Testing and Discovery: Responding to Challenges to Digital Philosophy of Science

Charles H. Pence[*]

Preprint, forthcoming in *Metaphilosophy* (please cite only published version)[†]

Abstract

For all that digital methods – including network visualization, text analysis, and others – have begun to show extensive promise in philosophical contexts, a tension remains between two uses of those tools that have often been taken to incompatible, or at least to engage in a kind of trade-off: the discovery of new hypotheses and the testing of already-formulated positions. I present this basic distinction, then explore ways to resolve this tension with the help of two interdisciplinary case studies, taken from preregistration in contemporary science and the debate over whig history in the history of science. These case studies, I argue, refocus our attention from a mutually exclusive testing/discovery binary to the relationship between our background data or philosophical views and the empirical generalizations that we might draw from that data. Finally, I develop a set of three challenges for philosophers and corresponding avenues for future work that will, I hope, allow us to better justify our use of these methods.

Keywords: digital philosophy; philosophy of science; hypothesis testing; preregistration; whig history; digital humanities

1 Introduction

While digital humanities is not a new discipline (precursors may be found as far back as Garfield 1955; de Solla Price 1965; or Busa 1980), its application has exploded in recent years, thanks in no small part to the advance of, on the one hand, digitzation methods that have enabled us to access large corpora of text, images, social media posts, and other source material that have the potential to radically reshape our understanding of a number of traditional questions in the humanities; and, on the other hand, increasing access to computational power and advances in algorithms for the analysis of this data that have enabled us to more easily draw relevant inferences

[*] corresponding author, Université catholique de Louvain, Institut supérieur de philosophie, charles@charlespence.net

[†] My thanks the audience at the New Directions in Metaphilosophy conference, especially Yafeng Shan, Stacie Friend, Jon Williamson, and Timothy Williamson; an audience at the University of Groningen Summer School on Methodology in the History of Philosophy, especially S. J. Donker, Ruth Boeker, Laura Georgescu, Christian Henkel, Hugo Hogenbirk, Nicola Ruschena, and Andrea Sangiacomo; and two anonymous reviewers for this journal.

for study in the humanities. This has been no less true in the case of philosophy. Numerous texts in the history of philosophy have now been digitized, in part through projects like Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO). To this can be added the increasing volume of journal literature in contemporary philosophy that is now available digitally, as well as native-digital philosophical resources such as the Stanford Encyclopedia of Philosophy. These have been matched by an increasing accessibility of sophisticated methods of textual analysis, network analysis, visualization, bibliometrics, and more, all of which can help us to find empirically grounded responses to traditional philosophical questions.

The possibilities for the use of digital methods are all the more apparent for branches of philosophy – most emblematically, and my focus here, the philosophy of science, though also moral philosophy, philosophy of religion, experimental philosophy, and others – that engage with bodies of knowledge generated in other disciplines, as well as for the history of philosophy, where the relevant body of knowledge would be the prior products of philosophers themselves.[1] For philosophers of science, these methods could provide new ways of looking at the work done by scientific practitioners, whether through the lens of their journal articles (Ramsey and Pence 2016), laboratory notebooks, social media accounts (Rogers 2013), or results in the form of large collections of scientific data (Leonelli 2016).[2]

For this promise to pay off, however, we need to critically consider exactly when and how these tools can be best applied. As with every technological fix in any domain of human inquiry, when we are possessed of such tools we run the serious danger of either a hasty rejection of novel methods, or an equally hasty inference that digital approaches will be a panacea for every question in the future of philosophy.

What exactly, then, are these tools supposed to be used for? In this paper, I will consider two very common answers to this question – that digital humanities should be used for the *testing of philosophical hypotheses,* and that they should be used for the *discovery of new hypotheses.* Each of these uses is obvious, and as I hope to demonstrate, both are essential for the full promise of digital philosophy to be realized. That said, there is a certain apparent tension between the two. As we will see, one cannot simultaneously use the same data to derive a hypothesis and test that same hypothesis – a fact which has led some to call for the wholesale rejection of discovery in favor of hypothesis testing.

Drawing on two case studies from elsewhere – in data-driven science and in history – I will attempt to offer a new way to think about, and potentially resolve, this apparent tension between testing

[1] I regret that I lack the space to extensively pursue the connections between this work and parts of philosophy beyond the philosophy of science. While my focus here will be thus be limited, I hope readers will be able to see how this might generalize to these other areas where digital work can provide a valuable "input" to philosophical reflection.

[2] I am using "philosophy of science" here to stand for a wide array of allied kinds of philosophical pursuits. The precise details of the ways in which digital methods will be useful will certainly depend on whether the effort is best understood as "contemporary philosophy of science (PoS)," "history of philosophy of science (HOPOS)," or "history and philosophy of science (HPS)." The data upon which we rely, for example, might be either the contemporary scientific process (PoS), the historical works of scientists (HPS), or the historical works of philosophers of science (HOPOS). I'll stick with "philosophy of science" for brevity's sake in the rest of this article, but I believe that the considerations that I raise here will be valid for any of these approaches to philosophical work. (My thanks to Laura Georgescu for encouraging me to bring out this point.)

and discovery, by more precisely illuminating exactly what it is that's at stake in this controversy. While the case studies that I will discuss seem at first to be relatively disperse, and are drawn from different fields with different concerns, I will argue that in fact they both derive from the same kind of underlying concern. That is, they are both, in the end, reminders that we must explore the relationship between our background, preexisting philosophical positions and the kind of inferences that we might hope to draw in digital philosophy. In claiming a mantle of "empiricism" or (more problematically) "objectivity" for these digital results – a claim that, whether explicit or implied, is practically inescapable in some form or another – we must be very sure that we are aware of the ways in which those empirical conclusions might depend upon our extant conceptual structure.

In doing so, I argue, we will hopefully find means to take advantage of *both* hypothesis-discovery and hypothesis-testing, without abandoning either one wholesale. I conclude by raising a handful of questions as targets for future work, which, while largely unresolved, will, I claim, form a crucial companion to applications of digital philosophy in the years to come.

2 What is Digital Philosophy?

While applications of digital philosophy are now becoming ubiquitous enough that many readers will be familiar with them already, it merits a brief pause here to describe the kinds of studies that I have in mind.[3] Perhaps the most common use of these tools thus far has been in what we might call "mapping" of the field of philosophy itself. What kinds of questions have philosophers been interested in, and when? How have these trends changed over time, and how has the reflection of them in our published articles matched or differed from the accounts of our own field that we tell in the history of philosophy? To point to just one particularly striking example of such work, Malaterre et al. (2019) have constructed a corpus of articles from the journal *Philosophy of Science* spanning seventy years (Figure 1). Their work uses a process known as *topic modeling,* on which an unsupervised algorithm (i.e., an algorithm that does not require its users to input significant amounts of domain knowledge in advance) breaks a corpus up into a collection of topics, roughly analogous to the subjects that each article might discuss (Blei 2012). Such a topic model shows us some classic features of the history of the philosophy of science that we might expect – for instance, the decreasing importance of logic and philosophy of language for philosophy of science (indicating their increasing independence as subdisciplines of philosophy), and the explosion of philosophy of biology beginning in the 1970s.

But these tools are not only useful for a retrospective understanding of philosophy – they also can help us test extant philosophical claims. To take one recent example, Moti Mizrahi (2020) has considered whether scientific publications tend to offer support for or counterexamples to some of the most common accounts of scientific progress. It should be clear enough that a philosophical theory about the nature of "progress" in the sciences (for instance, that science progresses in terms of approach to the truth, or increased understanding, or accumulation of empirical knowledge) should have at least some kind of empirical upshot – if it is right, scientists should tend to describe the lasting consequences of their work in certain ways (say, using terms connected to knowledge,

[3] Anyone not in need of such a refresher may feel free to move on to the next section.

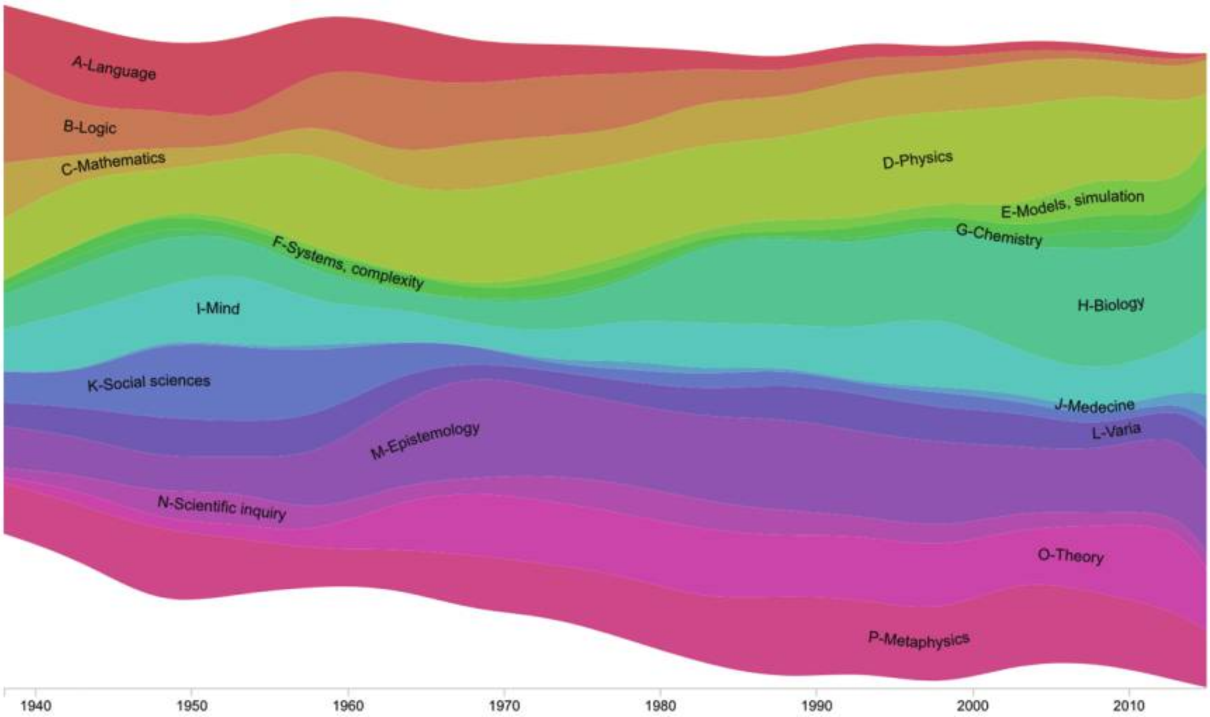

Figure 1: The evolution over time of large categories of topics in the journal *Philosophy of Science*, from 1934 until 2015. Figure 2 from Malaterre, Chartier, and Pulizzotto (2019). (**NOTE:** Need to secure reproduction rights for this image prior to publication of article.)

understanding, or truth), and avoid describing those consequences in other ways. While Mizrahi's analyses are not definitive, they seem to offer evidence against truth-based views of scientific progress and for knowledge- or understanding-based views.

Third and finally, we might see these tools as useful for the generation of new philosophical hypotheses. In my own prior work, I have evaluated what I call the "network of discourse" surrounding a debate over the nature of heredity in the history of biology at the turn of the twentieth century (Pence in press). This discourse network can be understood as the picture we get when we consider authors to be "connected" each time one author mentions another in a particular context – here a few decades from the correspondence pages of the journal *Nature.* As it turns out, such a network does not merely recapitulate the already-known networks of "allies" in this debate, or networks of training and mentorship, also already well understood (Kim 1994). On the contrary, what appears is a different network (Figure 2), on which authors that are heavily invested in debating one another actually seem to *remove themselves* from the mainstream of scientific discourse, leading me to propose, at least tentatively (more case studies being certainly required) a sort of "professional debater" or "paradigm warrior" category of social actor in the development of scientific theories.

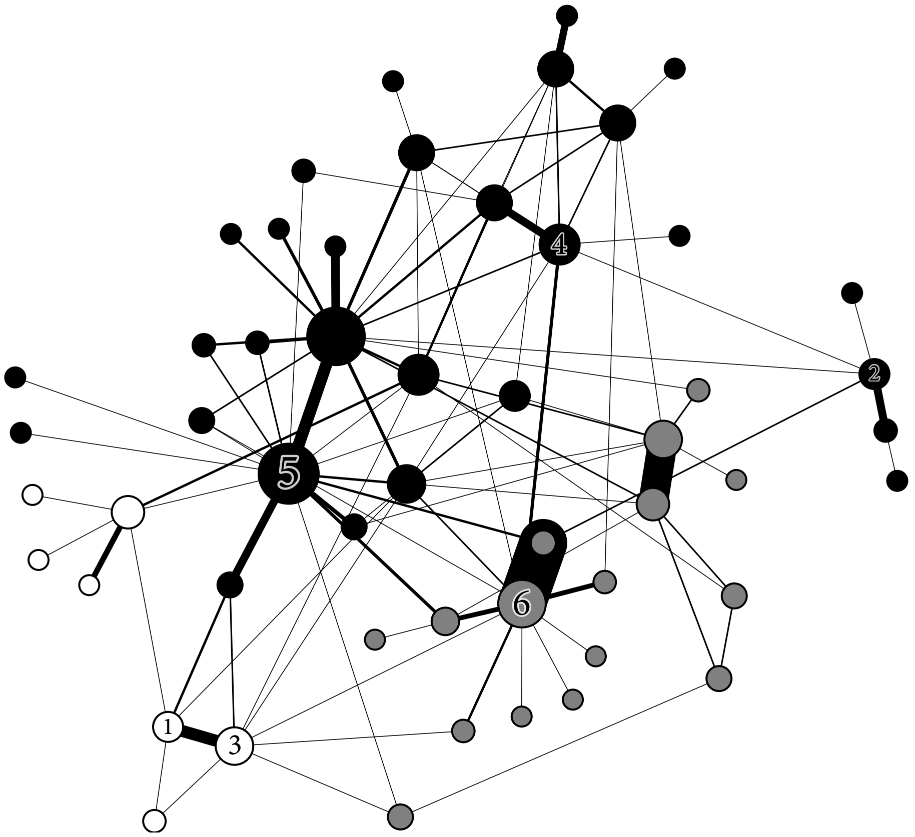

Figure 2: The network of discourse in *Nature,* from 1900 to 1904. W.F.R. Weldon and William Bateson, the two central players in the debate at issue, are labelled 1 and 3, and we can see them relatively isolated from the rest of discussion in these letters. Figure 4 from Pence (in press).

Again, it is not my purpose here to evaluate the merits of these particular examples. But I hope that this brief tour of some recent examples of digital philosophy of science can provide a window into the potential of these methods for philosophical investigation. In all these cases, the idea is to begin treating the outputs of the scientific or philosophical process as empirical data, aided in each case by our access to massive databases of those outputs and the kinds of analytical tools needed to understand them in compelling and useful ways.

3 Hypothesis Testing and Discovery

Even this short presentation, however, gives rise to a very important question that is in some sense an obvious one to pose to any new methodology: What is it that these methods are good for?

The answer to this question is less straightforward than it might seem, perhaps highlighted by the different aims of Mizrahi's and my own work as just discussed above. Clearly, these digital tools are extremely useful for the discovery of new and unexpected trends in the data, which might readily lead us to propose novel philosophical hypotheses in order to explain them. I tentatively offered an amendment to the standard account of the social structure of the historical controversy that I discussed, adding to it another class of scientific actor that could, at least potentially, be revealing in the analysis of a number of other vitriolic and public scientific controversies.

Indeed, the cultivation of what we might call "serendipitous discovery" is often an explicit target of scholars working in the digital humanities.[4] This focus derives from a related worry that one might have about the proliferation of digital technologies, hyperlinked references, and so forth: the process of searching for and finding academic information has radically changed. In the process, we have lost some of the randomness that at times led to important academic insight. Think, for instance, of looking for a book in the library, only to come away with several more nearby on the shelf, or of picking up a paper issue of a journal and reading more than the one sought-for article. With this in mind, dedicated efforts to design browsing or analysis systems that encourage such serendipity – think of an academic version of the "more items like this one" found at many online retailers – have been undertaken.

That said, any formulation of a new hypothesis *after* consulting a particularly large dataset comes with a serious problem. If one has a dataset of sufficient size, it's essentially guaranteed that it will be filled with what have become known as "spurious correlations." It has already been noted in a digital humanities context by Manovich *et al.* that any methodologies targeted at sampling of large datasets (in their case, of images) run the risk of remaining, nonetheless, non-representative (Manovich 2012, 259).[5] And as Calude and Longo have formally demonstrated, "the more data, the more arbitrary, meaningless and useless (for future action) correlations will be found in them," where here a meaningless correlation is defined as one that could be produced in a randomly generated dataset with no relation to real-world, empirical fact (Calude and Longo 2017, 600). The

[4] For instance, Deb Verhoeven, following Martin Weller, discusses the expansion of discovery as an explicit role for digital humanities work (Arthur and Bode 2014, 210); in the same volume, Sydney Shep underlines "search and discovery" as two of the key challenges "in big data sources that often deliver masses of unfiltered hits and whose subsequent systematic reorganization mirrors existing knowledge structures or assumptions" (Arthur and Bode 2014, 79).

[5] Strikingly, Manovich presents similar problems not only for the identification of trends within such a corpus of images, but also for the visualization over time of such corpora.

presence of such spurious correlations has as an obvious consequence that one cannot simply "read off" theoretical commitments from a dataset, no matter its size or how carefully it was constructed. But more subtly than this, it also entails that we have an obligation to engage in at least some degree of active prevention: how can we be sure that the promising and novel hypothesis our analysis has shown us is not just the result of a fortuitous coincidence of bits?[6]

There are, to be sure, many ways in which one might respond to this challenge – but one of the most common propositions has been a fairly extreme one. Many have called for a wholesale turn to hypothesis-driven research in these kinds of big-data contexts.[7] That is, if we simply *abandon* the use of digital tools as serendipitous avenues for discovery, and do not employ such large datasets until and unless we have formulated a prior hypothesis, then we can honestly claim that any results we produce serve as unbiased tests of that hypothesis.[8] Think here of the work of Mizrahi mentioned above, which took as its starting point the collection of existing philosophical theories concerning scientific progress, and used textual analysis as a way to offer evidence for or against each one.

The foregoing has been too vague, however, and these problems deserve a further degree of exploration. What exactly is it that's taken to be actually (or potentially) the flaw in these cases of serendipitous discovery, such that the generalizations produced are invalid, and how is it that hypothesis-driven research might avoid them? To close this section, I want to lay out three ways in which we might imagine what "goes wrong" in problematic cases, and then, in the next two sections, I will consider connections with two other fields that might give us tools useful for elaborating and better understanding them.

First, we might be worried that we are interpreting our data through a pre-existing theoretical frame or on the basis of an already-formulated hypothesis (or even just a "hunch"). In that case, we would run the serious risk of conflating *theory construction* with *theory testing.* After all, it is no surprise if the very data that we had in mind when we *derived* a given theory would go on to be compatible with that theory; we cannot then claim that those same data serve to test or confirm it. This is an old problem in the philosophy of science; there is a long debate concerning whether there is a difference between "predicted" evidence and merely "accommodated" evidence, with no real consensus concerning just exactly what the difference between the two consists in (Douglas 2009 contains a nice summary). However it is worked out in the details, it seems that the construction/testing relationship here is at best extremely complex, and at worst could threaten to undermine any work that does not take this distinction seriously.

Second, we might fear that being in some sense *too flexible* in our methods of analysis could lead to

[6] It is also important to note that, in addition to the size of our datasets leading to problems of spurious correlation – the clearest and least escapable such problem, and hence the example that I detail here – we also need to keep in mind that no dataset can possibly be constructed in a "neutral" manner (boyd and Crawford 2012), and no analysis tool is "free" of assumptions about the data which it analyzes (for one striking recent example, see Buolamwini and Gebru 2018), leading to yet more sources of skepticism about the validity of such generalizations.

[7] The arguments that I present for this claim in this section come largely from the sciences, but they have been echoed in the digital humanities literature as well; see, for instance, Ted Underwood (2017).

[8] This is not to say that interpreting the results of those analyses automatically becomes straightforward, but at least we cannot be accused of cherry-picking favorable evidence. We will see an example of an extreme view that entirely rejects discovery in the next section.

biased conclusions. It is increasingly recognized that even absent cases of outright academic fraud, researchers of good faith still operate in an environment of extreme external pressure – to publish, obtain positions and grants, and so forth. These kinds of pressures can lead even scrupulous researchers to take advantage of the freedom present within standard scientific practices to make methodological "tweaks" to render their work more attractive (Smaldino and McElreath 2016). For instance, one might stop an analysis short as soon as one finds a result that looks "interesting," not evaluate a full range of parameters for a given algorithm because an early attempt met expectations, or make a series of methodological choices with the goal of rendering a hoped-for result "clearer" or "more perspicuous." Again, such tweaks don't (necessarily) constitute outright fraud – but they can nonetheless provide a path by which biases could enter into research practice via *prima facie* innocuous methodological decisions.

Third, perhaps most nebulously, we might be worried that we're not approaching our source material with an apt set of concepts – that in engaging in digital analysis at all, we're failing to engage with that material on its own terms. This a relatively slippery idea, but we might imagine it in line with what Christia Mercer has called the Getting Things Right Constraint in the history of philosophy: the injunction that "historians of philosophy should not attribute claims or ideas to historical figures without concern for whether or not they are ones the figures would recognize as their own" (2019, 530). As she notes, such a principle is essentially second nature in contemporary work on early modern philosophy (her case study). While it is perhaps not quite as ubiquitous in the philosophy of science, one can certainly find all of the undercurrents of, for instance, distrust in rational reconstruction and concern for the views of practitioners across contemporary philosophy of science, especially in areas such as the burgeoning philosophy of science in practice movement (Soler et al. 2014). We might thus find at least some apparent support for a modified version of the Getting Things Right Constraint – philosophers of science should not attribute claims or ideas to practicing scientists without concern for whether or not they are ones those scientists would recognize as their own.[9] If digital methodologies do really threaten this increasingly widely accepted principle, then this would constitute good reason to be skeptical of them.

4 The "Preregistration Revolution"

How might we begin to approach these three interrelated problems surrounding hypothesis and discovery in digital philosophy of science? In this section and the next, I want to turn to two case studies, taken from disciplines outside of philosophy, to draw out some ways in which philosophers might begin to engage constructively with these kinds of worries, moving past a view of testing and discovery as mutually exclusive.

To begin, let's consider an example from contemporary, data-driven empirical science: *preregistration.* The concept of preregistration first began to take hold in the life and psychological sciences, both because of the massive amounts of data that these sciences began to generate in the 2000s (especially with the advent of inexpensive, fast DNA sequencing; for a humorous take on the situation, see Sagoff 2019), and the failure of several high-profile results to replicate when tested by

[9] As Mercer herself underlines (2019, 300), because this principle is phrased in negative terms (a kind of philosophy *not* to do), there are numerous ways of practicing philosophy consistent with such a principle.

multiple groups (Munafò et al. 2017). The basic idea is this: if scientists begin a research project by publicly stating the hypotheses to be tested, the ways in which those tests will be undertaken, and the empirical results that would indicate either confirmation or refutation of those hypotheses, then many of the worries surrounding the first two problems I mentioned in the last section (conflation of theory construction and testing, and flexibility in analysis) would be circumvented.

Preregistration has been approached by scientists in surprisingly philosophical terms, and at times with an almost religious fervor. Nosek et al., for instance, begin a theoretical introduction to preregistration by introducing the distinction between prediction and accommodation (which they call postdiction). They continue:

> To make confident inferences, it is important to know which is which. Preregistration *solves* this challenge by requiring researchers to state how they will analyze the data before they observe it, allowing them to confront a prediction with the possibility of being wrong. (Nosek et al. 2018, 2605, emphasis added)

Notably, the advantages of preregistration will vary depending upon what exactly it is that we are preregistering. On the one hand, preregistration of the hypothesis we hope to test targets the first problem I mentioned above. As Alison Ledgerwood puts it, this sort of preregistration ensures that "we should only adjust our confidence in a theory in response to evidence that was not itself used to construct the theoretical prediction in question" (Ledgerwood 2018, E10516). Preregistration of a complete plan of experimental analysis, on the other hand, targets the second worry above, that the data have not unduly influenced our choice of algorithm or methodology of analysis; as "flexibility in researcher decisions can inflate the risk of false positives" (Ledgerwood 2018, E10516).

While preregistration might seem more difficult to envisage in a digital-humanities context than a scientific one, the practice could certainly be adapted and utilized to good effect. Preregistering the scope of a corpus, for instance, would guard against the addition of further documents after initial analyses if expected results failed to materialize. Preregistering the plan for applying a given algorithm – think, for instance, of detailing the ways in which the parameters for a topic model would be tuned, the number of topics selected, their quality or coherence evaluated, etc. – could potentially alleviate worries that these choices rely too much on personal preference. Despite the fact that human interpretation of topic model coherence, among many other examples, remains "the gold standard" (Röder, Both, and Hinneburg 2015), it can be difficult to convince peer reviewers that such human-driven choices are well-motivated. Making the limits and role of subjective evaluation particularly clear, and declaring them in advance – rendering moot any objection that these choices could have been made with the intent of influencing study outcomes – could make this justificatory process easier.

That said, I do not want to argue here that preregistration is necessarily a silver bullet for these two problems in either scientific or philosophical practice. Perhaps the largest issue for the application of preregistration to the digital humanities concerns the very nature of *replication* itself. To the extent that preregistration in the natural sciences is often targeted at the resolution of failures of replucation, this is simply not an issue that is relevant for research in the humanities. There is only one sequence of historical events in the sciences, and only one Kuhnian theory of scientific revolutions – we cannot perform an experiment to see if perhaps a different way of understanding the natural world (say, one with no Newton) could have yielded a different evaluation of Kuhn's approach to theory change. Further, we can see that preregistration has often been presented in precisely the way that I have already cautioned against above, namely, as rejecting entirely the use of digital tools for discovery, in favor of pure hypothesis testing.

I think two potential responses to these critiques are important to highlight here. First, as I already noted above, acknowledging that there is no clear crisis of "replication" in the humanities is not to say that there are not *other* uses for preregistration that would be useful in philosophical contexts. Philosophers using digital methods are just as subject as our scientific colleagues to worries surrounding, as I discussed in the last section, for instance, the potential for our unforced choices in methodology to constitute a path for the introduction of bias. To return to the example I discussed just above, imagine that in preparing a topic model, I am interested at least in part in the dynamics of a specific concept within my corpus. If the first model I generate has a reasonable coherence score and seems to be free of meaningless or duplicate topics, and my concept of interest is particularly well picked out by one specific topic, there will be a natural pressure to accept this model and move on. This wouldn't in any straightforward sense be a fraudulent practice – this kind of judgment call is found throughout this kind of work. But it is worth questioning whether the presence of my desired concept in the model in fact exerted an undue influence on my choice, one that could have been avoided had I preregistered a plan for the evaluation and selection of models.

Second, I think that a careful look at the motivation behind preregistration begins to distill an important way to frame any potential response to all three of the problems that I detailed in section 3. Preregistration draws our attention to the importance of the relationship between the data that we have derived and the generalizations and analyses that we might produce from them, in at least two different ways. First, in what ways have the very data themselves influenced the generalization drawn from them? Can those data be said to be testing that generalization, or not? And second, have those data influenced the choice of analysis method or the parameters used to produce the generalization itself? As I will reconstruct the state of play a bit later on, I believe this question of the influences or relationship that holds between the data and these other aspects of our philosophical work is precisely the one that we need to ask – it opens space for a variety of more complex and nuanced answers that can better allow us to explore the use of both serendipitous discovery and hypothesis testing.

5 The Whig Interpretation of History

In 1931, the British historian of politics and science Herbert Butterfield, otherwise primarily known for works on the history of Christianity in England, published a short volume entitled *The Whig Interpretation of History.* Butterfield was bothered by what he took to be a destructive characteristic shared by many political histories of the day,

> the tendency in many historians to write on the side of Protestants and Whigs, to praise revolutions provided they have been successful, to emphasize certain principles of progress in the past and to produce a story which is the ratification if not the glorification of the present. (Butterfield 1931, v)

Such histories – *whiggish* histories – take both the success and the moral rectitude of our present moment for granted. The job of the historian, Butterfield decries, becomes to write a history that can explain how it is that we are now so right, and how it was that so many historical actors could have been so wrong for so long. The problems with any such story are manifold, but perhaps the most succinct way to put the problem is that laid out by the environmental historian William Cronon:

> Thanks in part to Butterfield, we now recognize such narratives as teleological, and we rightly suspect them of doing violence to the past by understanding and judging it with reference to anachronistic values in the present, however dear those values may be to our own hearts. (Cronon 2012, 5)

History is a deeply contingent affair; any approach that identifies within it "goals," whether descriptive states of affairs or normative clusters of values, should be kept at arm's length.

Butterfield himself is a difficult figure to parse, and exactly how historians across the twentieth century responded to his work makes for a fairly complex story (Sewell 2003). But the way in which Butterfield's caution was taken up by one particular discipline – the history of science – is not a complex story. The perils of whiggism are perhaps the most serious in crafting the history of science, where the pressure of contemporary theory weighs heavily and a realist epistemology could lead one to think that we are, in fact, steadily approximating the truth in the long run (about which more later). In that sense, the rejection of whiggism has been widely and deeply adopted by the community of historians of science. As Michael Gordin puts it, "in some ways, a militant hostility to whiggish narratives *defines* the history of science against other fields, and one can often spot historians of science at a talk when they query the potentially Whiggish approach of a speaker in, say, military or legal history" (Gordin 2014, 417).

Such an adherence makes a great deal of sense – one could surely point to numerous examples of whiggish histories that entirely obscure the actual practice of historical scientists.[10] But as David Hull rightly notes, an opposition to whiggism is only a negative constraint on the practice of the history of science, and just what is supposed to replace it is far less clear. The choice simply to accumulate historical facts and interpret them as minimally as possible – a sort of maximally "anti-whiggish" interpretation – seems to be no better solution. In Hull's words, "an inductivist philosophy of history is no less a philosophy of history because it is inductivist and widely shared by other historians" (Hull 1979, 2). At the very least, such an interpretation *is still an interpretation,* and thus owes us a justification every bit as much as any whiggish interpretation would.

This was not lost on Butterfield in his original critique. He wrote there that

> Our assumptions do not matter if we are conscious that they are assumptions, but the most fallacious thing in the world is to organize our historical knowledge upon an assumption without realizing what we are doing, and then to make inferences from that organization and claim that these are the voice of history. It is at this point that we tend to fall into what I have nicknamed the whig fallacy. (Butterfield 1931, 23–24)

Again, it is not the presence of assumptions *per se* that is the problem – it is the relationship

[10] The first chapters of introductory science textbooks are particularly likely to be offensive in this regard.

between those background assumptions and the generalizations which follow after them that is what counts. We need not search for "a dispassionate scientific understanding of the past" (Jardine 2003, 132) or attempt to deny the fact that "the histories we write typically end somewhere different from where they begin" (Cronon 2012, 5). Rather, we have no choice but to embrace and defend the interpretive choices that we make.

It is here, I claim, that we find a commonality between preregistration and the example of whiggish history. Again, our attention is drawn to the question of what kinds of background theories we should "let into" our narratives and how we should justify our having done so. In section 3, I mentioned Mercer's introduction of the Getting Things Right Constraint in the history of early modern philosophy. The argument that she goes on to make, however, is not to indict people for failing to adhere to this constraint. On the contrary (rephrased for our case here), she argues that the second-order question of whether or not to be "whiggish" is far less interesting, as it happens, than the first-order question of which kinds of context or background should in fact inform a particular explanation of interest (Mercer 2019).

We could see a number of ways in which this concern about whiggishness might manifest itself in work in digital philosophy. Any time that textual analyses are extended across a long timespan, we run the risk that, whether for technical reasons (e.g., the dramatically larger number of articles and books published in recent decades) or conceptual ones (the analysis of conceptual or disciplinary structures, say, that only make sense in a contemporary context), the categories that we use to analyze sometimes quite historically remote texts will apply only with some degree of infelicity to the texts analyzed. The same could be said for citation-network or other scientometric analyses across large temporal, subject-area, or geographic scale – systems of publication, collaboration, mentoring, and training are highly situated both in time and space, and it requires careful attention to detail to avoid precisely the kind of unjust extrapolation that Butterfield has in mind.

Of course, a host of ways to respond to this charge have been deployed in the historical literature, and I want to close this section by picking up on a provocative idea from the biologist and historian of science Ernst Mayr. Mayr spent much ink in the 1980s and 1990s defending himself against (entirely deserved) charges of whiggishness in his historical works on the life sciences. To justify his work, however, he pointed to particular features of the relationship between the subject matter of the history of science and the generalizations drawn from that subject matter:

> [The charge of whiggishness] was based on the erroneous assumption that a sequence of theory changes in science is of the same nature as a sequence of political changes. Actually the two kinds of changes are in many respects very different from each other. …[I]n a succession of theories dealing with the same scientific problem each step benefits from the new insights acquired by the preceding step and builds on it. (Mayr 1990, 302)

Put differently, Mayr is appealing to the very nature of science itself – for all that he is here using a naive, "accumulation of facts" picture of the scientific process that is today rather discredited – in order to claim that a certain kind of relationship between his background knowledge (namely, his knowledge that the future history of science would turn out in a particular way) is indeed relevant and acceptable to draw upon in telling that historical tale (rather than perniciously whiggish). I'll consider a potential philosophical analogue to this kind of move in the next section.

To sum up the last two sections, then, I have argued that we can interpret the apparent tension between hypothesis and discovery in digital methods in the philosophy of science instead as a set of demands for the explanation of relationships or influences between our philosophical presuppositions and data on the one hand, and the generalizations that we draw from those data on the other. With Butterfield, we can question which of the influences that our theoretical or philosophical background might have on those results are in fact legitimate. With Nosek and Ledgerwood, we can interrogate the relationship between the data themselves and the inferences drawn from them. And with Mayr, we can more speculatively imagine what characteristics of the material that we're aiming to describe might be relevant for our conclusions.

6 From Demands to Open Challenges

In this section, I want to draw on another piece of recent work in which several co-authors and I laid out an approach to understanding the place of digital analyses in a more general framework for doing empirical philosophy of science (Lean, Rivelli, and Pence 2021). While the full structure of that paper's argument will not be necessary to my purposes here, I want to borrow its central reconstruction of what exactly that process looks like – that is, as a three-step procedure beginning with the scientific literature, moving to generalizations about that literature, and finally using those generalizations to inform conclusions in empirical philosophy of science (Figure 3).[11] The essential idea is that Figure 3 represents the "core" of an empirical approach to the philosophy of science. One begins with a body of products of the scientific process – be it journal articles, datasets, laboratory notebooks, or other such traces. One then attempts to construct generalizations from that empirical corpus about how it is that science works, whether in that particular laboratory or, more often, in that field, or in science as a whole. This generalizing step is often the key move in such research, as we attempt to demonstrate that the scientific process shows certain kinds of reliable features that we can use to make philosophical inferences. Finally, we have to figure out how to construct such philosophical claims based upon those generalizations – how exactly can they be shown to be relevant for philosophical concerns?

In laying out where these demands for further explication fit into that three-part structure, I believe we will find ways in which we can transform those demands into challenges for future work on digital philosophy of science – in short, to set some positive goals for scholars invested, as I am, in advancing this field and its prospects.

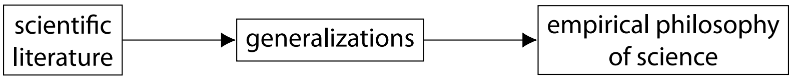

Figure 3: An excerpted portion of the central figure from Lean, Rivelli, and Pence (2021), representing schematically how scientific literature might inform empirical philosophy of science.

Let's begin with the common thread picked out by both the examples of preregistration and Whig

[11] In phrasing these in terms of scientific literature, I am reflecting my own group's focus on textual analysis; I believe the same considerations clearly apply to other parts of digital philosophy.

history. In both cases, we found evidence that Figure 3 needs at least one more box and arrow. In addition to being informed directly by the scientific literature, the generalizations that we draw will, of course, also be informed by our extant philosophical commitments, adding another "input" to the middle node in Figure 3. No reading of any text can occur in a philosophical vacuum. But this also turns into the first open challenge for philosophers wishing to use these methods. How exactly should we evaluate the potential impacts of our prior philosophical commitments on these empirical analyses? What kinds of such biases might exist, and how could we detect their presence or absence? Should we want to produce an analysis independent of one such presupposition, how would we do so?

In general, because the methods of digital philosophy are so new, there is very little work directed at answering these questions in a philosophical context. Digital humanists, who have produced a fair bit of sustained critical analysis of their own tools and methods (e.g., Rogers 2013; Arthur and Bode 2014; Estrada 2014; Berry and Fagerjord 2017; or the discussion of the role of stylometrics in the analysis of Henry James present in Hoover 2007), have of course not done so with the peculiar concerns of philosophers in mind.[12] There is thus a significant space here for further work. We should encourage the pursuit of systematic study of the relationship between philosophical commitments and the very analysis tools of digital philosophy themselves. For example, it seems likely that differing views about the very nature of the scientific process will lead to different conclusions about how we ought to interpret the empirical signals coming from our study of scientific texts (Lean, Rivelli, and Pence 2021). It is also worth considering whether this extends not only to our views of social or community epistemology, but also to ontology and metaphysics. Will adopting a realist or an anti-realist ontology, for instance, alter the appropriate epistemic attitude toward the results of these empirical analyses? More remotely, could our commitments to theories of causation have downstream impacts on how we think about the nature of these scientific products and the digital inferences we draw from them? These kinds of questions deserve to be explored in greater detail.

Second, consider the challenge raised by Nosek, inspired by preregistration. As philosophy isn't faced with a replication crisis, the fervor present in preregistration advocates in the sciences would likely be misplaced in a digital-philosophy context. But that said, one can still defend the use of preregistration in cases where there is a chance for our unforced methodological choices to have undue influence on the results of our analyses. More broadly, one might be inspired by these kinds of worries to focus in a more dedicated way on the development of *best practices* regimes for digital analyses in philosophy. In much the same way that scientific data analysis is often governed by informal norms surrounding which software packages to use, default settings that should be respected, and so forth, we should encourage the development of these same sorts of norms in philosophical practice. As anyone who has been to a scientific journal club (reading group) meeting can attest, the very first action of any scientist confronted with a new article is to turn to the methods section and make sure that these choices all pass muster. In the language of Figure 3, we could consider this a reinforced emphasis on the arrow connecting the scientific literature to those generalizations – the technical details of this process are clearly extremely important.

[12] J. T. Burman (2018) has also considered similar questions from the perspective of digital methods in the history of psychology.

Some such work already takes place, of course, in digital humanities journals, or in the particular technical papers that describe methodological advances, and these papers are fairly routinely cited by philosophers doing digital work. But spaces for dedicated philosophical discussion of this kind remain somewhat rare, with these methodological and technical points often being pushed into appendices or online-only content rather than developed and discussed as integral parts of our analytic work. Here, perhaps most of all, we have a need for sociological or professional change in order to build spaces for this kind of work. Some of these issues surround publication and credit. Broader digital humanities journals often aren't welcoming venues for philosophers, who in this case would largely be reworking questions already tackled by scholars in literary studies, history, or library science years or decades ago. Another issue involves interdisciplinarity. We have a great opportunity here to learn from other colleagues in digital humanities, though often this requires building bridges with communities that have historically been relatively remote from philosophy.

Third, and finally, consider the Mayr-inspired point about the very nature of the philosophy of science itself. Does and should our subject matter itself constrain the kinds of questions that digital methods might ask, or the kinds of answers that we might expect those questions to receive? One might call this a question of the *internal structuring* present within the philosophy of science: are there certain kinds of relationships between our methodological choices, the empirical facts on the ground, and our philosophical views, such that analyses taking some of these connections for granted are justified while others are not? We might envision this as an arrow moving backwards from the third to the second box in Figure 3 – that is, a connection between the nature of empirical philosophy of science itself and the generalizations that we might draw from the scientific literature (see the final resulting diagram in Figure 4).

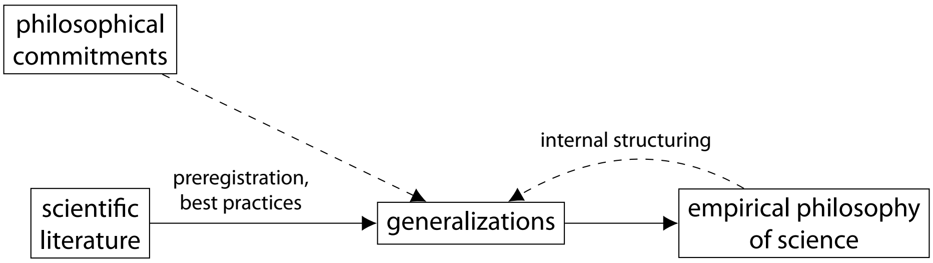
philosophical commitments
internal structuring
preregistration, best practices
scientific literature
generalizations
empirical philosophy of science

Figure 4: A modified version of Figure 3, adding in the kinds of concerns that might be raised by taking seriously the worries about practice in digital philosophy that I have raised here.

An extreme example of this kind of structuring, for instance, might be found in radical skepticism about the external world.[13] Obviously, a view such as this will have far-reaching consequences for our other philosophical positions, as well as what we might expect to find (or not) in the "empirical" record. Less extreme, we might consider Richard Boyd's argument for scientific realism. As quickly surfaced in arguments between Boyd and authors like Bas van Fraassen (1980), one of the points at issue between the scientific realist and the scientific anti-realist is the legitimacy of argument via inference to the best explanation. And yet, it is not only in science that such arguments occur – perhaps the most common way of interpreting the argument for scientific

[13] Thanks to Timothy Williamson for raising this example.

realism itself in philosophy of science is as a (philosophical, rather than scientific) inference to the best explanation. In that sense, our first-order views about ontology are directly tied to our second-order views about appropriate methodology. As Boyd himself puts it,

> If what is at issue is the legitimacy of abductive inferences to theoretical explanations in general, then there is a kind of circularity in the appeal to a particular abduction of this sort in the defense of scientific realism. [. . .] I suggest that our assessment of the import of the circularity in question should focus not on the legitimacy of the realist's abductive inference considered in isolation, but rather on the relative merits of the overall accounts of scientific knowledge which the empiricst and the realist defend. (Boyd 1983, 80–81)

That is, Boyd argues that the threat of circularity present here (using abduction to argue for the legitimacy of abduction) can be viewed as virtuous rather than vicious, if only we step back from the details of the fine-grained arguments for and against abduction itself and instead target the overall coherence of the systemic approaches that the realist and anti-realist offer us.

To be sure, there is no invocation here of digital methods in philosophy of science. But I think exactly this kind of relationship or "feedback" between different portions of our philosophical perspectives is what we should be on the lookout for given a "whiggish" worry about the ways in which our philosophical positions might either support or undermine the use of some empirical tools. I have no clear predictions about where this kind of work might lead – but a clear demonstration of the *lack* of such internal structure would be no less valuable.

7 Conclusion: Future Steps

I began my argument here by pointing out an apparent tension in the reasons for which we might use digital methods in the philosophy of science. On the one hand, these methods can show us a host of unexpected and interesting features of the scientific process, features that might be exactly the kind of inspiration needed to develop new philosophical views. But on the other hand, and following on the concerns about spurious correlation and implicit assumptions present in the analysis of big data, it has been argued that the use of these two tools for discovery is dangerous, and that we might instead be better off considering them as tools for the restricted testing of particular hypotheses, not their generation.

Such a tension puts us in an unenviable position, as taking either approach alone seems to deny us some of the real power of digital philosophy. The generation of novel hypotheses has already been important both in scientific (e.g., Wilkinson and Huberman 2004; Altman et al. 2008) and humanities contexts (e.g., see the discussion of the birth of "*nouveaux observables*" by means of digital analysis in Rastier 2010). We thus would be ill served by the outright rejection of either approach in favor of the other. A nuanced way to extract the advantages of both hypothesis testing and novel hypothesis generation is required.

This same tension has appeared in fields beyond philosophy – I thus then turned to two such examples, one from contemporary data-driven science and one from the history of science. These other views of this same problem helped to shift our frame from one of testing versus discovery to a less binary view of the relationship between our data or background philosophical views, and the empirical results that we might draw from them. In the last section, I made this frame yet more precise, by splitting that question of relationships into three particular challenges for future digital philosophy of science.

To conclude, I want to consider how we might start to address these challenges. As I briefly noted in the last section, I believe that this largely turns on building institutional and professional spaces in which philosophers can discuss the kinds of questions that I have raised here. Three kinds of considerations are, I have argued, especially important. Philosophers need ways in which we can:

1. Work to illuminate the influences of our philosophical commitments on our empirical work,
2. Discuss methodological questions and best practices in detail, perhaps with the aid of preregistration, and
3. Explore whether the nature of of philosophical questions themselves will alter that work.

As I mentioned, it's also unclear whether and when such work would be acceptable for publication in philosophy journals, and thus we may have community-level reforms to undertake as well, building opportunities for sharing and discussion among practitioners in this area.

In short, while digital methods in philosophy of science have much promise, I believe that promise has to be tempered by careful and reflective work about where, when, and how such methods will be most useful, as well as whether the kinds of uses that we envision for them will actually enable us to produce higher-quality philosophy. Much exciting work remains to be done.

References

Altman, Russ B., Casey M. Bergman, Judith Blake, Christian Blaschke, Aaron Cohen, Frank Gannon, Les Grivell, et al. 2008. "Open Access Text Mining for Biology - the Way Forward: Opinions from Leading Scientists." *Genome Biology* 9 (Suppl 2): S7. https://doi.org/10.1186/gb-2008-9-S2-S7.

Arthur, Paul Longley, and Katherine Bode, eds. 2014. *Advancing Digital Humanities: Research, Methods, Theories*. Basingstoke, Hampshire: Palgrave Macmillan.

Berry, David M., and Anders Fagerjord. 2017. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge: Polity.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77. https://doi.org/10.1145/2133806.2133826.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15 (5): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Boyd, Richard N. 1983. "On the Current Status of the Issue of Scientific Realism." *Erkenntnis* 19 (1-3): 45–90. https://doi.org/10.1007/BF00174775.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of Machine Learning Research*, 81:1–15.

Burman, Jeremy Trevelyan. 2018. "Digital Methods Can Help You… If You're Careful, Critical, and Not Historiographically Naïve." *History of Psychology* 21 (4): 297–301. https://doi.org/10.1037/hop0000112.

Busa, R. 1980. “The Annals of Humanities Computing: The Index Thomisticus.” *Computers and the Humanities* 14 (2): 83–90.

Butterfield, Herbert. 1931. *The Whig Interpretation of History*. London: Bell.

Calude, Cristian S., and Giuseppe Longo. 2017. “The Deluge of Spurious Correlations in Big Data.” *Foundations of Science* 22 (3): 595–612. https://doi.org/10.1007/s10699-016-9489-4.

Cronon, William. 2012. “Two Cheers for the Whig Interpretation of History.” *Perspectives on History* 50 (6): 5.

de Solla Price, D. J. 1965. “Networks of Scientific Papers.” *Science* 149 (3683): 510–15. https://doi.org/10.1126/science.149.3683.510.

Douglas, Heather E. 2009. “Reintroducing Prediction to Explanation.” *Philosophy of Science* 76 (4): 444–63. https://doi.org/10.1086/648111.

Estrada, Daniel. 2014. “In Defense of the Digital Humanities as a Science.” *Academia.edu*.

Garfield, Eugene. 1955. “Citation Indexes for Science: A New Dimension in Documentation Through Association of Ideas.” *Science* 122 (3159): 108–11. https://doi.org/10.1126/science.122.3159.108.

Gordin, Michael D. 2014. “The Tory Interpretation of History [Review of Chang, Hasok. *Is Water H_2O?: Evidence, Realism and Pluralism*].” *Historical Studies in the Natural Sciences* 44 (4): 413–23. https://doi.org/10.1525/hsns.2014.44.4.413.

Hoover, David L. 2007. “Corpus Stylistics, Stylometry, and the Styles of Henry James.” *Style* 41 (2): 174–203.

Hull, David L. 1979. “In Defense of Presentism.” *History and Theory* 18 (1): 1–15.

Jardine, Nick. 2003. “Whigs and Stories: Herbert Butterfield and the Historiography of Science.” *History of Science* 41 (2): 125–40. https://doi.org/10.1177/007327530304100201.

Kim, Kyung-Man. 1994. *Explaining Scientific Consensus: The Case of Mendelian Genetics*. New York: The Guilford Press.

Lean, Oliver M., Luca Rivelli, and Charles H. Pence. 2021. “Digital Literature Analysis for Empirical Philosophy of Science.” *British Journal for the Philosophy of Science*, April. https://doi.org/10.1086/715049.

Ledgerwood, Alison. 2018. “The Preregistration Revolution Needs to Distinguish Between Predictions and Analyses.” *Proceedings of the National Academy of Sciences* 115 (45): E10516–17. https://doi.org/10.1073/pnas.1812592115.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

Malaterre, Christophe, Jean-François Chartier, and Davide Pulizzotto. 2019. “What Is This Thing Called Philosophy of Science? A Computational Topic-Modeling Perspective, 1934–2015.” *HOPOS* 9 (2): 215–49. https://doi.org/10.1086/704372.

Manovich, Lev. 2012. “How to Compare One Million Images?” In *Understanding Digital Humanities*, edited by David M. Berry, 249–78. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_14.

Mayr, Ernst. 1990. “When Is Historiography Whiggish?” *Journal of the History of Ideas* 51 (2): 301–9.

Mercer, Christia. 2019. “The Contextualist Revolution in Early Modern Philosophy.” *Journal of the History of Philosophy* 57 (3): 529–48. https://doi.org/10.1353/hph.2019.0057.

Mizrahi, Moti. 2020. “Hypothesis Testing in Scientific Practice: An Empirical Study.” *International*

Studies in the Philosophy of Science 33 (1): 1–21. https://doi.org/10.1080/02698595.2020.1788348.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1: 0021. https://doi.org/10.1038/s41562-016-0021.

Nosek, Brian A., Charles B. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606. https://doi.org/10.1073/pnas.1708274114.

Pence, Charles H. in press. "How Not to Fight about Theory: The Debate Between Biometry and Mendelism in *Nature*, 1890–1915." In *The Evolution of Science*, edited by Andreas De Block and Grant Ramsey. Pittsburgh, PA: University of Pittsburgh Press.

Ramsey, Grant, and Charles H. Pence. 2016. "evoText: A New Tool for Analyzing the Biological Sciences." *Studies in History and Philosophy of Biological and Biomedical Sciences* 57: 83–87. https://doi.org/10.1016/j.shpsc.2016.04.003.

Rastier, François. 2010. "Sémiotique et linguistique de corpus." *Signata. Annales des sémiotiques / Annals of Semiotics*, no. 1 (December): 13–38. https://doi.org/10.4000/signata.278.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. Shanghai China: ACM. https://doi.org/10.1145/2684822.2685324.

Rogers, Richard. 2013. *Digital Methods*. Cambridge, MA: The MIT Press.

Sagoff, Mark. 2019. "Can Hypothesis-Driven Research Survive the Sequence-Data Deluge?" *Microbial Biotechnology* 12 (3): 414–20. https://doi.org/10.1111/1751-7915.13377.

Sewell, Keith C. 2003. "The 'Herbert Butterfield Problem' and Its Resolution." *Journal of the History of Ideas* 64 (4): 599–618. https://doi.org/10.1353/jhi.2004.0010.

Smaldino, Paul E., and Richard McElreath. 2016. "The Natural Selection of Bad Science." *Royal Society Open Science* 3 (9): 160384. https://doi.org/10.1098/rsos.160384.

Soler, Léna, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost, eds. 2014. *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*. New York: Routledge.

Underwood, Ted. 2017. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 011 (2).

van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford: Clarendon Press.

Wilkinson, Dennis M., and Bernardo A. Huberman. 2004. "A Method for Finding Communities of Related Genes." *Proceedings of the National Academy of Sciences* 101 (suppl 1): 5241–48. https://doi.org/10.1073/pnas.0307740100.