Is there anything special about the ignorance involved in big data practices?

María del Rosario Martínez-Ordaz
martinezordazm@gmail.com

Forthcoming in Lundgren, Björn, L. and N. Nuñez-Hernández (Eds.) Philosophy of Computing, Philosophical Studies Series, Vol. 143.

Abstract

It is a fact that the larger the amount of defective (vague, partial, conflicting, inconsistent) information is, the more challenges scientists face when working with it. Here, I address the question of whether there is anything special about the ignorance involved in big data practices. I submit that the ignorance that emerges when using big data in the empirical sciences is *ignorance of theoretical structure with reliable consequences* and I explain how this ignorance relates to different epistemic achievements such as knowledge and understanding. I illustrate this with a case study from observational cosmology.

Keywords— Epistemology of big data, ignorance, ignorance of theoretical structure, epistemic opacity, modal understanding, Bullet Cluster.

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 a long time, cosmology had been regarded to be very different from other empirical disciplines. Despite its successful predictions and observational discoveries, cosmology 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 their models more accurate and grants cosmologists access to novel phenomena.[1] But, cosmology

[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

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 to geology, climatology, biology, and other areas of scientific enquiry that had a long history of working with large datasets.

Unfortunately, and despite its positive outcomes, the incorporation of big data into scientific practice has come with some problems. Associated to the increase in the amount of data there is a significant increase in the scientists' ignorance regarding the ways in which such data hangs together. For example, observational cosmology has made much progress accessing phenomena that were initially considered to be unreachable for us, like galaxies that are million light years away, and that now can be "photographed" by us. However, much of this observational success depends on computational processes that cannot be fully scrutinized, examined and justified by human agents (see Humphreys 2009). This is, we could look at pictures of two galaxies colliding and rely on them as visual representations of the actual phenomena, yet we might not be able to rationally justify such a reliance.[2] This lack of epistemic access to the ways in which the received data holds together when generating certain outputs, such as pictures cosmological phenomena, is called *ignorance of theoretical structure* and it limits the understanding of the inference patterns that hold within a set (or a collection of sets) of data (Cf. Martínez-Ordaz 2020).

Here, I address the question of whether there is anything special about the ignorance involved in big data practices. I submit that the ignorance that emerges when using big data in the empirical sciences is *ignorance of theoretical structure with reliable consequences* and I explain how this ignorance relates to different epistemic achievements such as knowledge and understanding. I illustrate this with a case study from observational cosmology.

While philosophy of science has already started discussing the different epistemic challenges and ethical consequences of 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.

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. 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 in Sec. 5, I illustrate this with a case study from observational cosmology. Sec. 6 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]These problems in observational cosmology are approached again and in more detail in Sec. 5.

2 Epistemological worries about big data

In this section, I discuss the most distinctive epistemological worries associated with the use of big data in the empirical sciences; I divide these worries into two main categories: the methodological and the understanding-directed. 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 two concerns about the scope of these worries. Finally, Sec. 2.4. the main concerns related to understanding.

2.1 Preliminaries

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 turning data into something useful).[3]

From the outset I want to be clear about the main purpose of the paper. From now on, I only focus on the epistemic challenges that individual agents face when working with big data in the sciences. My aim in the rest of this section is to show that big data practices have introduced important challenges to the scientific activity −leaving 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.

2.2 The methodological worries

Traditionally, scientific knowledge has been regarded as hierarchical, explanatory and at least partly unified. First, according "to hierarchical models of

[3]I am fully aware of the fact that there is an ongoing philosophical debate about the status of the different characterizations of big data, however, I believe this will suffice for the purposes of the paper. If interested in comprehensive philosophical analyses of the inferential mechanisms used when faced with immensely large amounts of data, see: [Floridi 2011] and [Floridi 2019], and for introductory discussions regarding the epistemology of big data see [Floridi 2012] and [Leonelli 2014].

science, our scientific knowledge (...) forms a knowledge system that has two properties: (i) it is stratified, and (ii) the items of some layer are or should be justified in terms of items of a higher layer" (Batens 1991: 1999). Second, the demand for explanatory power has at least two sources: pragmatic strand in terms of the power to predict and manipulate reality and a epistemic in terms of understanding what reality is like. Pragmatism and manipulability require simplicity and optimization whereas understanding reality through science requires accurate representations. These aims can conflict, but whenever they go together happily our explanatory ambitions tend to be satisfied. Third, the hierarchical spirit of scientific knowledge aided by its explanatory character enable scientists to look for unification, at least, in particular domains. When different theories satisfactorily explain the same system at different levels, the explanations that they provide are compatible, interconnected and mutually reinforcing.[4]

But big data has affected information gathering processes making them more comprehensive and faster than ever before. The incorporation of big data into the empirical sciences has modified the ways in which scientific knowledge was traditionally pursued and scientific methodology followed. The three main methodological worries associated to big data practices that have caught the philosophers' attentions are: the lack of clarity about purposes and uses of data, the reliance on correlations, and the epistemic opacity that surrounds the results of big data (Cf. Humphreys 2009, Floridi 2012, Leonelli 2014).

The first two worries come from analyzing the actual novelty of big data practices. First, 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 the 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)

The first methodological worry concerns (1) and the need for determining the

[4]However, hierarchical models face important difficulties, like lacking stable justificatory mechanisms that can avoid infinite regress or the absence of (robust) relations that can explain how to increase the order of our knowledge system. This considered, contextual models tend to be more satisfactory in both respects specially when explaining the ways in which scientists rationally deal with incomplete, incompatible and even inconsistent information in their day-to-day practice (Cf. Batens 1991). This has had an important effect on the unificatory character of science, as contextual approaches seem to be the only effective resource to address scientific practice, unification should remain only as a regulatory ideal. Despite the weakening of the hierarchical and unificatory nature of scientific knowledge, it is still expected for science to be mainly an explanation-seeking activity.

purposes and uses of data. Nowadays, the key epistemological problem for the use of big data in the sciences is to identify which questions are interesting, or even essential, to answer at a certain moment, as well as the production and selection of the relevant answers (Cf. Floridi 2012).

Moreover, 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. Correlations being "the statistical relationship between two data values, are notoriously useful as heuristic devices within the sciences" (Leonelli 2014: 3). For a long time, they 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. Nowadays, correlations are 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). While part of our current scientific practices are explanatory, there is another significant part that takes correlations to be keystones for scientific development. Thus, big data practices have shown that the explanatory character of science is not as strong as the traditional view had suggested, and 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.

The second methodological worry is to determine under which circumstances can scientists rationally trust correlations as legitimate instances of scientific knowledge. This concern is motivated from two sides. First, correlations leave unexplained why and how something is the case, and therefore, it seems hard to trust them as grounds of any epistemic enterprise. Second, given the large amount of information involved in big data practices, it is inevitable that it exceeds our cognitive capacities. Since scientists cannot ever process such quantities of data by themselves, whenever wanting to access and manage this data, they require technological implementation; making any epistemic product linked to those datasets, necessarily constrained by specific technological resources. This is, the rational trust of correlations requires a previous defense of the rational trust on the technological resources involved in their discovery.

The third worry concerns the scientists' trust on the products of big data despite their lack of epistemic access to the way in which such products are achieved. Due to our physical limitations, the only way in which scientists can approach certain phenomena is through technological implementation. But, once the data is gathered, due to their cognitive constraints, scientists must now rely on computer resources that store, filter, classify and structure the data in the shortest possible time. The combination of these factors causes that the observational reports depend on not necessarily interconnected layers of technological implementation. And despite them being hardly scrutinizable by humans, these technological resources are indispensable for approaching phenomena that were initially inaccessible to us and for structuring data that we would have never been able to compute by ourselves. Such a lack of scrutiniz-

ability associated to big data processes has often been explained as a case of *epistemic opacity*.

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: 618). 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 *strongly* opaque to human agents. The third worry captures two main cases of epistemic opacity:

- **Opacity regarding the status of the products:** Nowadays, 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: Chap. 7). 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.[5]

- **Opacity regarding the procedures:** In big data practices, "no human can examine and justify every element of the computational processes that produce the output of a computer simulation or other artifacts of computational science" (Humphreys 2009: 618).[6] While the patterns that emerge through these computational processes are often necessary for the scientific enterprise, they are obtained in such unique ways that traditional human modeling techniques would not be able to generate (Cf. Bedau 1997). This has as a consequence that some of the procedures that underlie the filtering, the selection and the leading to specific outputs become not-reproducible by, and even inaccessible to, human agents.

  Some of the elements that remain strongly opaque to us include privileged inference patterns that have been produced via machine learning algorithms and that are now significantly distant from human programmers' initial inputs. The fact that scientists might not be able to scrutinize all the steps through which some outputs were obtained, leaves them lacking inferential explanations for such outputs.

The second and the third worries are interconnected in the following way: while big data practices concern primarily the identification of new correlations, those

[5] The reader might object that it is not clear if the status of the products of these computer-based methods is epistemically opaque, or if it is simply the case the their philosophical significance is not completely understood. However, while the question of whether a model counts as an experiment or an abstraction seems to be more philosophically oriented −rather than focused on our epistemic access to the world−, the epistemic opacity described here concerns only the status of the outputs of these computer-based methods, and not the methodology in itself. This is, this epistemic opacity concerns the question of whether the measurements, the descriptions and the visualizations of the data should count as observational reports, or only as theoretical expectations within a specific model.

[6] Here I am concerned with processes associated to the realization of algorithms in code as well as to the ways in which programs are actually run in particular instances. If interested in the conditions under which these processes can be made transparent see [Creel 2020].

supposed correlations rest on a multitude of sources −where there is often opacity of the workings of computer systems and reasoning, or research and observational techniques. There is a lot that is not known but that is trusted in big data practices; and therefore, the need for explaining how this ignorance does not affect the scientists' rationality.[7]

2.3 But, why big data?

At this point, the reader might wonder whether the second and, specially, the third worries actually address consequences of big data −and not common phenomena in any scientific discipline. Following this intuition, one might consider that phenomena such as reliance on technology are not immediately connected to the use of large datasets, and therefore do not need to be approached by studying the epistemic practices of big data.

This concern is properly grounded: nowadays, different types of computer simulations are key resources in astrophysics similarly as they are in genomics or in fluid dynamics, they are handy when using immensely large datasets but also when working with ordinary-size sets of information (Cf. Morrison 2015: Chap. 6 and 7). Even more problematically, some procedures of computer simulations are not strongly opaque to us, because they are only used to make mathematical operations go smoother but not to do work that exceeds human capacities. However, while there are elements that traverse scientific practice regardless the amount of data that scientists deal with, the main difference −datawise− that exists between big data practices and other scientific practices is that the former constitute limit cases of the amount of data that is processed for scientific purposes.

While many epistemic practices in science might not require to implement methodological networks of high performance computing and deep learning algorithms, it is a fact that any research that aims at working with immensely large datasets would need to do so. And when this happens, at least, some crucial parts of the processes will remain opaque for the scientists. What is characteristic of big data practices is that such an epistemic opacity, necessarily, would surround at least some of the main products and procedures of these practices, and despite this, some of them will have surprisingly novel and seemingly reliable outputs. These outputs would often be of the form of reliable correlations that are susceptible to ground part of the future research.

Furthermore, the number of scientific applications of big data has increased substantially in the last decade, and it is expected to only keep growing in the following years; thus, study of the epistemic successes and disadvantages of big data practices can only shed light on the grounds of our current, and future, science. In addition, if big data practices generate limit cases of epistemic opacity −mostly due to the amount of computational challenges that they deal with−, the ways in which this opacity is sorted out might be illuminating of ways in which similar problems can be tackled within similar ordinary-sized

[7] I am indebted to a referee who helped to give a better phrasing of my ideas on this point.

data practices.

2.4 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).[8] As a matter of fact, for achieving understanding, "the ability 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). In what follows, I present three worries that concern the achievement of scientific understanding in big data practices.

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.[9]

The 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. While this worry is clearly close to the ones presented in Sec. 2.2, it takes the problem a bit further and consider those cases in which a strong epistemic opacity does not prevent the achievement of knowledge, but conflicts with the pursuit of understanding.

A third worry concerns the quality of the data that inform the agents' beliefs. Scientific data is, and has been, often *defective* (vague, partial, conflicting

[8] 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.

[9] 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 interested 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].

or even inconsistent). 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 to 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 data, the satisfaction of this factive condition does not seem so straight forward, and therefore, neither does understanding.

I take this section to have shown that, when science incorporates big data to its epistemic practices there are, at least, six important epistemological worries to address. The next section is devoted to explain how these worries have a common ground: ignorance.

3 Types of ignorance

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 six worries associated with big data epistemic practices. While these worries might seem very different from each other, they have a common ground: ignorance.

Assume for a moment the intuitive characterization of ignorance as *lack of knowledge about something*. 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. The trust of correlations when also aspiring to explanation reveals the previous acknowledgment of ignorance of a causal link. 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.

As epistemic opacity indicates ignorance, when it conflicts with understanding, it can be said that because there is still a blank that should 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.

3.2 Ignorance(s)

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* (Cf. Martínez-Ordaz 2020). Orthogonally, one can also recognize (v) *erotetic ignorance* −absence of answers to questions.

Let's look at these types of ignorance by paying special attention to corresponding challenges that they (might) impose to the scientific activity:[10]

(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 ⌜The speed of light, in vacuum, is 299792458 metres per second⌝. 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 or S' cognitive limitations prevent her from knowing a particular fact.

[10]Because 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.

This type of ignorance conflicts with scientific reasoning by limiting the application of certain inferential rules. For example, if S fails at assigning an alethic value to p, S will not be able to detach the consequent of every conditional of the form $p \rightarrow q$.

(ii) **Objectual ignorance:** this ignorance requires absence of knowledge of a particular object. The main characteristic of this ignorance is that one ignores a whole set of properties that an object possesses and that are regarded to be indicative of such an object.

This ignorance conflicts with scientific activity by troubling preventing agents to connect lists of properties to a particular object. 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 a common object, preventing scientists from identifying (new) 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[11] Most of the time, an agent is considered to be procedurally ignorant when she cannot neither explain nor perform a specific task.

This ignorance conflicts with scientific practice especially in experimental contexts. For example, 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)

Ignorance of theoretical structure is often the cause of persistent instances of any of the other types of ignorance. Lacking access to a relevant part

[11] According to some epistemologists, this type of ignorance also resembles factual ignorance; Specifically, it can be translated into ignoring lists of causal relations, this is, not knowing what has to be done to obtain certain outcome (Cf. Williamson 2001, Snowdon 2004).

of the structural conditions of a theory prevents scientists from either inferring the value of certain proposition (causing factual ignorance), identifying whether distinct sequences of properties refer to the same object (causing objectual ignorance) or explaining inferential procedures (causing procedural knowledge).

The partial overcoming of this type of ignorance within a specific set of data, consists in identifying ways to inferentially secure particular regions of the logical space associated to such a set.[12] Determining ways in which reliable outputs are obtained and logical harm is avoided.[13]

Going back to the connection between ignorance and big data, one should still wonder to which extend the analysis of ignorance would be revealing of the epistemic grounds of big data practices. This, in light of the fact that as humans are epistemically limited, 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? With this in mind, in the next section I discuss the ways in which ignorance challenges scientific rationality in big data contexts.

4 Big data, big Ignorance?

Here I claim that the ignorance that underlies big data practices is, often, ignorance of theoretical structure *with reliable consequences*, and that this ignorance does not prevent scientific understanding from being achieved.

The section is divided in four parts: Sec. 4.1. I briefly acknowledge the importance of identifying the ignorance that is involved in big data practices and the ways to overcome it. Sec. 4.2. addresses the type of ignorance that underlies the big data practices and Sec. 4.3. sketches the type of understanding that is achievable through these practices.

[12] *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). It is important to remind that as theoretical structure is a dynamic entity, changing as a theory evolves, new findings can lead to changes in how inferences are made. For example, new findings in methods of approximation greatly affect the inferences made in many sciences.

[13] Faced with $p \rightarrow q$ and $\neg p \rightarrow q$, if S is ignorant of the truth of p (and its negation), many would be happy to detach q; however, when being ignorant of the theoretical structure that relates ps, $\neg q$s and qs this would not be necessarily possible. Ignoring the theoretical structure that relates a set of data means ignoring how negation works within that particular set, what can be inferred, what is not a consequence of the dataset, among others.

But once this ignorance is partially overcome, propositions will have a specific value in a world like the one described by a specific theory (or model); and this does not necessarily extend to the actual world.

4.1 The landscape

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 −that 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 data comes with the price of different degrees of ignorance (Cf. 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.[14]

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

In big data practices, 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 as facilitating 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). This resulted in aiding the identification and scrutiny of newly detected objects. For instance, the fact that the NASA Chandra X-ray Observatory constantly receives, stores, filters, classifies and integrates enormously large amounts of optical data and X-ray data, among others, has enabled the detection, and the later scrutiny, of the so called *Bullet Cluster*, a phenomenon that was expected to occur but which detection would have been impossible without the help of observatories that do not only receive data but also process it (see Sec. 5.1.). These practices have allowed scientists to acquire knowledge regarding the objects that were initially inaccessible; this is, to attain objectual knowledge.

[14] While there is not a uniform view on what *scientific rationality* 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 experimentation, 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.

However, as the selection of data sometimes depends on epistemically opaque processes, scientists end up ignoring how certain outputs were obtained as well as other possible outputs that were disregarded by the algorithms −meaning that there are going to be some important bits of information about phenomena that will remain ignored by the scientists despite having being initially captured. At this point, it only seems fair to say that what scientists ignore is the way in which particular sets of data hang together in order to entail certain outputs, this is, they ignore the relevant part of the theoretical structure that glues the received data and the products of computational processes.

In addition to reliance on technology, there is another element in data-driven sciences that should also be considered as causing epistemic opacity about processes and products, this element is the *increasing collaboration*.

Big data practices possess a massively cooperative nature which makes the transmission and acquisition of knowledge very opaque as well. Very often scientific communities rely on the quality of the datasets that were initially processed and now shared by other communities, making these practices based on a new type of epistemic trust. What is being at stake here is a reliance not on the individuals that integrate the communities, but on their technological choices and the procedures that such choices imply −regardless of how opaque they are for the individuals who have chosen and employ them. The result is that, if what is transmitted and acquired is a type of knowledge, it is not of the kind of knowledge by (expert) testimony (Cf. Sullivan 2019).[15]

In big data practices, the source of knowledge is not always an individual that can provide better explanations to support her claims if asked to do so; it is often a combination of methodologies plus machine implementation over inputs that come from very diverse sources in very different formats, and which interconnections are not always clear to us. In the long run, this has the effect of scientists being unable to provide explanations about procedures that might have lead to the discovery of novel phenomena. Yet, should this be understood as a case of procedural ignorance? not necessarily. When they cannot provide inferential explanations about why an output obtains, they are not ignoring only a specific recipe, they are ignorant of how the bits of data relate to one another −at least, inferentially; and this is indicative of ignorance of theoretical structure.

Consequently, when working with big data, scientists are trading knowledge of some parts of theoretical structure in exchange for access to inaccessible objects. As a matter of fact, the incorporation of big data to the empirical sciences has created a new epistemic preference: "answers are found through a process of

[15]This, especially when adopting a standpoint similar to the so-called assurance view of testimony, according to which "testimony is restricted to speech acts that come with the speaker's assurance that the statement is true, constituting an invitation for the hearer to trust the speaker. Such views highlight the intention of the speaker and the normative character of testimony where we rebuke the testifier in the instance of false testimony (Tollefsen 2009)" (Sullivan 2019: 21). Because testimony is often perceived as a highly intentional speech act that assumes that the expert can offer explanations to back up her assertions, it is not clear how technological implementations could provide us with something like that; in particular, in cases in which there is a strong epistemic opacity surrounding the outputs of such implementations.

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]).

Nonetheless, if the ignorance that underlies big data practices is ignorance of theoretical structure, one should not overlook the fact that, in the corresponding literature, this ignorance is described as the main source of negative epistemic outputs such as resilient anomalies, mutually conflicting inferential products, among others (Cf. Martínez-Ordaz 2020). In light of its negative impact on the pursuit of knowledge and understanding, one should worry that this ignorance causes a larger epistemic harm in big data contexts. In particular, by preventing scientists from determining the inference patterns that govern specific datasets and the selection mechanisms for inferential products from these sets, ignorance of theoretical structure might put in danger the epistemological basis of big data practices.

The challenge seems more complicated when, scientists might not be able to satisfactorily get rid of this ignorance due to the combination of (a) the fact that big data is often used to study phenomena that is hard to verify or intervene without heavy technological implementation, and (b) the strong presence of different types of epistemic opacity and diverse instance of epistemic trust. (a) and (b) might just be enough to undermine the epistemological basis of big data practices in the empirical sciences. This concern can be formulated in the following way:

1. If one cannot explain how certain (heavily) mediated observational reports are generated, one must weaken one's belief of them being evidence of something being the case.

2. In empirical data-driven sciences, certain outputs of big data procedures are expected to count as observational evidence of something occurring in a particular way.

3. But, when present, ignorance of theoretical structure messes up with the scientists' capability to explain how these outputs are generated; without necessarily affecting the scientists' capability to explain why a particular phenomenon would occur in a specific way.

In light of the above, scientists might have to chose between either rejecting these outputs as evidence about a specific phenomenon or, at the risk of being irrational, trusting products that they ignore where they come from and how they were obtained, and use them for testing empirical hypothesis.

The first option might look appealing in cases in which there are alternative methods for gathering evidence about phenomena which do not require the use of big data and opaque computational methods. However, if these phenomena are essentially inaccessible to us without heavy technological implementation, to take the first option would mean to lose our epistemic access to entire empirical domains. This makes the second option more appealing, but also rises the

question of how to make outputs of processes that are opaque to us more reliable when using them as evidence in the empirical sciences.

This concern has not been overlooked by the scientists, as a matter of fact, they have constantly sought for methodologies that allow them to preserve and justify the reliability of the data −regardless if they are ignorant of the processes that generated 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 include accurate predictions, measurements and descriptions. This considered, there is the need to explain the continued trust that scientists posit on big data practices despite the epistemic opacity that surrounds them.

While scientists might be ignorant of the way in which data hangs together in order to generate certain outputs, some of the results that are reached through different computational processes would be extremely novel and accurate. And is in light of such successful results that scientists are justified in trusting the processes that generated them −this justification is of a *reliabilist* nature. Scientists trust big data practices in a similar way than they trust their vision, mainly because they can recognize that the outputs of both big data procedures and processes carried out by the visual system produce more successful than ineffective consequences.

But, in big data contexts in the empirical sciences, what counts as a *successful* consequence? A successful consequence of a computational process would be an output (prediction, description, representation, etc.) that grants access to empirical phenomena −specially if that phenomena that wouldn't be accessible to humans without the aid of big data and computational processes−, and enhances the achievement of objectual knowledge regarding such phenomena. In addition, this output should be:

- novel in its field,
- empirically adequate, [16]
- fruitful − the output seems to be crucial for the development of related research programs, and

[16] Even if ignoring the status of the output; this is, not knowing if it should be accepted as a substitute for empirical experiment or a theoretical abstraction of a specific empirical domain.

- the output holds a possible evidential relation with a model or theory within the discipline.[17]

Note that, when the output concerns an empirical domain that is physically inaccessible to us, the empirical adequacy of such an output might be hard to evaluate observationally; therefore, the success of such a result can be graded considering both its connections with accepted models or theories, and its impact on measuring, predicting and explaining other empirical phenomena. But what is it about the volume, velocity, variety, etc. – what is it about big data's features? While in Sec. 2, I emphasised the epistemically negative consequences of these features and their connection with proprietary and opaque software −causing different types of epistemic opacity; here, I want to draw the reader's attention to the benefits of these features for the novelty of the outputs of big data practices.

Gathering largely immense amounts of data of different kinds at an extremely high speed, makes the resulting sets of data, when finally integrated, extremely informative. For instance, the representations that are obtained through big data are not only representations that, computationally and observably, we could not have constructed ourselves alone; but they are also exceptionally comprehensive representations of highly complex phenomena. The fine grained detail that is possible to recognize in big data products constitutes one of the most important parts of their novelty, and this is mostly caused by the fact that the data that is received comes not only in different formats but also refers, in great detail, to the different layers of the studied phenomenon. So, when this data is satisfactorily integrated and structured −even if the integration process is extremely opaque to us−, the result will be a highly detailed map of the phenomenon, in which scientists would be able to zoom in and zoom out, in such a way that will enhance their grasping of the phenomenon at different levels and in different scenarios.

The combination of all the above gives the impression of, while the ignorance that underlies big data practices in the empirical sciences is ignorance of theoretical structure, its main characteristic is that it possesses (significantly) reliable consequences.

4.3 Understanding big data

According to what has being discussed in previous sections, big data practices have granted scientists access to new phenomena, and have provided them with the opportunity of accurately identifying, measuring and predicting their behavior. In particular, the use of big data has allowed empirical scientists to achieve objectual knowledge of things that, for centuries, were considered to be

[17] I am fully aware of the fact that there are many ongoing debates regarding the sufficient (and necessary) conditions for the evaluation of scientific success in big data practices. And, for that reason, the criteria listed in this section are not expected to be taken as neither universal nor definitive, but only intuitive enough to consider them as indicative of success when satisfied.

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 all we can get from the implementation of big data practices?

In what follows, I argue that there is a way to interpret the epistemic profits of big data practices as a keystone for the achievement of scientific understanding. In particular, the type of understanding that can be gained through these practices is *modal understanding*.

First, 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 the data points out 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 grasping of specific theoretical structures −both in terms of inferences and properties, and in terms of experimental and operational procedures for its use.

Second, it might happen that scientists are able to, from big data analysis, obtain important information regarding an object. 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). That 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.

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 an 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 substantiates. An important remark is that the understanding that is obtained is understanding of the relations that hold within (a segment of) a theoretical structure given the presence of a newly discovered object −not to be confused with dependence relations between objects in the actual world.

However, at this point, the reader might wonder whether modal understanding does not assume the previous achievement of knowledge of theoretical struc-

ture −which we are supposedly ignorant of in these cases. Such knowledge could be expressed in the form of laws, grounding relations, or any kind of dependence relation within the structure. This considered, it is not clear how big data can aid the scientists to produce such modal understanding, or to even draw out a possibility space without directly positing causal relations and laws, among others. It seems that if there are correlations, even if they be reliable, either no possibility space can be generated solely from them or the possibility space that is obtained is caused by our previous knowledge of the relevant part of the theoretical structure. And therefore, big data practices have no impact on the achievement of modal understanding.

While the reader's concern might sound appealing, it is grounded on a misunderstanding about of the scope of both knowledge of theoretical structure and modal understanding. On the one hand, due to our cognitive limitations, it seems implausible that we can achieve full knowledge of the theoretical structure of a specific set of data −whether it is a theory, a model, or only a set of collected information about a specific phenomenon. But the same happens with ignorance, no scientist working within a specific field will be absolutely ignorant of the theoretical structures of the sets of data that she is working with. Knowing parts of such structure and satisfactorily drawing some inferences when using these sets, is compatible with ignoring other parts of the structure and failing at identifying which inferences are correct within it. So, yes, the scientist that achieves modal understanding can rely on both her previous knowledge of certain segments of the structure and the outputs of big data processes. These two elements can constrain of the possibility space, helping the scientist to identify the specific inference patterns that might govern the newly identified objects. On the other hand, modal understanding does not require the possibility space to be constrained by any type of metaphysical assumptions. In particular, while the possibility space that is being understood could reveal dependence relations of the form of grounding relations, in the large majority of cases, the relations that enhance this understanding are only inferential −making the possibility space, logical space.

Big data processes integrate immensely large amounts of data in such a way that they enhance extremely comprehensive representations of new objects; this grants scientists with objectual knowledge of the things that have been newly identified. After the acquisition of such knowledge, and thanks to the comprehensiveness of these representations, scientists are able to incorporate the factuality of these objects into theoretical structures that could explain why they occur the way they do. The comprehensiveness of these representations helps to narrow down the set of alternative structures that scientists can navigate, at least, at an inferential level; gaining understanding of the specific possible worlds that are compatible with what we now know about these new objects −without necessarily recognizing whether any of the alternative structures is isomorphic with a specific chunk of the world. This is, while scientists might ignore the theoretical structure of immensely large datasets and the ways in which certain outputs are produced within the sets, they can still gain understanding of the worlds in which, at least, the most salient outputs of big data processes are true.

In sum, despite the presence of ignorance of theoretical structure in big data practices, there are two main epistemic products that are obtained through them: objectual knowledge regarding objects that were initially unreachable to us, and modal understanding of how such newly identified objects fit in theoretical descriptions of the world. The conquest of modal understanding, allows the scientists to have a clear picture of the set of possible worlds that correspond to the structural connections that are relevant only with respect to some domain of the possibility space associated with the phenomena in question. In the next section, I illustrate this with a case study from cosmology.

5 Cosmology and big data

This section offers a case study from observational cosmology that illustrates the role that ignorance of theoretical structure plays in big data practices. Furthermore, this case brings the attention on three main phenomena: the successful consequences of computational processes, the acquisition of objectual knowledge and furthermore, the achievement of (some) modal understanding.

5.1 The story

Observational cosmology aims at providing precise agreement between large scale-physical theories, cosmological models and observation. On a daily basis, high-throughput detectors and (land and space based) telescopes generate terabytes of raw data about objects in specific regions of the sky. The data that is gathered comes in heterogeneous formats, once received, it has to go through different isolating, filtering and integrating processes; and only after the data is processed, merely a fraction of it is saved and put to the service of cosmologists.

In light of its observational character, the reduction of data into images constitute one of the best received outputs of technological implementation in cosmology.

> 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)

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. Once this is done, the constructed images should be reduced into catalogues. Some of the most successful results of these processes 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*, officially named 1E 0657-558, is one of the most energetic known galaxy clusters in the universe (Cf. Schramm 2017: 13). The 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 that most of the mass consists of dark matter" (Riess 2017). The Bullet Cluster was first discovered in 1998, later on, it was registered by The Chandra x-ray observatory in 2004. And it was only in 2004 when optical images of the Bullet Cluster were integrated by the Magellan telescope and the Chandra x-ray (Cf. Markevitch *et al.* 2004; Clowe, Gonzalez and Markevich 2004).

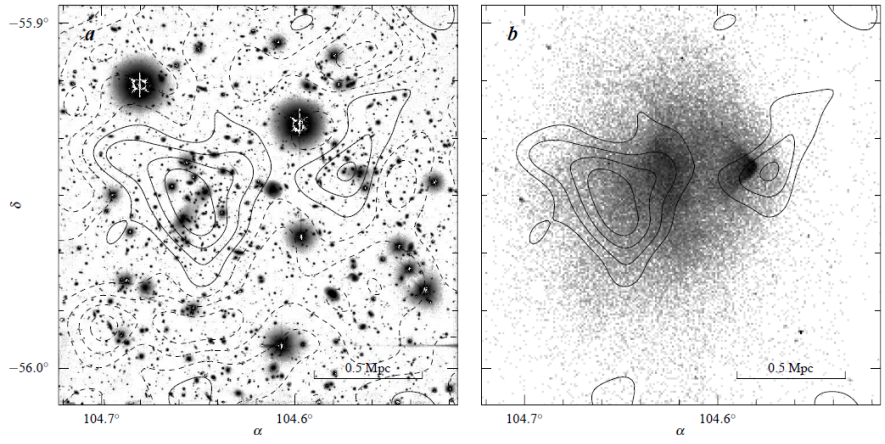

Image 1. First visual representation of the Bullet Cluster 2004 (gray-scale I-band VLT image). From Markevitch *et al.* (2004): 820.

In the following years, it was possible to provide a picture of the bullet cluster which comprehensively integrated optical data, X-ray data, and a reconstructed mass map, becoming one of the most famous and informative images in all of astronomy.

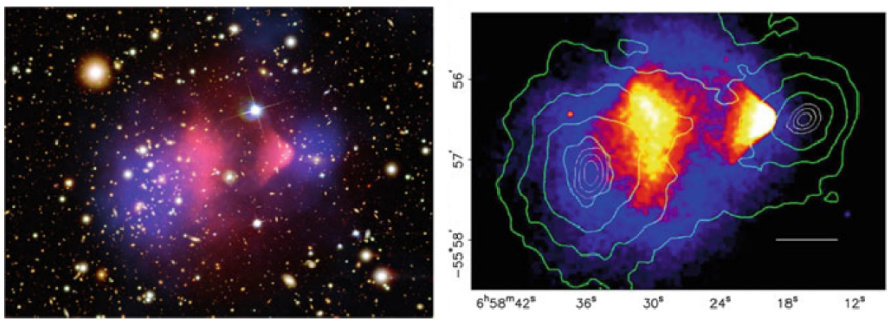

Image 2. Visual representations of the Bullet Cluster 2006.
Left: From [Clowe, D.*et al.* 2006], *Right*: From [NASA Chandra X-ray Observatory 2006]

5.2 Evaluating the case study

This case study illustrates two main things: (1) the ignorance that underlay the epistemic practices in the discovery of the Bullet Cluster is ignorance of theoretical structure with *reliable consequences*. And, (2) despite their ignorance, scientists were able to achieve objectual ignorance and modal understanding of the world(s) in which the Bullet Cluster could exists.

First of all, as it has been argued in Sec. 4.2., due to their methodological basis, many big data practices are underlain by ignorance of theoretical structure, specially the practices in which different instances of epistemic opacity are combined. The epistemic practices associated to the use of big data in observational cosmology are not an exception to this general claim. As a matter of fact, due to the nature of the discipline's main object of study, these practices are the result of the union of reliance on technology and increasing collaboration. The former is clearly linked to the common use of different layers of machine implementation and computational processes −aided by deep learning algorithms− which scope exceeds programmers' original input and human's cognitive capacities. The latter, is mostly connected to the way in which raw data is filtered, structure and reduced into catalogues by one community, and later on reused by another community. The choices that the first community made are taken by the second to be at their best correct and at their worst, at least, non-problematic; but, considering that those initial choices were constrained by ignorance makes the foundations of this trust shaky.

Yet, this ignorance does not necessarily undermine the scientists' rationality when relying on big data practices and some of their *successful* outputs; in this particular case, the combined models that enhance (visual) representations of the Bullet Cluster. Intuitively, the criteria that when satisfied, could be indicative of an output being successful includes evidence of the way in which such output grants scientists access to empirical phenomena while enhancing the acquisition of objectual knowledge and modal understanding of such phenomena.

In addition, the output should be novel, empirically adequate, fruitful, and there should exist a possible evidential relation between the output and a model or theory within the discipline.

In this respect, the study of the cluster produced by the NASA Chandra X-ray Observatory has provided

> sufficient precision to determine the mass distribution of the underlying galaxies through weak gravitational lensing (...) they made four main observations relating to the mass distributions:
>
> 1. Due to the large distance scales in question, stellar matter was only moderately affected. For the most part, the stars from each galaxy simply passed through the other galaxy without any inelastic interactions. The only visible effect is a velocity reduction through gravitational forces, with the occasional inelastic meeting of stars.
> 2. As usual, the gaseous component of the galaxies is much more spread out. The meeting of two gas clouds results in a significant interaction under the electromagnetic force, due to the shorter length scales (...)
> 3. The centre of total mass of the galaxies, observed through weak gravitational lensing, is offset from the stellar and gaseous matter. This suggests the presence of additional invisible matter.
> 4. The dark matter distributions can be inferred from the total mass contours, and they remain mostly spherical in shape. (Schramm 2017: 13,14)

(1)-(3) indicate (purely) observational discoveries that were achieved only thanks to the observation of the bullet cluster −regardless how heavily meditated this observation was. As a matter of fact, the comprehensiveness, the accuracy and fine granularity of the visual representations of the Bullet Cluster −and the models that ground them−, combined with the impact that they have had in the study of the universe in general, are indicative of the scientists' acquisition of objectual knowledge of, at least, the Bullet Cluster. This reinforces the intuition that, even if cosmologists cannot have sufficient epistemic access to the inferential processes that generated these images, they can consider the outputs of such processes to be evidence of something occurring in a particular way.

Another sign of success is the existence of a possible evidencial relation between the results of big data computational processes and a conjecture, a model or a theory within the discipline. (4) can be interpreted as indicative of the fulfillment of this condition. 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).[18]

[18]This model is a parametrization of the Big Bang cosmological model according to which

But the Bullet Cluster does not only hold a possible evidencial relation with ΛCDM, 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.

At this point, there is a weak form of underdetermination of theory by data surrounding the Bullet Cluster: As with the mass discrepancies in galactic structures (that were originally explained by the dark matter hypothesis but were eventually explained by MOND), it has been proved that the Bullet Cluster can be explained by both ΛCDM and MOND (Cf. Angus *et al.* 2006).[19] But, while this underdetemination is problematic for explaining the phenomenon in itself, it reinforces the idea of the description and representations of the Bullet Cluster being sufficiently empirically adequate to count as observational evidence that *must* be explained. And at the same time, the fact that nowadays, the discovery and observation of the Bullet Cluster is driving the modifications of ΛCDM and MOND is indicative of its fruitfulness within the discipline.

Finally, the fact that objectual knowledge has been gained with respect to the Bullet Cluster is not enough for neither achieving explanatory knowledge of what causes this phenomenon nor for deciding the truth value of the dark matter hypotheses or the Modified Newtonian dynamics. Yet, what has being gained through the observation of the Bullet Cluster is the opportunity of incorporating its factuality into alternative theoretical structures that could explain why it occurs the way it does and when doing so, exploring the logical space described by such structures.

Thanks to the observation of the Bullet Cluster, cosmologists have been able to develop modal understanding of what the Bullet Cluster might be and how it would behave in, at least, both a ΛCDM-constrained world (Cf. Kraljic and Sarkar 2014) and a MOND-constrained world (Cf. Angus *et al.* 2006). While, at least, part of the computational processes associated to the gathering, filtering and structuring of the data might remain opaque to the cosmologists, they have now been equipped with fine grain detailed representations of the Bullet Cluster and its behaviour in different contexts. The value of this structured information is that it provides the scientists with a possibility space constrained by the existence of the Bullet Cluster, which they can navigate in either a ΛCDM-direction or a MOND-direction. This is, while cosmologists might remain ignorant of the theoretical structure that underlies the set of raw data about the Bullet Cluster, they now have access to the inference patterns that the representation of the Bullet Cluster allows for −without necessarily knowing which of these patterns, if any, is satisfied in the actual world.

---

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.

[19] I am greatly indebted to an anonymous referee for pointing me to this problem.

6 Final remarks

When incorporating big data into the empirical sciences, scientists are able to reach objects that were initially inaccessible to them. However, the outcomes of big data applications often involve high degrees of epistemic opacity about how such outcomes were generated. This leaves scientists having to choose between rejecting these outputs as observational evidence or, at the risk of being irrational, relying on them −even if ignoring where they come from and how they were obtained.

Here I argued that the ignorance associated to the epistemic opacities found in big data practices is ignorance of theoretical structure *with reliable consequences*. Scientists might ignore the structural particularities of how the observational outputs are identified and generated, but, at the same time, they have evidence in favor of the reliability of these outputs −and therefore, of the processes that generated them. Such reliability has made possible that scientists achieve objectual knowledge of initially inaccessible objects as well as modal understanding of how these objects (could) behave and relate to one another, all this while being ignorant of the inference patterns that govern the datasets from which the access to these objects is constructed.

Acknowledgments

I am indebted to Moisés Macías-Bustos for his extremely valuable and constant feedback on different versions of this paper. Thanks also to Atocha Aliseda-Llera, Otávio Bueno, Gabrielle Ramos-García, Felipe Rocha, Alexandre Meyer and Carlos César-Jiménez for fruitful discussions on these issues. Thanks to Paweł Pawłowski for his challenging questions on this project. I owe special thanks to the four anonymous referees for their valuable comments and suggestions. Thanks to the audiences at the *IACAP2019* and the *Colóquio Virtual SELF - Problemas Filosóficos*.

References

1. Angus, G., Famaey, B., and H. Zhao (2006): "Can MOND take a bullet? Analytical comparisons of three versions of MOND beyond spherical symmetry", *Monthly Notices of the Royal Astronomical Society* 371(1): 138-146.

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

3. Bedau, M. (1997): "Weak emergence", *Philosophical Perspectives*, 11, 375–399.

4. Batens, D. (1991): "Do We Need a Hierarchical Model of Science?" in Earman (ed) *Inference, Explanation and Other Frustrations. Essays in the*

Philosophy of Science, Berkeley-Los Angeles-Oxford University of California Press: 199 – 215.

5. 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.

6. 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.

7. Baumberger, C. and G. Brun (2017): "Dimensions of Objectual Understanding" in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, Routledge: 76-91.

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

9. Boone, K., G. Aldering, Y. Copin, S. Dixon, R. S. Domagalski, E. Gangler, E. Pecontal and S. Perlmutter (2018): "Binary Offset Effect in CCD Readout and Its Impact on Astronomical Data", *Publications of the Astronomical Society of the Pacific*, Vol. 130, No. 988 (June):1-16.

10. 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.

11. 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.

12. Clowe, D., M. Bradač, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones and D. Zaritsky (2006): "A Direct Empirical Proof of the Existence of Dark Matter", *The Astrophysical Journal*, The American Astronomical Society, Vol. 648 (2): L109-L113. `https://doi.org/10.1086/508162`.

13. Creel, K.A. (2020): "Transparency in Complex Computational Systems", *Philosophy of Science*, 87 (4).

14. 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.

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

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

17. 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.

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

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

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

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

22. 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.

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

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

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

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

27. 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.

28. 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.

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

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

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

32. 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.

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

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

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

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

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

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

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

40. 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.

41. 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.

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

43. 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.

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

45. 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/

46. Markevitch1, M., A. H. Gonzalez, D. Clowe, A. Vikhlinin, W. Forman, C. Jones, S. Murray, and W. Tucker (2004): "Direct Constraints on the Dark Matter Self-Interaction Cross Section from the Merging Galaxy Cluster 1E 0657–56", *The Astrophysical Journal*, The American Astronomical Society, Vol. 606 (2): 819-824. https://doi.org/10.1086/383178

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

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

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

50. NASA/CXC/SAO (2006): X-ray: NASA/CXC/CfA/ M. Markevitch *et al.*; Optical: NASA/ STScI; Magellan/ U.Arizona/ D.Clowe et al.; Lensing Map: NASA/STScI; ESO WFI; Magellan/U.Arizona/D.Clowe et al.

    https://chandra.harvard.edu/photo/2006/1e0657/

51. Napoletani, D., Panza M. 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.

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

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

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

54. Schramm, S. (2017): *Searching for Dark Matter with the ATLAS Detector*, Springer Theses, Springer.

55. Sullivan, E. (2019): "Beyond Testimony: When Online Information Sharing is not Testifying", *Social Epistemology Review and Reply Collective* 8 (10): 20-24.

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

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

58. 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.

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

60. 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.

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

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