Forthcoming in *Philosophy and Technology*

The Fate of Explanatory Reasoning in the Age of Big Data

Abstract: In this paper, I critically evaluate several related, provocative claims made by proponents of data-intensive science and "Big Data" which bear on scientific methodology, especially the claim that scientists will soon no longer have any use for familiar concepts like causation and explanation. After introducing the issue, in section 2, I elaborate on the alleged changes to scientific method that feature prominently in discussions of Big Data. In section 3, I argue that these methodological claims are in tension with a prominent account of scientific method, often called "Inference to the Best Explanation" (IBE). Later on, in section 3, I consider an argument against IBE that will be congenial to proponents of Big Data, namely the argument due to Roche and Sober (2013) that "explanatoriness is evidentially irrelevant". This argument is based on Bayesianism, one of the most prominent general accounts of theory-confirmation. In section 4, I consider some extant responses to this argument, especially that of Climenhaga (2017). In section 5, I argue that Roche and Sober's argument does not show that explanatory reasoning is dispensable. In section 6, I argue that there is good reason to think explanatory reasoning will continue to prove indispensable in scientific practice. Drawing on Cicero's oft-neglected *De Divinatione*, I formulate what I call the "Ciceronian Causal-nomological Requirement", (CCR), which states roughly that causal-nomological knowledge is essential for relying on correlations in predictive inference. I defend a version of the CCR by appealing to the challenge of "spurious correlations", chance correlations which we should not rely upon for predictive inference. In section 7, I offer some concluding remarks.

Keywords: Inference to the Best Explanation; Bayesianism; Scientific Inference; Cicero; Big Data; Data-Intensive Science; Data-Driven Science

1. Introduction

In this paper, I critically evaluate several related, provocative claims made by proponents of data-intensive science which bear on scientific methodology. According to these "Big Data" enthusiasts, as our ability to gather and analyze data increases, the nature of scientific practice will change dramatically. In the future, theorizing will to a significant degree become obsolete. As a result, future science will be data-driven, rather than hypothesis-driven. Moreover, scientists will be able to dispense with theoretical background assumptions, and in particular, scientists will no longer have any use for familiar concepts like causation and explanation. Given that these methodological claims seem to fly in the face of current scientific practice, all of them are worthy of philosophical scrutiny. Unfortunately, however, general philosophy of science has, for the most part, had little to say about the Big Data movement and the new data-intensive science.[1] Accordingly, it is my goal in this paper

[1] However, see Pietsch (2016) who critically evaluates some of these claims, and with whose views I am broadly sympathetic, as well as Leonelli (2012) who considers the impact of Big Data on biological practice.

to connect some of these epistemological and methodological concerns about data-intensive science to relevant debates in general philosophy of science.

The plan for the rest of the paper is as follows. In section 2, I elaborate on the alleged changes to scientific method that feature prominently in discussions of Big Data. In section 3, I argue that these methodological claims are in tension with a prominent account of scientific method, often called "Inference to the Best Explanation" (IBE). Later on, in section 3, I consider an argument against IBE that will be congenial to proponents of Big Data, namely the argument due to Roche and Sober (2013) that "explanatoriness is evidentially irrelevant". This argument is based on Bayesianism, one of the most prominent general accounts of theory-confirmation. In section 4, I consider some extant responses to this argument, especially that of Climenhaga (2017). In section 5, I argue that Roche and Sober's argument does not show that explanatory reasoning is dispensable—at best, their argument demonstrates that in certain cases, i.e. inferences from frequency data, explanation does not play a direct role in theory-confirmation. However, there are compelling cases of what seems like direct explanatory reasoning that do not have the same structure as the example that Roche and Sober exploit in their argument. In section 6, I argue that there is good reason to think explanatory reasoning will continue to prove indispensable in scientific practice. Drawing on Cicero's oft-neglected *De Divinatione*, I formulate what I call the "Ciceronian Causal-nomological Requirement", (CCR), which states roughly that causal-nomological knowledge is essential for relying on correlations in predictive inference. I defend a version of the CCR by appealing to the challenge of "spurious correlations"—chance correlations which we should not rely upon for predictive inference. In section 7, I offer some concluding remarks.

2. The Alleged Impact of Big Data on Scientific Method

The last few decades have witnessed the rapid growth of technologies that permit the collection, integration, and analysis of massive quantities of data. The character of this data-collection is unprecedented, not only its volume, but also in its variety and the velocity at which such data is being amassed (Kitchin 2014). As a result, this so-called "Big Data" movement promises to bring about lasting changes to the way that we "live, work, and think" (Mayer-Schönberger and Cukier 2013). Since much of the current data-collection is conducted by private enterprises, consists of information about consumers and their purchasing decisions, and is used for marketing and advertisement, this new movement obviously raises several important ethical concerns, especially with regards to privacy and consumer autonomy (e.g. Mittelstadt and Floridi 2016). Equally important,

however, are the epistemological implications of these sophisticated modes of data-gathering. According to some enthusiastic commentators, our newfound abilities to collect and analyze massive quantities of data will, with the help of computational tools, profoundly alter the way that science is done. Along with the rise of Big Data comes the new "paradigm" (Hey et al. 2009) of "data-intensive" science, which is said to differ markedly from traditional conceptions of scientific method.

First, whereas the orthodox account of scientific method consists, more or less, in formulating some hypothesis, deriving test predictions from that hypothesis, and then running experiments to verify those predictions, the model of scientific method suggested by Big Data is one that is "data-driven" rather than "hypothesis-driven" (Mazzocchi 2015). Instead of commencing inquiry with some pre-conceived hypothesis whose confirmation or disconfirmation is sought by means of an experimental test, scientists can turn to data-mining and data-analysis software. A data-driven approach rather than a hypothesis-driven approach has the potential to lessen the impact of cognitive biases, such as confirmation bias, on scientific research and to reveal patterns in the data, such as hidden interesting correlations, that might not have been noticed by the human researcher.

Second, proponents of data-intensive science frequently make the related claim that soon scientists will be able to dispense with theoretical background assumptions entirely (Kitchin 2014). While it is commonplace for scientific models to include substantive background assumptions, often requiring adjustable parameters whose values need to be estimated in some way, the new paradigm of data-intensive science points the way toward simple, "theory-free", algorithmic models, which need not specify any nomological relationship between dependent and independent variables. In this way, proponents of the new data-intensive science allege that theory (and, eventually, the human scientist) will gradually recede from the scene, at last allowing the data to "speak for themselves".[2]

Finally, data-intensive science involves "a move away from the age-old search for causality", where instead the focus will be on correlations, which in many contexts are "good enough" (Mayer-Schönberger and Cukier 2013, p. 14). Indeed, the prospect of predicting, with a high degree of accuracy, certain outcomes of interest simply on the basis of observational data, without knowing the

[2] See Kitchin (2014, pp. 5-7) for a more extensive discussion and critical examination of such claims. One moderate and less provocative thesis that Kitchin advances is that, rather than replacing human scientists, new data-mining techniques should be used to supplement traditional scientific methods by "reveal[ing] information which will be of potential interest and is worthy of further research" (2014, p. 6). Here, Kitchin locates the primary epistemological significance of Big Data analytics in the "context of discovery" (or the "context of pursuit") rather than the "context of justification".

causal mechanism that explains the data, is perhaps *the* hallmark of data-intensive science.[3] According to one Big Data proponent, Eric Siegel, "[w]e usually don't know about causation, and we often don't necessarily care…the objective is more to predict than it is to understand the world…It just needs to work; prediction trumps explanation" (2013, p. 90). As Siegel goes on to claim, it is often the case that "[w]e know the *what*, but we don't know the *why*," (2013, p. 90). Given that causation is one of the most important explanatory relations, what this third claim of data-intensive science amounts to is that *explanation*, which is often regarded as one of the "cardinal aims of science" (Strevens 2006, p. 516), will no longer prove to be an important goal of scientific inquiry. In short, according to proponents of data-intensive science, we will no longer need to ask *why* certain patterns hold in nature.

3. Inference to the Best Explanation Meets Roche & Sober's Screening-off Thesis

On its surface, the claim that data-intensive science will be able to dispense with explanatory hypotheses or theoretical causal models is puzzling. This is because on one prominent account of scientific inference, often called "inference to the best explanation" (IBE), or "abduction", or "explanatory reasoning", appeals to explanation are essential for the justification of scientific theories. According to IBE, contemporary discussion of which traces back to Harman (1965), much scientific inference manifests the taking of an "explanatory detour" (Lipton 2004, p. 65). A hypothesis H is upheld as rationally justified by showing how well H would, if true, explain some set of facts. That H would explain these facts better than its competitors is taken to ground our justification for believing H to be true. Often, IBE is formalized as a four-step argument pattern (e.g. Psillos 2002):

(i) F is some fact or collection of facts
(ii) Hypothesis H_1, if true, would explain F
(iii) H_1 is a better explanation of F than its competitors H_2, H_3,...H_n
(iv) Therefore, probably, H_1 is true

According to Douven (2011), our use of IBE in everyday life is "so routine and automatic that it easily goes unnoticed". Moreover, there is a long list of prominent episodes from the history of science in which philosophers have argued that IBE was the central form of reasoning at play.[4] Given the epistemically privileged role that IBE affords to distinctively explanatory factors, one might argue then

[3] See Northcott (forthcoming) for some helpful case studies on the extent to which Big Data analytics actually improves predictive performance.
[4] In addition to Darwin's theory of natural selection and common ancestry (Okasha 2000), this list includes the Copernican argument for the heliocentric model of the solar system (Gauch 2012), and Huygens' argument for the wave theory of light over Newton's particle theory (Thagard 1978).

that the provocative claims made by proponents of the new data-intensive science are in tension with one highly intuitive and attractive account of the nature of scientific method.

However, Big Data enthusiasts who advocate or predict a methodological shift in the sciences might take inspiration from an argument by Roche and Sober (2013), which bears on the epistemic status of IBE. According to Roche and Sober—henceforth R&S—"explanatoriness is evidentially irrelevant." In their argument that explanatoriness is evidentially irrelevant, by "evidential relevance", R&S have in mind the standard qualitative Bayesian account of confirmation, according to which O confirms, or is evidence for, H if and only if $Pr(H|O) > Pr(H)$. The proposition that R&S subject to scrutiny and which is supposed to capture the explanatoriness of H with respect to O is:

(E) If H and O were true, H would explain O

What R&S consider is whether E gives H any confirmational boost on top of O, assuming that O is true. In particular, R&S ask whether the Bayesian ought to endorse the following inequality:

(EER) $Pr(H|O\&E) > Pr(H|O)$

This inequality codifies the claim that explanatoriness is evidentially relevant from the Bayesian perspective.[5] According to R&S, however, EER is false. Instead, R&S endorse the following:

(SOT) $Pr(H|O\&E) = Pr(H|O)$

In subsequent discussion, this equality is referred to as the "Screening-off Thesis" (R&S 2017). Here is how R&S summarize in words their central anti-explanationist claim:

> This equality says that the observation O screens-off E from H; according to [SOT] the explanatoriness of H is evidentially idle, once the truth of O is taken into account. If you already know that O is true and you have computed Pr(H|O), learning E does not change how confident you should be in H (2013, p. 660).

Thus, according to R&S— and in keeping with the claims of Big Data proponents—the Bayesian has no reason to care about the explanatoriness of H, at least in the sense captured by the proposition E.

The primary way that R&S argue for SOT is by first offering an example in which SOT is alleged to be true, and then by arguing that this example should convince us that SOT is likewise true in a "wide range of realistic cases" (R&S 2017, p. 582). The case that R&S discuss concerns smoking and lung cancer. Following R&S (2017), let Sm be "S was a heavy smoker before the age of 50" and let Ca be "S gets lung cancer after the age of 50". Now, consider the following pair of statements:

---

[5] The reason that R&S choose to evaluate EER instead of the similar inequality: $Pr(H|O\&E) > Pr(H)$ is that O might confirm H by itself, and E might be irrelevant to H once joined with O, neither lowering the probability of H nor raising H's probability. Thus, it is not enough to simply show that $Pr(H|O\&E) > Pr(H)$ to demonstrate that explanatoriness is evidentially relevant for the Bayesian. What must be shown is EER.

(1) $Pr(Sm) = a$

(2) $Pr(Sm|Ca) = b$

Suppose, based on some sample frequency data, and some statistical estimation technique, that the best estimate for $a = .3$ and the best estimate for $b = .7$. In this case, getting lung cancer after the age of 50 is evidentially relevant to having been a heavy smoker because:

(3) $Pr(Sm|Ca) > Pr(Sm)$

This inequality is established purely by assembling the relevant frequency data and by employing some statistical estimator, such as maximum likelihood, which allows us to move from observed associations in our sample data, to claims about population frequencies, and eventually to justified claims about the relevant probabilities, which in this context are to be understood as "rational credences." In general, Pr(A | B) is determined by $\text{Freq}_{\text{POPULATION}}$(A | B), which is estimated from $\text{Freq}_{\text{SAMPLE}}$(A | B).

Now, given that (3) has been established, suppose we consider the particular instance of EER in the smoking-and-cancer case. Here, let E_{Sm} be "S's being a heavy smoker before the age of 50 would explain S's getting lung cancer after the age of 50, if S's being a heavy smoker before the age of 50 and S's getting lung cancer after the age of 50 were true." Thus, the inequality at issue is this one:

(EER_{Sm}) $Pr(Sm|Ca\&E_{Sm}) > Pr(Sm|Ca)$

According to R&S, this instance of EER is false. Rather, the particular instance of SOT is true:

(SOT_{Sm}) $Pr(Sm \mid Ca\&E_{Sm}) = Pr(Sm|Ca)$

The reason that SOT_{Sm} is true according to R&S is simple: an estimation of Pr(Sm | Ca) using statistical techniques applied to the same sample frequency data will yield the same value for Pr(Sm | Ca&E_{Sm}). Since there are many cases that are similar to this one, in which the values for Pr(H | O) and Pr(O) are estimated from sample frequency data, the argument from the single case generalizes widely. Moreover, as methods of data collection/analysis become more advanced, the availability of the sort of maneuver that R&S exploit, will increase. R&S's argument thus dovetails nicely with the claims made by Big Data enthusiasts. Accordingly, then, it is not just SOT_{Sm} but a multitude of other specific screening-off theses of a similar form that can be defended, undermining EER as a general claim.

In light of R&S's SOT, we can construct a simple argument against IBE. The argument has as its conclusion that IBE and Bayesianism are incompatible, which is supposed to be a problem for IBE, given that Bayesianism, despite various well-known difficulties (e.g. Sober 2002), remains the

most prominent account of scientific reasoning among philosophers of science (Douven 2011). I'll refer to this as the "New Argument for Incompatibilism", or the "New Argument" for short[6]:

> **Premise 1 (P1)**: If SOT is true in a wide range of realistic cases, then IBE and Bayesianism are incompatible.
> **Premise 2 (P2)**: SOT is true in a wide range of realistic cases.
> ∴ So, IBE and Bayesianism are incompatible.[7]

P1 is true if IBE is committed to the denial of SOT as a general claim, or at least the denial of the specific instance SOT_{Sm}. P2 is supported by R&S's analysis of the smoking-and-cancer case and the claim that this analysis will generalize to other structurally similar cases involving an estimation of probabilities from sample frequency data. The significance of R&S's argument then for our purposes is that it purports to show, as Big Data proponents have begun to insist, that explanation is ultimately dispensable for scientific inference.

4. Some Extant Objections to the Screening-off Thesis

Before turning to how I think the proponent of IBE ought to respond to R&S's argument, which I will discuss in section 5, it is worth considering briefly what some of R&S's critics have had to say about SOT. Climenhaga (2017) has disputed P2 in the above argument for incompatibilism, i.e. the claim that SOT is true in the smoking-and-cancer case, and thus in a wide variety of cases. According to Climenhaga (2017, p. 366), Pr(Sm|Ca) will not be high unless Ca is conjoined with E_{Sm}, and as a result "the existence of an explanatory connection between cancer and smoking is precisely what licenses the inference from S's smoking to S's cancer."[8] With that said, according to Climenhaga, it is an objective fact about the probabilistic relations that obtain between the propositions, Sm, Ca, and E_{Sm} that, relative to our background K: $Pr(Sm|Ca\&E_{Sm}) > Pr(Sm|Ca)$—that is, SOT_{Sm} is false. Here, I will focus on Climenhaga's response, which attempts to show with formal rigor that EER_{Sm} is

[6] This is in contrast to the "old argument" for incompatibilism laid out by van Fraassen (1989).

[7] A few points of clarification are in order. First, it should be noted that R&S do not explicitly make this argument, and so they do not defend P1 of the New Argument. Second, in their defense of SOT, R&S presuppose a Bayesian conception of evidential irrelevance; however, they leave open the possibility that "there are alternative senses of evidential irrelevance on which explanatoriness is evidentially relevant" (2017, p. 582). As we will see below, this possibility can be used to push back against the New Argument for Incompatibilism.

[8] Climenhaga explicitly puts the point in terms of "epistemic" probabilities, the sort traditionally defended by Keynes (1921) and Carnap (1950). These probability statements are said to codify supposed objective relationships between propositions, and are such that if $Pr(H|O\&K) = r$, then our degree of belief in H given that our evidence is O&K ought to be set equal to *r*. Thus, the sort of Bayesianism assumed here is one in between the purely subjective and the purely the objective view. The objectivity of epistemic probabilities is supposed to be analogous to the way in which the entailment relation between propositions is objective.

true. The argument for EER_{Sm} and against SOT_{Sm} is short, relying on a few premises, and so can be summarized quickly.[9] My exposition relies heavily on R&S (2017), but see Climenhaga (2017) for the original statement.

First, following Climenhaga, let C_1 be "sometimes lung cancer after the age of 50 is caused by heavy smoking before the age of 50". Then, the argument relies on the following premises:

(16) $Pr(Sm|Ca\&E_{Sm}) = Pr(C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&C_1) + Pr(\sim C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&\sim C_1)$

(17) $Pr(Sm|Ca) = Pr(C_1|Ca)Pr(Sm|Ca\&C_1) + Pr(\sim C_1|Ca)Pr(Sm|Ca\&\sim C_1)$

(18) $Pr(C_1|Ca\&E_{Sm}) > Pr(C_1|Ca)$

(19) $Pr(Sm|Ca\&E_{Sm}\&C_1) \geq Pr(Sm|Ca\&C_1)$

(20) $Pr(Sm|Ca\&E_{Sm}\&\sim C_1) = 0$

Here, (16) and (17) follow from an application the law of total probability. (18) is true because given that S gets cancer and that S's smoking would explain S's cancer, then clearly this increases the probability that at least one instance of cancer is caused by smoking. (19) makes the plausible claim that E_{Sm} does not make Sm less probable once conjoined with $Ca\&C_1$. Finally, according to Climenhaga, (20) is true because $Ca\&E_{Sm}\&\sim C_1$ entails ~Sm, since if Sm were true, then it would follow that C_1 is true, which by hypothesis is false. In light of (20), it is evident that (16) reduces to:

(16)* $Pr(Sm|Ca\&E_{Sm}) = Pr(C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&C_1)$

Using the equivalency of Pr(Sm | Ca) in (17), it follows that SOT_{Sm} is true just in case:

(21) $Pr(C_1|Ca\&E_{Sm})Pr(Sm|Ca\&E_{Sm}\&C_1) = Pr(C_1|Ca)Pr(Sm|Ca\&C_1) + Pr(\sim C_1|Ca)Pr(Sm|Ca\&\sim C_1)$

Now then, we know by (18) that $Pr(C_1|Ca\&E_{Sm}) > Pr(C_1|Ca)$ and we know by (19) that $Pr(Sm|Ca\&E_{Sm}\&C_1) \geq Pr(Sm|Ca\&C_1)$, and that thus the left side of (21) is bigger than the first summand on the right side of (21). Therefore, we know that SOT_{Sm} is true just in case the arithmetic difference of the left side of (21) and the first summand on the right side of (21) exactly equals the second summand of the right side of (21).

On the matter of whether (21) is true, Climenhaga remarks that "while this could be the case, there is no reason to expect it a priori. Hence, far from being a general truth, if [SOT] is true in this case it is only by fortuitous coincidence" (2017, p. 368). The point seems to be that we have no reason

[9] Here, I will maintain Climenhaga's premise numbering scheme. Moreover, I ignore Climenhaga's argument against a logically independent variant of SOT, which he calls SOT*.

to think that the relevant complex probability statements will balance out so neatly as to make (21) true, and in fact it would unlikely if they did. Thus, we have no reason to expect that SOT_{Sm} will be true, which, if right, undercuts the argument that SOT is true in many situations.

R&S (2017), however, are not convinced by the above argument against (21). Instead, R&S argue that (21) is true. If perhaps we had no reliable way of estimating the relevant probabilistic expressions that feature in (21), then perhaps one might wonder why it should be the case that the values of these complex expressions should balance out so nicely such that (21) is true. If (21) were an equation that contained four random, empirical propositions A, B, C, and D, and we were in the dark concerning the values of the complex expressions that contain A, B, C, and D, then perhaps we might think it unreasonable for the values on the left side to exactly equal the values on the right.

But that we are in an analogous situation with respect to (21) is precisely what R&S deny, and so Climenhaga's argument against SOT ultimately begs the question. From R&S's perspective, we are not in the dark with respect to the values of the complex probability statements that feature in (21). Rather, according to R&S, the truth of (21) is easily established because, as noted above, the values of Pr(Sm|Ca) and Pr(Sm) are estimated from sample frequency data in some way or another. It is a central claim of R&S's that whatever estimation technique is used to infer the value of $Freq_{POPULATION}$(Sm|Ca) and thereby determine Pr(Sm|Ca), this same estimation technique will yield the same result when estimating the value of $Pr(Sm|Ca\&E_{Sm})$. Since the relevant probabilities are estimated by the same frequency data, according to R&S, SOT_{Sm} will be true. Since SOT_{Sm} is established in this way, and since SOT_{Sm} entails (21), assuming the truth of (16)-(20), it follows that (21) will also be true as a matter of mathematical necessity. If R&S are right that $Freq_{POPULATION}$(Sm|Ca) ought to be used for both Pr(Sm|Ca) and $Pr(Sm|Ca\&E_{Sm})$, then it will hardly be a fortuitous coincidence that (21) is true. Thus, despite the attempt to formalize an explanationist rebuttal to R&S, it doesn't seem that Climenhaga has said enough to refute SOT.

5. Against the New Argument for Incompatibilism

The dispute over SOT has concerned the evidential relevance of a single proposition E_{Sm} and its more general form E. Consider though, that the standard formalization of IBE is that of a four-step argument schema presented in section 3. Furthermore, notice that E—and by extension E_{Sm}—in effect captures only the content that is conveyed by the first and second steps of the schema. According to IBE as construed here, that a hypothesis would, if true, explain the data is epistemically relevant only if H_1 is sufficiently good and is better than its rival explanations H_2, H_3,...,H_n. Whether

H_1 is the better explanation depends on how well H_1 does with respect to the explanatory virtues, e.g. *simplicity*, *scope*, *precision*, etc. However, effectively, what E says is *just* that H is a potential explanation of O, where a potential explanation has all the features of an actual explanation, with the only exception being that we are not sure whether H is true. In other words, then, the thesis that R&S endorse—SOT—is that H's being a potential explanation of O *in itself* is not evidentially relevant, once the confirmatory import of O is taken into account.

But if *that* is what SOT says, then an explanationist who endorses the four-step argument pattern above *can* accept SOT, and thus P1 of the New Argument is false. Therefore, I don't dispute SOT, or P2 of the New Argument. Rather, I dispute the scope and significance of SOT, or P1 of the New Argument.[10] On standard formulations of IBE, H's being a potential explanation is not enough to be epistemically relevant, or at least appreciably epistemically relevant given a set of facts F. There are many implausible explanations that are potential explanations. For instance, that aliens visited me in the night and burglarized my home is a potential explanation of the missing beer from my refrigerator. But the explanationist need not be committed to the view that H's being a potential explanation of F *alone* is a feature that is epistemically relevant in our evaluation of H. The truth of SOT is thus not a threat to IBE. Rather, only if "H is the best explanation of O" is evidentially irrelevant in the Bayesian sense might this be a cause for concern.

However, rejecting the New Argument by pointing out that SOT is compatible with IBE construed as the four-step argument pattern in section 3 is not as significant a victory as it might seem at first blush. This is because, according to R&S, a related screening-off thesis that takes into account the fact that H is the best explanation of O is just as defensible as SOT. In the concluding section of their response to Climenhaga (2017), R&S briefly address how their argument bears on IBE as I have treated it here. There, R&S (2017, p. 589) articulate the following extension of SOT:

> **Comparative Screening-Off Thesis** (CSOT): Let H be a hypothesis, O an observation, and C the proposition that H would explain O if H and O were true, where H is better than the alternatives as a potential explanation of E. Then $Pr(H|O\&C) = Pr(H|O)$

With that said, R&S endorse the following specific instance of CSOT:

CSOT_{Sm} $\quad Pr(Sm|Ca\&C_{Sm}) = Pr(Sm|Ca)$

[10] My response to R&S is thus similar in kind, though different in its details, to that which has already been explored by McCain and Poston (2014, 2018), who can also be interpreted as rejecting P1 of the New Argument. I'll have more to say about McCain and Poston's response to R&S and its relation to my main argument below.

Here, C_{Sm} means "Sm would explain Ca if Sm and Ca were true, and Sm is the best explanation". According to R&S, the same reason that SOT_{Sm} is true applies to $CSOT_{Sm}$ as well. As before, the value of Pr(Sm|Ca) is determined by $Freq_{POPULATION}$(Sm|Ca), which is estimated from $Freq_{SAMPLE}$(Sm|Ca). The truth of the C_{Sm} will neither change the estimate of the population frequency, nor does C_{Sm} play a role in the inference from population frequencies to probabilities. The latter inference is mediated by some chance-credence calibration principle, such as the Principal Principle (Lewis 1980), in which it's not clear how C_{Sm}, or any other explanatory consideration, is at all relevant.

If R&S are right that their defense of SOT transfers to CSOT, then this again seems to pose a problem for IBE. Although the explanationist can easily rebut P1 of the New Argument by pointing out that IBE does not say that merely being a *potential explanation* is sufficient for conferring epistemic merit on a hypothesis, certainly, the explanationist will want to say that the fact that H is the best explanation of O has a great deal of evidential bearing on the truth of H. But as R&S (2017, p. 589) conclude: "If IBE is meant to apply to cases of causal explanation in which probabilities are estimated from sample frequencies, it follows that CSOT conflicts with IBE".

The proponent of IBE, however, might simply deny the antecedent of the above conditional. Whether R&S's CSOT is a problem for the explanationist depends on the "grade" of explanationism that one adopts. Lycan (2002, p. 417) helpfully delineates four different versions of explanationism of varying degrees of strength. First, there is "Weak Explanationism", which is the view that IBE can sometimes rationally justify a conclusion. The Weak view leaves open whether IBE is derived from some more fundamental form of reasoning. Second, "Sturdy Explanationism" says, like the Weak view, that IBE can rationally justify a conclusion. But the Sturdy view says in addition that IBE is a fundamental rational inference method—one which cannot be justified by a deeper inference method, such as Bayesian updating, or enumerative induction. There *may* be other fundamental inference methods, but according to the Sturdy view, IBE is at least one of them. Third, "Ferocious Explanationism" is the view according to which IBE is the *only* fundamental rational, *non-demonstrative* inference method. On this view, all other non-demonstrative reasoning requires IBE for its justification. Finally, Lycan gives the name "Holocaust Explanationism" to the view according to which IBE is the *only* fundamental, rational inference method. On this view, all reasoning, including deductive reasoning is justified, at bottom, by explanatory considerations.

In arguing for SOT, R&S do not explicitly discuss these different versions of explanationism[11], but we should distinguish the controversial view that IBE is the *only* fundamental, non-demonstrative inference method, from the more modest view that IBE is a justified way of reasoning. It is clear that R&S's argument for SOT, if sound, is an immediate threat only to the stronger versions of explanationism. Only those versions of explanationism that insist that all ampliative inference can be reduced to a type of explanatory inference will be undermined by R&S's SOT. While there are defenders of the stronger versions of explanationism[12], many adopt either the Weak view or the Sturdy view, both of which are much more plausible. For instance, Psillos acknowledges that "[n]ot all changes in the background knowledge will be based on explanatory considerations" (2002, p. 620). In the same spirit, Lipton admits that "any sensible version of [IBE] should acknowledge that there are aspects of inference that cannot be captured in these terms" (2004, p. 1). The availability of these weaker explanationist views undermines P1 of the New Argument and leaves open the possibility—contra R&S and Big Data proponents—that explanation is *sometimes* essential for scientific reasoning.

To see this point more clearly, let's suppose one accepts the Sturdy view. Then one could accept R&S's claim that explanatory considerations are evidentially irrelevant in the smoking-and-cancer case. That is, one could admit that purely statistical, non-explanatory inductions are possible, and since there are a lot of cases like this, R&S are right that often explanation is evidentially irrelevant. But, if one accepts the Sturdy view, one also thinks that there are other cases in which IBE sometimes can justify a conclusion without relying for its justification on some other inference method. Provided there are cases of this sort, then R&S's argument won't threaten the Sturdy view.

Of course, to fully defend the Sturdy view would amount to giving a full defense of IBE. Still, though, we can appeal to some illustrative examples. Take, for instance, the case of Lavoisier who, in remarking on the support for his oxygen theory of combustion, writes:

> I have deduced all the explanations from a simple principle, that pure or vital air is composed of a principle particular to it, which forms its base, and which I have named the oxygen principle, combined with the matter of fire and heat. Once this principle was admitted, the main difficulties of chemistry appeared to dissipate and vanish, and all the phenomena were explained with an astonishing simplicity [quoted in Thagard (1978, pp. 77-8)].

---

[11] Although no explicit discussion of the different versions of IBE has hitherto appeared in connection to the debate over SOT, it should be noted that at one point, R&S briefly allude to a weaker version of explanationism, admitting that their argument, of course, won't undermine views according to which, "IBE is entirely parasitic on a Bayesian calculation of posterior probabilities"(2013, p. 665, fn. 3).

[12] As Lycan (2002, p. 417) notes, the Ferocious view, although having its defenders, is "disputed by almost everyone". Surprisingly, this version of explanationism has a number of proponents, including Harman (1986), Lycan (1988), Conee and Feldman (2008), and Poston (2014).

Or consider an oft-cited passage from the end of the sixth edition of Darwin's *On the Origin of Species*:

> I have now recapitulated the chief facts and considerations which have thoroughly convinced me that species have been modified, during a long course of descent, by the preservation or the natural selection of many successive slight favourable variations. I cannot believe that a false theory would explain, as it seems to me that the theory of natural selection does explain, the several large classes of facts above specified (1872, p. 421).

The "large classes of facts" to which Darwin refers include the existence of vestigial structures—"organs bearing the stamp of inutility"—such as the useless teeth of an embryonic calf and withered beetle wings (1872, p. 420), which are best explained by the theory of common ancestry.

Without going into much detail, we can already see that these cases are sufficiently dissimilar to the smoking-and-cancer case that R&S introduce to support SOT. Here, Lavoisier does not infer some population frequency from some sample frequency and then set his rational credences accordingly. Rather, he upholds the oxygen theory of combustion over the then-rival phlogiston theory because the former could explain the phenomena much better than the latter—specifically, in a way that was markedly simpler. Likewise, Darwin appears to appeal to purely explanatory considerations in his argument against creationism. Any attempt to try to recast Lavoisier's or Darwin's argument as a simple statistical inference would prove both uncharitable and artificial.

It is doubtful that R&S would be sympathetic toward even the Sturdy explanationist view, since, in their response to McCain and Poston (2014)—i.e. "M&P"—, R&S question whether explanatoriness is ever epistemically fundamental (2014, p. 198). It is worth mentioning that M&P (2014, p. 146) make a similar point as the one pursued here, noting that certain cases of explanatory reasoning—e.g. Newton's gravitational theory explaining the orbits of the planets and Einstein's theory of general relativity explaining the precession of Mercury's perihelion—are different in kind from the smoking-and-cancer case exploited by R&S in defense of SOT. The existence of such cases, according to M&P, casts doubt on the generality of SOT. In response, R&S (2014, p. 195), argue that such cases are not convincing because "explanatoriness has no confirmational significance, once purely logical and mathematical facts are taken into account." So, it is likely that with respect to the two cases I've discussed above—and all the other cases from the history of science commonly upheld as paradigm applications of IBE—R&S would say the same thing: the confirmational import of explanatory considerations can be analyzed in terms of deeper, non-explanatory concepts, such as purely logical or probabilistic relations. [13] Even so, the argument that R&S provide for SOT in cases

[13] Given the many different ways in which explanatory reasoning manifests itself, clearly this task is easier said than done. R&S (2014, p. 195) attempt to dispatch with the Einstein and Newton cases put forward by M&P

involving frequency data is not sufficient to refute the Sturdy view or the Weak view. Thus, there are weaker explanationist views that are not shown to be untenable by R&S's argument alone. For this reason, it does not seem that advocates of the methodological revolution owing to the rise of Big Data will be able to help themselves to R&S argument against IBE. Even if there are cases in which explanatoriness is evidentially irrelevant, as I have argued, it is not clear that all cases fit this mold.

6. The Ciceronian Causal-nomological Requirement and the Necessity of Explanation[14]

Now one might wonder whether the defense of explanatory reasoning in the previous section *really* serves as a convincing response to the provocative claims made by Big Data proponents. If we grant that the case discussed by R&S is one for which justified beliefs can be acquired purely on the basis of frequency data, and if such cases are common, which does appear to be the case, then won't appeals to theory, causation, explanation, etc. at least take a back seat to purely data-driven inference, as techniques to collect and process data continue to become more and more sophisticated? The proponent of data-intensive science could grant, for example, that Sturdy Explanationism is true, and that sometimes explanatory considerations play a justificatory role, while maintaining that such inferences will become rarer and less important as data-intensive science continues to develop. We should still count data-intensive science as leading to a radical methodological shift, if it indeed turns out that explanatory reasoning is eclipsed by purely data-driven inferences, and even if it happens that explanatory reasoning does not become entirely extinct.

However, there are further reasons to be skeptical of anti-explanationist attitudes in Big Data circles, for even in the case of inferences that *appear* purely data-driven, certain explanatory hypotheses play an indispensable *enabling* role in the inference process. In the course of their critiques of R&S's argument for SOT, both Climenhaga (2017) and M&P(2014, 2018), have made similar claims on behalf of IBE. For instance, although M&P attempt to refute SOT, they argue additionally that even if SOT is true, explanatory considerations can still be evidentially relevant in a wider sense, by increasing the "resiliency" of a probability function, which "concerns how [a probability function] changes in response to new information" (2014, p. 148).[15] So too, Climenhaga points out in his critique of SOT,

---

by pointing out that these are cases of hypothetico-deductive reasoning, which can be given a straightforward Bayesian rationale; however, since Lavoisier's reasoning makes essential reference to some form of *simplicity*, it is not obvious how to analyze this instance of explanatory reasoning in non-explanationist terms. See, Cabrera (2017), for a discussion of the relationship between Bayesianism and the various explanatory virtues.

[14] In this section, I draw heavily on my analysis of Cicero's *De Divinatione* (Cabrera 2019).

[15] See R&S (2014, pp. 196-7) for a response to M&P's proposal regarding the connection between explanatory considerations and the resiliency of a probability function. R&S (2014, p. 197) remark that their SOT is "neutral

with respect to the smoking-and-cancer case, that if there is "no explanatory connection between smoking and cancer, the observed frequency data are a huge fluke. But we should not expect huge flukes to continue" (2017, p. 363). The central claim I will develop and defend in this section complements these claims about the enabling role of explanatory considerations. In particular, as I will argue, even if R&S are right that sometimes explanatoriness is irrelevant toward theory-confirmation in the Bayesian sense, nevertheless, it remains true that it is rational to rely upon correlations in *any* predictive inference *only if* one possesses certain *explanatory knowledge*.

Interestingly, this is a point that the great Roman statesman and philosopher Marcus Tullius Cicero (106-43 BCE) insisted upon in his *De Divinatione* ("On Divination"). Written between 45-44 BCE, *De Divinatione* is a philosophical dialogue containing both the author and his brother Quintus as characters examining the rationality of the practice of divination. In Book I, Quintus attempts to establish the legitimacy of divination in several ways, most notably by appealing to a number of *exempla*—examples of divination, both Roman and non-Roman, which for Quintus are unquestionable. In Book II, Marcus[16] attacks his brother's arguments with a mixture of philosophical argument and the kind of rhetorical *vituperatio* ("invective") that characterized Roman oratory.

In Book II, Marcus provides several arguments against Quintus's main thesis. The most obvious objection to Quintus' argument for divination is that haruspices (i.e. readers of animal entrails) astrologers, and other diviners simply have not been reliable predictors of the future (*Div.* 2.53). However, in addition to doubting the correlations between alleged divine signs and future outcomes, Marcus appeals to another more interesting argument. The interest of this argument lies in the fact that Marcus, in spite of his previous objections, grants Quintus his claims about the historical track record of divinatory practice. But even with this point granted, Marcus still argues that there is sufficient reason to reject divination. The reason is that there is no plausible causal-nomological connection between the signs identified by diviners and the outcomes foretold. There are several instances of this argument given by Marcus in Book II. Consider the following representative passage:

> Surely if entrails have any prophetic force, necessarily that force either is in accord with the laws of nature, or is fashioned in some way by the will and power of the gods...what possible connexion can there be with—I shall not say the gall of a chicken, whose entrails, some men assert, give very clear indications of the future, but—the liver, heart, and lungs of a sacrificial

on M&P's thesis regarding explanatoriness and evidential relevance". It is likely that they will respond similarly to the proposal I develop below. However, one worry is that if R&S admit other senses in which explanatoriness can be evidentially relevant, then the slogan that is used to characterize their thesis may end up misleading.

[16] Here, I follow Beard (1986) in using "Cicero" to denote the author of the dialogue and "Marcus" to denote the character in the dialogue.

> ox? And what natural quality is there in the entrails which enables them to indicate the future? (*Div.* 2.29).

According to this argument, even supposing that there is a high correlation between, say, diseased livers in sacrificial animals and negative outcomes for the Roman state, one should *not* employ haruspicy to predict future events. This is because there is *no* plausible causal-nomological connection, given our background knowledge, between diseased livers and politico-military missteps. How is it possible for diseased livers to cause a Roman army to be defeated? What possible law of nature could there be which links the feeding habits of chickens with the Roman army's victory or defeat?

In presenting this argument against the legitimacy of divination, Marcus seems to tacitly rely upon a substantive, philosophically interesting principle governing scientific inference. Let's call the principle that underlies the above argument the "Ciceronian Causal-nomological Requirement" (CCR). We can formulate the CCR more precisely as follows:

> (CCR) For any two logically distinct event-types *A* and *B*, it is rational to predict some token-event *b* of type *B*, on the basis of the presence of some token-event *a* of type *A*, only if, given one's background knowledge, there is a plausible causal-nomological connection between *A* and *B*.[17]

According to the CCR, even if the correlations between events of type *A* and events of type *B* is almost perfect, it is irrational to predict *b* to be highly probable on the basis of *a*, unless there is some plausible causal-nomological connection between *A* and *B*. To clarify the CCR, we need to say a bit more about what it means for there to be causal-nomological connection between *A* and *B*.

A paradigmatic instance of the sort of connection required by the CCR would be if one of the event-types, *A* and *B*, were the cause of the other. Granted, this causal connection need not be direct. It is entirely consistent with the spirit of the CCR if *A* causes some *C* which causes *B*. To put the point in the language of contemporary causal modeling, *A* need not be a *parent* of *B*, i.e. a direct cause; *A* can be an *ancestor* of *B*, i.e. an indirect cause. Perhaps, smoking cigarettes causes one to perform some further action, which is the more direct cause of lung cancer. Of course, if it were possible for *B* to be the cause of *A*, i.e. developing lung cancer after the age of 50 causes smoking a lot of cigarettes before the age of 50, then this connection would also satisfy the CCR. However, *B*'s being the cause of *A* here would violate the prohibition on "backwards causation"; future events cannot cause past events.

[17] In order to avoid trivial falsity, *A* and *B* need to be logically distinct event-types. For instance, if *A* = "is a bachelor" and *B* = "is an adult, unmarried, male", then, obviously, one should infer *B* on the basis of *A*, even though there is no causal connection between *A* and *B*. In this case, the connection between *A* and *B* is logical rather than causal; knowledge of this logical connection ensures that the inference from *A* to *B* is rational.

Now, *A*'s being the direct or indirect cause of *B*, or vice versa, are not the only conceivable ways in which there could be a causal-nomological connection between *A* and *B*. We should also allow at least a third possibility. Let's suppose there is some further event-type *C*, which happens to be the cause of both *A* and *B*. If there is some such *C*, then there will be a causal connection between *A* and *B*; but this is not because *A* is the cause of *B*, or vice versa, but rather because *A* and *B* are the joint effects of a common cause, i.e. *C*. Let *C* be the possession of some discernible genetic profile which predisposes one to be attracted to nicotine and which also tends to cause one to develop lung cancer. In the smoking-and-cancer case then, *C* is a cause of both *A* and *B*, which ultimately accounts for why we observe a correlation between *A* and *B*, i.e. between smoking and lung cancer.

In my formulation of the CCR, I include the phrase, "there exists a plausible causal-nomological connection". This phrasing suggests both a strong and a weak reading of the CCR. According to the strong reading, we need adequate evidence to justify belief in a *particular* causal explanation of the correlation. If we adopted the strong reading, then we could infer that Joe the heavy smoker will develop lung cancer *only if* we have already determined either that smoking causes lung cancer, lung cancer causes smoking, or that some common cause, e.g. a genetic predisposition, is causally responsible for both effects. According to the weak reading, we don't need strong evidence for any *particular* causal explanation. Instead, what needs to be the case is only that *some* causal story or other connecting *A* and *B* remains a live option given our background knowledge. The weak reading demands only that we be able to specify *some* potential causal-nomological connection between *A* and *B* that fits well enough with our background knowledge. Such a constraint is much easier to satisfy than what is required by the strong reading of the CCR. We might understand the constraint enshrined in the weak version of the CCR as stating that there must be some accessible, causal explanation of the correlation that exists between *A* and *B*, given our background knowledge, with a "sufficiently high" prior probability. The weakness of the weak reading varies, of course, with how exactly one understands "sufficiently high prior probability." If one's threshold *r* is high, e.g. >.5, then the weak CCR will be more difficult to satisfy than if *r* instead were low, e.g. > .1.

It is clear that the strong version of the CCR is extremely implausible. The strong version of the CCR would significantly undermine our ordinary inductive practice, much of which depends upon inferences from systematic correlations between *Fs* and *Gs*, without necessarily having strong evidence for some detailed causal explanation of those correlations. Even if we do not understand exactly how the causal mechanism between, say, smoking and lung cancer should be fleshed out, that doesn't seem to make it irrational to predict that Joe the heavy smoker will get lung cancer after the age of 50, once

we learn that he has smoked a lot of cigarettes. While the strong version of the CCR is an untenable constraint on scientific inference, there are good reasons to accept the weak version of the CCR.

To see why, first, consider Reichenbach's (1956) Principle of the Common Cause, which states, roughly, that if there is a probabilistic correlation between *A* and *B*, then either *A* caused *B*, *B* caused *A*, or *A* and *B* are the joint effects of some common cause *C*. More formally, if there exists a probabilistic correlation between *A* and *B*, i.e. $(4)\ Pr(A\&B) > Pr(A)P(B)$, and if neither *A* nor *B* is the cause of each other, then according to Reichenbach, there exists some common cause *C*, which satisfies the following probabilistic conditions:

(5) $Pr(A\&B|C) = Pr(A|C)P(B|C)$

(6) $Pr(A\&B|{\sim}C) = Pr(A|{\sim}C)P(B|{\sim}C)$

(7) $Pr(A|C) > Pr(A|{\sim}C)$

(8) $Pr(B|C) > Pr(B|{\sim}C)$

Conditions (5) and (6) ensure that *C* "screens-off" the correlation between *A* and *B*. That is, conditional on *C*, it follows that *A* and *B* are probabilistically independent. Conditions (7) and (8) ensure that *C*, being the common cause of *A* and *B*, raises the probability of *A* and *B* respectively, thereby honoring the intuition that causes raise the probability of their effects. From conditions (5)-(8) it is possible to logically deduce (4), which for Reichenbach suffices to explain (4).

While clearly something like Reichenbach's Principle of the Common Cause plays an important role in scientific reasoning, what sort of status and scope the principle has remains a matter of controversy (Arntzenius 2010). One challenge to Reichenbach's principle derives from "spurious correlations". Indeed, our ability to gather massive amounts of data has led to a "deluge" of spurious correlations (Calude and Longo 2016), those which we should not rely on for prediction. Consider, for instance, the correlation between the rise in British bread prices and Venice sea-levels over the past few centuries (Sober 2001). Although we have observed a strong correlation between British bread prices and Venice sea-levels, it is overwhelmingly likely that this odd series of trends is a fluke. As such, the British bread prices/Venice sea-levels example is often invoked as a counterexample to Reichenbach's principle, at least if the principle is interpreted as always demanding that we infer a common cause to account for some probabilistic correlation, which cannot be accounted for by separate causes. In this case, it is likely that there is no common cause that satisfies the four conditions that are part of Reichenbach's Principle of the Common Cause. Even though there is a probabilistic correlation between Venice sea-levels and British bread prices, we shouldn't postulate some intricate mechanism to explain the observations, in the sense of providing a common cause to screen-off the

correlation. We shouldn't posit, say, some nefarious international conspiracy involving the Illuminati covertly fixing the prices of British bread to match up with increases in Venice sea-levels. This is because such a causal explanation would be extremely implausible given our background knowledge.[18]

Additionally, we ought not employ our knowledge of British bread prices in the present year to predict increases in Venice sea-levels in the future, or vice versa. It is highly probable that the observed correlation is simply coincidental and thus unstable. In this case, there is no available, plausible causal-nomological connection between British bread prices and Venice sea-levels, and so the weak version of the CCR goes unsatisfied. Hence, the weak version of the CCR affords us a lucid and attractive account of why we ought not to rely on such a correlation in predictive inference.

In fact, it is along similar lines that some philosophers have objected to the radical claims often made by proponents of Big Data, according to which in the future scientists will be able to do away with causal models. As Pietsch and Wernecke (2017) argue, we ought to reject the claim, commonly advocated in data science circles, that causality will become extinct in future scientific inquiry. On their view, there's a distinction between those correlations "that can be attributed to a common cause, and then those which have arisen purely by chance"; what's more, "[c]orrelations can establish reliable predictions only in the former case" (Pietsch and Wernecke 2017, p. 49). Here, we can consider the weak version of the CCR as a more precise specification of such objections to the pretensions of data-intensive science. Those correlations that "can be attributed to a common cause" are the ones for which there exists a plausible causal-nomological connection given our background knowledge, and those correlations which "have arisen purely by chance" are the ones for which there exists no such connection. Thus, the CCR helps elaborate on some of the push-back against the claim that causal-explanatory notions will become dispensable given future developments in data-intensive science.

It is worth noting that for the sake of simplicity, I have focused on causal relations as that which turns what would otherwise be a coincidental correlation into one that is suitable for predictive inference; but what matters most is not that the relation between event-types A and B is causal *per se*, but merely that the correlation is non-accidental. This leaves open the possibility that there are nomological relations that are non-causal, which nevertheless would rationally permit one to exploit correlations in predictive inference. And there very well might be such relations. For example, it seems

[18] Drawing on Steel (2003), Climenhaga considers the possibility that *time* could be a common cause of the correlation between British bread prices and Venetian sea-levels (2017, p. 364); however, it is unclear if time *itself* can stand in causal relations, given the plausible assumption that causation is a relation between *events*. On a standard account of events (e.g. Kim 1993), events are objects instantiating a property at a time. Since time is necessary but not sufficient for being an event, it is unclear how time itself could be the cause of anything.

right to regard the principle of the constancy of the speed of light as a law of nature, although it is not obvious that this is a causal law. So too, there are well-known difficulties in regarding the laws of quantum mechanics as expressing causal relations (Norton 2003). So, provided there are plausible nomological relations (causal or otherwise), between event-types *A* and *B*, then such relations would make predictions based on a strong correlation between *A* and *B* rationally permissible.

Returning the smoking-and-cancer case, and R&S's defense of the SOT, we can say then that, while it is true that "explanatoriness" is evidentially irrelevant in the Bayesian sense, nevertheless, if it turned out that there was no plausible causal-nomological mechanism connecting smoking and lung cancer, then, according to the weak version of the CCR, it would not be rational to infer that Joe the heavy smoker will get lung cancer after the age of 50. If the CCR is not satisfied in the smoking-and-cancer example, then it would be structurally analogous to the Venice sea-levels/British bread prices example. In the Venice sea-levels/British bread prices example it does not seem warranted to predict future Venice sea-levels on the basis of present British bread prices, or vice versa, because the correlation is clearly some chance coincidence. For this reason, we should expect the correlation to be unstable and unreliable. Likewise, if the CCR is not satisfied in the smoking-and-cancer case, then we should, on pain of inconsistency, say the same thing. It is only because we think that there is some plausible causal-nomological mechanism, given our background knowledge, which connects smoking a lot of cigarettes with getting lung cancer, i.e. the former causes the latter, that it seems rational to infer that Joe the heavy smoker will probably get lung cancer. Thus, causal-explanatory factors are indeed relevant in the inferential process, even if they are irrelevant in the sense captured by SOT.[19]

Thus, it will do no good for proponents of Big Data analytics to appeal to an R&S-style argument in order to show that explanatory factors are largely irrelevant to the scientific process. Even if explanatory considerations are not directly relevant in the sense specified by the strong version of the CCR, they remain indirectly relevant, in the sense specified by the weak version of the CCR. Unless we think that some causal-nomological connection between *A* and *B* is a live option given our background knowledge, then it is irrational to rely upon correlations for predictive inference, even if the correlation is quite high. Consequently, while the development of a new data-intensive science may lead to interesting changes in scientific method, one change that we should not expect to occur

[19] Something like the CCR may be what McCain and Poston have in mind when they write that the inference in the smoking-and-cancer cases is justified only if we have a "justified belief in an unknown explanatory story" (2014, p. 150).

is the total elimination of explanatory factors from the inference process. The central insight captured by IBE, according to which explanation guides confirmation, thus survives the rise of Big Data.

7. Concluding Remarks

In this paper, I have considered to what extent future data-intensive science will be able to dispense with causal-explanatory background assumptions and hypotheses. While such claims are often made by proponents of the Big Data movement, as I have argued, we have good reason to be skeptical. For one thing, IBE remains a plausible account of the way in which science is currently practiced, and moreover, as I have argued, we should not expect explanatory reasoning to go entirely extinct any time soon. To be sure, the central insight of Roche and Sober's SOT ought not to be neglected: not all scientific inference is explanatory inference. Straightforward statistical reasoning does not exploit explanatory considerations, at least not directly. However, even if Roche and Sober's SOT is correct, this will not justify the methodological claims made by Big Data enthusiasts. As I've argued, predictive inference ought to respect the "Ciceronian Causal-nomological Requirement" (CCR), according to which in order to make use of some correlation between *A* and *B* for future predictive inference, it is necessary that there be some plausible causal-nomological connection between *A* and *B*. According to the CCR, unless there is some plausible causal-nomological connection undergirding the correlation, one may not rationally rely on the correlation for predictive purposes, even if the correlation between *A* and *B* is nearly perfect. The primary attraction of the CCR is that it helps prevent us from relying on chance, spurious correlations in making predictions, correlations such as that which exists between increases in Venice sea-levels and British bread prices. As a result of the Big Data movement, the number of spurious correlations that we have identified has dramatically increased. It is important therefore, as data-intensive science continues to develop, that we keep in mind the insights enshrined in the CCR; otherwise we might unwittingly end up practicing what is tantamount to a form divination for the technological age.

References

Arntzenius, F. (2010). "Reichenbach's Common Cause Principle," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2010/entries/physics-Rpcc/>.

Beard, M.T. (1986). "Cicero and Divination: The Formation of a Latin Discourse." *Journal of Roman Studies* 76: 33–46.

Cabrera, F. (2017). "Can there be a Bayesian Explanationism? On the Prospects of a Productive Partnership," *Synthese*, 194(4):1245–1272

Cabrera, F. (2019). "Evidence and Explanation in Cicero's On Divination", *Studies in History and Philosophy of Science Part A*, 1-25

Calude, C.S. & Longo, G. (2016). "The Deluge of Spurious Correlations in Big Data", *Foundations of Science*, 1-18, doi: 10.1007/s10699-016-9489-4.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Cicero. *On Old Age. On Friendship. On Divination.* Translated by W. A. Falconer. Loeb Classical Library 154. Cambridge, MA: Harvard University Press, 1923.

Climenhaga, N. (2017). "How Explanation Guides Confirmation", *Philosophy of Science*, 84*:* 359-368.

Conee, E., & Feldman, R. (2004). *Evidentialism: Essays in Epistemology*. Oxford University Press.

Darwin, C. (1872). *On the Origin of Species*. London: John Murray.

Douven, I. (2011). "Abduction", *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), ed. E. N. Zalta, URL = <http://plato.stanford.edu/archives/spr2011/entries/abduction/>.

Gauch, H.G. (2012). *Scientific Method In Brief*. Cambridge, UK: Cambridge University Press.

Harman, G. (1965). "The Inference to the Best Explanation," *Philosophical Review*, 74: 88–95.

Harman, G. (1986). *Change in View: Principles of Reasoning*. MIT Press.

Hey, T., Tansley, S., and Tolle, K. (2009). "Jim Grey on eScience: A transformed scientific method." In: Hey T, Tansley S and Tolle, K. (eds.) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, pp. xvii–xxxi.

Keynes, J. (1921). *A Treatise on Probability*. London: Macmillan.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. New York: Cambridge University Press.

Kitchin, R. (2014). "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1 (1): 1-12.

Leonelli, S. (2012). "Introduction: Making sense of data-driven research in the biological and biomedical sciences,"*Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1): 1-3.

Lewis, D. (1980). "A Subjectivist's Guide to Objective Chance," *The University of Western Ontario Series in Philosophy of Science*, 15: 267-297.

Lipton, P. (2004). *Inference to the Best Explanation*, *2nd ed.* New York: Routledge.

Lycan, W.G. (1988). *Judgement and justification*. Cambridge: Cambridge University Press.

Lycan, W.G. (2002). "Explanation and Epistemology", in Paul Moser (ed.), *The Oxford Handbook of Epistemology*. Oxford: Oxford University Press.

Mayer-Schonberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray Publisher.

Mazzocchi F. (2015). "Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science," *EMBO Rep.*, 16(10): 1250–5.

McCain, K. & Poston, T. (2014). "Why Explanatoriness is Evidentially Relevant," *Thought* 3: 145-53.

McCain, K., & Poston, T. (2018). "The Evidential Impact of Explanatory Considerations", in McCain and Poston (eds.), *Best Explanations: New Essays on Inference to the Best Explanation.* Oxford: Oxford University Press, 121-29.

Mittelstadt, B., & Floridi, L. (2016). "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts," *Science and Engineering Ethics*, 22 (2): 303-341.

Northcott, E. (forthcoming). "Big data and prediction: Four case studies", *Studies in History and Philosophy of Science*, https://doi.org/10.1016/j.shpsa.2019.09.002.

Norton, J.D. "Causation as Folk Science," *Philosophers' Imprint*, 3(4): 1-22.

Okasha, S. (2000). "Van Fraassen's Critique of Inference to the Best Explanation," *Studies in the History and Philosophy of Science,* 31: 691-710.

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

Pietsch W., & Wernecke, J. (2017). "Introduction: Ten Theses on Big Data and Computability," In: Pietsch W., Wernecke J., Ott M.(eds.), *Berechenbarkeit der Welt?*. Springer VS, Wiesbaden.

Poston, T. (2014). *Reason & Explanation: A Defense of Explanatory Coherentism*. New York: Palgrave-MacMillan.

Psillos, S. (2002). "Simply the Best: A Case for Abduction," in A. C. Kakas and F. Sadri (eds.), *Computational Logic: Logic Programming and Beyond*. Berlin: Springer-Verlag, 605-26.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

Roche, W. & Sober, E. (2013). "Explanatoriness is evidentially irrelevant, or inference to the best explanation meets Bayesian confirmation theory," *Analysis*, 73: 659-668.

Roche, W. & Sober, E. (2014). "Explanatoriness and Evidence: A Reply to McCain and Poston," *Thought*, 3(3):193–199.

Roche, W. & Sober, E. (2017). "Is explanatoriness a guide to confirmation? A reply to Climenhaga," *Journal for General Philosophy of Science*, 48(4): 581–590.

Siegel, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken: Wiley.

Sober, E. (2001). "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause," *Brit. J. Phil. Sci.*, 52(2):331–346.

Sober, E. (2002). "Bayesianism—Its Scope and Limits" in R. Swinburne, (ed.), *Bayes' Theorem*, *Proceedings of the British Academy Press*, 113: 21–38.

Steel, D. (2003). "Making Time Stand Still: A Response to Sober's Counter-Example to the Principle of the Common Cause", *The British Journal for the Philosophy of Science*, 54(2): 309–17.

Strevens, M. (2006). "Explanation," in D.M. Borchert (ed.), *Encyclopedia of Philosophy*, *2nd ed.* Detroit: Macmillan, 518–27.

Thagard, P. (1978). The Best Explanation: Criteria for Theory Choice," *The Journal of Philosophy*, 75(2): 76-92.

van Fraassen, B.C. (1989). *Laws and Symmetry*. Oxford: Oxford University Press.