# Learning More and Knowing Less:

# Big Data, Spurious Correlations, and the Problem of Ignorance

by

George Stroubakis

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Arts

in

Philosophy

Waterloo, Ontario, Canada, 2022

## Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

# Abstract

There are few things we can name that surpass big data's boon for our empirical endeavours. Big data promises an ever-growing source of data that could be readily accessed by researchers in virtually any field of study. Data, as we know, are the rudiments of our knowledge claims about the world. Big data multiplied the number of data collected and stored to unfathomable numbers. But it is well known that our tools that help us sort through the datasets bring up spurious correlations. This project explores the implications of spurious correlations on our epistemic endeavours through the lens of ignorance. I will draw on recent scholarship in agnotology, the study of ignorance, namely by Nancy Tuana, Robert Proctor, and Nicholas Rescher.

I will argue that spurious correlations, especially those that are less obvious to the epistemic agent can mislead the latter in forging false beliefs. In effect, big data, through the presence of spurious correlations can create ignorance. I will also argue that the epistemic agent shares in the responsibility of promoting and disseminating the ignorance that can arise from too much data. The promise of big data to provide the building blocks upon which only knowledge can stand is mistaken, for it also provides the building blocks for a new type of ignorance that is not an absence of knowledge, but the presence of something masquerading as it.

## Acknowledgements

I owe my deep gratitude to Professor Patricia Marino for her invaluable advice that will serve me for a lifetime or thinking and writing about philosophy, and her patience in supervising this thesis. I am also most grateful to Professors Doreen Fraser and Carla Fehr, not only for being my readers and providing valuable feedback, but also for their guidance throughout the last few years to help me fulfill the lifelong dream of completing my graduate studies.

# Dedication

I dedicate this to my children, Zaynab and Ilyas, who are a source of immeasurable joy and pride. They presence in my life motivates me to always strive in becoming a better person.

# Table of Contents

# Introduction

Recent technological advances have made recording and storing data in vast databases incredibly accessible, giving us the ability to carry our data along everywhere we go and access it wherever we are. Big data is the collection and electronic storage of vast and complex datasets. With roots in the scientific revolution when the foundations of modern statistical methodology were being laid, we now have more data than we have ever seen before in the history of record-keeping (Holmes, 2017, p. 22). The size of our datasets is so great that we need tools to help us sort through it. As a result, we deploy new algorithmic techniques to help us extract useful information from them (ibid.). But the results have often been surprising. The more our tools are refined and, in our datasets, get bigger, we see a staggering emergence of puzzling correlations.

When scientists and natural philosophers during the Scientific Revolution set out to gather more data about phenomena under study, they would have had little reason to foresee the potential problems that spurious correlations could pose for our epistemic projects, since the need to gather more information is an obvious step in creating more knowledge. This same assumption motivated technologists of recent times to create products that could store even more data and deploy algorithms to help us sort through the ever-growing datasets. Although the presence of spurious correlations is known, proponents of big data are largely under the impression that we can rely on big data for our inquiries without much thought given to the disagreeable problem of spurious correlations, which have the potential to distract from our epistemic objectives. In fact, it is more than just a question of distraction—they also have the ability to mislead the inquirer into thinking that what they have discovered in the data is a correlation that represents something true about the world, when in fact it does not.

The aim of the current project is to expose how spurious correlations that emerge in big data have an adverse effect on our epistemic aims, by creating ignorance rather than knowledge. With this in mind, one wonders then how one can make sense of these correlations in view of what we already know about the world. This poses a problem with our epistemic aims because, on the one hand, we increasingly rely on data to understand the world and to get more information about it and, on the other hand, because spurious correlations emerge (as studies included in this work will show) infinitely. It seems, therefore, that big data, a boon for all the areas of knowledge, may just be delivering more than we expected.

First, we need to examine what has been said about ignorance in recent scholarship to understand ignorance emerges in other contexts before we examine how it can emerge in the context of big data. Chapter 1 will provide background in the study of ignorance, also referred to as agnotology. This will help us understand what kind of ignorance emerges from big data and where it fits within the broader context of agnotology. Thus, we will benefit from the analyses of Nancy Tuana, Robert Proctor and Nicholas Rescher, whose work is seminal in recent agnotology, in defining the key components of my claim.

Chapter 2 will show how big data does not necessarily and inevitably always lead to more knowledge. Rather, there are good chances that the opposite can happen, by leaving the inquirer misled and thus more ignorant. In this chapter, I will define what a spurious correlation is and what is its opposite. To do this, I will present two examples of spurious correlations, one obvious and one less so, in order to evaluate them and their inability to represent something about the world. We will also see what effects spurious correlations have, namely, of misleading the knower. As mentioned, I argue that misleading the inquirer results in the latter becoming more ignorant than

knowledgeable. Misleading is a type of ignorance because the knower is misled to think that something is knowledge when it really is ignorance that is masquerading as knowledge.

Then, I will make a first attempt to define what kind of ignorance emerges from big data and how this type of ignorance differs from the general understanding we have about the term ignorance (i.e., the absence of knowledge). While Proctor and Tuana point out that ignorance is at times useful or even necessary, ignorance is antithetical to the direct aims of big data. The essential point made here is that while the ignorance that is dealt with in the current literature is defined as a neglect or an absence of knowledge, the ignorance I will expose is the kind of ignorance that masquerades as knowledge.

In Chapter 3, I will further develop the new kind of ignorance that I have identified and tried to compare it and position it within the categories proposed by Tuana and Proctor. While their categories broadly align, the new kind of ignorance that I identify in this work differs in key ways that I will exposit in this chapter. Finally, I argue that possessing ignorance, neglecting to eliminate it, or choosing to maintain it are blameworthy in varying degrees. Knowing that ignorance can emerge is not enough as knowers we ought to know how to deal with it. While I will not provide prescriptions of how to deal with it, I will use Rescher's criteria of culpability, namely, reasonable effort and ample opportunity to eliminate ignorance, to evaluate whether big data ignorance is culpable, especially taken in view by how widely big data is used and how easily the ignorance that arises from it can be disseminated.

My conclusion will not present a solution for data scientists because they are best equipped to deal with the technical issues of the discipline, although I point out some possible avenues of inquiry for the more philosophical issues. More than anything, this project is a word of caution. I would consider it a failure on my part if my reader thought that I was opposed to big data. Quite

the contrary, the boon for our empirical methods is clear. But we should also not let our guard down when confronted with unexpected correlations, for although they can either point to new and surprising discoveries, they are far more capable of turning out to be spurious, which can lead the inquirer down the wrong path of inquiry, wasting valuable time and resources, but also potentially disseminating ignorance rather than knowledge. Because I believe that all philosophical writing should end on positive notes, I will propose some potential avenues of investigation that may leave researchers and inquirers able to avoid big data's traps confidently and cautiously.

The present work is one of epistemology. The main subject is the inquirer or the knower and both are used interchangeably. There is a subtle difference between them: the inquirer denotes activity, as in what one does when actively inquiring and researching, whereas the knower is a state of being, which will be employed in more broader contexts.

# Chapter 1 - Introducing Agnotology

Since the Scientific Revolution, recording observations has been seen as central to doing science. The imperative of gathering empirical data served to support and evaluate the plethora of emerging hypotheses. In the 21st century, we developed technologies to help us gather what would have been unimaginable sets of data, not only about natural phenomena but also about facts that relate to our personal lives, economic performance, and human and animal movements across the globe. Big data is a recent, yet powerful way in which we gather, record, and store data about our observations and interactions with the world. The industry behind gathering and storing data seeks to further refine technology in order to achieve even greater accuracy and precision of data. Accumulating over the last couple of decades, we are now sitting on a deeply rich source of information, that continues to grow at a pace with which we can hardly keep up. Common wisdom tells us that although we concede not having access to all the possible data in the world, having access to more data is better than having access to less. But to deal with this rich source of digitized information, we had to invent a new science to help us sort through it. But with every new scientific endeavour, we confront problems. One such problem that confronts data scientists is distinguishing useful from useless data. Useful data may be strong evidence for or against a claim about the world. They inform us about the object of study, its properties and may even allow us to establish either causal connections or lawlike predictions. Useless data, in contrast, offer none of these benefits and are considered simply background noise. Hence the expression garbage in garbage out. Therefore, while solutions for these problems may fall under the purview of data scientists and statisticians, other problems, however, are of a philosophical nature.

When we look for correlations in big data, we expect to function as evidence of causality or lawhood. One can call these correlations fruitful although this term is vague. But for every

fruitful correlation, there are more that are misleading because they do not provide the same benefits: causal connections or law-like behaviour. Confronting spurious correlations in big data is common and a hindrance to our endeavours in gleaning knowledge from that technological medium. From this point of view, big data may not be contributing to knowledge as much as we had hoped it would. To fully appreciate the claim that more data can lead to ignorance, we need to review what has been said about ignorance by scholars in recent times.

Epistemology is the study of our theories of knowledge, while agnotology (from the Greek *agnoia*) is a study of ignorance. The term agnotology was proposed by Robert Proctor [insert reference], although other alternative names have been proposed in the past. The entry in the Cambridge Dictionary of Philosophy on agnotology describes the latter as an "increasingly growing area in epistemology," implying a growing recent interest in the subdiscipline. However, no one will dispute the claim that the inquiry into ignorance has deep roots, having been discussed by philosophers both ancient and medieval.

Early medieval sources in the Western and Islamic traditions dealt with the question of ignorance as either the absence of knowledge or, a competing view that seems to imply that ignorance could also be a thing in itself (al-Tawḥīdī et al., 2021, p. 51-52). In other words, ignorance can be a thing which can coexist alongside knowledge, not only when knowledge is not existent. In the following chapter, we will delve deeper into the definition of ignorance that arises from big data, which aligns well with the latter view articulated by medieval philosophers. It is insufficient to discuss what kind of thing ignorance is without also understanding the different types that exist and when they occur. In this latter regard, the study of ignorance has been taken up by scholars working in different disciplines and subdisciplines.

Nancy Tuana approaches agnotology through the lens of feminist epistemology. Robert Proctor is a historian of science who has long worked on ignorance. Finally, Nicholas Rescher is a pragmatist philosopher with many contributions in logic and epistemology and has written extensively of the subject of ignorance. Tuana and Proctor, respectively, offer four and three categories of ignorance that are useful in helping us determine where ignorance that emerges from big data belongs in the spectrum of ignorance. Rescher's analysis of culpable versus non-culpable ignorance, as we'll discuss in more depth in Chapter 3, is useful in determining if and when ignorance resulting from too much data is considered blameworthy.

Feminist philosopher Nancy Tuana presents a helpful set of distinctions between the kinds of ignorance that arise, dividing them into four domains. The first she puts forward is "knowing that we do not know, yet do not care to know." She relates this category of ignorance to the historical study of the women's health movement (WHM) (Tuana, 2006, p. 2). Prior to the WHM, when scientists conducted research, the latter judged what was "worthy" of their attention according to the values of the scientific community at that time (ibid., p. 4). When it is left up to the scientific community to judge whether a topic of research is worthy of being pursued or not, what is not being pursued remains underdeveloped (under theorized and under experimented). That kind of neglected knowledge may be left to languish in the dark until there is a concerted effort to shed light on it. This is what happened with the WHM. The movement called for public inquiries on questions such as the curious non-existence of the male birth-control pill, a call to recover an area of knowledge that remained underdeveloped. The lack of a male pill is not explained by its non-viability, instead what explains its non-existence is the lack of interest in pursuing the male birth control pill. The lack of interest pushed not only the discovery of the male pill but also the potential knowledge of the male pill outside the realm of our knowledge (ibid.).

Tuana describes the interest of the scientific community and what it chooses to exclude as knowledge worthy of pursuit, as the "*the mirror image"* values in science. Tuana further states that the "*question of whose interests are being served sheds light not only on how values impact what we know but also how they impact what we do not know and why*" (Tuana, 2006, p. 6). In other words, the values espoused by a community also defined the knowledge that was worth pursuing.

Tuana identifies a second kind of ignorance: *Not even knowing that we do not know*. This is a double ignorance, the unknown unknown, which is central to my analysis in Chapter 3, where I argue the need for adding ignorance caused by too much data within the literature of agnotology. Tuana's double ignorance is pernicious because it veils the knower from her own ignorance. Tuana argues that it emerges by being led astray by prior beliefs and adopted theories (Tuana, 2006, p. 6). Tuana further argues that adopting wrong beliefs and faulty theories will generate double ignorance. Take, for instance, Tuana's example of our knowledge of the female reproductive system and genitalia. Tuana explains that the divisions that exist in describing female genitalia are neither as precise nor as numerous as in divisions describing male genitalia (ibid.). This gap seems arbitrary at first but is a rather pointed example of what happens when a double ignorance sets in (by predominantly a male medical establishment) in studying the female reproductive system. Medical experts grew ignorant of the complex structure of female genitalia because they did "not know what they did not know" that they had become ignorant of parts of the female anatomy (ibid.). Put differently, their theories blinded them from producing pertinent knowledge. We see this second category aptly illustrated in medical theorizing. Previous theory stipulated that female orgasm was necessary to conceive a child. When that theory was later corrected, dismissing orgasm as necessary for conception, the interest in female orgasm as part of the human

reproductive process waned, and scientific research in the area dwindled. From that point forward, any developments in reproductive theory may well leave out any knowledge specific to female orgasm, its causes, and its role in the reproductive process. Consequently, what could have been discovered through medical research remained in the dark to the extent that it is no longer known that this knowledge remained unknown.

Tuana's third type of ignorance involves one privileged party preventing another less privileged party from knowing, which she describes as a "systematic cultivation" of ignorance in other groups (Tuana, 2006, p. 9). Privileged parties can come in different kinds. They can include for example, corporations protecting trade secrets "where selected groups of people are purposefully kept ignorant" (ibid.). The privileged party wants to control access to certain kinds of knowledge. More concerning are the privileged parties that also include clusters of medical practitioners who did not disseminate knowledge about birth control pills and the nefarious effects they have on women (ibid.). At a time when we were experimenting with different methods of contraception, one of the most successful has been the female birth control pill. But it was accompanied by side effects. While the impact the pill has had on women's bodies was significant, no such risks needed to be assumed by men since they could dispense with taking a hormonal contraceptive, since their female partners took them instead. When knowledge about the associated risks became known to the medical community, efforts were made to keep that knowledge under wraps. Naturally, such risks to women's health would have spurred women to call for a safer pill or demanded alternatives such as having men share the risk by taking a male hormonal contraceptive (ibid., p. 4). By keeping society and in particular women in the dark about the effects of birth control was a way of controlling access to knowledge, at the expense of women's health. The reasons that motivated the privileged group gives rise to an important philosophical

discussion, but one that is not central to the present work. What is important to know for the present discussion is that one group has the privilege to control what knowledge is to be shared or not with a group that is less privileged.

Tuana describes willful ignorance, her fourth category of ignorance, as a "*systematic process of self-deception, a willful embrace of ignorance that infects those who are in positions of privilege, an active ignoring of the oppression of others and one's role in that exploitation*." (ibid., p. 11). This kind of ignorance, Tuana argues, figures prominently in racial theory (ibid., p. 10). It is worth taking a moment to highlight that while Proctor talked about agnotology, Tuana draws from Charles Mills' epistemology of ignorance, a term that he coined in his 1997 book, *Racial Contract* (Tuana, 2006, p. 10). Mills' term predates the coining of "agnotology" by more than 10 years. While I have no intention of conflating the two into a single subdiscipline, what I draw from both is the valuable and consistent ways in which they approach the question of ignorance creation, the central focus of the present work.

Willful ignorance is self-imposed with the aim of "ignoring … the oppression of others and one's role in that exploitation" (ibid.). A helpful formula summarizing an example of willful ignorance is offered by Spelman, whom Tuana cites. "W [White Americans] ignores g [the claim that Black America's grievances are real], avoids as much as he can think about g. He wants g to be false, ***but if he treats g as something that could be false, then he would also have to regard it as something that could be true***. Better to ignore g altogether, given the fearful consequences of its being true. Better not to have thought at all, than to have thought and lost … ignoring g, not thinking about it, allows W to stand by g's being false, to be committed to g's being false, without believing g is false." (Tuana, 2006, p. 11). The need to conceal knowledge is itself a testimony to its existence. To dismiss a piece of information when it has the potential of being true is to embrace

ignorance because it helps keep the perceived privilege intact. But the perceived privilege is an illusion, for unless the knowledge is tightly sealed, an arrangement that can quickly fall apart since the oppressed group is bound to revolt against its oppressors, the truth will eventually surface. Regrettably, the ignorance will be maintained for as long as it can be, but so is the oppression.

Nancy Tuana's four categories cover a broad range of instances of ignorance, ranging from indifferent ignorance to double ignorance, to oppressive ignorance, and finally willful ignorance. Her analysis offers potential candidates for what kind of ignorance big data produces; a topic discussed later. It is easy to see, from these examples and countless others that the lack of interest in a topic and its subsequent exclusion from research programs, how ignorance is created to the detriment of the knower. But by showing how something that was once known, which is now forgotten on account of our false beliefs about it, it is no wonder that these same false beliefs prevent us from discovering knowledge that we ought to have. We overlook knowledge either by a lack of interest in looking for it, or by being misled into believing that something else is that knowledge, or both simultaneously. A lengthier discussion in Chapter 3 will examine how well big data ignorance fits in Tuana's four categories.

Robert Proctor, a historian of science, edited a seminal work in agnotology produced in recent years. He contributed a chapter explaining how he arrived at the term agnotology and recounts in detail different case studies that illustrate his threefold division of ignorance. This classification differs slightly from Nancy Tuana's but both accounts align in such a way as to confirm how we can reasonably divide the concept of ignorance into different categories. Proctor builds his case by examining in thorough detail historical accounts of the way ignorance was disseminated. Featured prominently, is the example of disinformation and the sowing of doubt in

the scientific community's findings that linked tobacco consumption and lung cancer (Proctor, 2008, p. 12).

Proctor offers three different categories of ignorance: native state, the lost realm, and the strategic ploy. He defines native state ignorance as a kind of "desolate deficit caused by the naïveté of youth or the faults of improper education," or, simply put, "the place where knowledge has not yet penetrated" (ibid., p. 4). It is this kind of ignorance that prompts one to inquiry, and to disabuse oneself from lack of knowledge, fulfilling an innate need to be less foolish. In this sense, it is an absence of knowledge. While we understand that to fight ignorance in the native state, we need proper education and learning and to avoid miseducation. But although one can get educated to reduce ignorance, Proctor emphasizes that rising to the challenge to gain knowledge is not sufficient in actually gaining knowledge, because it can lead one along the path of faulty learning (ibid.). But this can still mislead the learner in thinking that they have achieved some level of knowledge when all they really have achieved is a compounded ignorance with false or misleading information. The miseducation can also occur from Proctor's second category. In some, native state ignorance is a state in which knowledge has not yet penetrated, but also the state where rather than having knowledge penetrate, we can witness an intrusion of the product of miseducation.

Proctor's second distinction of ignorance is the lost realm, that involves selective choice (ibid., p. 6). If we think of the native state as a vacuum that needs to be continually filled, we can think of the lost realm as a shifting body, propelled by selective choice, that encompasses knowledge within its perimeter and ignores whatever resides outside it. Wherever this amorphous perimeter moves, whatever it contains in its perimeter is where the domain of knowledge exists. If it moves again, parts of whatever was erstwhile considered knowledge, is now plunged into oblivion. It is lost knowledge. The lost realm's perimeter is also finite, while knowledge is

theoretically infinite (Calude & Longo, 2017, p. 8), making it inevitable that some knowledge will either remain out of reach or some will become forgotten to make room for other knowledge. It would be ideal if only confirmed knowledge remained in the realm of knowledge, and whatever is lost would include only that knowledge that is no longer valid (such as outdated theories), but that is not the point the Proctor is making. Rather the point he is making is that the lost realm will swallow up knowledge that ought to remain known.

An example of this kind of ignorance involves the treatment of indigenous medical practices during the Spanish conquest of the Americas. According to Proctor, indigenous American cultures had refined a process that allowed women to end an unwanted pregnancy (Proctor, 2008, p. 8). Upon discovery by the Spanish, this knowledge receded into the dark corners beyond the perimeter of the shifting body of knowledge, effectively eliminating it as a viable option for female colonists, who might have feared bringing a child into a very hostile world. He writes:

> *"The potato was fine, as was quinine from the bark of the Cinchona tree (for malaria), but not the means by which (white) women might have prevented conception or caused abortion. European governments were trying to grow their populations and conquer new territories, for which they needed quinine but not the peacock flower (the abortifacient described by Sibylla Maria Merian in 1710). Methods of contraception or abortion were low on the list of priorities, and the plants used for such purposes by the indigenes were simply ignored"* (ibid.).

The newly colonized lands needed all the births they could muster to compete with the indigenous population and there were surely strong religious injunctions against abortions that, if attempted, would have attracted the ire of the clergy that joined the expeditions of the Spanish conquistadors.

It would have been impractical at that time to consider any effort spent in the way of gathering this kind of knowledge because it was antithetical to the aims of the colonists. Knowledge for the sake of knowledge, and especially knowledge that can hamper social planning efforts, was not a luxury the Spanish colonists could afford.

Each era has concerns that rank above others and expending effort in gathering and maintaining knowledge that does not address their pressing concerns, no matter how benign or criminal, is impractical, a waste of time, and even dangerous. In these cases, creating ignorance is preferable, the shifting body of knowledge is adjusted to fit the needs of the regulators of knowledge of that specific time.

The third type of ignorance discussed by Proctor is the strategic ploy, the type, according to Proctor, that is the most widespread and under-theorized (ibid.). An immediate feature of the strategic ploy is that it is wholly intentional. When I discussed Tuana's third category, I mentioned that in it she included the need to intentionally keep some knowledge or information secret, hence the need for privacy laws and confidentiality clauses in contracts. The situations that are interesting to our discussion, and which Proctor elaborates, are those situations where ignorance is deliberately produced. And when that intentional ignorance fails to overshadow knowledge, disingenuous scepticism is deployed to cast doubt over the knowledge. Proctor illustrates this point by recounting the tactics used by the tobacco industry, and in particular, Philip Morris, the executive of which declared that the company's chief product was the "manufacturing of doubt" (ibid., p. 11). The strategic ploy encompasses both deliberate overshadowing of knowledge and, at the very least, scepticism about it.

A generally accepted aim of science is the notion that our scientific theories provide reliable inferences about the world. As powerful evidence about the harms of tobacco products surfaced,

it was enough to persuade consumers to give up smoking. The prospect of seeing their consumer base erode, executives of powerful tobacco companies went on the offensive, funding research projects to question and refute the findings that posed a threat to their industry. They resorted to reducing the study of epidemiology to "mere statistics," and implying that correlations emerging from research do not imply causation—that they were, in short, spurious (ibid.). Instead, they argued for leaving the question of causation (that tobacco causes disease) open, a tactic that eliminates causation (tobacco causing cancer) and leaves correlation (a weaker evidence for causality), which can be more easily attacked. By championing this approach and casting doubt on the stronger claim of causation, the scientific community, and, more importantly, the consumer, are left with relying on a weaker knowledge because the stronger claim has doubts cast upon it, thanks to the unscrupulous efforts of tobacco industry executives (ibid.). As such, the stronger and widely supported causal claim was undermined in favour of the weaker claim which can be more easily attacked. Proctor highlights that a parallel argument the pro-tobacco lobby made was that forcing the issue that tobacco causes disease is tantamount to putting an end to scientific inquiry, which is an enterprise that aims towards having more certain knowledge. An overzealous embrace, then, of causal justification may sound the death knell to the scientific enterprise by paving the way towards dogmatism (ibid., p. 12). By paying lip service to the ideals of science, pro-tobacco proponents distorted strong evidence in the service of reducing losses sustained to corporate revenues.

By sowing doubt, the intent of the tobacco companies was to leave opportunities for scepticism, a state in which one tends to suspend judgment and decision-making. The non-sceptical attitude, on the contrary, adopts a less risk-averse approach in view of the acquired knowledge. In the interest of tobacco executives to take fewer risks by dismissing claims that could bring about

catastrophic financial consequences for the tobacco industry. To mitigate the latter risk, it was imperative to mitigate the epistemic risk first. The pro-tobacco camp did this by calling for more "rigour" and "precision," stalling tactics to be sure, but also "dumping" tactics, which, Proctor reminds us, was a well-known strategy used by lawyers who wished to hamper the opposition by "dumping" paperwork to scatter and exhaust their efforts (ibid., p. 25). This same worry, as we shall see, arises with an overabundance of data. Although the tactic may not be willfully deployed, the situation in which having a dump of data can still scatter efforts by deploying spurious correlations and muddying the epistemic field.

We see an example of how one party, to protect its interests, willfully manufactures a situation to create doubt, but, more pertinent to the current discussion is that they managed to do that with calls for more information, not with less. That is to say that rather than seeking to *suppress* the knowledge that posed the greatest risk to their industry (which implies that they have something to hide), they found that an ostensibly more cooperative and thus less suspicious tactic was to gather and highlight more data pointing to other conclusions, to confuse the issue and spread scepticism.

Proctor includes a different case showing what happens when too much information is shared with the public, demonstrating that sometimes ignorance is better than "total information awareness" (ibid., p. 23). That is because sometimes the costs of knowing are much too great. He offers the example of the flood of information issued from media sources after the 9/11 attacks in the US, that exposed a myriad of "vulnerabilities" in infrastructure across the country (ibid.). Reporters and experts elaborated harrowing scenarios in the event of another attack, sending the message to the public that they were not safe, effectively sending the public into a hysteria, obsessing over security. While no one would dispute the importance of gathering this data, the

example of exposing national security vulnerabilities and the hysteria that followed is a clear example of a negative consequence in sharing too much information. But excessive information also has another, more subtle consequence. Although too much data exposing knowledge about vulnerabilities can have detrimental effects on social cohesion, I will argue that within the scope of big data, ignorance of data and of the facts is antithetical to the aims of gathering data. What we decide to do with the knowledge (the data and the facts) is an entirely separate issue altogether.

In sum, Proctor's account of ignorance in its three realms is supported in well-detailed historical case studies where each realm emerges. The native state is generally the starting point of ignorance, where one is prompted to fill it with knowledge. The lost realm sheds light on some kinds of knowledge while obscuring others based on the interests and aims of a given social group. The strategic ploy intends to subvert knowledge by deploying either disinformation or unjustified scepticism. With respect to big data, we will see that the ignorance that stems from too much information overlaps with Proctor's realms, which I will further discuss in the following chapters.

Now that we have substantive background information on what kinds of ignorance we can encounter, we will address the further question of whether it can be blameworthy to espouse ignorance or to allow oneself to remain in a state of ignorance, a discussion that will be central to the present work. The expression that comes to mind "Ignorance is bliss," implies that not having a certain kind of knowledge will exculpate one from indulging ignorantly, or, as we have seen, ignorance of certain unnerving tasks or knowledge can spare the knower of unwanted mental and psychological distress. However, Rescher's analysis gives perspective to the question of blameworthiness of the knower's ignorance.

In the Western scholastic tradition, Thomas Aquinas, and William of Ockham name two types of ignorance: vincible and invincible. The former is a type of ignorance that can be

eliminated by acquiring knowledge, and therefore it is inexcusable if left uncorrected. That is to say that vincible ignorance left unaddressed is blameworthy. Invincible ignorance is the type of ignorance that cannot be remedied—or, if it can be remedied, it can only be with great difficulty and laborious effort (Langston, 2011, p. 28). Because it would be considered too demanding of knowers to disabuse themselves of their ignorance under these conditions, Aquinas and William of Ockham deemed it non-culpable (ibid., p. 35). That is to say that the knowers could not be blamed if they were ignorant of that knowledge.

In *Ignorance: On The Wider Implications of Deficient Knowledge*, Nicholas Rescher borrows the distinctions articulated by the Medieval philosophers. He writes:

*"There are many different types of ignorance, prominently including that of the stupidity of one who cannot learn and that of the foolishness of one who will not learn. Neither of these is at issue here. For our present concern is specifically with the sort of unavoidable ignorance that besets man notwithstanding his best efforts and intentions."* (Rescher, 2009, p. 8).

Whether we refer to ignorance from incompetence, foolishness, or the kind of ignorance that persists despite our "best efforts," Rescher argues that each comes with a varying degree of culpability (ibid.). For example, incompetence is less culpable when one knows not that they do not know. If one does not know how to use a device, one cannot be faulted for being unable to wield it. However, when there is an opportunity to replace ignorance with knowledge, one should avail oneself of it. But when the opportunity does not arise, or when it does, it comes at an immense cost to the knower, then, culpability once again decreases, for it would be too high an expectation of the knower to undertake such a pursuit.

With respect to culpability, Rescher seems to be drawing from medieval terminology on errors committed from ignorance. Ignorance stemming from incompetence is vincible and

therefore ought to be remedied because the effort required is reasonable and plentiful opportunities exist. Refusing to do this is blameworthy and inexcusable. Rescher adds an additional distinction which is not widely discussed in the agnotology literature, in reference to the inexcusable kind of ignorance of someone's craft or profession (ibid., p. 34). A doctor who is ignorant of parts of her profession would be more culpable than the less qualified person who is not a doctor but for whom the knowledge is nonetheless reasonably accessible. This distinction is important because, with the advent of the information age, much of our information and knowledge is no longer exclusively codified in expensive textbooks or in the minds of professors but is widely available on the Internet. But we would be hard-pressed to expect that anyone with an Internet connection ought to have the knowledge that a doctor would have. While the opportunity to eliminate ignorance may appear to exist, it is not a simple matter to go from ignorant to a qualified physician.

Rescher distinguishes the above category of technical knowledge with common sense knowledge—available to the layperson, a distinction that reinforces how culpability can be attributed to ignorance that emerges from either category. An example he offers is that being submerged in water for a long time will cause someone to drown (ibid.). This is hardly specialized medical advice of the former kind described earlier. Furthermore, although much common knowledge is codified in bytes across the Internet, much of it is not, for much common knowledge is obvious to the knower and acquired through direct experience rather than by reading about it. In sum, common knowledge is acquired easily through everyday experiences, while highly technical knowledge is acquired through effort.

Since specialized knowledge is acquired with laborious effort, Rescher argues that the knower of this kind of knowledge is more responsible for ensuring thoroughness and rigour because one's oversight can have more severe consequences if there are fewer knowers with the

same level of knowledge around to point out and prevent the oversight. A lack of thoroughness is less excused, because of the rarity of the knower in comparison to the abundance and availability of common knowledge and the knowers of common knowledge. The ignorant doctor cannot be so easily excused because she trained in the profession and invested time and effort in becoming qualified. Therefore, it is less excusable that they have failed in it if they remain ignorant about certain types of information and knowledge that they ought to have. For example, when a physician misreads vital signs or miscalculates the dosage of a prescription is more serious than if it were left up to a non-physician. An obvious objection here arises: with the inexorable specialization in most domains and professions, could we reasonably expect that physicians keep abreast of everything related to their specific area of practice? This opens a highly controversial philosophical debate, better left for another time.

In contrast to Tuana and Proctor, Nicholas Rescher chooses to distinguish between two broad categories of knowledge rather than ignorance. Rescher juxtaposes common knowledge, on the one hand, and highly specialized knowledge on the other. But with regard to the discussion on ignorance, which is the central topic of the present work, Rescher's distinctions are helpful in determining criteria of when ignorance of common knowledge, technical knowledge or even knowledge barely possible to obtain become culpable or non-culpable, in line with the medieval tradition. Non-culpable ignorance is a result of incompetence (or "stupidity") and "is in general remediable by adequate effort" (ibid., p. 36). It deserves censure "only when it is culpably willful" (ibid.). Non-culpable (or at least excusable) ignorance is the kind of knowledge that could only be acquired with laborious effort, the kind that would be too demanding for the knower, or the kind that is inevitable or whose factuality is unavailable, like knowledge of the future, which persists notwithstanding our efforts to predict it.

In the preceding pages, I presented work done by three scholars already working in the field of agnotology. Nancy Tuana's categories and Robert Proctor's realms of ignorance, as I will show in the following chapters are helpful in determining what kind of ignorance is created by too much data. Rescher's analysis is helpful in determining how culpability applies to false beliefs that stem from ignorance masquerading as knowledge.

To sum up, Tuana's first and second categories align with Proctor's native state and lost realm (his first and second category as well). Whereas Tuana's third and fourth categories overlap with Proctor's strategic ploy, his third category. Rescher has categories of knowledge instead of ignorance (common knowledge and specialized knowledge and difficult to obtain knowledge). What is interesting from his account is the criteria of culpability of the kind of ignorance that we have.

The latest trends in big data have forced us to revise our categories and make amendments by including another kind of ignorance, which is the topic of the following chapter. Thanks to the sum of these contributions to agnotology by the scholars mentioned above, we now have sufficient background to move forward with an analysis of how big data can lead to ignorance.

In the following chapter, I will present two examples of correlations that I deem spurious, one which is obviously so, and another less. Spurious correlations arise from an excess of information processed by algorithms, which, although intended to help produce more knowledge, paradoxically, they can leave us with less. First, I will show how big data can lead to ignorance through well-camouflaged spurious correlations masquerading as knowledge after which I will consider what kind of ignorance big data can lead to.

# Chapter 2 – Spurious Correlations and Ignorance

In the previous chapter, I discussed the contributions of scholars Tuana, Proctor and Rescher to the study of ignorance and introduced my claim that too much data can lead to ignorance. The claim first appears counterintuitive. How can more data lead to ignorance and not just more knowledge? In this chapter, I will argue that big data generates a potentially infinite number of spurious correlations that can mislead the knower into thinking that the correlations have a causal connection or a lawlike explanation and predictability. First, I will define spurious correlations and provide two examples. Then I will divide spurious correlations between two further categories based on their degree of obviousness. My analysis will show that spurious correlations in big data can lead to ignorance, the kind of ignorance that masquerades as knowledge—not the kind of ignorance that is an absence of knowledge.

To support my argument that too much data can lead to ignorance, I first need to identify what about big data leads to ignorance. In collecting vast datasets, we deploy algorithms to help us sort through and make sense of them. A main objective is to detect correlations that ideally have epistemic value. But not all do. In fact, relatively few correlations have epistemic value compared to the astonishingly high occurrence of spurious correlations (Calude & Longo, 2017, p. 16). A recent study by Cristian Calude and Giuseppe Longo shows that the ratio of spurious versus fruitful correlations (ibid.) is remarkably high. Spurious correlations' preponderance, however, is only part of the worry. More importantly, they have the tendency to mislead the inquirer because they emerge from data and are identified by big data algorithms in ways indistinguishable from non-spurious correlations. The fact that they can occur more often than non-spurious correlations only exacerbates the problem.

Common synonyms of spurious are "not genuine" and "illegitimate." To eliminate any ambiguity about what I mean by spurious, I will briefly point out what is not a spurious correlation (for our purposes, a correlation is the association of two objects and events). I do not mean to say that the correlation itself is illegitimate. When a correlation emerges from the data, it is a true correlation; otherwise, it would not have been detected by the algorithm, which identifies it in the same way as it identifies a non-spurious correlation. As all correlations are true correlations, what I mean by spurious is not that they result from a mistake—at least the kind of mistake of interest to this discussion. Correlations occurring from faulty data or a bug in the software is a practical problem that computer engineers can resolve. If the algorithms are working properly, spurious correlations nonetheless occur frequently (Calude & Longo, 2017, p. 16). The correlations that I am interested in are the ones that occur even when the software and algorithms are in working order. In this sense, they are true correlations because they indeed correlate to objects or events, and they have had, in retrospect, a semblance of predictability over the time that the data was gathered, a point that I discuss below.

Some say that what makes a correlation spurious is that is its absence of a causal connection. Judea Pearl, in the Book of Why, makes a strong case that the reason we dismiss spurious correlations is because they lack causation (Pearl & Mackenzie, 2018, p. 69). I define spurious correlations more broadly. Realists about causation seek causal connections between objects or events. A correlation, once sufficiently investigated, should suggest causation between the two or suggest a third cause linking the two objects or events. Not all epistemic knowers, especially within the scientific community seek causal connections. Leaving causation aside, they prefer to look for other regularities, such as lawlike explanations, reliable predictability, and testability. But even in contexts where the inquirer is not seeking to establish causal connections,

spurious correlations still lack attributes such explanatory power, reliable predictability, and testability. That is to say that the knower is unable to propose an explanation that is plausible, and which accords with our knowledge of the world. Establishing a reliable predictability and testability (i.e., our ability to test the components of the correlation and confirm if they occur as expected) also seems unlikely. In the examples that will follow, I will address some questions that spring to mind when confronted with spurious correlations, thus showing that the latter fail to account for any of the above attributes. Therefore, having dismissed causal connection, and now having to give up a plausible explanation, prediction, and testability, what remains is that spurious correlations are merely coincidences. What we ought to retain is that spurious correlations are not only a problem for epistemic projects that assume causation; they are problematic for those projects of a more instrumentalist variety.

Spurious correlations are a known problem in humankind's inquiry of the natural world, although perhaps under different names. In logic, as spurious correlations are at the heart of an informal fallacy called *post hoc, ergo propter hoc* (after this, therefore, because of this) (Audi, 2015). The Cambridge Dictionary of Philosophy's entry describes it as the "fallacy of false cause." The fallacy is described as the *"error of arguing that because two events are correlated, especially when they vary together, the one is the cause of the other"* (ibid.). The example used to illustrate is to present the spurious correlation between the presence of storks and the number of babies born as evidence for causation, which is an erroneous conclusion. They concede, nonetheless, that "*[i]n general … correlation is good, if sometimes weak, evidence for causation.* ***The problem comes in when the evidential strength of the correlation is exaggerated as causal evidence***. *The apparent connection could be coincidental or due to other factors that have not been taken into account,*

*e.g., some third factor that causes both the events that are correlated with each other."* [emphasis added] (Audi, 2015)

Therefore, while correlations may offer evidence for causation, spurious correlations (the kinds that masquerade as epistemically valuable correlations) do not offer causal evidence. That means that they should not confuse the two kinds of correlations. To clarify, to say that an epistemic agent uncritically accepts that a spurious correlation implies causation is different from saying that a spurious correlation masquerades as non-spurious one. They are separate problems. The former involves the belief of an epistemic agent, the latter simply states a fact about a spurious correlation. I argue that the epistemic agent has the opportunity of addressing the problem of spurious correlations masquerading as fruitful ones in order to prevent the correlation in question from being used as an inference to causation.

What is illuminating in this definition is that a correlation may not be sufficient to offer a causal connection or explanatory power and that if it is exaggerated to provide causal explanation, it would be a mistake, hence a fallacy. Instead, the definition of the informal fallacy cautions us that the correlation could be coincidental or hint at a "third factor," or a common cause, that could explain why the two objects or events coincide (Audi, 2015). In this sense, the problem of spurious correlations is similar because it is likely a coincidence emerging in the data. But it differs in that given its ability to appear like an epistemically valuable correlation, it does not offer any causal relation or lawlike predictability but misleads the knower into thinking that it does. The Cambridge Dictionary of Philosophy's definition states that correlations can be good evidence for causation, but not if they are exaggerated to the status of causal evidence. My claim agrees with correlations potentially offering causal evidence, but, as we see with big data, spurious correlations can also

masquerade as epistemically valuable. Because, at best, they offer weak evidence, or, at worst, no evidence for causality.

So far, I have claimed that spurious correlations are correlations that can be statistically robust (as we will see below) and, because they truly are correlations, they a have retrodictive power (retrospective predictability), over the period the data has been gathered. It is "retrospective" because the predictive power is only noticed in after the spurious correlation is discovered, not before (I discuss predictability in more detail below). I also define spurious correlations as not having causal relations – although we could dispense with causality if we had an instrumentalist approach. We will discuss the instrumentalist approach in more detail below. But since they are virtually indistinguishable from non-spurious correlations, they can mislead the knower into thinking that they have causal properties, a misdirection that can bear negative outcomes on our epistemic practices. Without causal connection or explanatory power, spurious correlations also lack reliable predictability.

Spurious correlations lack predictive mainstay. When one identifies a correlation, it is thanks to the cooccurrence of two objects or events with remarkable frequency over a period of time. At first glance, it appears that the two objects or events might indeed have some causal connection. Assuming they do, then the one who identified it should be able to predict the next occurrence of the objects or events. However, this is an illusion. Indeed, the data shows that the occurrence of the objects or events over a period (it would have to have such cooccurrence if it is to be identified as a correlation) however the cooccurrence has only been tracked over a set period. There is no evidence that the cooccurrence preceded the period in which the data was gathered and analyzed and there is no telling how long it would last after that period. My mention of retrospective predictability highlights this feature of spurious correlations: because they lack any

causal connection, they are unlikely to be predicted in the future. Their predictive power is only noticed *after* the spurious correlation is discovered.

It is important to note spurious correlations are not the same as poorly understood correlations. Poorly understood correlations can turn out to be epistemically fruitful correlations, giving us new information about the world. When I refer to spurious correlations, my claim does not imply the possibility of their being epistemically fruitful—on the contrary, they provide no epistemic benefit at all. But my definition is more specific. What makes a (true) correlation spurious is that it has no underlying causal relation or underlying law between the two objects or events. But spurious correlations masquerade as non-spurious ones well enough to give the knower the impression that the objects or events that the correlations link are causally related.

In the example I present below, the correlations are tracked for nine years. But if the examples suggest an actual link found in nature, nine years may seem insufficient to establish if it is lawlike or causal. It is reasonable in this case to challenge the data by wanting to extend the period for much longer, to adequately dispel any doubts whether the link suggested by the correlation is one that has causal relations or has lawlike predictability. In fact, I contend that without causal or lawlike properties and sustained predictability is unlikely. All we can say about it is that it is a coincidence.

In default of these attributes—causal connection, lawlike explanatory potential, and sustained predictability and testability—spurious correlations, however remarkable their statistical significance may be, can lead the inquirer to ignorance by suggesting that what they represent is part of reality, when in fact, they do nothing of the sort. Every so often, big data will detect a correlation that has epistemic value, and which will stand in opposition to spurious correlations.

Non-spurious correlations, in contrast, have what spurious ones lack. Non-spurious correlations make causal connections and possess explanatory potential. In this sense, they are epistemically valuable, in that, exploring them further could be an opportunity to gain more knowledge and allow the inquirer to explain something about the two correlated objects or events. Spurious correlations, on the other hand, take the knower from *not knowing a particular thing* to *espousing a false belief.* There exists a parallel between being misinformed from disinformation and espousing false beliefs from spurious correlations. When one happens upon a non-spurious correlation, one can make more plausible causal connections, or at least plausibly explain why the objects or events correlate.

Non-spurious correlations, on the contrary, are established on a greater scale of predictability, capable of extending predictability further into the past, should such data exist, and beyond the time parameters of the algorithms, into the future. In other words, non-spurious correlation predictability is not bound to a specific period. If we extended the dataset in question into the past and into the future, a non-spurious correlation would have a correlation factor that is stable across time. In fact, all things being equal, non-spurious correlations that point to some kind of law can be predicted reliably all the time. If spurious correlations happen to have predictability over a period, it means that they were coincidences over that period, and ceased to be so after that period. Non-spurious correlations occur because of an underlying causal connection or law of nature that proffers an enduring predictability.

Now that we defined a spurious correlation as being devoid of any causal connection or lawlike explanation, we can therefore briefly describe it as a coincidence. Coincidences are not only hard to predict, but they are also hard to test. Testability, in turns out, also results from

sustained predictability. In the examples discussed below, I suggest possible tests to elucidate the spurious nature of the correlations.

## Examples of Spurious Correlations

Below are two examples of spurious correlations, one obvious and another less obvious, a further distinction I will discuss later. Whatever their degree of obviousness, they are nonetheless spurious correlations for the same reasons: they lack causal connections, lawlike explanation, reliable predictability, and testability. I will evaluate how each of them might mislead the knower.

The first is an example of an obvious spurious correlation that links per capita cheese consumption in the United States and the number of people who died by getting entangled in their bedsheets. We see that highest peak of the chart, around the year 2008, that the per capita cheese consumption surpasses 33 pounds, correlating with over 800 deaths by bedsheet entanglement in the same year. This correlation has a remarkable statistical significance (0.98). Retrospectively, we could have attributed an astonishing rate of prediction over a period of almost 10 years, had one suspected that a causal connection truly existed between the events. That is to say that if one had to bet on people dying by getting entangled in their bedsheets by how much cheese was being consumed, the odds would have been favourable for nearly a decade.

The example aptly illustrates how difficult it is to plausibly explain the correlation, or attribute to it some underlying law of nature. The only explanation that can be offered is nothing short of imaginary. We could not justify pursuing such a correlation, hoping to discover something about the world, because it is misleading. It misleads the knower by suggesting that there is a link between the two events when there is none. Its obviousness becomes even more striking when we attempt to follow through with what it appears to be suggesting.

Suppose that death by bedsheet entanglement increased to epidemic proportions. Would it make sense to institute a policy requiring that our per capita cheese consumption be reduced, in the aim of reducing the number of deaths by bedsheet entanglement? Enacting such a policy would imply that there is a causal relation between cheese consumption and people dying entangled in their bedsheets. But nothing suggested any such causal relationship, nor is it likely that one would have imagined that such a relationship existed, had it not been linked by algorithms sorting through vast datasets. Just because an algorithm detected a correlation, we are under no obligation to accept that there is a causal connection between the two events, especially when no other observation of the world indicates a link between the two events. In other words, even though the algorithm picked up a correlation of the two events over a period of time, nothing in human experience indicates that eating cheese can cause people to meet such a strange and untimely death.

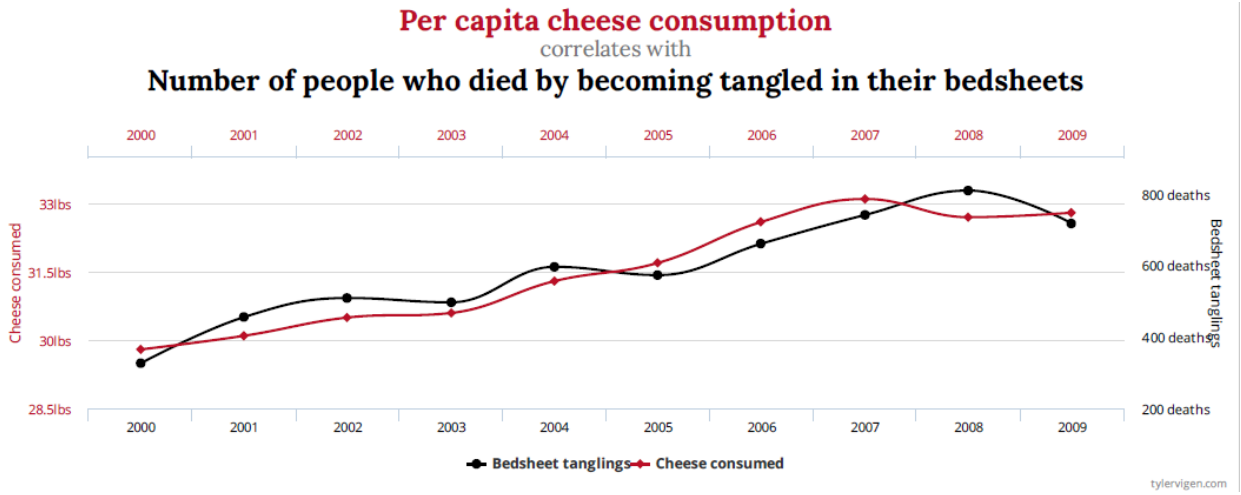


**Figure 1 (tylervigen.com)**

Next, we have a correlation that is less obviously spurious than the previous one. In this example, revenue that is generated by arcades, establishments where electronic and computer games are played, correlates with the number of computer science doctorates awarded in the

United States. In hearing about it for the first time, one can credit a flourishing interest in pursuing computer science studies to the countless hours playing computer-generated games, a mostly childhood and adolescent pastime. One could, of course, offer the explanation that since this kind of establishment is a prosperous venture, the interest in getting computer science degrees increases allowing people to capitalize on the market growth. Its statistical significance is equally high (0.94), over a period of nine years. Once again, the remarkable length of the time of this correlation retrospectively gives it the allure that there might be some underlying cause for it. But there are questions that need to be answered before we take this correlation seriously.

For example, does interest in playing games likely translate into an increase of *doctoral studies* in computer science, an endeavour that requires substantial commitment, investment of time and money, and effort to complete? If there is a market for programmers working on computer games, what are the educational and training requirements? Would not an undergraduate degree in computer science suffice? Furthermore, games are not the culmination of the work of programmers alone, but also of graphic designers and project managers who manage the development and delivery of the computer game product. Is there a similar uptick in demand for individuals holding the requisite educational degrees to perform these duties?

Accepting this correlation at face value can lead to more obviously implausible conclusions. For example, if there was a need for more computer science PhDs, one must judge whether it makes sense for a jurisdiction to sanction the opening of more arcades in every neighbourhood to increase university enrolments in that program. Such a move would be justifiably considered a preposterous solution because there is no explicit evidence of a law or causal connection other than a correlation produced by an algorithm. It is insufficient and

misleading to accept a correlation as epistemically valuable based solely on the fact that it emerged from big data.

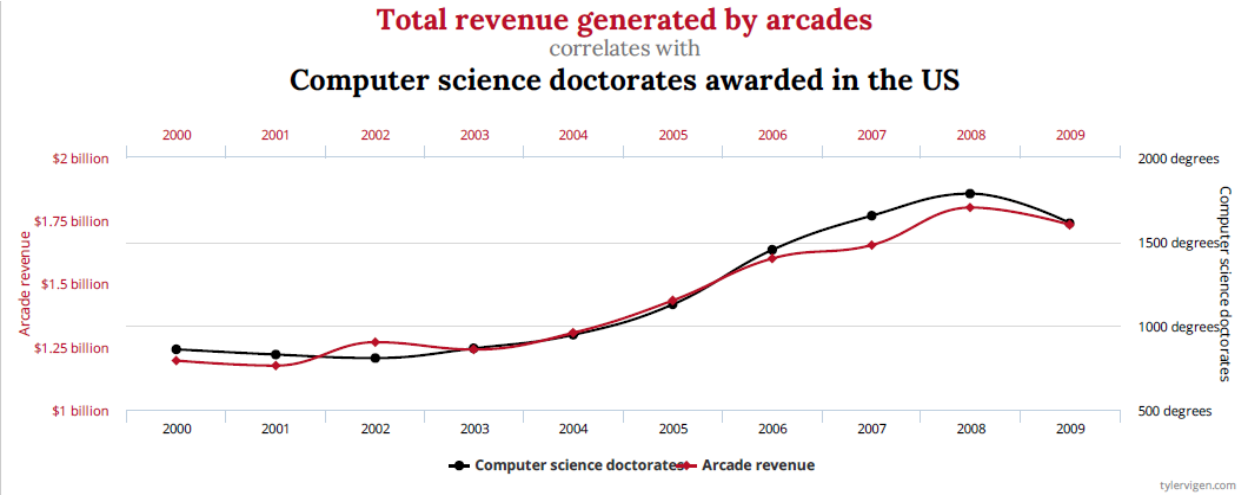


**Figure 2 (tylervigen.com)**

My hope is that I have established the preponderance of spurious correlations and the varieties in which they can appear, and the degree to which they fail to answer fundamental questions about what they tell us about the world. Scientists have been confronting this issue before the advent of big data. J. Cohen and C. Callendar propose ways in which we can identify and codify natural laws in a system. A main goal of such a system is to distinguish genuine laws from "cheap ceteris paribus [cp] generalizations." (Cohen & Callender, 2009, p. 26). Ceteris paribus generalizations are exceptions to law or generalization but are nonetheless accepted as "legitimate" (ibid. p. 25). The latter are like our spurious correlations because, like cheap generalizations, they are not "intrinsically deficient," in fact, they are "*syntactically and otherwise essentially the same as some perfectly respectable cp-laws."* That is to say that the statistical regularity, on the surface, *appears* like a "perfectly respectable cp-law," much in the same way spurious correlations bear all

the statistical resemblance of non-spurious correlations, but in reality, they are neither laws nor respectable. That is because the *"cheap [generalization] simply fails to play a role in the [system] of the field of interest*." (Cohen & Callender, 2009, p. 26). They fail to do so because they fail on other counts where robust laws do not. For example, when a law makes it into a system for thermodynamic laws *"then that means it is tied to* ***testing, prediction, explanation and all the other facets of a proper scientific enterprise***. *Cheap cp-generalizations, by contrast, don't make it in"* (ibid.). In other words, just because cheap generalizations or spurious correlations look like robust generalizations or non-spurious correlations, they are dissimilar in that they lack at least the three facts of "proper" scientific practice: testing, prediction, and explanation—all attributes that define non-spurious correlations, as described above. Similarly, spurious correlations appear to suggest that a link exists between two objects or events, and although they have a remarkable regularity over a specific period, they ultimately fail in testability, reliable prediction, and explanation.

The analogy of cheap generalizations is helpful in cementing our understanding about how spurious correlations work. Cheap generalizations lack epistemic value and once ascertained, are treated as such. They are analogous to spurious correlations emerging from big data because they lack the attributes that valuable generalizations and correlations have, and they are also able to mislead the knower into thinking that they are valuable. Scientific practice may be more adept in quickly identifying cheap generalizations, mitigating any negative influences they can bring to the epistemic aims of sciences. When it comes to spurious correlations, in non-obvious cases, detection is crucial for they can create ignorance, as I will show in the coming pages.

## Obvious vs. Non-Obvious Correlations

Obvious spurious correlations are more easily detectable because they effectively defy what we expect them to be: evidence of something causal or lawlike. They resist attempts to explain and to connect them to our knowledge of the world. Some might call them ludicrous or ridiculous. Non-obvious spurious correlations all are less resistant to our attempts to explain them and are more pliant to consider them as evidence of something causal or lawlike. While the inquirer might be tempted to admit non-obvious correlations as evidence, they require more effort in debunking them. Since they are less easily detectable, they have the power to distract or mislead the inquirer. As a result, non-obvious spurious correlations pose a greater threat than obvious spurious ones, the threat being that they can lead to ignorance if they are taken as evidence of something causal or lawlike. A well-camouflaged spurious correlation can spur new theories based on falsehood, require revisions to existing theories, amend research questions and require updates to textbooks. What we could see at play here is ignorance masquerading as knowledge, leading epistemic agents to adopt false beliefs.

Returning to our definition of spurious correlations posing difficulties for knowers who are either realists about causality or instrumentalists, we can plainly see that the questions prompted in the previous paragraph appear impossible to answer. These correlations cannot be evidence for the existence of a causal connection or assume the minimal instrumentalist requirement of going "from a given set of observations to a predicted set of observations" (Audi, 2015). That is why the definition contains either causal connection or lawlike explanation and predictability. Whatever one's philosophical commitments regarding causation, spurious correlations still lack the coveted explanatory power and reliable predictability without any reference to causality. If one with instrumentalist commitments is only interested in the ability to predict when two events coincide

without reference to causation, spurious correlations provide none of these guarantees. In other words, there is no reason to believe that two possibly disparate events that coincided over a period have any intention of coinciding for longer—nor is the dataset able to provide such a guarantee

The two examples, one obvious, one less so, illustrate how too much data, accruing exponentially are inexhaustible in their ability to produce spurious correlations, which lead to ignorance. I will explain how this happens, but first we need to briefly discuss what kind of ignorance is relevant here.

## Ignorance Not as an Absence of Knowledge

The generic definition of ignorance is a void that is filled by knowledge. Whereas knowledge is considered a thing-in-itself, ignorance—according to this definition—is the absence of knowledge. As we have seen in the recent scholarship in the previous chapter, ignorance, we could argue confidently, can be itself a thing. We recall that in Proctor's strategic ploy, the tobacco industry deployed doubt to undermine well-supported evidence and overshadow it with disinformation. The parallel in the case of big data is central. If the correlations were indeed genuine and lobbyist used them opportunistically to sow doubt, the correlation in question is nonetheless genuine. It not creating ignorance, but only doubt. However, if the correlation is in fact spurious, then the lobbyist would be sowing doubt *and* spreading ignorance.

It is not the absence of knowledge; rather, ignorance stands *in* for knowledge, masquerading with all the apparent attributes of knowledge, even though it is the opposite of knowledge. While something qualifies as knowledge when it represents something true, ignorance, in this sense, represents something that is false. It is analogous to disinformation, which is a thing that appears as information, but in fact misinforms rather than informs one about the facts of the

world. Ignorance stands in for knowledge but serves to mislead the knower into thinking that it is knowledge. In the end, this ignorance leads the knower to espouse false beliefs.

By masquerading as knowledge, spurious correlations spread ignorance, by misleading the knower to think that she has gained knowledge. With big data, the knower is overwhelmed with information, which has more chances of originating from spurious correlations than non-spurious ones. Moreover, the deluge of data is compounded by the fact that spurious correlations resist our attempts to explain them reasonably. The obvious kind may be more easily detected by the knower, but when spurious correlations are less obvious, they can masquerade more effectively as knowledge misleading the knower in adopting false beliefs.

In sum, ignorance emerging from big data differs from the general definition of ignorance as "the absence of knowledge." In the case of spurious correlations, ignorance masquerades as knowledge. They do not withhold knowledge from the knower; rather, they convey ignorance leading to false beliefs. It is, therefore, cause for concern when big data enthusiasts embrace a data-centric approach to epistemology without acknowledging the danger of being led astray by lurking spurious correlations within data.

A famous data enthusiast once wrote: "Correlation supersedes causation"[1] and that "with enough data, the numbers speak for themselves" (Anderson, 2008). These words were written by C. Anderson, who was the editor-in-chief of Wired Magazine. The article in question, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,* appeared in the magazine in 2008. By relying solely on vast amounts of data, Anderson implied that scientists could read off the data and dispense altogether with hypotheses, models and simulations, staples of scientific theorizing (Calude & Longo, 2017, p. 16). But if scientific research relies on big data in the same

[1] https://www.wired.com/2008/06/pb-theory/

way other epistemic endeavours rely on data, then the risk of being misled by spurious correlations applies in the case of science as well, leading researchers down dead-end trails rather than fruitful pursuits for scientific discovery.

The accumulation of data and the development of increasingly sophisticated algorithms to sort through it can yield valuable results for researchers as long as they are equally aware of the danger of spurious correlations that lurk within the datasets. Taken to the extreme, what is thought to be in the service of the scientific enterprise's aims (empirical data) can now quickly work against its aims by flooding research with obvious and less obvious spurious correlations.

Sauro Succi and Peter V. Coveney take the argument a step further. In their article entitled, *Big Data: the End of the Scientific Method?,* they claim that too much data overwhelms our ability to make any sense of it without the guiding light of theoretical reasoning.

> *"[I]n science, we strive to go from data starved to data rich, yet a blind data driven procedure, as often advocated by the most enthusiastic big data neophytes, may well take us from data rich to data buried science, unless a just dose of theoretical reasoning is used as an antidote"* (Succi & Coveney, 2019, p. 10)

Without intending to broach an unavoidably lengthy discussion on the scientific method, their claim of "overwhelming our ability to make sense" of data is consistent with my claim that inquirers relying on big data will frequently encounter spurious correlations (ibid.).

What is in large part to blame, Succi and Coveney state, is a misconception of the pyramid of information. The pyramid is based on the oft-quoted, oversimplified conception of the linear

progression of knowledge: data, making up the base of the pyramid, coalesces into information, which rests on top of the base. The layer above information is where knowledge is formed and, at the apex of the pyramid, sits wisdom. The pyramid analogy is faulty, because it does not account for the uncontrollable growth of spurious correlations (ibid., p. 8). Succi and Coveney criticize this conception because it assumes a natural (read inevitable) progression from data to wisdom, by overlooking the much more complex processes involved in producing knowledge—a question left for another time. Suffice it to say, the proliferation of data makes it more difficult to find the "nuggets" that have the potential to turn into information, then knowledge, and finally wisdom (ibid.). The authors argue that although the "nuggets" may be more numerous as we gather more data, they will become harder to find in vast datasets riddled with spurious correlations. This is because, as Succi and Coveney show mathematically, that as data grows exponentially, spurious correlations will always outnumber non-spurious ones (ibid.).

In the preceding pages, I have made the case that spurious correlations lead to ignorance because, in addition to their misleading nature, they tend to occur with astonishingly high frequency in big data. Big data enthusiasts, such as Anderson quoted above, profess that the more data we have, the more knowledge we can produce. This hope is misguided because it ignores the potentially infinite number of spurious correlations and the risks they represent for our epistemic aims. Without acknowledging their existence and the problem they pose for epistemology, the cornerstone of the present work, one wonders how successful one can be in producing knowledge by simply "reading off the data"(Anderson, 2008).

Given this state of affairs, spurious correlations' detrimental effects on our epistemic aims can also carry on into the social sphere where they can inflict harm and thwart scientific progress. In Chapter 1, we looked at cases by Tuana and Proctor where ignorance had direct effects on

society. Since spurious correlations emerging from big data lead to a type of ignorance, in Chapter 3 we will look at how this kind of ignorance fits in and contributes to agnotology more broadly.

# Chapter 3 – What Kind of Ignorance Emerges from Big Data?

In the previous chapter, I defined spurious correlations as a correlation that offer no causal connection or suggest lawlike properties. I argued that spurious correlations mislead the knower, resulting in a kind of ignorance. The ignorance of spurious correlations is not the ignorance which denotes the absence of knowledge; rather, it is a presence of something pretending to be knowledge, but in fact is ignorance because it does not increase the knower's knowledge. When one accepts an ignorance masquerading as knowledge, one is led to form false beliefs, the opposite of true beliefs from knowledge. In the pages that follow, I will argue that the kind of ignorance that results from too much data is different from but overlaps with the categories of ignorance that Proctor and Tuana describe. The only kind that most closely resembles big data ignorance is Tuana's "not knowing that we do not know," (Tuana, 2006, p. 6), the second of the four summarized in Chapter 1. I will contrast big data ignorance with the different kinds put forward by Proctor and Tuana and then evaluate whether we can hold the epistemic agent responsible for mistaking big data ignorance as genuine knowledge. To accomplish this evaluation, I draw on Rescher's criteria of culpability, namely, ample opportunity and reasonable effort required to eliminate the ignorance.

My hope is to make plain the contribution of the present work in the study of ignorance, by building on the edifice of scholarship put in place by Tuana, Proctor and Rescher, whose work has been central to my study. This work's contribution is the addition of a new kind of ignorance that departs from the usual sense of an absence of knowledge. It is by no means unheard of; most of us have encountered terms such as disinformation and fake news, both of which suggest that they are things that masquerade as something else. In comparing the categories of ignorance set out the study of agnotology by Tuana, Proctor and Rescher, this way of thinking about ignorance—

as something masquerading as knowledge, not the absence of it—is new. It is new because big data technology is new, and whatever the promise of big data, it is also revealing unexpected results that can cause problems for the main reason that we collect data: producing knowledge. I support my argument by comparing big data ignorance with the different kinds proposed by the aforementioned scholars, and I will argue that we should include this new kind of ignorance as part of agnotology, especially since big data is here to stay.

## Big Data Ignorance compared to Proctor's Realms

Earlier, I defined big data ignorance in terms different from the standard definition of ignorance, a void that is filled by knowledge. Ignorance will never cease being an absence of knowledge. But *big data* ignorance, specifically, is a thing in itself, an ignorance that masquerades as knowledge—not the absence of knowledge. To put it succinctly, the spurious correlations that emerge from big data can make us more ignorant by misleading us into thinking that the correlations have an underlying causal connection or point to some law of nature. Ignorance in the sense of a void arises in a state in which one finds an untutored child, for example, who is then sent to school to assimilate the rudiments of literacy. The untutored state is a native state of knowledge.

As you will recall in Chapter 1, Proctor's first kind of ignorance is the native state. In the native state, one moves from incompetence to competence via the acquisition of knowledge and life experience (Proctor, 2008, p. 4). Proctor described it an ignorance that beckons the knower to eliminate it by learning. Big data shares this aspect with the native state: both are states of ignorance whose motivations are to learn more. Undertaking the collection of discrete pieces of data could not have been accomplished without the motivation to produce more knowledge.

However, big data ignorance is a different kind of ignorance from the native state because, when we collect data, our starting point is not entirely untutored. We must at least know what kind of data we are seeking, which presupposes having foundational knowledge about the problem that motivates the inquirer to seek answers in the data. More importantly, they differ in that native state moves from a state of ignorance to a state of knowledge, whereas in the context of big data, one can add to their baseline of knowledge more ignorance that they think is knowledge. A scenario in which both kinds of ignorance may overlap exists.

One can be in a native state of ignorance and become "informed" or "educated" with "knowledge" that can give rise to false beliefs. The untutored child, who does not have prior knowledge about a topic, may be instructed to assimilate ignorance masquerading as knowledge—the same kind of ignorance that emerges from big data. Therefore, although the native state is not identical with big data ignorance, both types of ignorance can occur simultaneously.

Proctor's lost realm, his second kind of ignorance, pertains to some knowledge becoming illuminated while plunging other knowledge into darkness. The lost realm guides and dictates the interests of knowledge and scholarship. Big data ignorance remains distinct from the lost realm in that big data is always accumulating in great datasets and does not operate in the withdrawal of knowledge. In the lost realm, knowledge becomes forgotten, as were the methods of aborting unwanted pregnancies practised by indigenous women of the Americas colonized by the Spanish (ibid., p. 8). Knowledge in the lost realm is forgotten, whereas knowledge it is not forgotten in the context of big data. Instead, big data can suggest that a correlation indicative of causality or a law of nature, when if fact they are just coincidences detected by an algorithm.

The lost realm and big data ignorance are different from one another but can be combined. The latter is an affirmation of something being knowledge when it is not. The former can occur

when strong false beliefs, resulting from big data ignorance, then sway the shifting body of interest away from actual knowledge. We see this happen with a research program that elicits much excitement and interest, the origin of which might be a non-obvious spurious correlation that was not recognized as such. Epistemic interest then shifts to pursue research on this correlation at the expense of more fruitful research. Although both kinds of ignorance, big data, and lost realm, remain different, they resemble each other by being (mis)led by spurious correlations and social interests, respectively.

The lost realm and big data ignorance can occur simultaneously, as they do in the native state. The shift of focus can settle on misleading knowledge, but it can also trigger a shift of interest that leads one to pursue misleading knowledge. In other words, the shifting body may apply focus to a part of knowledge, which can also include misleading knowledge arising from spurious correlations. This is a similar scenario that we encounter with Tuana's first type of ignorance which is dictated by the values of an epistemic community (Tuana, 2006, p. 6).

Proctor's strategic ploy, the third of the "three realms" summarized in Chapter 1, does not provide a befitting definition for big data ignorance either. Proctor describes the strategic ploy as the realm where knowledge is withheld or can be attacked by extreme scepticism (Proctor, 2008, p. 9). Misleading knowledge and disinformation have a long history in the tactics of subterfuge and deception in military and political strategizing. Behind these tactics lies ill intent. Big data ignorance, in contrast, is rather motivated by the need to gather more data and create knowledge and the occurrence of spurious correlations are perhaps the unfortunate by-product of our technology and algorithmic techniques, not our intentions.

However, both types of ignorance can coexist and overlap. Spurious correlations can become convenient for the ill-intentioned researcher as "evidence" that runs contrary to the

stronger claim. The strategy could be a way of sowing doubt or promote a false belief in order to keep people in the dark or, at the very least, dismiss the stronger claim. Suppose that there is strong evidence supporting the hypothesis that consuming product A will develop illness B. Since this poses a risk to product A, researchers, ostensibly trying to confirm the evidence, seek instead a spurious correlation that undermines the strong evidence. This sown doubts in the minds of the consumers who like product A, leading them to believe that product A is probably safe. In short, albeit different, strategic ploy ignorance can be combined with big data ignorance, the latter used in furtherance of the goal of the former.

Big data ignorance cannot causally explain a phenomenon or suggest lawlike properties for it. Cheese consumption has nothing to do with people dying by getting entangled in their bedsheets, no matter how powerful that statistical correlation is. The strategic ploy is when one group uses a spurious correlation such as this to cast doubt on the strongly supported finding of an opposite claim (or at least a strongly supported claim that runs contrary to the interests of the promoters of the strategic ploy). But big data ignorance does not need for one group to purposefully mislead another. The spurious correlations emerging from big data mislead the knower without the possibility of applying intent on anyone. The algorithmic techniques or the software cannot form the intent of purposefully misleading the inquirer—at least until we invent general artificial intelligence that then takes a dystopic turn. Therefore, while the strategic ploy requires an agent with a motivation to mislead, big data ignorance has no such agent with dubious motivations.

Taken together, big data ignorance interacts and overlaps with the three categories proposed by Proctor but remains distinct from them because Proctor's categories a) presuppose an agent behind the ignorance and b) presuppose that ignorance is an absence of knowledge or knowledge that is undermined by doubt. My argument is that big data ignorance is a new type of

ignorance that is created, as a thing in itself, not as a void of knowledge, with the purpose of deceiving the knower that this entity is knowledge when it is actual ignorance masquerading as knowledge. It deceives the knower by presenting them with statistically powerful correlations as evidence for causal explanations or possessing lawlike properties. In other words, the void of knowledge can be filled with an impostor. Instead of withholding knowledge *tout court*, big data ignorance can lead one to espouse false beliefs based on that knowledge.

## Big Data Ignorance compared to Tuana's Categories

Nancy Tuana's first kind of ignorance, as described in chapter 1, is "not knowing and not caring to know." Tuana points out that what motivates this ignorance ultimately stems from the values of the community of knowers (Tuana, 2006, p. 6). When less value or no value is assigned to a certain kind of knowledge, it gets left in the dark, as happens with Proctor's lost realm. However, it differs markedly from the kind of ignorance that could emerge from big data. Big data ignorance departs from Tuana's "not knowing and not caring to know" in that big data is primarily a reflection of the value to learn more about the world by collecting more data. I will set aside any other possible motivations large corporations had to develop technologies to support the collection and storage of data. Big data ignorance, however, can overlap with Tuana's first kind of ignorance even though both kinds are different.

Tuana's first kind of ignorance can be reflected in the way we use big data. For example, we might choose to collect data about knowledge that we value and not collect any data for areas of knowledge to which we assign no value. As a result, big data strengthens the empirical force behind the value of a specific kind of knowledge, while ignoring the potential for gathering data areas that are less valued, but ought to be. Therefore, while both kinds of ignorance can overlap, they remain distinct in that big data ignorance occurs after the collection of data—not before the

collection. In other words, big data collection implies the decision to accumulate data, not to withdraw it.

Tuana's third kind of ignorance, as you will recall from Chapter 1, is ignorance that arises when one privileged group controls what knowledge to convey to a less privileged group (Tuana, 2006, p. 13). In this respect, Tuana's third kind of ignorance also differs markedly from big data ignorance. It differs in that big data ignorance arises from spurious correlations unexpectedly and if left undetected, they can mislead the knower. But the ignorance that is created by one privileged group is not an unexpected occurrence, but a calculated one. In other words, we cannot blame spurious correlations for intentionally misleading the knower in the same way that we can blame a privileged group for withholding knowledge or misleading a less privileged group from knowing.

This point merits some elaboration. Most societies have power structures in place, overt or covert, whereby one group has privileges beyond what another group enjoys. These come in the privileges of wealth, status, or knowledge. Homogenous societies may have less considered thoughts and attitudes towards correlations about race, for example. Heterogenous epistemic communities may be better equipped to deal with spurious correlations in this regard, possibly because they might detect a spurious correlation more easily because they have competing intuitions about what the correlation seems to suggest, making the agents more ready to challenge it. For example, a correlation linking the occurrence of a disease recorded in a social group, might slip by undetected and mislead the knowers into thinking that there is something causal underlying the disease and its apparent tendency to appear in a certain population.

Albeit different, Tuana's third kind of ignorance can also overlap with big data ignorance. The overlap occurs when spurious correlations can be deliberately used by a privileged group to distort or undermine knowledge disseminated to a less privileged group. In other words, the

privileged group can use the occurrence of spurious correlations to their advantage to serve their interests. Therefore, while ignorance created by one privileged group in order to keep another less privileged group ignorant occurs with deliberation, it resembles Proctor's strategic ploy more than it does ignorance emerging from big data. Both Tuana's third kind of ignorance promoted by a privileged group, and Proctor's strategic ploy ignorance involve intent and deliberate action to cause the ignorance, whereas big data ignorance occurs unexpectedly, and blame can only be placed on the technological tools and techniques used to detect them.

Tuana's fourth kind of ignorance is willful ignorance, also discussed in Chapter 1. While this type of ignorance differs from Tuana's third kind of ignorance promoted by privileged groups (it will occur to the reader that both are in fact willful), willful ignorance is a refusal to let go of ignorance when knowledge has been made known (Tuana, 2006, p. 10). Willful ignorance, Tuana explains, is the agent's deliberate acceptance and willingness to maintain a state of ignorance of the agent himself. It differs from the third kind because the latter targets a group of knowers outside oneself. Unless willful ignorance makes use of spurious correlations in big data to help maintain the ignorance, it differs from the ignorance caused by too much data. In the case of big data, it is unlikely that one would resort to collecting and analyzing more data with the specific aim of remaining ignorant. Researching in big data assumes that the inquirer is looking for information and knowledge in data. Here too, as in the case of ignorance promoted by a privileged group, intent to deceive or mislead (either oneself or another group) is central to the kinds of ignorance they are, but intent to deceive or mislead is not central to big data ignorance.

Nevertheless, we cannot ignore that even here willful ignorance and big data ignorance overlap. The overlap occurs when the inquirer fosters the intention of deceiving and misleading either himself or another group, as mentioned above. By adding the factor of ill intent, the agent

now may want to look through vast datasets to find correlations that support the harboured false beliefs. When the algorithmic techniques are left alone to sort through the data, they are perfectly capable of finding spurious correlations that can deceive or mislead without the help of a malevolent agent.

It is Tuana's second kind of ignorance — "not knowing that we do not even know"—that resembles most closely big data ignorance. What Tuana points out about this type of ignorance is that our theories and beliefs about the world create barriers preventing us from the potential knowledge that can be produced and cause knowledge to recede into oblivion (ibid. p. 6). Where this kind of ignorance resembles big data, ignorance is in the faith we place in our new powerful methods of gathering data, a belief that can ultimately prevents us from acknowledging the existence of spurious correlations. So much for their resemblance. The difference between the two kinds of ignorance is in what they do with knowledge. Tuana's formulation, on the one hand, negates the knowledge by neglecting (not pursuing because it is no longer acknowledged). Big data ignorance, on the other hand, posits an entity (the opposite of knowledge). Indeed, it puts forward an ignorance masquerading as knowledge. Big data, therefore, does not withdraw or neglect knowledge, but adds alongside knowledge something else that is not knowledge, but ignorance.

Summing up, as with Proctor's "three realms," we find that Tuana's "categories" of ignorance overlap with big data ignorance, with the exception of her second type of ignorance (not knowing that we do not know), which resembles closely big data ignorance. By showing that the variety of ignorance that we discussed thus far, and how none is identical to the ignorance produced by big data, I conclude that the latter is a new kind of ignorance that as epistemic agents (especially those of us handling big data) need to contend with.

## Is Big Data Ignorance Culpable? Rescher's Criteria

The question that is perhaps triggered in the reader's mind is whether big data ignorance is caused by human agency, placing the blame on the agents that promote it. If I am required to trace back where the ignorance, which is at the level of the spurious correlation, and then trace back the correlation's beginning to the algorithms that produced from the pool of data, where the algorithm was programmed by a human ingenuity, and what data to gather was decided by a human, then the answer is that yes, human agency is involved in creating big data ignorance. I will not be arguing this point because it is too simplistic. Instead, my argument is that although human agency is at the source of where big data ignorance starts, human intent is still required to make the case of direct culpability. The programmers and software engineers behind big data creation are likely not hatching diabolical plans to mislead big data users. I first started this discussion that spurious correlations were a surprising product of big data. They were surprising because no one intended to produce them. In fact, we know spurious correlations can arise accidentally at remarkable frequency, as the analysis of Calude and Longo has conclusively indicated, and which was described in Chapter 2 (Calude & Longo, 2017). But knowing about the existence of this curious product of spurious correlations, one can still use big data irresponsibly. This is precisely where culpability begins.

Using big data without regard to the possibility of facilitating the adoption of false beliefs, based on spurious correlations, is irresponsible and therefore culpable. But determining the degree of culpability of big data is a more complicated matter. To facilitate this task, I will use Rescher's criteria of ample opportunity to dispel the ignorance and reasonable effort required to acquire the knowledge (Rescher, 2009). With big data, we have a potentially infinite source of spurious correlations because they are devoid of any causal explanation or law-like properties. A user of data must therefore use caution to identify spurious correlations, in particular the non-obvious

ones, and dismiss them. Obvious spurious correlations require less effort compared to non-obvious correlations, but in either case, the effort required is reasonable. An inquirer seeking information for whatever epistemic endeavour is essentially an investigator. Part of the investigation should include evaluating correlations.

With regard to ample opportunity, the inquirer relying on big data is required to evaluate the data and interpret it in order to glean any knowledge from it. The data itself does not contain ready-made answers to questions; the answers need to be found. In this sense, opportunity always exists for the inquirer on account of the nature of the activity. What Rescher meant by not having the opportunity to disabuse oneself of ignorance is precisely the point where one was never given the opportunity to access the knowledge that would eliminate the ignorance. Knowledge of the future is a case in point. When the inquirer relies on big data, the latter has no problem with accessing the information, for it is within their reach. Assuming that the inquirer began their journey intending to acquire knowledge, using big data irresponsibly is culpable when they can use reasonable effort to eliminate the ignorance, and when ample opportunities to do so present themselves.

In Chapter 1, I summarized Proctor's recounting of the tactics used by tobacco-industry backed researchers digging up contradicting evidence to undermine the strong link between tobacco consumption and lung cancer. In contrast to the well-intentioned inquirer above, these researchers differ in their culpability because their intent was to discredit and distort knowledge by sowing doubt. A separate discussion is needed to deal with the degrees of culpability and the consequences that ensue. For the present work, it suffices to identify big data ignorance as vincible—that is, ignorance that can be identified and contained or eliminated with reasonable effort and ample opportunity. Failing to do so is blameworthy.

Consider our non-obvious spurious correlation from Chapter 2, relating arcade revenues with the number of doctoral degrees in computer science. A researcher discovers the correlation between the two events over a period of nine years and sees a plausible connection between computer science and the creation of computer games. The researcher also knows that the data is accurate and that such a strong correlation (0.94) for nine years (at least, because that data was gathered for that period) seems to be saying something about the world. Otherwise, the researcher might ask, 'could such a high correlation last for nine years? It couldn't possibly be a coincidence!' But before concluding that there is a causal or lawlike connection between the events, and presenting the findings to a community of knowers, the researcher will need to confirm the findings. Nothing prevents the researcher from deliberating (ample opportunity) on the findings before accepting them as a true conclusion, and the effort required to do this is not beyond the capabilities of the researcher (i.e., it is reasonable). We can take any number of scenarios involving spurious correlations and they unfold in similar ways: an inquirer accessing datasets to seek answers to a question already creates the opportunity to deliberate on any correlation that may arise and already requires effort, which the inquirer applies. If process sounds familiar because it is the largely the way the scientific method works. This method encourages that correlations be investigated in order to determine whether they are fruitful in making a scientific discovery.

Made aware of the existence of an abundance of spurious correlations lurking beneath the data, ideally, the knower needs to exercise caution when dealing with data-centric inquiry. If not, the knower could be blamed for dealing with data irresponsibly and forming false beliefs. When the knower seeks to gain knowledge by accepting spurious correlations uncritically, the knower is still culpable but without ill intent, for they started off with the intention of gaining true knowledge, even though they ended up gaining false belief. Some big data proponents, like Anderson from

Chapter 2, who take a less critical approach about the benefits of relying heavily of data and the technology that collects and stores, may belong to this class. That is because they are, for the most part, seeking to glean true knowledge from the mistaken assumption that correlations are reliably capable of allowing the knower to do so without impediment, while obvious and non-obvious spurious correlations lurk beneath the surface. However, while we cannot begrudge their intentions, blame is nonetheless assigned to their uncritical approach for it can spread false belief, and, if I take the argument further, cause harmful downstream effects, which often follow from ill-intent. Therefore, false beliefs following from spurious correlations can lead to harm. Taking our arcade example further, a false belief about it may lead to promoting more arcades to be open and encouraging patrons to frequent them more, in order to increase computer science graduates. This may come at the expense of the patrons, who are better off spending time in rigorous physical activity or socializing with people, rather than playing computer games in arcades. We can take turns in placing blame either on the researcher or the over-credulous arcade investor for espousing the belief—a separate discussion for another time—but what we can agree by thinking about this example is that inquirers who have both opportunity and reasonable effort can and ought to eliminate the ignorance that is caused by big data in the same way one takes advantage of opportunities and makes the necessary effort to eliminate other kinds of ignorance.

One aim of the current project is to create awareness of a new issue that we should be cautious about when researching. The recent and powerful technologies and algorithmic techniques we are using to collect, store and sort through data reveal the presence of spurious correlations, capable of misleading the researcher, by suggesting causality or lawhood. To accept what the spurious correlations are suggesting as meaningful about the world is to form a false belief. Such an uncritical acceptance is blameworthy and wrongheaded. It is wrongheaded because,

again, spurious correlations are misleading and, therefore, counterproductive to epistemic endeavours. Uncritical acceptance is also blameworthy because there is ample opportunity to critically assess the suggestions arising from big data.

A fairly recent case that illustrates how spurious correlations lead to false beliefs is how Google Flu Trends, a now-defunct Google product, tried to predict future flu outbreaks. The algorithm deployed to deal with the incoming data collected by searches on flu symptoms correlated the searches with actual cases arising from specific locations. However, Google's initial success was short-lived. The algorithm did not take into account that people carried out searches on Google to learn more about flu symptoms out of curiosity—not because they were experiencing the symptoms (Holmes, 2017, p. 76). By the following flu outbreak, their predictions fell apart. Insofar as they rely on correlations that turn out to be spurious, the engineers—or any researcher for that matter—risk being misled even after critically assessing where the algorithms go wrong. The risk persists because addressing the issues with their algorithms is not the only problem. The engineers would need to address their assumptions about big data as a rich source of spurious correlations that may likely be the root of their problem.

If in the future, inquirers investigating a problem, like the Google engineers, could be blamed for having the opportunity to deal with the spurious correlations and the ensuing ignorance. In retrospect, the effort required to dispel the ignorance that was setting in, was arguably reasonable. We can add to the blame on account of their expertise. If you recall in Chapter 2, Rescher distinguished between common knowledge that the layperson has and the specialized knowledge of the expert (we can replace the example of the medical practitioner with the Google engineer) (Rescher, 2009, p. 20). In this view, we severely condemn the engineers' neglect of the

existence of spurious correlations, since they are experts in implementing algorithmic techniques that collect and sort through the data.

It should be noted that, even if, by chance, the engineers were still able to trend the flu, which was the intended outcome, by relying on correlations that turned out to be spurious, they would have still been operating ignorantly—not knowledgeably. The ignorance would have eventually revealed itself, as it had done in 2009, when the swine flu broke out and the "Google's big data algorithm famously failed to deliver" (Holmes, 2017, p. 76) . Simply relying on predictions without the causal or lawlike frameworks that allow us to explain the predictions is insufficient. This is consistent with the claim I made about the spurious correlations in Chapter 2, which enjoyed an astonishingly high statistical correlation over a period of nine years. Eventually, predictions for death by bedsheet entanglements and computer science doctoral degrees were bound to crumble.

In sum, it is too great an expectation that no spurious correlations ever slip by unnoticed. In fact, we ought to expect, in data-intensive epistemic endeavours, that spurious correlations will inevitably creep in and may even mislead us. The answer is not to cease relying on big data, but simply to keep in mind that encountering spurious correlations is highly likely given the frequency with which they occur.

I compared big data ignorance with the kinds of ignorance set out by Proctor and Tuana, highlighted the differences and similarities and where they can overlap with ignorance emerging from big data. I hope that the verdict clearly shows that by proposing an ignorance that is not the absence of knowledge, but a belief that masquerades as knowledge, we admit it as a kind of ignorance that we need to acknowledge as a threat to our epistemic projects, in particular those relying heavily on data. The rate at which technologies are moving to meet our increasing demand

of data collection will only increase the number of spurious correlations we encounter. This much has been made clear by Calude and Longo and many others who are seeing the potential risks along with the advantages of big data (Calude & Longo, 2017, 16-17).

Given this state of affairs, it was only appropriate to evaluate the knower's culpability vis-à-vis big data ignorance by using Rescher's criteria of vincibility (Rescher, 2009, p. 12). On this count, the verdict also supports culpability on the part of the knower, in light of the ample opportunity and reasonable effort to eliminate the ignorance.

There may be situations where ignorance is more desirable than knowledge, but, as I argued in Chapter 2, in the context of big data, ignorance is antithetical to the former's aims. We will consider it a failure to plunge into data, only to produce more ignorance than knowledge. We would leave bemused by such a possibility, for the progression of knowledge, epitomized in the simplistic conception of the pyramid referred to in the previous chapter, does not allow this retrogressive movement. I remind the reader what this simplistic and rather idealistic pyramid posits: Data is supposed to cluster together to create information, which in turns serves as the basis for creating knowledge, which finally turns into wisdom (Succi & Coveney, 2019, p. 8). But the retrogression does not mean that the pyramid of knowledge is broken (although certainly it is simplistic), rather it is indicative of our lack of understanding of the complexity of knowledge and ignorance creation. Spurious correlations are just the latest factor adding a layer of complexity to how we theorize about knowledge and, as it turns out, ignorance.

# Chapter 4 - Conclusion

When we want to learn something about a subject, we begin by collecting more information about it. It is obvious that having access to more data should provide more knowledge. This is made plain in our everyday practices from time immemorial. We created libraries to house manuscripts and books and invented methods to record information. In the 21st century, we save countless digital documents, images, and videos on our handheld devices. We digitize and store electronic copies of files in ever more capacious and powerful servers. Seen from outside, one might think that we have an innate fear of not only losing data, but also not having access to more of it. More data, we assume, will give us a bigger picture and more knowledge. Less data, we fear, might mean missing something important that could have a significant impact on our epistemic outcomes. With this project, I tried to show that our fears about a lack of data and our expectations of it might be unwarranted. The methods we use in collecting and storing and organizing data have the unfortunate tendency of creating spurious correlations.

In chapter 1, I presented recent scholarship on ignorance and the conditions under which it can permeate our society by drawing on the pertinent taxonomies of Nancy Tuana and Robert Proctor. I also reviewed the apt criteria proposed by Nicholas Rescher to help determine the cases in which big data ignorance can be blameworthy.

In chapter 2, I defined spurious correlations as not having any underlying causal or lawlike attributes. Obvious ones are easily detectable as epistemically useless or mere coincidences, but the less obvious ones can cause problems for knowers because they appear as epistemically fruitful correlations. Once we are misled in accepting a spurious correlation, we cannot claim to have gained more knowledge. What we have instead gained is ignorance masquerading as knowledge. This ignorance is different from the kinds of ignorance described by Tuana, Proctor and Rescher.

In chapter 3, I support my claim that this new type of ignorance is different by comparing it in more detail with the types of ignorance set out by the three scholars. I also show that albeit different, big data ignorance can occur simultaneously or be combined with other kinds of ignorance. Finally, I show that big data ignorance involves culpability, because using big data irresponsibly may entail disseminating ignorance that is believed to be knowledge.

The conclusions we draw from the previous chapters may illuminate avenues that we can pursue to resolve the tension between needing data to create knowledge and the fact that too much data can create ignorance. Furthermore, as our technological tools get better and our data continues to increase exponentially, we will only see the presence of spurious correlations increase along with it.

The solution to this problem could be a straightforward technological one. Perhaps some invention will make better use of the algorithms that sort through the data. Then again, the problem could be with our algorithms themselves. One might also ask whether we put too much stock in them. We use them daily during our commutes when we check for road traffic, when we are Google-searching for information, or when we are shopping online. Some may be questioning whether we rely on them excessively.

However, the solution might lie outside technology altogether. At least with regard to scientific research, we may want to find answers in revised approaches to epistemology and the scientific method. The answer may lie in rethinking the role theorizing has in the different disciplines to help guide us with how we deal with the deluge of data. Admittedly, this is not a new solution—creating theories has been part of the way we have been doing science for a very long time. If our reliance on big data mires scientists in an endless series of dead ends, wasting time and resources, it might renew our commitment to theorizing and hypothesizing. One thing

needs to become abundantly clear: when it comes to big data, we may be getting more than what we bargained for. Ignorance produced by spurious correlations is antithetical to the aims of data collection in the first place.

This study on big data and ignorance has prompted other troubling concerns about how big a role technology plays in our epistemic endeavours. If big data technology and the algorithmic techniques it uses have the ability to produce wrong results, could not other technologies also have the ability to lead us astray? With this worry in the back of our minds, we might also greet with some scepticism the "hype" surrounding "ground-breaking" technologies with the promise of revolutionizing science. While being sceptical is an unlikely stance for creating knowledge, it nevertheless appears to be a dependable one for preventing ignorance.

# Bibliography


al-Tawḥīdī, A. Ḥ, Miskawayh, A. A., Vasalou, S., Montgomery, J. E., & Rée, J. (2021). *The Philosopher Responds: An Intellectual Correspondence from the Tenth Century*. NYU Press. https://books.google.ca/books?id=ZWwDEAAAQBAJ

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, *16*(07).

Audi, R. (2015). The Cambridge Dictionary of Philosophy, 3rd Ed. In *Cambridge University Press*.

Calude, C. S., & Longo, G. (2017). The Deluge of Spurious Correlations in Big Data. *Foundations of Science*, *22*(3). https://doi.org/10.1007/s10699-016-9489-4

Cohen, J., & Callender, C. (2009). A better best system account of lawhood. *Philosophical Studies*, *145*(1). https://doi.org/10.1007/s11098-009-9389-3

Holmes, D. E. (2017). *Big Data: A Very Short Introduction*. Oxford University Press. https://books.google.ca/books?id=NXw7DwAAQBAJ

Langston, D. (2011). Medieval Theories of Conscience. In *Stanford Encyclopedia of Philosophy*.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books. https://books.google.ca/books?id=9H0dDQAAQBAJ

Proctor, R. N. (2008). Agnotology: A Missing Term to Describe the Cultural Production of Ignorance (and Its Study). In *Agnotology: the making and unmaking of ignorance*.

Rescher, N. (2009). Ignorance: (On the wider implications of deficient knowledge). In *Ignorance: (On the Wider Implications of Deficient Knowledge)*.

Succi, S., & Coveney, P. v. (2019). Big data: The end of the scientific method? In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 377, Issue 2142). https://doi.org/10.1098/rsta.2018.0145

Tuana, N. (2006). The speculum of ignorance: The women's health movement and epistemologies of ignorance. *Hypatia*, *21*(3), 1–19.