Positively Misleading Errors

Matthew H. Haber
Department of Philosophy, School for Biological Sciences,
and Center for Quantitative Biology
University of Utah
Salt Lake City, UT, USA
matt.haber@utah.edu
ORCID: 0000-0001-6559-3802

Abstract

Positively misleading errors are errors of statistical reasoning in which adding data to an analysis will systematically and reliably strengthen support for an erroneous hypothesis over a correct one. This pattern distinguishes them from other errors of statistical inference and pattern recognition. Here I provide a general account of positively misleading errors by describing an exemplar case from biology along with a candidate case from clinical medicine. Though well known in biology (phylogenetic systematics, to be precise), positively misleading errors are likely more widespread and deserve to be brought to the attention of the wider research community. This will facilitate a better understanding of them and sharpen our ability to assess statistical and probabilistic methods, providing resources for researchers to more effectively identify, diagnose, and dislodge these errors of statistical inference. This reflects the way we have gained a better understanding of scientific reasoning from studying other errors of statistical and probabilistic reasoning.

1 Introduction

In 1978 the biologist Joseph Felsenstein published the paper, "Cases In Which Parsimony Or Compatibility Methods Will Be Positively Misleading" in the journal *Systematic Zoology*. In it he demonstrates that popular methods for reconstructing evolutionary histories ('phylogenies' or 'evolutionary trees') will, under certain specified conditions, systematically yet erroneously lump together taxa as closely related when they are, instead, separated by long evolutionary branches. This erroneous behavior has been dubbed *long-branch attraction* and identified as a distinctively challenging kind of statistical inconsistency, which Felsenstein called a *positively misleading error*.

Statistical methods have the property of *consistency* when they converge on the correct outcome as more data accumulate (Felsenstein, 1978); *statistical inconsistency* is the failure to converge on that outcome. Felsenstein observed that inconsistency may be expressed as convergence on an incorrect outcome (as opposed to just failing to converge at all). The positive support attributed to the erroneous outcome comes at the expense of the correct one, mimicking statistical consistency, and, left undiagnosed, can mislead researchers about the systems they are studying. Hence, these errors are positively misleading.

My goal is to provide a more general account of positively misleading errors (PMEs), arguing that they are a distinctive and important category of statistical or probabilistic reasoning. Outside of a few fields of biology, researchers are generally less familiar with PMEs than other errors of statistical reasoning, e.g., type I errors or base-rate fallacies. Drawing attention to PMEs will help us gain a better understanding of them and provide resources to researchers to more effectively identify and dislodge these errors, much as we have gained a better understanding of good scientific reasoning from studying other errors of statistical reasoning (Mayo, 1996).

Here, I introduce PMEs (§2) with an idealized, cartoon case (§2.1). This is intended as an entry point into phylogenetics—the field of biology that includes reconstructing evolutionary history (§2.2)—setting the stage for a presentation of the case from which Felsenstein first identified and described positively misleading errors (§2.2.1). Following a brief discussion of why PMEs are distinct from more familiar errors of statistical reasoning (§2.3), I explore some of the philosophically interesting features of PMEs (§2.4). That includes how PMEs suggest adopting a minimal methodological pluralism and the way PMEs can generate *epistemic traps* for researchers.

PMEs are unlikely to be limited to phylogenetics (§3). To demonstrate this I consider a candidate case from clinical medicine (§3.1) and submit other possible cases (§3.2). This suggests that PMEs may be more widespread than appreciated and highlights what is at stake in understanding them. There are important and pressing consequences of this error of statistical and probabilistic reasoning and we ought to seek ways to improve our ability to identify, diagnose, and dislodge PMEs. A good starting point is naming the problem and identifying exemplar cases we may study and learn from.

2 Positively Misleading Errors

Positively misleading errors were introduced by formal proof in Felsenstein (1978). This is part of a series of papers and larger debates over treating methods for reconstructing phylogeny (evolutionary history) as methods of statistical inference. Though not without controversy, this stance permits an evaluation of competing phylogenetic methods by the properties and behaviors they may exhibit *as* statistical methods. For our purpose, this stance is notable in the way it promoted innovation and discovery about statistical inference. §2.2 will review Felsenstein's paper and long-branch attraction as an exemplar case. §2.3 discusses why PMEs are distinctive errors of statistical reasoning that should not be confused with type I errors or false positives—which, in turn, should also not be conflated.

Though Felsenstein's paper is straightforward, a bit of an entry point into phylogenetics may be useful for readers unfamiliar with that field. So let's begin by introducing PMEs and some key concepts from phylogenetics through an idealized, cartoon example.

2.1 Introducing PMEs

Imagine a population of wild sheep living in a valley splitting into two distinct populations, eventually diverging into two distinct species. The details of the split may be left aside, though let's suppose they remain morphologically similar despite the speciation event. Imagine, further, that soon after this initial divergence each of the two new species subsequently split again into two distinct populations, each eventually forming into distinct species. These latter splits follow a similar pattern, each involving a population of sheep moving from the valley up into the surrounding mountains. After this period of divergence there is relative stasis, though the four species continue evolving in response to their respective environments. The ecological pressures of living at high altitude on the sides of mountains result in both mountain species transforming (morphologically, physiologically, etc.) away from their respective sister species in the valley, who both still closely resemble their ancestral populations (and each other). Due to the similar ecological pressures faced by the mountain species, they end up closely resembling each other (morphologically, physiologically, etc.). Figure 1 shows one way biologists might represent the patterns of descent described here, displaying the patterns of divergence, transformation and genealogy of the extant species.

Suppose, now, that you are a biologist studying these sheep. Central to this task is reconstructing the evolutionary history of these species, namely, the pattern of descent from a common ancestor. Which species are most closely related? The events described in the previous paragraph happened long before we were able to directly observe them, so the best we can do is construct hypotheses of that history.[1] One method you might employ is to compare the morphological characters of these species, hypothesizing that those species that are most similar to each other are also the most closely related. The mountain species are more similar to each other in this regard than either is to the valley species; likewise for the two valley species. So you hypothesize that the two mountain species are sister species, as are the two valley species. You might treat as a research question whether the ancestral species was a mountain or valley species, but we can leave that aside here. As our investigative tools improve, we might gather more data,

[1] This is the problem of *phylogenetic inference* (Hull, 1988; Sober, 1988, 2008; Haber, 2009; Velasco, 2013; Haber and Velasco, 2021).

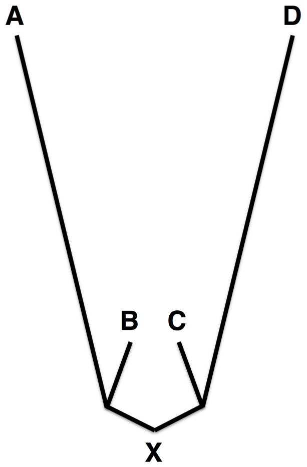


Figure 1: **Patterns of Descent.** X is an ancestral species that split into two distinct species. Both of these new species subsequently speciated as well, resulting in four extant species, A, B, C, and D. Species B and C resemble the ancestral form, while A and D have undergone similar transformations in response to environmental conditions. Nodes in this figure represent common ancestors; splits represent speciation events. This tree displays that A and B are sister species, more closely related to each other than either is to C and D. The length of branches represents the amount of transformation or accumulation of characters in a species from a speciation event. A longer branch means more transformation of characters than a shorter branch. Here that means that species A and D have seen more evolutionary transformations than either B or C. Figures like this are called *phylogenetic* or *evolutionary trees*. Biologists use these to display evolutionary hypotheses about genealogy, character evolution, etc.

accumulating ever finer descriptions of morphological, ultrastructural, and physiological characters (and perhaps even molecular data) of these sheep. Accumulating more data strengthens support for your hypothesis of evolutionary history (at the expense of alternative hypotheses), establishing ever greater similarities between what you suppose are sister species. Of course, your hypothesis is wrong, despite strong evidential support. Yet, the data are good and adding more of it increasingly provides support for your hypothesis at the expense of other, more accurate, hypotheses. Unfortunately, you are caught in an epistemic trap, generated by a *positively misleading error*.

Positively Misleading Errors (PMEs) are a kind of statistical inconsistency well-known and well-studied in certain fields of biology, such as phylogenetics. A lot more will be said about this case and why the inference goes wrong, but for now look past what is *causing* the error and focus instead on its *pattern*. More precisely, this is a case where adding data systematically and reliably produces increasingly greater strength of support for an erroneous hypothesis at the expense of the correct one. As the number of data approach infinity, the strength of support for the erroneous hypothesis will approach a maximal level. Convergence, in other words, is on an erroneous conclusion. It is this pattern of misplaced support that is characteristic of PMEs.

2.2 Phylogenetic Systematics

The basic idea of what constitutes a positively misleading error comes from Felsenstein's (1978) demonstration that a popular method for phylogeny reconstruction "is guaranteed to converge" on the wrong evolutionary tree as more data are added. The formal proof he offers is straightforward, yet it will be helpful to provide some context about the phylogenetic inference problem space to frame discussion of the exemplar case in §2.2.1.

Sitting at the intersection of evolutionary biology, statistics, and computer science (Sterner and Lidgard, 2014, 2018; Haber and Velasco, 2021), the scope of phylogenetics includes reconstructing patterns of evolutionary history, represented as *phylogenies* or *evolutionary trees*. Exponentially many conflicting phylogenies may be consistent with a dataset, and part of the phylogeneticist's job is to identify the best of these competing hypotheses (Felsenstein, 1988, though see Nakhleh, 2013 and Haber, 2019). Biologists disagree about many aspects of this, from what constitutes 'best' in this context, to what data are most informative, and whether and how to incorporate other evolutionary hypotheses or models in phylogenetic analyses. In the 1970's, whether phylogenetic methods should be regarded as methods of statistical inference was a major point of debate.[2] This, in part, is the problem of phylogenetic inference.

In addition to being the first field to recognize PMEs as distinct, phylogenetics is a science of *patterns* (of evolutionary history), as opposed to processes or mechanisms (of evolution). PMEs are errors of pattern recognition, which phylogeneticists are especially well attuned to notice.

The systems being studied in phylogenetics are composed of evolutionary lineages and the data are distributions of characters across those lineages. For example, phylogeneticists seeking to reconstruct the pattern of descent from a common ancestor of bovids (antelopes, cattle, sheep, and relatives) might generate a dataset by sampling gene sequences across bovid species, each of which is taken to be a distinct evolutionary lineage. Hypotheses of this pattern of descent—phylogenetic trees—are typically displayed as branching diagrams (similar to that in figure 1), and may be tested by re-analyzing the data used to construct those hypotheses. Those tests may include employing new models describing the relationship between the data and the system being studied. These new

[2] This proposition is now more widely accepted, though there remain important and influential holdouts.

models may be more complex ('realistic') than previous ones,[3] take advantage of more powerful computing resources,[4] reflect new hypotheses of the underlying biology (e.g., processes or mechanisms of evolution), or use different modes of inference to extract underlying patterns and test prior hypotheses.

Phylogeneticists recognize PMEs as distinct and important errors of inference. In retrospect, this should hardly be surprising. The advent of molecular sequencing techniques has provided vast amounts of data for phylogenetic analysis. Conflicting patterns are routinely strongly supported by those data, and identifying the relevant pattern is a difficult task.[5] Access to increasingly powerful desktop computing and progressively richer datasets have permitted phylogeneticists to extract ever more subtle patterns, and to challenge entrenched hypotheses of phylogeny. In some cases, support for these hypotheses proved to be positively misleading (Yang and Rannala, 2012).

Furthermore, the backdrop against which this takes place primes phylogeneticists to be attuned to these sorts of errors of pattern recognition. From at least the 1960's, a central research problem in the field concerned whether, and, if so, how evolutionary history may be justifiably inferred, infusing the field with a strong philosophical tradition. An upshot of this is that a strongly critical focus on methods of inference is often a driver of innovation in the field, and there is intense pressure on the details of inferential support.[6] Researchers devise, test, and experiment with modes and methods of phylogenetic inference, and discovery of systematic errors in well-known or popular phylogenetic techniques constitutes a major discovery (Haber, 2009). Let's turn our attention to an important example of this, long-branch attraction.

2.2.1 An Exemplar: Long-Branch Attraction

Felsenstein's "Cases in which Parsimony or Compatibility Methods Will be Positively Misleading" (1978) introduced the phrase *positively misleading* to describe the systematic errors of statistical inference characterized above. The abstract cases he describes exhibit long-branch attraction.

[3] A simple evolutionary model that may be used in phylogenetic analyses is the Jukes-Cantor model, JC69 (Jukes and Cantor, 1969), which treats the substitution rate of nucleotides as uniform, i.e., the probability that any of the four nucleotides (A, G, C, T) will be substituted for any other is equal. A more realistic, or complex, model is HKY85 (Hasegawa et al., 1985), which distinguishes substitution rates between transitions (purine/purine or pyrimidine/pyrimidine swaps) and transversions (purine/pyrimidine swaps), and parameterizes other JC69 assumptions., e.g., base rate frequencies of the four nucleotides.

[4] These permit more sophisticated or effective heuristic search methods for exploring probability space.

[5] This is a variant of underdetermination of theory by evidence, resembling what Mayo (1997) calls the *alternative hypothesis objection*. Phylogeneticists routinely dispute what criteria ought to be used to determine which hypothesis is *best* supported by the evidence. For philosophical accounts of this problem see Hull (1988); Sober (1988, 2008); Haber (2009); Velasco (2013); Haber and Velasco (2021).

[6] Often without focus on any specific hypothesis of phylogenetic relationship. This abstract feature of the focus on method is one reason phylogenetics is such a philosophically rich field.

To understand the significance of this, recall that a central research problem in 1970's phylogenetics was justifying phylogenetic inference as a *scientific* hypothesis, as opposed to speculation, intuition or metaphysics. This reflects the central conceptual arguments that arose in systematics as phylogenetics took hold (Hull, 1988; Sober, 1988; Haber, 2009; Sterner and Lidgard, 2018). One popular approach, dubbed *parsimony*, was justified as scientific by appeal to falsificationism (Wiley, 1975).[7] The idea is simple enough: the phylogenetic tree requiring the fewest evolutionary events to explain the data is considered a testable, bold hypothesis of the pattern of evolution of those taxa. As more data (characters) are considered, that hypothesis is tested. If a new phylogenetic hypothesis (tree topology) can explain the data with fewer evolutionary events, that constitutes a rejection (falsification) of the previous phylogenetic tree as the most parsimonious. Otherwise it will have survived the test and be further corroborated. An attractive feature of this justification is that each subsequent addition of datum constitutes a test of the most parsimonious phylogenetic hypothesis.[8] Even on shedding the falsificationist spin, the appeal here is evident (Sober, 1988, 2008).[9] More good data confer greater evidential support or corroboration for the most parsimonious hypothesis.

Only it doesn't. Or, rather, it only does on pain of committing a PME dubbed *long-branch attraction* (LBA). The proof is straightforward,[10] demonstrating not merely statistical inconsistency, but that under certain conditions parsimony methods are guaranteed to converge on an incorrect hypothesis. LBA describes the tendency for terminal taxa at the end of long branches[11] to be incorrectly grouped together as most closely related, rather than correctly grouped with their actual sister taxa related by a short branch. This is reminiscent of the toy example above, in which the mountain sheep were incorrectly grouped as most closely related, rather than with their respective valley sister species (Figure 2).

The reason for this incorrect grouping is simple, if not readily apparent: evolution need not be parsimonious. Convergent or parallel evolution, for example, may produce patterns of biodiversity that could have emerged more parsimoniously.[12] In such cases,

[7] Leaving aside whether or not this constitutes Popperian falsificationism (Hull, 1983; Sober, 1983).

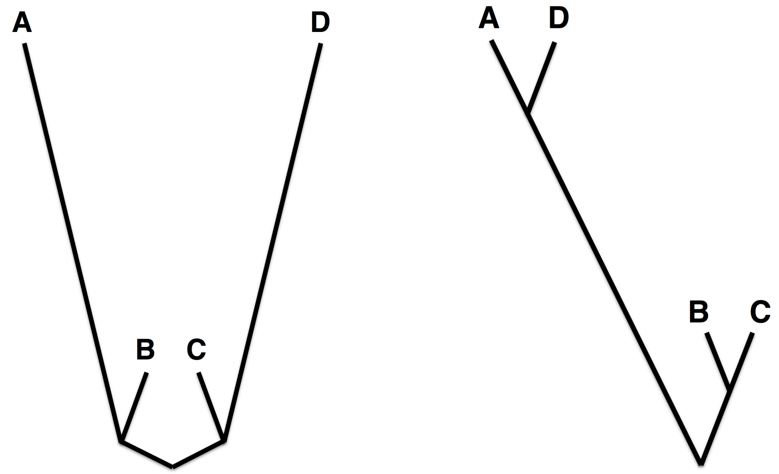
[8] This characterization is a bit idealized. More precisely, it is not *evolutionary events* that are counted, but *hypotheses of homology* that are tested against one another (Wiley, 1975).

[9] A referee astutely noted the connection of phylogenetic reasoning to abductive reasoning, i.e., inference to the best explanation, wondering about the "potential trade-offs among different explanatory virtues." Sober noticed the same, explicitly connecting parsimony methods to abduction. Fitzhugh (2006) and Quinn (2016) made similar connections, though in very different ways, that reflect choices about different trade-offs that might be made in service of phylogenetic inference.

[10] See Felsenstein (1978).

[11] *Branch length* may be understood as corresponding to amount of evolutionary change, number of evolutionary novelties, or branch transition probabilities. See figure 1.

[12] Other conditions may also produce patterns of descent susceptible to LBA, e.g., differential divergence (Ward, 2014; Ward et al., 2015).



(a) Evolutionary History (b) Parsimony Reconstruction

Figure 2: **Long-branch attraction.** The pattern of descent of taxa A-D (a) may include long branches with a significant amount of convergent or parallel evolution. Though these 'long branch taxa' (A and D) are most closely related to taxa that have seen relatively few evolutionary changes (B and C), parsimony methods will incorrectly identify (b) as the best supported hypothesis of evolutionary history. This incorrect clustering of long branches is a positively misleading error called long-branch attraction. Adding more data to a parsimony analysis increases corroborative support of (b) at the expense of phylogenetic hypotheses displaying patterns concordant with (a). Adapted from Felsenstein (1978).

parsimony methods will increasingly support more parsimonious (though erroneous) hypotheses of evolutionary history over less parsimonious (though correct) ones. This is exactly what Felsentein demonstrates:

> . . . as we accumulate more and more information the . . . parsimony method is *increasingly certain to give the wrong answer*. (Felsenstein, 1978, p. 404)

and:

> [Under these conditions the] parsimony method is *guaranteed* to converge to the wrong estimate of the [phylogenetic] tree as we accumulate more and more data. (Felsenstein, 1978, p. 405, emphasis added)

Felsenstein offers a diagnosis of what has gone wrong: parsimony methods implicitly assume too simplistic a model of evolution,[13] corrupting the analysis. Patterns in the data

[13] Though not, importantly, too *simple* a model. Indeed, if we count *parameters* rather than *evolutionary novelties* (or hypotheses of homologies), parsimony methods may be less parsimonious than statistical ap-

are obscured and conflated, and implicit erroneous hypotheses about the relation of the data to the system of study are instead captured and displayed. Felsenstein concludes that LBA undermines both the implicit theoretical assumptions and the conceptual justification of using parsimony methods for phylogenetic inference.[14]

Setting aside Felsenstein's discussion of *why* things went wrong, the upshot is clear: parsimony methods return positively misleading results under biologically significant conditions. As more data are added to a parsimony analysis, erroneous hypotheses receive increasingly positive support over correct ones. This is an error of statistical inference. The cause may ultimately be understood as the result of underdetermination, over-simplistic assumptions, or some other combination of confounders of scientific reasoning. But it is not the *cause* of the error that is of concern here; after all, under different conditions parsimony may outperform other phylogenetic methods (Siddall, 1998). What matters is the *pattern* of error that arises. As the case in §3 demonstrates, recognizing (or *failing* to recognize) that pattern for what it is may carry serious consequences.

2.3 Distinction From Other Errors of Statistical and Probabilistic Reasoning

With erroneous conclusions receiving support at the expense of correct ones, PMEs are similar to false positives and type I statistical errors. Yet there are important differences that distinctively mark PMEs.

It will be helpful to first distinguish false positives from type I statistical errors. Though sometimes treated as synonyms, these are more properly understood as analogues sharing important and relevant similarities concerning slightly different scopes.[15] A type I error concerns statistical testing of hypotheses, and, along with type II errors, describes how the role of chance may be misattributed in explaining a pattern found in a dataset. Consider a dataset drawn from a population under study. If the sampled data display

proaches that incorporate standard models of evolutionary change, e.g., maximum likelihood (ML) methods that stipulate a JC69 (Jukes and Cantor, 1969), K80 (Kimura, 1980) or HKY85 (Hasegawa et al., 1985) model of transformation. Forcing a tree with the fewest number of homologies may implicitly require far more complex (highly parameterized) underlying evolutionary models or treating rate of character transformation as highly dynamic and variably complex across a phylogenetic tree (Tuffley and Steel, 1997; Steel and Penny, 2000; Swofford and Sullivan, 2009). This is also a nice reminder of why what counts as most parsimonious depends on what is being counted—which itself is a question reflecting conceptual commitments (Haber, 2009).

[14] This does not, however, mean that parsimony techniques ought to be abandoned. The correct response is a bit more complicated and subtle than that. One interpretation might be described as pragmatic; that these techniques ought to be regarded heuristically and as part of the toolbox of methods available to phylogeneticists that can play both a direct and indirect role in supporting good phylogenetic inference (Hillis et al., 1994; Felsenstein, 2004; Haber, 2009; Haber and Velasco, 2021).

[15] Spanos (2010) argues that even this may be too strong, and that treating these as analogues in all cases leads to mistaken conclusions about the susceptibility of frequentist testing to base-rate fallacies.

a particular pattern, we might ask whether that pattern is present because it reflects a pattern in the larger population from which it was drawn, or whether the pattern is the product of unlucky or unrepresentative sampling ('chance'). A type I error is the incorrect rejection of chance as the source of the pattern; a type II error incorrectly attributes chance as the source of the pattern (see Table 1).

	No Difference/ Pattern due to chance (H_0)	Significant Difference/ Pattern *not* due to chance ($\sim H_0$)
Reject H_0	**Type I Error**	Correctly Reject Chance
Accept H_0	Correctly Accept Chance	**Type II Error**

Table 1: Type I & II Statistical Errors. A type I error incorrectly rejects chance as the source of a pattern in a sampled dataset. A type II error incorrectly attributes chance as the source of that pattern. The framing here displays how those errors may occur in the context of hypothesis testing.

False positives, on the other hand, concern failure rates, i.e., incorrect reports of the features of a sample. A diagnostic test, for example, might report, incorrectly, the presence of a protein in a blood sample that lacks that protein.[16] This concerns the *sensitivity* and *specificity* of that diagnostic test; the capacity, respectively, to correctly identify positive and negative results. Knowing the sensitivity and specifity of a diagnostic test, along with its negative and positive predictive values, helps researchers and clinicians make decisions about how much and what kind of testing is needed before an actionable conclusion may be drawn (or how to balance the results of those tests against other diagnostic evidence) (Altman and Bland, 1994a,b; see Watson et al., 2020 for an example of this in clinical practice).

Type I errors are analogous to false positives, insofar as they are incorrect reports about an underlying pattern or condition of the population or system from which a sample was drawn. They differ in that type I errors faithfully report the pattern or properties of a dataset, but of a set that fails to resemble the study population in the relevant regard. False positives, on the other hand, fail to faithfully report an underlying condition in the sample or dataset that the test was designed to detect. The former is a product of poor, insufficient, or simply unlucky datasets, the latter a failure rate.

Type I errors are predictably diagnosable—a predictability that is valuable for scientific reasoning (Mayo, 1996).[17] The conclusions drawn from hypothesis testing may be strengthened or themselves tested, e.g., serially testing the hypotheses using distinct

[16] A false negative would incorrectly report the absence of that protein when it was present (see Casscells et al., 1978; Watson et al., 2020).

[17] Mayo uses this fact, among others, to develop an account of scientific reasoning that incorporates learning from error. Learning from PMEs is another example of the utility of the predictable diagnosability of statistical reasoning.

datasets, or increasing the size of the dataset. These will either tighten confidence limits or allow us to place confidence limits on the overall set of analyses. That is, more good data will tend to reveal and resolve type I errors. The limit case drives this home. If sampling size was increased to include the entire population under study, type I errors go away; the dataset would faithfully reflect the study population because they would be identical.

Analogously, recognizing the specificity and sensitivity of diagnostic tools informs researchers and clinicians of the epistemic value of those tests (Watson et al., 2020). So if a clinical diagnostic test is known to have a false positive rate of 1%, a positive result may not yet be good evidence for treatment. Knowing or refining pre-test probabilities (base rates), serially repeating the test, or adding data to the analysis will provide clinicians with better information, and increase the probability of revealing a false positive result.

In contrast, PMEs are not the result of bad data, lack of data, poor or unlucky sampling, ignoring a base-rate, or failure within a range of confidence limits; they are not rates of failure to correctly diagnose the underlying condition of sampled data. Neither repeating the analysis, nor using independent (or distinct) datasets, nor adding more data will reveal a PME. Just the opposite. As more data are included, the relative strength of support for the erroneous hypothesis will increase—even if the underlying model's confidence limits are tightening up! In contrast with type I errors, even on the limit case—where all data are collected and included in the analysis—the PME persists. This strongly suggests that the problem is not a function of the data or diagnostic failure, but is symptomatic of something else gone wrong.

Positively misleading errors are neither false positives nor type I errors. They are a type of statistical inconsistency, i.e., the failure to be statistically consistent, or to converge on the correct value as data accumulate:

> A statistical estimation method has the property of consistency when the estimate of a quantity is certain to converge to its true value as more and more data are accumulated. (Felsenstein, 1978, p. 402).

Yet something logically stronger than mere inconsistency is being asserted here. PMEs are not simply failures to converge, they guarantee convergence on the *wrong* value:

> Saying that a method is positively misleading is logically stronger than saying it is inconsistent: The failure of an estimator to converge to the true [hypothesis] does not imply that the estimator converges to an incorrect [hypothesis]—it might not converge at all. (Chang, 1996, p. 191-2)

PMEs are the opposite of statistical consistency. The latter guarantees convergence on a correct value as data accumulate; on PMEs, convergence is on an incorrect value. And it is this convergence that generates such a vexing epistemic position.

2.4 Discussion

2.4.1 Multiple Methods and the Richness of Scientific Reasoning

Ultimately, long-branch attraction was diagnosed and resolved by employing other modes of inference and incorporating ever-more sophisticated models of evolution into phylogenetic techniques (Felsenstein, 2004). Long-branch attraction, for example, has been diagnosed using maximum likelihood and Bayesian methods. However, we cannot know ahead of time which methods will be statistically consistent and which will generate positively misleading errors. Diagnosing this may require contrasting competing methodological approaches under a variety of specified conditions. Systematists use both simulation and empirical studies to test and study the relative performance of their phylogenetic methods under different conditions (Hillis et al., 1994; Degnan and Rosenberg, 2006).[18]

Even if we could know ahead of time (or come to learn) that some method was positively misleading under certain conditions, there is no guarantee that there will be a single best method that outperforms all others under all relevant conditions. If different methods outperform each other under different conditions, then researchers will need to identify the latter to select the former. This may, in fact, be the case in phylogenetics, e.g., Mark Siddall (1998) considers the mirror-image case of Felsenstein (1978), arguing that parsimony may outperform statistical methods under very specific conditions, described as *long-branch repulsion*.[19]

Testing how competing methods perform under different conditions help us identify systematic ways these methods may succeed or fail under those conditions. That variable yet systematic susceptibility of competing methods is, in turn, used to identify the underlying conditions of a system and provide guidance for determining which methods ought to be employed or may be subject to PMEs under those conditions. This suggests that phylogeneticists ought to adopt at least a minimal methodological pluralism, and

[18] A referee observed that phylogenetics is about reconstructing an historical event, asking about possible differences between PMEs in research design for retrospective and prospective cases. This is a good question! Hillis et al. go some way towards answering this, maintaining bacteriophage lines to generate and manipulate known phylogenies for experiments designed to assess competing phylogenetic methods. See Sober (2004), Haber (2005), Velasco (2008), and Autzen (2011) for more on the use of probabilistic reasoning for inferring historical phylogenies.

[19] Autzen (2018) describes a slightly different way that Bayesian approaches might fail to converge on the correct outcome, i.e., if it is excluded from the analysis. One of the challenges in phylogenetics is that the probability space (of possible phylogenies) is much larger than can be explored by heuristic methods, leading to just the sort of conditions that Autzen considers. See Yang (2007) and Quinn (2016) for related discussions.

be wary of narrowly monist accounts of scientific reasoning (Haber, 2005, 2009).[20]

Furthermore, there is often epistemic value not merely in identifying a PME, but in carefully diagnosing and understanding the conditions that give rise to it. Exploiting those conditions may lead to a richer, deeper understanding of that system (e.g., horizontal gene transfer in bacteria, Galtier and Daubin, 2008).

Though the discovery of LBA was deeply disruptive, it also spurred a highly fruitful and productive period in systematics, and the field is richer for it, e.g., it stimulated a thorough study of the application of models of evolution to phylogenetic methods (Huelsenbeck and Crandall, 1997) that ultimately provided systematists with the resources for, say, studying both the horizontal and vertical components of phylogenetic variance and biodiversity (Maddison, 1997; Degnan and Rosenberg, 2006; Galtier and Daubin, 2008).[21] I call these sorts of episodes *productive disruptions*. There is a rich story waiting to be mined about this ongoing period of history in systematics that carries important implications for scientific reasoning (Felsenstein, 2001), e.g., how communities of scientists shift from one set of commitments to another; and the constructive role that a fine understanding of errors plays in good scientific reasoning (building on Mayo 1996).[22]

2.4.2 Epistemic Traps and Patterns of Errors

Positively misleading errors are not a function of bad data or poor sampling; adding more data or increasing sample size will not resolve the error, only exacerbate it. Furthermore, PMEs may result from several different causes; there is no single reason they occur. They are errors of statistical inference or pattern recognition and may be symptomatic or effects of familiar confounders, and, thus, a powerful tool for assessing and describing reasoning and error. In the cartoon case described in §2.1, the PME was caused by intuitively plausible assumptions that turned out to be overly simplistic, including that morphological similarity reliably tracks relatedness;[23] conflating competing senses of similarity; an unrealistic[24] model; a clustering algorithm that conflicts with

[20] The selection of evolutionary models in statistical phylogenetics may also be sensitive to underlying conditions (Sullivan and Joyce, 2005). It is an empirical question whether multiple methods ought to be employed given the various conditions under which a system might be studied. Notably, more parameter rich (i.e., complex) models do not always perform better than simpler ones. See Sullivan and Joyce (2005) for further discussion on the risk of overparameterization and why simply selecting the most complex model available is a poor strategy. (Thanks to an anonymous referee that encouraged noting the philosophically rich topic of overparameterization.)

[21] This coincides and overlaps with similar developments in microbial systematics, though these developments are also at times regarded as competing views (Doolittle, 2000).

[22] Rice and Khalifa (2025) offer a related account of how *valuable misunderstandings* can generate fruitful responses through a community's *corrective processes*.

[23] Cichlids provide an example of when this assumption is violated (Stiassny and Meyer, 1999); so too that termites are eusocial cockroaches, and not closely related to ants (Inward et al., 2007).

[24] See footnote 3.

model assumptions; and confusing an effect for a cause, among other problems (see Felsenstein, 2004 for more details). Undiagnosed common causes, underdetermination, conflating correlation for causation, and other familiar confounders may also produce PMEs.

What makes PMEs particularly vexing is the epistemic position they generate. Positively reinforced convergence conveys confidence—especially if the results reinforce intuitively appealing entrenched commitments and are supported by a large stock of good data. This can make dislodging a method or commitment with positively misleading support particularly challenging. After all, this is just one of the patterns that scientists are (rightly) trained to identify as conveying good support for a hypothesis. "Good methods of scientific inference are expected to have desirable limiting features as the data size goes to infinity" (Autzen, 2018, 261; see also Howson and Urbach, 1991; Mayo, 1996, among others).[25]

PMEs are, thus, *epistemic traps*. Dislodging a PME means rejecting a hypothesis for which there is strong support based on good data. In this way, they are contrapositives of Gettier cases. In the latter, poor reasoning nonetheless produces the correct outcome (Gettier, 1963). In contrast, PMEs are cases of good reasoning leading us astray. So we have an interest in better understanding PMEs in order to effectively identify and distinguish them from those cases they resemble yet succeed on converging on the best (or correct, or what have you) hypothesis.[26]

It may be tempting to frame the problem in terms of converging on *positive* support, and view this as an argument for placing primacy on rejection rather than confirmation, e.g., falsificationism (Popper, 1959). This would be a mistake for at least two reasons. First, falsificationist methods are as susceptible to generating PMEs as any other (see §2.2.1). Second, statistical methods comparing strength of support of competing hypotheses may be used, under certain conditions, to diagnose and resolve PMEs. PMEs are errors of pattern recognition expressed as convergence, be it corroboration, verification, or probabilistic support. Luckily, at least in the case of long-branch attraction, under conditions in which one mode of inference[27] generates a PME, other modes generate statistically consistent results on the same dataset (i.e., they will converge on the correct hypothesis). Which modes will generate PMEs will depend on the details of the system, etc., and different modes may perform better or worse under different conditions (Hillis et al., 1994; Felsenstein, 2004). This constitutes a good reason for embracing at least a minimal notion of methodological pluralism.

Any large complex dataset will likely be susceptible to PMEs, though this is, ultimately,

[25] Though this is hardly unique to Bayesians or frequentists.

[26] Here we can be agnostic about what counts as 'best' in this context; see also Quinn (2016).

[27] Phylogenetic modes of inference might include, but not be limited to, maximum likelihood comparative analysis, Bayesian methodology, parsimony analysis, or abductive reasoning.

an empirical question. This is one of the challenges of so-called 'big data'. Though it has been an epistemic feature of phylogenetics for quite some time, it is becoming more familiar to other researchers as high throughput data grows more common across a wide range of disciplines (Nature Editorial, 2008; Bell et al., 2009; Leonelli and Ankeny, 2012; Vogt, 2013; Wu et al., 2014).[28] Successfully managing and drawing inferences from this generation of large amounts of data is, arguably, one of the central challenges facing 21st century researchers and clinicians (Howe et al., 2008; Callebaut, 2012; Leonelli, 2013; Pietsch, 2013; Wang and Krishnan, 2014). Understanding PMEs will be increasingly relevant for meeting this challenge, and a good first step is putting a name to the problem and building up a set of exemplar cases.

3 Candidate PMEs

Positively misleading errors are unlikely to be constrained to phylogenetics. Any field using convergence methods for extracting patterns from sets of data will be susceptible; all the more so as the systems of study are complex, and constituted, like evolutionary lineages, of levels or parts that may produce conflicting or confounding patterns.[29] PMEs may also manifest themselves in more informal reasoning as well, especially if a field includes intuitively appealing entrenched commitments that may be overly simplistic yet have historically generated successful outcomes (or, at least, have been *perceived* as generating those outcomes). In those cases, success may counterintuitively mask or obscure underlying patterns, better explanations, treatments, methods, etc., for the task at hand.

So there is a strong incentive to better diagnose PMEs and learn how to escape the epistemic traps they generate. One way to accomplish this better understanding is to find more cases to study, drawn from a broad range of disciplines. Below I describe a candidate case from clinical medicine (§3.1), then suggest others that look promising but need more careful analysis (§3.2).

3.1 Clinical Medicine: Septic Lactic Acidosis

In phylogenetics, *methods* or *analyses* are positively misleading when they converge on an erroneous hypothesis as more data accumulate.[30] In clinical medicine it is *treatment protocols* or *standards-of-care* that may be positively misleading when they converge on the

[28] This is closely related to the challenges of extracting patterns from complex systems (Mitchell, 2009).

[29] One sort of complexity I have in mind is Wimsatt's (1974) *multiple decomposability*, which Haber (2012) argues is exhibited by evolutionary lineages.

[30] "A tree reconstruction method is called consistent if the probability of reconstructing the correct tree converges to certainty as the sequence length tends to infinity." (Galla et al., 2019, 1180)

wrong treatment as more data accumulate as evidence.[31] Though the reasoning here is not always expressed in formal statistical or probabilistic terms, we should still consider them positively misleading.

Research on treatment protocols in clinical cases are likely to present many of the features that produce PMEs. The systems being studied are highly complex, with hard to discern feedback loops, conflating and confounding patterns emerging from interacting levels and shared parts, and a history of data from which patterns of treatment are extracted.[32] Convergence will be on a standard treatment protocol (of a possible set of responses/treatments) to administer under particular conditions. Entrenched standards-of-care can be difficult to dislodge—especially if new treatment protocols are considered counterintuitive or cut against tradition. Furthermore, using evidence-based approaches may fail to identify or resolve positively misleading treatment protocols. Merely adding high quality evidence to a study may simply exacerbate the error.

Let's consider one candidate case: the treatment of septic lactic acidosis with sodium bicarbonate (Forsythe and Schmidt, 2000). Though not expressed formally, the pattern of reasoning here mirrors that of long-branch attraction, suggesting the presence of a PME and characteristic epistemic trap.

Acidosis is the condition of blood becoming acidic. Infections may cause this drop in arterial pH, i.e., septic lactic acidosis. For patients in critical conditions this situation is life-threatening, with a 60-90% mortality rate in critical care units (Stacpoole et al., 1992). An entrenched standard-of-care treatment for patients in this condition was to attempt to neutralize the blood by raising the pH with intravenous administration of sodium bicarbonate ($NaHCO_3$) (a base). The justification for this rests on intuitive presumptions (e.g., you neutralize an acid with a base) and case histories of patient improvement on this treatment. That is, accumulation of data (case histories) produced convergence on entrenched support for treatment of acidosis with $NaHCO_3$.

Treatment with $NaHCO_3$ is no longer the standard-of-care for septic lactic acidosis, though dislodging it as the entrenched protocol proved a difficult task. The underlying assumptions, though intuitive, proved to be overly simplistic characterizations of physiology. This recognition helped reveal that support from case histories were hardly positive, but may have been positively misleading; re-interpreting those data on a more realistic physiological model revealed that $NaHCO_3$ was often detrimental to patients,

[31] Evidence in medicine may be expert opinion, case histories, non-randomized controlled studies, randomized controlled studies, etc. Institutional organizations rank different kinds of evidence by quality, e.g., the National Health Service in the UK ranks evidence on a scale from A to D. See Worrall (2002, 2007); Ashcroft (2004); Cartwright (2007) and Cartwright and Stegenga (2011) for critical consideration of evidence in evidence-based medicine.

[32] Social contexts and other non-epistemic factors can also play important roles. Though not discussed here, this extends a moral and ethical imperative for considering PMEs in clinical treatment protocols (e.g., Liao and Carbonell, 2023).

while a simple intravenous saline bolus proved beneficial (Cooper et al., 1990; Mathieu et al., 1991).[33] This is reminiscent of long branch attraction, where new, more realistic, models revealed that previous ones were returning positively misleading results on the datasets. The underlying *cause* of error in the two cases is different; what is analogous is the pattern of error and the difficulty of diagnosing and dislodging it.

Let's take a look at arguments offered against the entrenched treatment protocol, using a prominent review (and forceful condemnation) as a representative example. Forsythe and Schmidt (2000) argue that the convergence on treating septic lactic acidosis with administration of $NaHCO_3$ was an error. Worse, rather than helping patients, it may harm them! Yet, despite strong reasons for updating standards of care, dislodging the entrenched treatment protocols was challenging. This was further hampered by the lack of familiarity with PMEs. Lacking good models or examples of how to assess, diagnose, and dislodge PMEs made Forsythe and Schmidt's job more difficult.

Let's extract the pattern of reasoning employed by Forsythe and Schmidt to start developing these models of reasoning from positively misleading errors. Fortunately, they carefully lay out the details, arguing that the following chain of reasoning was either explicitly or implicitly used to justify $NaHCO_3$ treatment for septic lactic acidosis (p. 260-1):

1. A low pH, in and of itself, is harmful (most notably by impairing cardiovascular function).
2. Sodium bicarbonate can increase the pH when infused intravenously.
3. Raising the pH with sodium bicarbonate improves cardiovascular function.
4. Any adverse effects of sodium bicarbonate are outweighed by its benefits.

Forsythe and Schmidt proceed to consider each premise, identifying why each is too simplistic, and how this can lead to ever-increasing support for an erroneous hypothesis and sub-optimal treatment at the expense of better ones. This is just the pattern of reasoning to be expected in the presence of PMEs, and mirrors that found in phylogenetics.

Consider Forsythe and Schmidt's treatment of (2): *sodium bicarbonate can increase the pH when infused intravenously*. They begin by noting that this relies on an overly simplistic yet intuitive model of physiology: "It seems straightforward that adding a base to acidic blood will raise the pH—the reality is more complex" (p. 262). For example, "because

[33] $NaHCO_3$ is administered in a saline solution. The saline was improving the patient's condition more than the $NaHCO_3$ was causing harm, though $NaHCO_3$ was attributed with causing the improvement. In other words, the PME obscured the saline-only treatment from the set of treatments under consideration. Though this is reminiscent of other cases of error in scientific reasoning, what is novel here is the pattern of error, not the cause.

adults with acidosis generally also have sepsis, hypoxemia, intoxication, or hypoperfusion, discerning the physiologic effects of low pH from those of endotoxemia, hypoxemia, and so on is a challenging task" (p. 261). Just as in phylogenetics, the complexity of the system is producing confounding and conflicting signals. The job of the clinician here is to extract the relevant signals and recognize when discordance may present a complicating factor (p. 262):

> Yet, the body has multiple compartments separated by membranes of differing permeabilities and systems of active transport. Even when sodium bicarbonate added to the central veins reliably elevates the arterial pH, its effects on the cerebrospinal fluid and intracellular spaces may not be concordant.

Even worse, though arterial pH may increase with the addition of $NaHCO_3$, the intracellular pH may drop (p. 262):

> [S]odium bicarbonate can raise the blood pH when given IV. In contrast, this therapy fails to augment reliably the intracellular pH. Indeed, intracellular pH falls in most animal models and in most organs studied, but the effect is variable.

So go the arguments for each of the premises described above. E.g., premise (1): Is a low pH harmful? Perhaps, but too simplistic an analysis will merely confuse matters: "it is overly simplistic to assume that the clinician's window on acid-base state, the arterial blood pH, reflects accurately the pH at a (likely more important) cellular level" (p. 261). Counterintuitively, a low arterial pH may even be beneficial (p. 261-2):

> Paradoxically, acidosis may have protective effects in critical illness. A low pH has been shown to delay the onset of cell death in isolated hepatocytes exposed to anoxia and to chemical hypoxia. Correcting the pH took away the protective effect and accelerated cell death.

As with long branch attraction, an intuitively plausible assumption turned out to be poorly supported. Only on dislodging this overly simplistic assumption about low pH levels were clinical researchers able to see that support for the $NaHCO_3$ protocol was positively misleading, that it was a sub-optimal response of those available to acidosis.

Likewise with premise (3): Does $NaHCO_3$ benefit patients? Though numerous animal and whole body studies on this question have been performed, the answer is not obvious and "one must take care not to interpret these studies too simplistically" (Forsythe and Schmidt, 2000, p. 263). Indeed, any benefits were likely derived from the saline solution in which $NaHCO_3$ is administered: "The hemodynamic effects were indistinguishable from those of saline solution" (see Forsythe and Schmidt, 2000, p. 263 for a list of empirical studies).

Forsythe and Schmidt conclude that though extensive case histories have been taken as support for treatment of lactic acidosis with $NaHCO_3$, reinterpreting these cases (data) through the lens of a more realistic (i.e., less simplistic) model suggests otherwise (p. 265):

> The oft-cited rationale for bicarbonate use, that it might ameliorate the hemodynamic depression of metabolic acidemia, has been disproved convincingly.

Furthermore, echoing Felsenstein's conclusion about long-branch attraction, Forsythe and Schmidt take this finding to undermine both the implicit theoretical assumptions and conceptual justification for treating septic lactic acidosis with $NaHCO_3$ (p. 265):

> Even theoretical arguments in favor of sodium bicarbonate administration rely on a naïve representation of acid-base physiology, ignoring the complex compartmentalization of pH, the second-level effects of bicarbonate infusion, the impact of carbon dioxide generation, or the negative consequences of hyperlactatemia.

Using an overly simplistic yet intuitive model of physiology produced positively misleading results. Based on support from a history of case studies, there was convergence on a sub-optimal treatment for septic lactic acidosis. Yet this support was positively misleading, and obscured better treatment options. More precisely, the PME effectively excluded the saline-only treatment from the set of treatments under consideration.[34]

Re-analyzing these same data using more sophisticated physiological models helped identify and diagnose this PME, revealing stronger support for other treatments and producing a richer, more fruitful understanding of the system and condition. Had PMEs been as well understood as false positives or type I errors, these might have been recognized more quickly. At the least, Forsythe and Schmidt would have been able to frame their argument as an example of a well understood statistical error, aiding their argument and providing more familiar modes of diagnosing and dislodging a PME and more effectively made their case.

Forsythe and Schmidt's reasoning mirrors how phylogeneticists evaluated established hypotheses by reinterpreting the very same data used to generate those hypotheses. Phylogeneticists used more realistic (i.e., less simplistic) models and multiple modes of inference to reveal positively misleading errors in the entrenched analyses that first generated the established hypotheses. Furthermore, the presence of positively misleading errors undermined the monistic theoretical justification underwriting those methods, and challenged the intuitively appealing but overly simplistic assumptions built into those methods.

As in phylogenetics, the PME generated by the entrenched standard-of-care generated

[34] Reminiscent of Autzen (2018) (see footnote 19).

an epistemic trap. Dislodging the $NaHCO_3$ treatment protocol was difficult for many of the same reasons other PME discoveries generated resistance, e.g., it had become an entrenched standard-of care, with a history of (what was at least perceived as) success; alternative treatment protocols struck many as counterintuitive, etc. (Forsythe and Schmidt, 2000). The underlying cause of the PME here is less important than the pattern of error on display. Once recognized, a better understanding of the system (and treatment of it) became available. New fruitful avenues of clinical research became apparent, and the attending commitments in the field were richer for it, e.g., a more effective treatment of acidosis. These are good reasons to develop techniques for identifying and diagnosing PMEs, and appears to be another case of productive disruption spurred by diagnosing a PME.

3.2 Other Candidate Cases

In addition to long-branch attraction and septic lactic acidosis, there are other candidate cases of positively misleading errors. These range from recently discovered PMEs to more speculative cases. Establishing that these, or other cases, are PMEs will require a more careful and detailed analysis, but effectively identifying whether, and, if so, what other fields have PMEs will be important for developing a good understanding of these errors—and how to avoid them.

A more recent PME has been identified in phylogenetics by Degnan and Rosenberg (2006), who convincingly argue that inferring species trees from gene trees can lead to positively misleading results under certain conditions. These 'wicked forests' (as they are called) are cases where the most frequent gene tree among a group of taxa is guaranteed to differ from the species tree; this challenges an entrenched method for inferring a species tree from the most frequent gene tree topology. Under these conditions, as more gene trees are included in an analysis, the incorrect species tree will receive increasingly strong support at the expense of support for the correct tree (Degnan and Rosenberg, 2006, p. 762):

> Consequently, use of the most commonly observed gene tree topology to estimate the species tree topology—the "democratic vote" procedure among gene trees—can be "positively misleading," that is, convergent upon an erroneous estimate as the number of genes increases.

Notably, Degnan and Rosenberg attribute both the phrase and definition of 'positively misleading' to Felsenstein (1978). Furthermore, diagnosing this PME, and the conditions under which it may be present, led to a fruitful fine-tuning of how to read support of hypotheses off of the data (Degnan and Rosenberg, 2006, p. 767):

> The counterintuitive result is that if two trees from the same wicked forest were considered as hypotheses for a phylogeny, observing a higher propor-

> tion of gene trees that match one species tree would be evidence in favor of the other species tree, and vice versa.

This is precisely the sort of outcome that should motivate a better understanding of PMEs.

Though a bit more speculative, other candidate PME cases are widespread in the clinical care literature. These include the diagnosis of brain death (Black, 1978a,b), treatment of acute renal failure with dopamine (Denton et al., 1996), diagnosis of neonatal necrotizing entercolitis (Holt and Friedland, 1974), and emphysematous bullae (Morgan et al., 1989). Liao and Carbonell (2023) present a possible case with explicit social and ethical elements regarding how biases can get encoded and embedded in medical equipment like pulse oximeters and spirometers.[35] Careful analysis of each of these would be required to demonstrate the presence of PMEs, but, at least superficially, many of the relevant patterns appear to be present. This is particularly pressing in clinical medicine, given the rapid increase in access to 'big data' and the challenges of extracting good information and patterns from those datasets (Wang and Krishnan, 2014).

Economics and the social sciences also include sub-fields susceptible to PMEs. The datasets are often complex, with individual datum dependent on each other in numerous overlapping ways, coupled with the inclusion of overly simplistic but intuitively appealing entrenched commitments. Methods might identify any number of patterns that fail to correspond to the feature of the system we hope to be studying. One candidate case includes the economics of professional sports, e.g., the valuation of professional athletes in North American baseball. Entrenched methods of evaluating players were sub-optimal, incorrectly valuing players' contributions to generating wins for their teams. Dislodging these entrenched methods for more sophisticated ones, and the ensuing debates, have the hallmark of a PME-generated epistemic trap (Hakes and Sauer, 2006; for a popular account, see Lewis, 2003).

4 Conclusion

Positively misleading errors are errors of statistical reasoning where adding more data systematically produces increasingly greater strength of support for an erroneous hypothesis over the correct one. It is this pattern that distinguishes them from other errors of inference and pattern recognition. They are not caused by poor data, lack of data, sampling errors, chance, or failure rates, but are symptomatic of some other confounder.

[35] Liao and Carbonell (2023) provides an exemplar of how *philosophical* methodology may be employed to identify and resolve PMEs when social contexts and non-epistemic factors are playing a confounding role. This will be especially important for addressing cases of algorithmic biases in data science (Fazelpour and Danks, 2021; Knox et al., 2023).

Though well-known in phylogenetics, PMEs are likely widespread—particularly in fields focused on discovering patterns in large, complex datasets.

Though the focus here has been on developing a general account of PMEs by looking at exemplar and candidate cases, there are numerous other avenues by which PMEs could be examined. A broad recognition of PMEs promises to facilitate a better understanding of reasoning and error about complex systems, and to spur a shift towards better explanations, methods, and treatment protocols in the face of entrenched sub-optimal commitments.

PMEs may be contrapositives of Gettier cases. In the latter, we draw the right conclusion for the wrong reasons; PMEs, in contrast, get you to the wrong conclusion but for the right reasons. They may thus generate *epistemic traps*. The pattern of support generated by PMEs replicates at least one form of good reasoning, and, thus, can be particularly difficult to dislodge.

In many cases PMEs will be diagnosable. One lesson of this from the exemplar case—long-branch attraction—is that we ought to be at least minimal methodological pluralists. Especially so if we cannot know ahead of time under which conditions various modes of inference will generate PMEs. Even when we can know, different methods may perform better under different conditions. More troubling is how we might be certain that a case of convergence is *not* positively misleading. Appealing to an argument by convergence to establish that we are not making a positively misleading error will simply raise a new question of whether that argument is itself positively misleading, threatening a recursive epistemic trap.

Type I errors and false positives are well known and well understood, with well studied responses for resolving or treating these errors. That may include increasing sample size or re-sampling, or serially testing to guard against base-rate fallacies, respectively. Though the treatment for resolving PMEs in both the phylogenetic and clinical medicine case shared some similarities, more work needs to be done to identify an analogous family of responses for resolving PMEs. In the case of phylogenetics, we have also learned that it may not be a *single* response but a multi-pronged approach, e.g., identifying and testing for specific conditions known to generate PMEs; adopting a context-sensitive assessment of multiple methodologies; etc. Both cases also suggest that rather than adding new data, it may be more informative to review analyses of old datasets when underlying assumptions are updated. If those results are discordant with prevailing views, that may be an indication that a deeper assessment for PMEs is needed.

Yet, to be a bit more speculative, if social contexts and non-epistemic factors can also generate (or contribute to) PMEs (as may be the case for the pulse oximeter and spirometer, Liao and Carbonell, 2023), then we may need more than just statistical or probabilistic tools to successfully identify and resolve PMEs. This adds a moral and ethical imperative to determining whether reasoning errors might be categorized as PMEs. One goal of this paper is to prompt this sort of research and more.

Nevertheless, PMEs *can* be dislodged, and the cases above provide insight into what we stand to gain from a better understanding of how this happens. Diagnosing a PME may not only permit us to avoid a potentially costly error, but to also enhance our reasoning about and understanding of complex systems. As such, we should seek to identify a wide array of PMEs in order to build a rich set of examples for further study. As more cases are diagnosed across disciplines, our understanding of them—and how to avoid them—will improve. Philosophers of science are especially well positioned to get out in front on this and explore the implications and special challenges that might accompany positively misleading errors.

References

Altman, D. G. and J. M. Bland (1994a). Statistics notes: Diagnostic tests 1: sensitivity and specificity. *BMJ 308*(6943), 1552. https://doi.org/10.1136/bmj.308.6943.1552.

Altman, D. G. and J. M. Bland (1994b). Statistics notes: Diagnostic tests 2: predictive values. *BMJ 309*(6947), 102. https://doi.org/10.1136/bmj.309.6947.102.

Ashcroft, R. E. (2004). Current epistemological problems in evidence based medicine. *Journal of Medical Ethics 30*(2), 131–135. https://doi.org/10.1136/jme.2003.007039.

Autzen, B. (2011). Constraining prior probabilities of phylogenetic trees. *Biology & Philosophy 26*(4), 567–581. https://doi.org/10.1007/s10539-011-9253-7.

Autzen, B. (2018). Bayesian convergence and the fair-balance paradox. *Erkenntnis 83*(2), 253–263. https://doi.org/10.1007/s10670-017-9888-0.

Bell, G., T. Hey, and A. Szalay (2009). Beyond the data deluge. *Science 323*(5919), 1297–1298. https://doi.org/10.1126/science.1170411.

Black, P. M. (1978a). Brain death (part 1). *New England Journal of Medicine 299*(7), 338–344. https://doi.org/10.1056/NEJM197808172990705.

Black, P. M. (1978b). Brain death (part 2). *New England Journal of Medicine 299*(8), 393–401. https://doi.org/10.1056/NEJM197808242990805.

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43*(1), 69 – 80. https://doi.org/10.1016/j.shpsc.2011.10.007.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties 2*(1), 11–20. https://doi.org/10.1017/S1745855207005029.

Cartwright, N. and J. Stegenga (2011). A theory of evidence for evidence-based policy. In W. Twining, P. Dawid, and D. Vasilaki (Eds.), *Evidence, inference and enquiry*, Chapter 11, pp. 290–322. Oxford University Press.

Casscells, W., A. Schoenberger, and T. B. Graboys (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine 299*(18), 999–1001. https://doi.org/10.1056/NEJM197811022991808.

Chang, J. T. (1996). Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences 134*(2), 189 – 215. https://doi.org/10.1016/0025-5564(95)00172-7.

Cooper, D. J., K. R. Walley, B. R. Wiggs, and J. A. Russell (1990). Bicarbonate does not improve hemodynamics in critically ill patients who have lactic acidosis. *Annals of Internal Medicine 112*(7), 492–498. https://doi.org/10.7326/0003-4819-112-7-492.

Degnan, J. H. and N. A. Rosenberg (2006, 05). Discordance of species trees with their most likely gene trees. *PLoS Genet 2*(5), e68. https://doi.org/10.1371/journal.pgen.0020068.

Denton, M. D., G. M. Chertow, and H. R. Brady (1996, 07). "Renal-dose" dopamine for the treatment of acute renal failure: Scientific rationale, experimental studies and clinical trials. *Kidney International 50*(1), 4–14. https://doi.org/10.1038/ki.1996.280.

Doolittle, W. F. (2000). Uprooting the tree of life. *Scientific American 282*(2), 90–95. http://www.jstor.org/stable/26058605.

Fazelpour, S. and D. Danks (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass 16*(8), e12760. https://doi.org/10.1111/phc3.12760.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology 27*(4), 401–410. https://doi.org/10.1093/sysbio/27.4.401.

Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics 22*(1), 521–565. https://doi.org/10.1146/annurev.ge.22.120188.002513.

Felsenstein, J. (2001). The troubled growth of statistical phylogenetics. *Systematic Biology 50*(4), 465–467. https://doi.org/10.1080/10635150119297.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, Inc.

Fitzhugh, K. (2006). The abduction of phylogenetic hypotheses. *Zootaxa 1145*(1), 1–110. https://doi.org/10.11646/zootaxa.1145.1.1.

Forsythe, S. M. and G. A. Schmidt (2000). Sodium bicarbonate for the treatment of lactic acidosis*. *Chest 117*(1), 260–267. https://doi.org/10.1378/chest.117.1.260.

Galla, M., K. Wicke, and M. Fischer (2019). Statistical inconsistency of maximum parsimony for k-tuple-site data. *Bulletin of Mathematical Biology 81*(4), 1173–1200. https://doi.org/10.1007/s11538-018-00552-2.

Galtier, N. and V. Daubin (2008). Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences 363*(1512), 4023–4029. https://doi.org/10.1098/rstb.2008.0144.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis 23*(6), pp. 121–123. https://doi.org/10.1093/analys/23.6.121.

Haber, M. H. (2005). On probability and systematics: Possibility, probability, and phylogenetic inference. *Systematic Biology 54*(5), 831–841. https://doi.org/10.1080/106351591007444.

Haber, M. H. (2009). Phylogenetic inference. In A. Tucker (Ed.), *A Companion to the Philosophy of History and Historiography*, Number 41 in Blackwell Companions to Philosophy, Chapter 20, pp. 231–242. Malden, MA: Wiley-Blackwell.

Haber, M. H. (2012). Multilevel lineages and multidimensional trees: The levels of lineage and phylogeny reconstruction. *Philosophy of Science 79*(5), 609–623. https://doi.org/10.1086/667849.

Haber, M. H. (2019). Species in the age of discordance. *Philosophy, Theory, and Practice in Biology 11*(21), 1–22. https://doi.org/10.3998/ptpbio.16039257.0011.021.

Haber, M. H. and J. Velasco (2021). Phylogenetic Inference. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/phylogenetic-inference/.

Hakes, J. K. and R. D. Sauer (2006). An economic evaluation of the *Moneyball* hypothesis. *Journal of Economic Perspectives 20*(3), 173–185. https://doi.org/10.1257/jep.20.3.173.

Hasegawa, M., H. Kishino, and T.-a. Yano (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22*, 160–174. https://doi.org/10.1007/BF02101694.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham (1994). Application and accuracy of molecular phylogenies. *Science 264*(5159), 671–677. https://doi.org/10.1126/science.8171318.

Holt, S. and G. Friedland (1974). Neonatal necrotizing enterocolitis: Clinical and radiological features. *Western Journal of Medicine 120*(2), 110–115.

Howe, D., M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, and S. Yon Rhee (2008, 09). Big data: The future of biocuration. *Nature 455*(7209), 47–50. https://doi.org/10.1038/455047a.

Howson, C. and P. Urbach (1991, 04). Bayesian reasoning in science. *Nature 350*(6317), 371–374. https://doi.org/10.1038/350371a0.

Huelsenbeck, J. P. and K. A. Crandall (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics 28*, pp. 437–466. https://doi.org/10.1146/annurev.ecolsys.28.1.437.

Hull, D. L. (1983). Karl Popper and Plato's metaphor. In N. Platnick and V. Funk (Eds.), *Advances in Cladistics*, Volume 2, pp. 177–189. Columbia University Press.

Hull, D. L. (1988). *Science As A Process*. Chicago, IL: The University of Chicago Press.

Inward, D., G. Beccaloni, and P. Eggleton (2007). Death of an order: a comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. *Biology Letters 3*(3), 331–335. https://doi.org/10.1098/rsbl.2007.0102.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In M. N. Munro (Ed.), *Mammalian protein metabolism*, Volume III, pp. 21–132. N. Y.: Academic Press.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution 16*, 111–120. https://doi.org/10.1007/BF01731581.

Knox, B., P. Christoffersen, K. Leggitt, Z. Woodruff, and M. H. Haber (2023). Justice, vulnerable populations, and the use of conversational AI in psychotherapy. *The American Journal of Bioethics 23*(5), 48–50. https://doi.org/10.1080/15265161.2023.2191040.

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 44*(4, Part A), 503 – 514. https://doi.org/10.1016/j.shpsc.2013.03.020.

Leonelli, S. and R. A. Ankeny (2012). Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43*(1), 29 – 36. https://doi.org/10.1016/j.shpsc.2011.10.003.

Lewis, M. (2003). *Moneyball: The Art of Winning An Unfair Game*. New York: Norton.

Liao, S.-y. and V. Carbonell (2023). Materialized oppression in medical tools and technologies. *The American Journal of Bioethics 23*(4), 9–23. https://doi.org/10.1080/15265161.2022.2044543.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology 46*(3), 523–536. https://doi.org/10.1093/sysbio/46.3.523.

Mathieu, D., R. Neviere, V. Billard, M. Fleyfel, and R. Wattel (1991). Effects of bicarbonate therapy on hemodynamics and tissue oxygenation in patients with lactic acidosis: A prospective, controlled clinical study. *Critical Care Medicine 19*(11), 1352–1356. https://doi.org/10.1097/00003246-199111000-00008.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.

Mayo, D. G. (1997). Severe tests, arguing from error, and methodological underdetermination. *Philosophical Studies 86*, 243–266. https://doi.org/10.1023/A:1017925128970.

Mitchell, S. D. (2009). *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.

Morgan, M. D. L., C. W. Edwards, J. Morris, and H. R. Matthews (1989). Origin and behaviour of emphysematous bullae. *Thorax 44*(7), 533–538. https://doi.org/10.1136/thx.44.7.533.

Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution 28*(12), 719 – 728. https://doi.org/10.1016/j.tree.2013.09.004.

Nature Editorial (2008). Community cleverness required. *Nature 455*(7209), 1. https://doi.org/10.1038/455001a.

Pietsch, W. (2013, August). Big data? the new science of complexity. In *6th Munich-Sydney-Tilburg Conference on Models and Decisions*, pp. 1–20. http://philsci-archive.pitt.edu/9944/.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson & Co.

Quinn, A. (2016). Phylogenetic inference to the best explanation and the bad lot argument. *Synthese 193*, 3025–30329. https://doi.org/10.1007/s11229-015-0908-9.

Rice, C. and K. Khalifa (2025). Thank you for misunderstanding! *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02311-1.

Siddall, M. E. (1998). Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris zone. *Cladistics 14*(3), 209–220. https://doi.org/10.1111/j.1096-0031.1998.tb00334.x.

Sober, E. (1983). Parsimony in systematics: Philosophical issues. *Annual Review of Ecology and Systematics 14*, pp. 335–357. https://doi.org/10.1146/annurev.es.14.110183.002003.

Sober, E. (1988). *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge, MA: MIT Press.

Sober, E. (2004). The contest between parsimony and likelihood. *Systematic Biology 53*(4), 644–653. https://doi.org/10.1080/10635150490468657.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Spanos, A. (2010). Is frequentist testing vulnerable to the base-rate fallacy? *Philosophy of Science 77*(4), 565–583. https://doi.org/10.1086/656009.

Stacpoole, P. W., E. C. Wright, T. G. Baumgartner, R. M. Bersin, S. Buchalter, S. H. Curry, C. A. Duncan, E. M. Harman, G. N. Henderson, S. Jenkinson, J. M. Lachin, A. Lorenz, S. H. Schneider, J. H. Siegel, W. R. Summer, D. Thompson, C. L. Wolfe, and B. Zorovich (1992). A controlled clinical trial of dichloroacetate for treatment of lactic acidosis in adults. *New England Journal of Medicine 327*(22), 1564–1569.

Steel, M. and D. Penny (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution 17*(6), 839–850. https://doi.org/10.1093/oxfordjournals.molbev.a026364.

Sterner, B. and S. Lidgard (2014). The normative structure of mathematization in systematic biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 46*(0), 44 – 54. https://doi.org/10.1016/j.shpsc.2014.03.001.

Sterner, B. and S. Lidgard (2018). Moving past the systematics wars. *Journal of the History of Biology 51*(1), 31–67.

Stiassny, M. L. J. and A. Meyer (1999). Cichlids of the rift lakes. *Scientific American* February, 64–69.

Sullivan, J. and P. Joyce (2005). Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics 36*(1), 445–466. https://doi.org/10.1146/annurev.ecolsys.36.102003.152633.

Swofford, D. L. and J. Sullivan (2009). Phylogeny inference based on parsimony and other methods using PAUP. In P. Lemey, M. Salemi, and A.-M. Vandamme (Eds.), *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, pp. 267–312. Cambridge: Cambridge University Press.

Tuffley, C. and M. Steel (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology 59*, 581–607. https://doi.org/10.1007/BF02459467.

Velasco, J. (2008). The prior probabilities of phylogenetic trees. *Biology and Philosophy 23*, 455–473. https://doi.org/10.1007/s10539-007-9105-7.

Velasco, J. D. (2013). Philosophy and phylogenetics. *Philosophy Compass 8*(10), 990–998. https://doi.org/10.1111/phc3.12070.

Vogt, L. (2013). eScience and the need for data standards in the life sciences: in pursuit of objectivity rather than truth. *Systematics and Biodiversity 11*(3), 257–270. https://doi.org/10.1080/14772000.2013.818588.

Wang, W. and E. Krishnan (2014, Jan). Big data and clinicians: A review on the state of the science. *JMIR Med Inform 2*(1), e1. https://doi.org/10.2196/medinform.2913.

Ward, P. S. (2014). The phylogeny and evolution of ants. *Annual Review of Ecology, Evolution, and Systematics 45*(1), 23–43. https://doi.org/10.1146/annurev-ecolsys-120213-091824.

Ward, P. S., S. G. Brady, B. L. Fisher, and T. R. Schultz (2015). The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). *Systematic Entomology 40*(1), 61–81. https://doi.org/10.1111/syen.12090.

Watson, J., P. F. Whiting, and J. E. Brush (2020). Interpreting a covid-19 test result. *BMJ 369*. https://doi.org/10.1136/bmj.m1808.

Wiley, E. O. (1975). Karl R. Popper, systematics, and classification: A reply to Walter Bock and other evolutionary taxonomists. *Systematic Zoology 24*(2), 233–243. https://doi.org/10.1093/sysbio/24.2.233.

Wimsatt, W. C. (1974). Complexity and organization. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1972*, 67–86.

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science 69*(S3), pp. S316–S330. https://doi.org/10.1086/341855.

Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass 2*(6), 981–1022. https://doi.org/10.1111/j.1747-9991.2007.00106.x.

Wu, X., X. Zhu, G.-Q. Wu, and W. Ding (2014, Jan). Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on 26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109.

Yang, Z. (2007). Fair-balance paradox, star-tree paradox, and bayesian phylogenetics. *Molecular Biology and Evolution 24*(8), 1639–1655. https://doi.org/10.1093/molbev/msm081.

Yang, Z. and B. Rannala (2012, 05). Molecular phylogenetics: principles and practice. *Nat Rev Genet 13*(5), 303–314. https://doi.org/10.1038/nrg3186