Correlation isn't good enough: Causal explanation and Big Data

Gary Smith and Jay Cordes: *The Phantom Pattern Problem: The Mirage of Big Data*. New York: Oxford University Press, 2020, 240 pp, 25.00 € HB

Frank Cabrera
cabrera@msoe.edu
Humanities, Social Science, and Communication Department
Milwaukee School of Engineering
Milwaukee, Wisconsin, USA

In their book *The Phantom Pattern Problem*: *The Mirage of Big Data*, economist Gary Smith and statistician Jay Cordes demonstrate with an arsenal of examples derived from such diverse areas as sports, finance, gambling, astronomy, medicine, etc., the pervasiveness of so-called "phantom patterns", i.e., coincidental past correlations that have little to no future predictive value. In addition to illustrating the ease with which humans can be fooled by these coincidental correlations, the authors provide many strategies throughout the book to avoid being misled by phantom patterns. The book is written in a highly engaging, conversational style, is largely non-technical, and is therefore suitable for a wide range of audiences. At the risk of mistakenly identifying phantom patterns within the book itself, I will attempt to characterize the general structure of *The Phantom Pattern Problem* and discuss what to my mind are some of its most important themes.

In the Introduction, Smith and Cordes (henceforth "S&C") begin with a case from 2007 in which the Urban Institute, "a highly regarded Washington think-tank" proposed that America was experiencing an "iCrime Wave" based on an observed correlation between the rise in reported murders and the rise in iPod sales between 2004 and 2006 (1). However, as S&C point out, this correlation (i) was based on only three years of data, (ii) was no stronger than the correlation between ice cream sales and violent crime in the same period, and (iii) vanished shortly thereafter as the murder rate dropped in 2007. This example serves as a cautionary tale: even experts can be bewitched by misleading patterns, a danger that has become more acute in the era of "Big Data", an age in which we now have the ability to gather, process, and analyze massive quantities of data.

In many ways, the example of the spurious correlation between murder rates and iPod sales sets the tone for the book. In the succeeding chapters, the reader will be confronted with myriad similar examples: cases in which someone mistakenly infers a causal explanation from *merely* correlative data, or perhaps what is worse, cases in which someone tries (and fails) to predict future events based on past correlations in the absence of any causal explanation. While S&C examine several themes that will be of interest to philosophers of science and epistemologists, e.g., the role of causal knowledge in predictive inference, the problem of over-fitting the data, the replication crisis, etc., those readers with mostly abstract, philosophical concerns may find the quantity of examples unnecessary and somewhat tedious. For instance, S&C spend about six pages (180-7) charting the rise and fall of the California-based web services provider *Yahoo! Inc.* to illustrate the simple point that a "CEO is seldom the reason for a company's success" (187). Others without a taste for sports might grow weary of the extensive discussion of the controversial "hot-hands" phenomenon in basketball (80-6).

When viewed in a certain light, however, the frequency with which S&C make recourse to examples of phantom patterns serves an important, salutary function. As S&C argue in Chapter 1, our general propensity to recognize patterns in the world is likely a crucial evolutionary adaptation. As a result, we are probably "hard-wired to notice patterns" (19). But since natural selection is not an optimization process, our pattern-recognition abilities will sometimes misfire, leading us to draw unfounded conclusions from useless, meaningless, and misleading patterns. As many of the examples in the book illustrate, even expert data scientists (26), financial consultants (35), medical researchers (58), astronomers (68), and economists (145) can fall into this epistemic trap. One can think of *The Phantom Pattern Problem*, then, partly as a kind of "epistemological self-help" book. To change bad habits, we often must resort to processes of repetition and reinforcement. So, even the antecedently skeptical reader—someone who is already aware of the danger of spurious correlations—will likely benefit from the many examples explored in the book.

Of course, S&C not only want to highlight how widespread these coincidental correlations are, but they also seek to defend several substantive methodological claims. One of the central targets of the book—made most explicit in Chapter 5—is the claim, sometimes heard in data science circles, that soon predictive analytics algorithms and data mining techniques will replace or radically alter traditional scientific methodology. Some of these alleged changes include: (i) a shift from a "hypothesis-driven" method to a purely "data-driven" method and (ii) "a move away from the age-old search for causality", with a focus instead on correlations, which are thought to be "good enough" (Mayer-Schönberger and Cukier 2013, 14). Perhaps as this new paradigm of data-intensive science develops, the scientist will gradually be replaced by algorithms, which will at last allow the data "to speak for themselves" unhindered by human biases.

Throughout the book, S&C come to the rescue of traditional scientific methodology by arguing against such lofty pretensions. While each chapter contains some helpful lessons—e.g., in Chapter 4, random processes will invariably generate *some* striking patterns—perhaps the most philosophically rich chapter of all is Chapter 2. There, S&C draw a distinction between "meaningful" and "meaningless" patterns, where meaningful patterns are ones that have "an underlying causal explanation" (23). To be sure, data mining techniques can uncover models that fit past data, but the real test of a model is the prediction of new data, and according to S&C, "consistently reliable predictions with fresh data require a causal structure" (24). Now, S&C do not claim that we need to know *exactly* what the correct causal explanation of the correlation is. Perhaps *A* causes *B*, or some third factor *C* may be the common cause of both *A* and *B*. But if there is no plausible explanation for some correlation, e.g., the correlation between "avocado prices in San Francisco in 2015 and Google searches for the Virgo zodiac sign", then the correlation should be regarded as "a fleeting coincidence that is useless for making predictions" (25).

On S&C's view, the best way to establish causality is through randomized controlled trials (RCTs), a method often regarded as the "gold standard" in medical research (33). Often, of course, RCTs prove impractical, in which case we must rely on observational data. In such cases, S&C appeal to the tried-and-true, hypothetico-deductive method. According to S&C, instead of resorting to data-mining techniques unmoored from human expertise in order to uncover patterns in past data, we ought to stick with traditional scientific method and begin with a "plausible theory" formulated in advance that we then go on to *test* by looking at *new* data (38). In Chapter 2, S&C motivate the hypothesis-driven approach over the data-driven approach primarily by discussing examples in which the latter fails to predict new data, whereas the former succeeds. Further support for the hypothesis-driven approach can be provided by a point that is explored most fully in Chapter 5, which S&C call the "paradox of big data": meaningless patterns invariably arise in large data sets, and since the number of meaningful patterns is likely fixed, as we gather more data, the probability that any randomly selected pattern is meaningful radically decreases. In the Epilogue, S&C introduce Bayesian reasoning, and argue based on such considerations that the prior probability that a randomly selected correlation is meaningful will be extremely low (215-16).

By and large, I agree with S&C's main methodological claims, as well as their critique of the provocative prognostications of Big Data enthusiasts (e.g., Cabrera 2020). However, I suspect that the philosopher of science or the epistemologist reading the *Phantom Pattern Problem* will find herself wishing for more clarification or defense at various junctures. For example, one of S&C's central points is that causal knowledge is indispensable for prediction. But what exactly constitutes *causation* is a matter of long-standing, contentious debate (Schaffer 2016). Similarly, as is well-known among philosophers of science, it is a non-trivial problem to distinguish, in a principled way, lawlike generalizations, which are suitable for prediction, from accidentally true generalizations. Unfortunately, S&C have little to say to about such matters, leaving the concept of causation largely intuitive. Additionally, in their defense of the hypothesis-driven approach over the data-driven approach, S&C implicitly appeal to the superiority of *novel predictions*. However, S&C do not provide much by way of theoretical justification for this assumption, something which has been questioned by several philosophers of science (Barnes 2018). Indeed, according to some prominent accounts of scientific reasoning, such as "abduction" or "inference to the best explanation" (Lipton 2004), hypotheses can gain support *precisely* because of their ability to explain *past* data. Those sympathetic with S&C's position will want such a justification, if only to ensure that their impression that prediction trumps accommodation is not *itself* a phantom pattern. Despite these philosophical quibbles, overall, I found the *Phantom Pattern Problem* a worthwhile and enjoyable read, and so I happily recommend the book to anyone interested in the epistemological issues raised by Big Data.

References

Barnes, E. (2018). "Prediction versus Accommodation," *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/ fall2018/ entries/prediction-accommodation/>.

Cabrera, F. (2020). "The Fate of Explanatory Reasoning in the Age of Big Data," *Philosophy & Technology*, pp. 1–21 doi:10.1007/s13347-020-00420-9

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

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

Schaffer, J. (2016). "The Metaphysics of Causation", *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2016/ entries/causation-metaphysics/>.