference patterns that allow them to successfully use the theory in question while avoiding logical triviality (Cf. [Martínez-Ordaz 2020]).

and cultivating illusions of comprehensiveness is a risky and potentially misleading strategy" (Leonelli 2020). This concern has not been overlooked by the scientists, as a matter of fact, curators, and researchesrs have constantly search for methodologies that allow them to preserve and justify the reliability of the data. An important instance of how these collaborative work succeeds are "taxonomic efforts to order and visualise data inform causal reasoning extracted from such data (Leonelli 2016, Sterner and Franz 2017), and can themselves constitute a bottom-up method—grounded in comparative reasoning—for assigning meaning to data models, particularly in situation where a full-blown theory or explanation for the phenomenon under investigation is not available (Sterner 2014)" (Leonelli 2020).

The positive outcomes of the joint work of researchers, curators and programmers for preserving the reliability of the analyzed data include accurate predictions, measurements and descriptions. This is, while scientists are losing detailed track of the inference paths that are actually followed by the programs that analyze the data, some of the results that these programs are reaching are extremely novel, accurate and trustworthy; this is quite unique and should be regarded as a special feature of the ignorance involved in Big Data practices.

The absence of knowledge of (relevant parts of) theoretical structure is normally explanatory of negative scenarios −such as the presence and tolerance of contradictions in the sciences (Cf. Martínez-Ordaz 2020). However, for the case of Big Data practices, it seems that when scientists accept to lose sight of the ways in which inference patterns are selected and followed, the results are clearly epistemically beneficial: acquisition of objectual knowledge (of objects that were initially thought to be inaccessible) as well as reliability preservation for the analyzed data.

4.3 Understanding Big Data

Big Data practices have granted scientists access to new phenomena, and have provided them with the opportunity of accurately identifying, measuring and predicting their behaviour. In particular, Big Data has allowed empirical scientists to achieve objectual knowledge of things that for centuries were considered to be too complex for the human mind. But this success is not without its downside, it comes with the loss of causal explanations −with respect to, at least, novel phenomena−, and therefore, the loss of explanatory understanding with regard to the newly discovered objects. However, is this knowledge all we can get?

I think there is a way to interpret the epistemic profits of Big Data practices as a keystone for the achievement of scientific understanding. Nonetheless, the type of understanding that can be gained is *modal understanding*. Let me press further this point.

Given the multiplicity of approaches undertaken by scientists to extract useful information from Big Data, it could be said that even in those cases where this information points out to the existence of an object and some of is properties, it does not do so in a way that suffices for full blown knowledge of theoret-

ical structure. Crucially, the disjointness in methods, types of information and models used to arrive at the object leaves gaps in the scientific understanding of theoretical structure, both in terms of inferences and properties and in terms of experimental and operational procedures for its use.

Nonetheless, it might happen that scientists are sometimes able to obtain important information regarding an object from big data analysis. If that information is put together with independent theoretical knowledge in the special sciences, it becomes possible for scientists to generate a representation of the object, a possible world or a proper part of one that represents how the object is embedded in a relevant theoretical domain. However, given the distinct sources of knowledge of the object and the lack of unity in methods and conceptual resources scientists cannot be sure that these possible worlds are actual, i.e., that these representations hold of some actual empirical domain.

In the corresponding literature, it has been argued that *false Theories can still Yield Genuine Understanding* (see De Regt and Gijsbers 2017). This is, for a given set of propositions, even if the veridicality condition is not satisfied, this would not necessarily prevent scientists from gaining understanding of such a set of data. According to De Regt and Gijsbers, what is needed for understanding is only the satisfaction of an 'effectiveness condition' −where, for this case, 'effectiveness' could be understood as the tendency to produce useful empirical outcomes of certain kinds, such as accurate descriptions and predictions.

The limitations that could come with gaining understanding via Big Data might include that the type of understanding that is gained is, only, *modal understanding*. "One has some modal understanding of some phenomena if and only if one knows how to navigate some of the possibility space associated with the phenomena" (Le Bihan, 2017: 112). In the case of Big Data practices, to achieve modal understanding of the behaviour of novel objects in a established theoretical domain would be to determine the set of possible worlds that correspond to the generic structural features assumed by the theoretical view that such a cluster of data substantiate (this is, if the theoretical description of that domain and the recently constructed objects were to be true, which type of scientific practices would it describe, how would these practices relate to one another, among others).

All this considered, I think that the downsides of using Big Data in the empirical sciences, does not prevent scientists from gaining modal understanding of how the new objects *could* relate to theoretical frameworks. In what follows, I illustrate this with a case study from cosmology.

4.4 Cosmology and Big Data

Observational cosmology is a branch of astronomy which aims at providing precise agreement between large scale-physical theories, cosmological models and observation. However, the type of observation that cosmologists address is not what the 'naked eye' can see, but the result of combining different types of data that are largely mediated by extremely technical processes of data analysis and data integration. On a daily basis, high-throughput detectors and (land

and space based) telescopes generate gigabytes of information about objects in specific regions of the sky; as the data that is gathered comes in heterogeneous formats, it has to go through different isolating, filtering and integrating processes before being of any use for cosmologists.

For instance, the raw data has to be reduced into images. In order to do so, the data that is received by a particular telescope should be treated by using different statistical methods for reducing the noise, remove particle events, and combine overlapping scans. While the selection of such methods is often described in internal technical memoranda, it commonly goes unnoticed and is almost never subject to public scrutiny. The fact that this information is kept mostly among astronomers and programmers working in astrostatistics and astroinformatics, feeds the cosmologists' epistemic opacity towards processes of data analysis (Cf. Feigelson and Babu 1997).[16] Once this is done, the constructed images should be reduced into catalogues.

> Much work has also been directed to the automated analysis and classification of objects on images, particularly the discrimination of stars from galaxies on optical band photographic plates and CCD images. Each object is characterized by a number of properties (e.g., moments of its spatial distribution, surface brightness, total brightness, concentration, asymmetry), which are then passed through a supervised classification procedure. Methods include multivariate clustering, Bayesian decision theory, neural networks, k-means partitioning, CART (Classification and Regression Trees) and oblique decision trees, mathematical morphology and related multiresolution methods (Bijaoui et al. 1997; White 1997). Such procedures are crucial to the creation of the largest astronomical databases with 1-2 billion objects derived from digitization of all-sky photographic surveys. (Feigelson and Babu 1997: 365)

Some of the most notorious (Big Data-related) results in observational cosmology include the reports of the microwave background from the COBE, WMAP and Planck satellites, the detection of gravitational lensing, the ensamble of surveys such as Kepler, Gaia and DES, SDSS, DESI, LSST, Euclid and WFIRST, and the observation of the *Bullet Cluster*; being the latter one of the most important contributions to the cosmology of this century.

The *Bullet Cluster* consists of "two merging galaxy clusters, that the hot gas (ordinary visible matter) is slowed by the drag effect of one cluster passing through the other. The mass of the clusters, however, is not affected, indicating

[16] When working with Big Data, cosmologists are pushed to heavily rely on other astronomy-fields such as astrostadistics and astroinformatics, their methodologies and their selection of (formal and instrumental) tools which might be mostly opaque for cosmologists themselves. In addition, they have to face scientific challenges about scalability, data integration, strategies for data analysis and data visualisation. And only, once the data is processed, cosmologists can interpret it in light of their models and use the resulting interpretation either for testing or for informing the models; however, whenever they reach such a point, their procedural ignorance about the process of selecting and relating data is quite solid.

that most of the mass consists of dark matter" (Riess 2017). The Bullet Cluster was fist discovered in 1998, later on, it was registered by The Chandra x-ray observatory in 200. But it was only in 2004 when optical images of the Bullet Cluster were integrated by the Magellan telescope and the Chandra x-ray; and in 2016, it was finally possible to provide a picture of the bullet cluster which integrated optical data, X-ray data, and a reconstructed mass map, becoming one of the most famous and informative images in all of astronomy.

The high quality and the degree of accuracy of these pictures as well as the impact that they have had in the study of the universe can be interpreted as indicative of scientists' objectual knowledge of the Bullet Cluster. Thanks to the satisfactory integration of extremely large amount of heterogeneous data that was gathered for many years, cosmologists have seen the collision of two galaxy clusters with the same sharpness than we can see an apple or our own reflection in the mirror. The pictures of the Bullet Cluster constitute an extremely reliable epistemic product that, even if cosmologists cannot track back all the inferential steps that generated them, they can trust the visual results of the simulations.

The Bullet Cluster has been taken by many cosmologists as evidence in favor of the existence of dark matter and, transitively, in favor of the model of cosmology- ΛCDM (*Lambda-Cold Dark Matter*) (Cf. Lage and Farrar 2015).[17] Nonetheless, despite its, alleged, strong evidential role in favor of ΛCDM, the images of the Bullet Cluster only suffice for proving the objectual knowledge that cosmologists have gained about this particular phenomenon. But considering the cosmologists' procedural ignorance with regard to the processes through which the gathered data from the Bullet Cluster was actually cleaned and integrated, they are still ignorant of some of the facts that might have been lost in the data analysis but that seem to be needed for grounding the truth of the model.

The above considered, some cosmologists have interpreted the existence of the Bullet Cluster as a challenge to modify alternative models in order to give account of this phenomenon without accepting the existence of dark matter; a good example of this are refined versions of the *Modified Newtonian dynamics* (MOND) −which is a hypothesis that proposes a modification of Newton's laws to account for observed properties of galaxies and constitutes an alternative to the hypothesis of dark matter.

The fact that objectual knowledge has been gained with respect of the Bullet Cluster has not been enough for neither having achieved explanatory knowledge about what causes this phenomenon nor for deciding the truth value of the dark matter hypotheses or the Modified Newtonian dynamics. However, what has being gained is the opportunity of incorporating the factuality of the Bullet Cluster into possible theoretical structures, for instance, ΛCDM-universe (Cf. Kraljic and Sarkar 2014) to explore the logical space of such universes; this is, the ignorance of theoretical structure that undermined the cosmologists epistemic practices has been partially overcome. Providing cosmologists with modal

[17]This model is a parametrization of the Big Bang cosmological model according to which the universe is composed mainly of three elements: a cosmological constant denoted by Λ and associated with dark energy; the postulated cold dark matter (CDM); and ordinary matter.

understanding of how could a ΛCDM-world be, which inferences would be allowed within its limits, how could information be transmitted and what could be inferred from the existence of the Bullet Cluster −even if ignoring whether the model is empirically adequate.

I take this brief case study to have shown that when empirical scientists incorporate the use of Big Data into their practices, they gain epistemic access to objects that were initially thought to be inaccessible. Yet, at the same time, they also become ignorant of the theoretical structure that underlies the data-analysis as well as the one underlies the connections between the tested cosmological model and the observational evidence. This ignorance, however, can be easily mistaken with ignorance of the facts that were initially accessible to the scientists but got lost in the machine implementation processes as well as ignorant about the followed inferential procedures that the data-analysis programs followed.

5 Final remarks

When incorporating Big Data into the empirical sciences, scientists are able to reach objects that were initially inaccessible to them, as well as to provide accurate measurements, descriptions and predictions with regard to such objects. However, as the outcomes of Big Data applications often come in the shape of patterns and correlations, the price that scientists pay for this access is the loss of causal explanations. This leaves scientists having to face a particular type of ignorance: ignorance of theoretical structure *with reliable consequences*. They ignore the structural particularities of how these correlations are identified and how these objects are constructed, but, at the same time, they have evidence in favor of the reliability of the mechanisms that underlie the identification of the patterns and objects. Such reliability has made possible that, contrary to what traditional literature on scientific understanding would suggest, scientists working with Big Data could achieve modal understanding of the domain that they are studying.

References

1. Baumberger, C. (2011): "Understanding and its Relation to Knowledge", in C. Jäger and W. Löffler (eds.) *Epistemology: Contexts, Values, Disagreement. Papers of the 34th International Wittgenstein Symposium*, Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society: 16–18.

2. Baumberger, C., C. Beisbart and G. Brun (2017): "What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science" in *Explaining Understanding: New Perspectives from Epistemologyand Philosophy of Science*, Routledge: 1-33.

3. Baumberger, C. and G. Brun (2017): "Dimensions of Objectual Understanding" in *Explaining Understanding: New Perspectives from Episte-*

mology and Philosophy of Science, Routledge: 76-91.

4. Bird, A. (2010): "The epistemology of science—a bird's-eye view", *Synthese*, 175(1), 5-16.

5. Barberousse A. and M. Vorms (2014): "About the warrants of computer-based empirical knowledge", *Synthese* 191(15): 3595–3620.

6. Brescia, M., Cavuoti, S., Amaro, V., Riccio, G., Angora, G., Vellucci, C., and Longo, G. (2017): "Data Deluge in Astrophysics: Photometric Redshifts as a Template Use Case". In *International Conference on Data Analytics and Management in Data Intensive Domains*, Springer, Cham: 61-72.

7. Chen, Y., Argentinis E. and G. Weber (2016): "IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research", *Clinical Therapeutics* 38 (4): 688-701.

8. De Regt, H. W. (2009): "Understanding and Scientic Explanation" in S*cientific Understanding: Philosophical Perspectives*, H. W. de Regt, S. Leonelli, and K Eigner(eds.), University of Pittsburgh Press: 21–42.

9. De Regt, H. W. (2015): "Scientic Understanding: Truth or Dare?", *Synthese* 192: 3781–97.

10. De Regt, H. W., and D. Dieks. (2005): "A Contextual Approach to Scientific Understanding", *Synthese* 144: 137–70.

11. De Regt, H. W. and V. Gijsbers (2017): "How False Theories Can Yield Genuine Understanding" in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, Routledge: 50-75.

12. Elgin, C. Z. (2004): "True Enough", *Philosophical Issues* 14: 113-131.

13. Elgin, C. Z. (2009): "Exemplification, Idealization, and Understanding", in Mauricio Suárez (ed.) *Fictions in Science: Essays on Idealization and Modeling*, Routledge: 77-90.

14. Elgin, C. Z. (2011): "Making Manifest: Exemplification in the sciences and the arts" *Principia* 15: 399-413.

15. Elgin, C. Z. (2017): "Exemplification in Understanding", in E*xplaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, Routledge: 76-91.

16. Feigelson, E.D. and G.J. Babu (1997): "Stadistical methodology for large astronomical surveys", in McLean, B., D. A. Golombek, J. J. E. Hayes, and H. E. Payne (eds.) *New Horizons from Multi-Wavelength Sky Surveys*: 363-370.

17. Floridi, L. (2011): *The Philosophy of Information*, Oxford UK: Oxford University Press.

18. Floridi, L. (2012): "Big Data and their epistemological challenge", *Philosophy & Technology* 25(4): 435–437.

19. Floridi, L., Fresco N. and G. Primiero (2015): "On malfunctioning software", *Synthese* 192(4): 1199–1220.

20. Fricke, M. (2015): "Big Data and its epistemology", *Journal of the Association for Information Science and Technology* 66(4): 651–661.

21. Fulton, B. J. and E.A.Petigura (2018): "The California-Kepler Survey. VII. Precise Planet Radii Leveraging Gaia DR2 Reveal the Stellar Mass Dependence of the Planet Radius Gap", *The Astronomical Journal*, 156 (6): 1-13.

22. Garofalo, M., A. Botta and G. Ventre (2016): "Astrophysics and Big Data: Challenges, Methods, and Tools", *Astroinformatics (AstroInfo16) Proceedings IAU Symposium No. 325*: 1-4.

23. Grimm, S. R. (2006): "Is understanding a species of knowledge?", *British Journal for the Philosophy of Science*, 57(3): 515–535.

24. Grimm, S. R. (2014): "Understanding as knowledge of causes", in A. Fairweather (Ed.), *Virtue epistemology naturalized*. New York, NY: Synthese Library: 329–345.

25. Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615-626.

26. Ivezić, Z., A. J. Connolly, J. T. VanderPlas and A. Gray (2019): *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (Updated Edition), Princeton University Press.

27. Khalifa, K. (2013): "Is understanding explanatory or objectual?", *Synthese*, 190(6), 1153–1171.

28. Kelp, C. (2014): "Knowledge, understanding and virtue", in A. Fairweather (Ed.), *Virtue epistemology naturalized*. New York, NY: Synthese Library. 366: 347–360

29. Kraljic, D. and S. Sarkar (2014): "How rare is the Bullet Cluster (in a ΛCDM universe)?", *Journal of Cosmology and Astroparticle Physics*

30. Kvanvig, J. (2003): *The value of knowledge and the pursuit of understanding*. Cambridge, UK: Cambridge University Press.

31. Lage, C., and G.R. Farrar (2015): "The bullet cluster is not a cosmological anomaly", *Journal of Cosmology and Astroparticle Physics*, 2015(2).

32. Lawler, I. (2016): "Reductionism about understanding why", *Proceedings of the Aristotelian Society*, 116(2): 229–236.

33. Lawler, I. (2018): "Understanding why, knowing why, and cognitive achievements", *Synthese*.

34. Le Bihan, S. (2017): "Enlightening Falsehoods: A Modal View of Scientific Understanding" in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, Routledge: 111-136.

35. Le Morvan, P. and R. Peels (2016): "The Nature of Ignorance: Two Views", in Peels, R. and M. Blaauw (eds.) *The Epistemic Dimensions of Ignorance*, Cambridge University Press: 12-32.

36. Leclercq, F., A. Pisani and B. Wandelt (2014): Cosmology: From theory to data, from data to theory.

37. Leonelli, S. (2012): "When humans are the exception: Cross-species databases at the interface of biological and clinical research", *Social Studies of Science*, 42(2): 214–236.

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

39. Leonelli, S. (2020): "Scientific Research and Big Data", in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* , `https://plato.stanford.edu/archives/sum2020/entries/science-big-data/`

40. Manyika, J., Chui, M., Brown, B., *et al.* (2011): "Big Data: The Next Frontier for Innovation, Competition, and Productivity". McKinsey Global Institute.

41. Martínez-Ordaz, M. del R. (2020): "The ignorance behind inconsistency toleration", S.I. Knowing the Unknown, *Synthese*.

42. Morrison M. (2015): *Reconstructing Reality: Models, Mathematics, and Simulation*, New York, NY: Oxford University Press.

43. Napoletani, D., M. Panza and D. C. Struppa (2014): "Is Big Data Enough? A Reflection on the Changing Role of Mathematics in Applications", *Notices of the American Mathematical Society*, 61(5): 485–490.

44. Norton, J. (2008): "Ignorance and Indifference", *Philosophy of Science* 75: 45–68.

45. Riess, A. (2017): "Dark matter", in *Encyclopædia Britannica*, Encyclopædia Britannica, inc.

    `https://www.britannica.com/science/dark-matter`

46. Sterner, B. (2014): "The Practical Value of Biological Information for Research", *Philosophy of Science*, 81(2): 175–194.

47. Sterner, B. and N. M. Franz (2017): "Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data", *Biological Theory*, 12(2): 99–111.

48. Swan, M.(2015): "Philosophy of Big Data: Expanding the Human-Data Relation with Big Data Science Services," 2015 *IEEE First International Conference on Big Data Computing Service and Applications*: 468-477.

49. Symons, J., and Alvarado, R. (2016): "Can we trust Big Data? Applying philosophy of science to software", *Big Data & Society*, 3(2), 2053951716664747.

50. Wheeler, G. (2016): "Machine Epistemology and Big Data", in McIntyre, L. and A. Rosenberg (Eds.): *The Routledge Companion to Philosophy of Social Science*. Routledge: 321-329.

51. Toumani, F. (2014): "When data management meets cosmology and astrophysics: some lessons learned from the Petasky project" (talk presented at *Journées de l'interdisciplinarité*).

52. Zhang, Y and Zhao, Y (2015): "Astronomy in the Big Data Era". *Data Science Journal*, 14(11): 1–9.