# Causal Inference From Noise

**Nevin Climenhaga**
Australian Catholic University
nevin.climenhaga@acu.edu.au
sites.google.com/site/nevinclimenhaga

**Lane DesAutels**
Missouri Western State University
lane.desautels@gmail.com
lanedesautels.com

**Grant Ramsey**
Institute of Philosophy
KU Leuven
grant@theramseylab.org
www.theramseylab.org

**Abstract** *Correlation is not causation* is one of the mantras of the sciences—a cautionary warning especially to fields like epidemiology and pharmacology where the seduction of compelling correlations naturally leads to causal hypotheses. The standard view from the epistemology of causation is that to tell whether one correlated variable is causing the other, one needs to intervene on the system—the best sort of intervention being a trial that is both randomized and controlled. In this paper, we argue that some purely correlational data contains information that allows us to draw causal inferences: statistical noise. Methods for extracting causal knowledge from noise provide us with an alternative to randomized controlled trials that allows us to reach causal conclusions from purely correlational data.

**1. Introduction.** Suppose you want to know whether diet soft drinks cause type-2 diabetes. You examine the historical data and find that cases of type-2 diabetes have occurred more often among individuals who report consuming one or more diet soft drink per day. Is this evidence sufficient to conclude that diet soft drinks cause type-2 diabetes? The answer, according to statistical orthodoxy, is an emphatic *no*. The historical data reveal a correlation between the consumption of diet soft drinks and type-2 diabetes, but this does not—and cannot—by itself show that the diet soft drinks are the relevant causal factor in the increased rates of type-2 diabetes. It might be that the direction of causation goes the other way: perhaps type-2 diabetes increases thirst, which increases diet soft drink consumption. Or perhaps individuals who consume large amounts of diet soda also tend to be lower in economic status, and people with fewer economic resources tend to consume more calorie dense processed foods. If this is the case, then the relatively high incidence of type-2 diabetes among diet soft drink consumers may be a result of other features of their diet. Processed food consumption, in this case, is a confounding causal factor hidden in the observed correlation between diet soft drinks and type-2 diabetes.

In light of these difficulties, it is commonly thought—and sometimes explicitly argued (Fisher 1947, Papineau 1994)—that the only way to eliminate confounding causal factors when testing the effect of an intervening factor on some outcome of interest is to conduct *randomized controlled trials* (RCTs). Inferring causation from historical or observational studies, for reasons illustrated above, is claimed to be epistemologically untenable. Fisher writes:

> The full procedure of randomization [is the method] by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated. (1947, p. 19)

Consider again our diet soft drink example. In order to meet the conditions required for an RCT, you would need to randomly assign test subjects into two groups: an *experimental* group provisioned with and assigned to drink diet sodas, and a *control* group who does not drink diet soda. By randomly selecting who goes into which group, we should end up with people of varying ages, socioeconomic backgrounds, levels of fitness, and any other potentially relevant factors in *both* groups. This aims to achieve Fisher's guarantee by ensuring that diet soda consumption is the sole causal difference between the two groups and, as such, eliminates any hidden causal influence from skewing the results. It also ensures that the subjects drinking more diet soda are doing so not because of type-2 diabetes, but because we assigned them to do so. If instances of type-2 diabetes are higher in the experimental group than in the control, then (and only then) can we conclude that diet sodas cause type-2 diabetes.

Philosophers and scientists continue to debate the viability of RCTs for determining causality (Urbach 1985, Papinau 1994, Worrall 2007). In what follows, we argue that there is an alternative to RCTs not appreciated in discussions of the epistemology of causation: drawing causal inferences from statistical noise. The philosophical literature generally treats statistical noise as a nuisance. It is thought to mask causal relations and to interfere with theory choice. Forster and Sober, for example, argue that "if we think of the true curve as the 'signal' and the deviation from the true curve generated by errors of observation as 'noise', then fitting the data perfectly involves confusing the noise with the signal" (1994, p. 5).

This view of noise—as merely obscuring causal information—has remained unchallenged. We challenge this orthodoxy by showing how, in some cases, not only does noise not obscure causal relations, it is an invaluable source of insight regarding them. If this is right, the signal-noise dichotomy offered by Forster and Sober, and generally held by the philosophical community, is in error. Noise itself can *carry a signal*, and an important one, one from which we can gain knowledge of the causal underpinnings of our world.

New statistical research helps to support our arguments by offering rigorous techniques for making reliable causal inferences from the statistical noise in historical, observational data. We call these techniques Noise Inference Methods (NIMs). If successful, NIMs may allow us to make causal inferences in domains in which experimentation carries prohibitive costs or is impossible. If these techniques could be applied to fields like epidemiology and drug development, the ethical payoffs could be tremendous. At the very least, these new techniques warrant careful consideration within the context of the ongoing philosophical debate regarding theory choice and the importance of RCTs for the epistemology of causation.

Our paper proceeds as follows. In Section 2, we demonstrate through a series of examples the in-principle possibility of causal inference from noise. In Section 3, we consider the application of NIMs to real-world datasets. In Section 4, we consider some theoretical implications of our argument to the epistemology of causation. In Section 5, we conclude.

**2. Causal Inferences from Noisy Data.** As our diet soft drink example shows, there are good reasons for thinking that we cannot rationally infer that X causes Y from mere knowledge that X and Y are correlated. For the most part, defenders of the statistical orthodoxy that one cannot draw particular causal inferences from correlational data will agree that an observed correlation between $X$ and $Y$ must be due either to some causal relation between $X$ and $Y$ or to chance.[1] But even in cases in which we have confidence that the correlation between $X$ and $Y$ is due to a causal relation between them, it appears that from mere historical data we cannot conclude that $X$ causes $Y$ rather than $Y$ causing $X$ or their having a common cause. Fortunately, for many correlations we know not merely *that* $X$ and $Y$ are correlated, but *how* they are correlated. And this information can serve as a key to unlock an overlooked source of information for inferring causal relations: noise. We will now consider how this can occur.

Let $X \rightarrow Y$ stand for the hypothesis that $X$ causes $Y$, $X \leftarrow Y$ stand for the hypothesis that $Y$ causes $X$, and $X \leftarrow Z \rightarrow Y$ stand for the hypothesis that $X$ and $Y$ have a common cause.[2] Where $E$ is the historical data on which $X$ and $Y$ are correlated, the worry that many have about inferring $X \rightarrow Y$ from historical trials is that $P(E|X \rightarrow Y) = P(E|X \leftarrow Y) = P(E|X \leftarrow Z \rightarrow Y)$. That is to say, we would expect $X$ and $Y$ to be correlated, *whatever* the causal connection is between them. Hence,

[1] Sober (2001) offers the following illustration of how events can be correlated over long spans of time in the absence of a causal link: although there is no causal connection between the price of bread in Britain and the height of the sea in Venice, they nevertheless both tend to increase over time. Correlation thus does not imply a causal connection. Nevertheless, as Sober points out, in many cases the hypothesis that there is no causal connection between $X$ and $Y$ predicts that there is no correlation between $X$ and $Y$. (Most events that are not causally connected will not be correlated.) The examples we consider in this paper are all ones in which correlation does appear to indicate some causal connection.

[2] Note that more than one of $X \rightarrow Y$, $X \leftarrow Y$, and $X \leftarrow Z \rightarrow Y$ could be true, and in many cases more than one is true. Nevertheless, we may be interested in determining whether $X \rightarrow Y$ is true, irrespective of whether $X \leftarrow Y$ or $X \leftarrow Z \rightarrow Y$ is also true. The objection to inferring $X \rightarrow Y$ from $E$ stated in this paragraph, and the response to it in the rest of this section, apply whether the hypotheses are compatible in a given case or not.

we cannot use the fact that X and Y are correlated to discriminate between the different possible causal connections between X and Y.

This argument has some plausibility. It is indeed true that, ordinarily, any of the above causal hypotheses will predict that X and Y will be correlated. However, ordinarily, our evidence E will not merely tell us that X and Y are correlated: rather, it will tell us that they are correlated in a specific way. And particular *kinds* of correlation, it turns out, will often be more likely on one causal hypothesis than on another. When we take into account all available historical evidence, then, we may indeed be able to legitimately infer (with high probability) that X→Y from solely historical data.

A familiar example of this is when our historical evidence includes temporal lags between X and Y. For example, circadian temperature and sunshine cycles are correlated. It is warmer during the day when the sun is shining and cooler at night when it is dark. But does the rising sun cause the earth to warm or does the warming earth cause the sun to rise? One piece of evidence supporting the former hypothesis is that there is a lag: if the sun crests at noon, we can observe that the temperatures are cooler at nine in the morning then they are at three in the afternoon, despite the solar intensity being the same. This lag in temperature values is more likely given that the rising sun causes the earth to warm than given that the warming earth causes the sun to rise—more likely, that is, given X→Y than given X←Y, where X is the rising sun and Y is the rising temperature.

While a temporal lag can be informative, in many cases there is no practical way to obtain data of sufficiently high grain to discern such lags. We will now show that even in the absence of temporal lag data, particular patterns of noise can be more likely given certain causal connections than other causal connections.

Our argument will proceed via a series of progressively more complicated cases. In this section, we will consider four thought experiments showing the in-principle possibility of using patterns of noise to discriminate between X→Y, X←Y, and X←Z→Y when these hypotheses are mutually exclusive. In the next section, we will look at the prospects of using Noise Inference Methods to discriminate between causal hypotheses using real-world datasets.

Consider first the following simple example.

### CASE 1

> Max the Mathematician has placed you in a room with two computers, X and Y. Each displays a number. Max tells you that he has a sliding dial that he uses to alter the number displayed on *one* of the computers, but he does not tell you which one. The number displayed on the *other* computer, he tells you, is determined by the number displayed on the one he controls, but it is modified by a mathematical function. If he controls X, then X sends a signal to Y to make its number twice X's; if he controls Y, then Y sends a signal to X to make its number half Y's. Your task is to observe the numbers displayed on the computers at various times, and to figure out which direction the causal influence is running: is X causing the output of Y? Or is Y causing the output of X?

Assuming there is no discernable temporal lag between the displays, this task is impossible if the causation from X to Y or Y to X works perfectly as described above—for these two hypotheses predict any observed pairs of numbers equally well. However, suppose that Max admits to you that there is an occasional glitch in the message sent from the input computer to the output computer. In particular, this glitch leads the output computer to display a number 1 greater or 1 less than the value it is supposed to display. If you know this, your task is now possible. Suppose you record the series of values in Table 1.

|  | ***X*** | ***Y*** |
|---|---|---|
| $t_1$ | .5 | 1 |
| $t_2$ | -11 | -22 |
| $t_3$ | 4 | 9 |

**Table 1**. Your recordings from your observations of the computers in the first case.

Observations $t_1$ and $t_2$ do not distinguish between the two causal hypotheses. Observation $t_3$, however, does. Where $x$ is the value displayed on computer X, and $y$ is the value on computer Y, it is consistent with $t_3$ that $y = 2x + \text{noise} = 2(4) + \text{noise} = 8 \pm 1$, but it is not consistent with $t_3$ that $x = y/2 + \text{noise} = 9/2 + \text{noise} = 4.5 \pm 1$.

In this first example, it was part of our background knowledge that either X→Y or X←Y. Consider now a second example, in which X←Z→Y is a third possibility.

### Case 2

The glitch functions as before, but Max tells you that one of *three* causal hypotheses is true: either X determines Y according to the function $y = 2x + \text{noise}$, Y determines X according to the function $x = y/2 + \text{noise}$, or both Y and X are determined by a third hidden computer Z, according to the functions $x = 2z + \text{noise}$ and $y = 4z + \text{noise}$. In this third case, the glitch might affect both output computers, it might affect just one of them, or it might affect neither.

|  | ***X*** | ***Y*** |
|---|---|---|
| $t_1$ | .5 | 1 |
| $t_2$ | -11 | -22 |
| $t_3$ | 4 | 9 |
| $t_4$ | 3 | 4 |

**Table 2**. Your recordings from your observations of the computers in the second case.

Suppose you record the same three values as before, and one more, as in Table 2. In this case, observation $t_3$, as before, lets you rule out $x = y/2 + \text{noise} = 9/2 + \text{noise} = 4.5 \pm 1$. This time, however, observation $t_4$ also lets you rule out $y = 2x + \text{noise} = 2(3) + \text{noise} = 6 \pm 1$. Thus, while $t_3$ is consistent with X→Y, and $t_4$ is consistent with X←Y, they are jointly inconsistent with both. The only remaining hypothesis compatible with the data is X←Z→Y, according to which $x = 2z + \text{noise}$ and $y = 4z + \text{noise}$. This hypothesis could result in $t_3$ when $z = 2$, if the glitch affects Y but not X, and $t_4$ when $z = 1$, if the glitch affects X but not Y.

If, instead of observing a result incompatible with X→Y, you continued to gather data that were consistent with both X→Y and X←Z→Y, this would be evidence that X is determining the output of Y. This is because, if Z is determining the output of both X and Y, you should expect that eventually you will see the glitch manifest itself in the value displayed on X; if you do not see this, this is evidence that Z is not the causal root.

These examples illustrate how "imperfect" causation sometimes allows us to draw causal inferences that would be impossible if the variables we are studying were perfectly related. In these examples, the glitch in the program is causal noise that makes the output of our causal function identifiably different than its input.

These cases are unrealistically simple since we knew the precise possible causal functions and we knew the precise magnitude of the noise. Can we still use noise as evidence for causal hypotheses in cases that are more realistic? Let's consider a more complicated example.

### CASE 3

> As before, we have the two computers, X and Y. This time, all that Max tells you is that the value on one of the computers determines the value on the other according to some mathematical function. This process is again glitchy, but the glitch is not as constant as before. The glitch either adds or subtracts from the output, and it can add or subtract any number. However, it is more likely to add or subtract a small amount than a large amount. In particular, the value on the output computer differs from the true value[3] by a particular amount with the probability represented by a normal Gaussian distribution. Graphically, this looks like a bell curve.

The Gaussian noise created by the glitch in this case is more realistic than the artificial noise pattern created by the glitch in Cases 1 and 2.[4] This noise, however, also leads to asymmetries in our observations of the cause and the effect. Suppose that, as before, you

[3] By 'true' value, we mean the value derived from the function alone.

[4] Gaussian noise is common in nature partly because non-Gaussian distributions approach Gaussian distributions under special circumstances. The well-known Central Limit Theorem is one case of this. See Jaynes 2003: ch. 7 for discussion.

examine the values on X and Y at various times, and record your observations. This time, you plot them on a graph, as in fig. 1.[5]

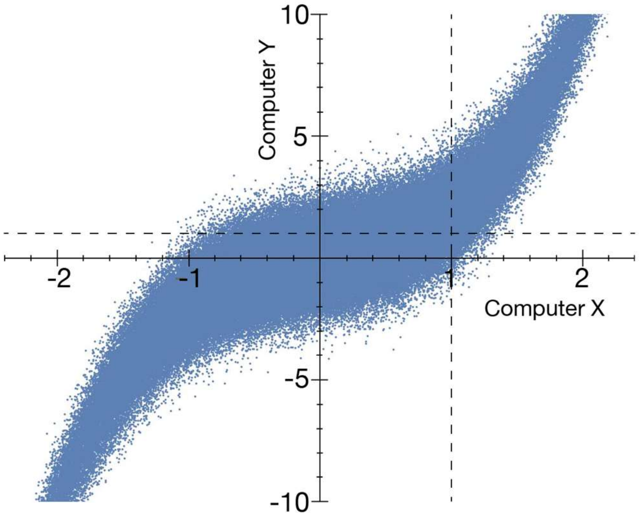

**Fig. 1.** Your observations from the two computers in Case 3.

After collecting a large number of data points, you are able to plot the frequency of different values for *y* given particular values for *x*, and vice-versa. Fig. 2a shows the observed distribution of 610 *y*-values when X = 1. This distribution shows the frequency of different *y*-values for the data points on the vertical dashed line in Fig. 1; the most frequent *y*-value is 2, with *y*-values further away from 2 occurring less frequently in a Gaussian manner. More generally, you find that as you collect more data, when X = *x*, the distribution of *y*-values looks roughly like a bell curve centered on $x + x^3$.

[5] The graphs in this and the next example are based on computer-generated data. For more detail on these data and how we constructed the graphs from them, see the Appendix.

However, in general, the shape approached by your distribution of *x*-values when Y = *y* is not a bell curve. Moreover, the shape is different for different values of Y. Fig. 2b shows the observed distribution of 446 *x*-values when Y = 1. This distribution shows the frequency of different *x*-values for the data points on the horizontal dashed line in Fig. 1.

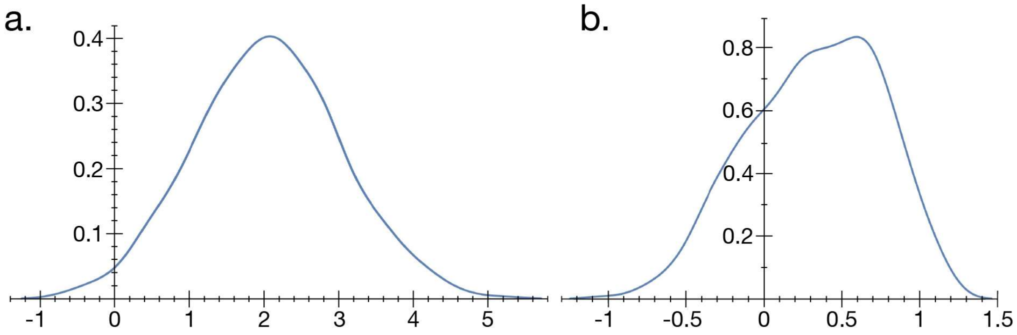

**Fig. 2a.** The observed distribution of *y*-values when X = 1 in Case 3.

**2b.** The observed distribution of *x*-values when Y = 1 in Case 3.

You can now reason as follows. If Y were causing X according to some function $x = f(y)$ + noise, then the frequency with which X takes on different values when Y has a particular value should be a bell curve centered on $f(y)$, since the glitch leads the output to deviate from its true value with Gaussian noise. However, this is not what you observe. Instead, you observe the converse: when X has a particular value, the frequency with which Y takes on different values is Gaussian. This is consistent with X causing Y according to the function $y = g(x)$ + noise. Moreover, in general the most frequent Y-value when X = $x$ is $x + x^3$, suggesting that $g(x) = x + x^3$.[6] The hypothesis that X→Y makes a prediction that you have found to be true, while the

[6] Hoyer et al. 2009 also use this example function.

hypothesis that X←Y makes a prediction that you have found to be false. So in this case, you can again conclude, solely from historical data, that X→Y.

Many causal functions are similar to this example in that adding Gaussian noise to them results in the probability distribution of the cause given the effect sometimes being non-Gaussian. Informally speaking, this non-Gaussianity is the result of the causal noise being permuted by Bayes' Theorem, where the precise character of this permutation depends on the prior distribution of the cause (e.g., a uniform distribution from 0 to 1), the causal function (e.g., squaring the cause and adding 1), and the kind of noise added to this function (e.g., Gaussian noise with mean 0 and variance 1).[7] Hence, in many cases, if you know that the causal noise is Gaussian, you will be able by observation to distinguish cause from effect in the above manner.

In Case 3, unlike in Cases 1 and 2, the noise was Gaussian and the possible causal functions were unknown. However, as in Case 1, you knew that either X→Y or X←Y. Let us finally consider a case with Gaussian noise and an unknown causal function, where there is also the possibility of a common cause.

[7] The precise shape of the distribution of cause given effect is determined by the continuous form of Bayes' Theorem, according to which $h(x|y) \propto h(x)h(y|x)$—that is, the relative frequency with which X takes on different $x$-values when Y = $y$ is proportional to the relative prior frequency with which X takes on those values multiplied by the relative frequency with which those values of X lead to Y = $y$. This latter factor is determined by the mathematical function from cause to effect and the noise added to this function. Most combinations of prior distribution of the cause, mathematical function from cause to effect, and Gaussian noise lead to the distribution of effect given cause sometimes being non-Gaussian. The main exception is when the prior distribution is Gaussian and the function is linear. In this case the distribution of effect given cause will always be Gaussian. (See Section 2.1.1 of Mooij et al. 2016 for the mathematical details.)

### CASE 4

As in the previous cases, you are able to make observations of the values displayed on two computers, X and Y, and your task is to determine the causal connection between these values. The glitch functions as in Case 3. However, unlike in Case 3, Max tells you that one of three causal hypotheses is true: either X determines Y according to some function plus Gaussian noise, Y determines X according to some function plus Gaussian noise, or both X and Y are determined by a third hidden computer Z, according to some functions plus Gaussian noise. In this third case, the glitch functions independently in the signals sent from Z to X and Z to Y; that is, the Gaussian noise added to the two functions is independent.

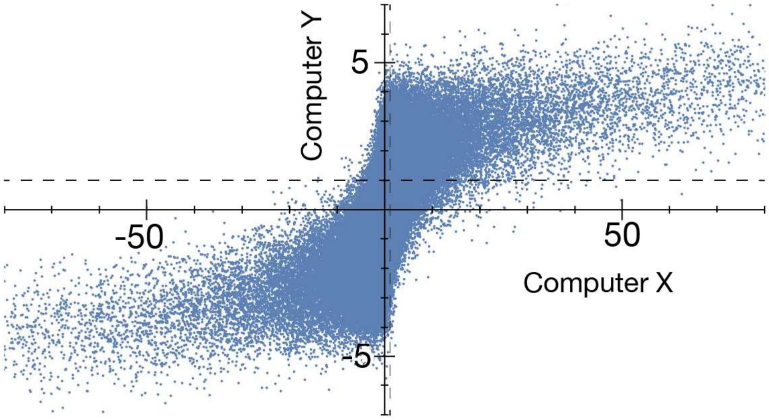

**Fig. 3.** Your observations from the two computers in Case 4.

As in Case 3, you plot your observations on a graph. As before, if one of X and Y is the causal root, you should expect the distribution of the values on the other computer given a value for that computer to approximate a bell curve. To see whether this is the case, you plot the frequency of different values for *y* given particular values for *x*, and vice-versa. Fig. 4a shows the observed distribution of 313 *y*-values when X = 1. This distribution shows the frequency of different *y*-values for the data points on the vertical dashed line in Fig. 3. Fig. 4b shows the observed distribution of 496 *x*-values when Y = 1. This distribution shows the frequency of different *x*-values for the data points on the horizontal dashed line in Fig. 3.

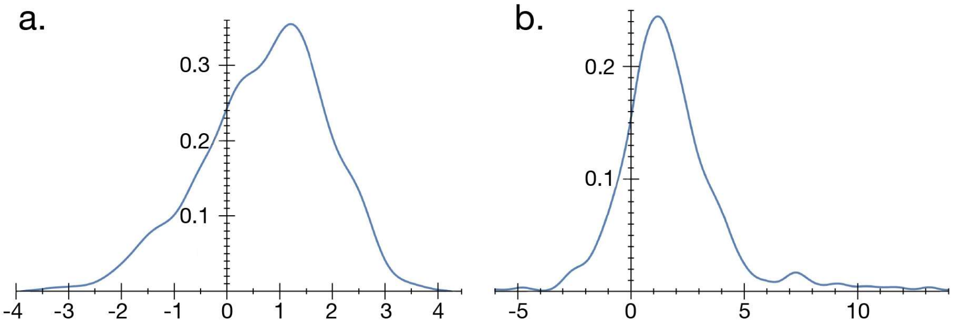

**Fig. 4a.** The observed distribution of *y*-values when X = 1 in Case 4.

**4b.** The observed distribution of *x*-values when Y = 1 in Case 4.

Neither of these distributions are very Gaussian. Unlike normal distributions, both have non-zero skewness (i.e., they are skewed to one side). Fig. 4b also has a much higher kurtosis than a normal distribution (i.e., it has very long tails). The non-Gaussianity of Fig. 4b is in conflict with the hypothesis that Y is causing X according to some function $x = g(y)$ + noise. According to this hypothesis, the frequency with which X takes on different values when Y has a particular value should be a bell curve centered on $g(y)$, since the glitch leads the output to deviate from its true value with Gaussian noise. And the non-Gaussianity of Fig. 4a is in conflict with the hypothesis that X is causing Y according to some function $y = f(x)$ + noise. According to this hypothesis, the frequency with which Y takes on different values when X has a particular value should be a bell curve centered on $f$(x), since the glitch leads the output to deviate from its true value with Gaussian noise.

On the third hypothesis, however, according to which X and Y are both caused by some hidden cause Z, according to some functions $x = h(z)$ + noise and $y = j(z)$ + noise, these observations are not surprising. Just as the distribution of a cause given an effect is often non-Gaussian even when the causal noise leading from the cause to the effect is Gaussian, so the distribution of one effect of a common cause given the other effect of a common cause is often non-Gaussian even when the causal noise leading from the common cause to each of the effects is Gaussian.

This case is dissimilar from Case 3 in that the precise causal function from the causal root to the effects is not evident merely from inspecting the conditional frequency distributions. Nevertheless, it is significant that you *are* able to tell that Z is the causal root, even without knowing what the causal functions are from Z to X and from Z to Y.[8]

The progression of cases above are sufficient to show the in-principle possibility of causal inferences from noise in a wide variety of circumstances. Table 3 summarizes the characteristics of the four cases.

| | ***Possible causal relationships*** | ***Possible causal functions*** | ***Linearity of causal function*** | ***Causal noise*** |
|---|---|---|---|---|
| ***Case 1*** | X→Y, X←Y | Known | Linear | Non-Gaussian |
| ***Case 2*** | X→Y, X←Y, X←Z→Y | Known | Linear | Non-Gaussian |
| ***Case 3*** | X→Y, X←Y | Unknown | Non-linear | Gaussian |
| ***Case 4*** | X→Y, X←Y, X←Z→Y | Unknown | Non-linear | Gaussian |

**Table 3**. Summary of the four Max cases.

[8] In fact, the data in Case 4 were generated according to the functions $x = 1/z$ + noise and $y = 1/z^{1/3}$ + noise.

Case 4 is particularly instructive. It exhibits causal inference in a case where there is the possibility of a common cause, the common cause is unobserved, and the causal function is unknown.

**3. NIMs in the Real World.** While the cases in Section 2 demonstrate the considerable power of NIMs for inferring causes from historical observation, all of the above cases involve a good deal of idealization. In this section, we will consider the possibility of applying NIMs to real-world problems. We will first examine a real-world dataset using the same kind of reasoning as in the cases above, which will let us see some ways in which the computer-generated data in our last two cases differ from real-world data. We will then briefly describe what more systematic NIMs look like, and the extent to which these more complicated statistical algorithms have been successful in analyzing other real-world datasets.

Figure 5 shows the relationship between water temperature (in Celsius) and ocean depth (in meters), from 1998 to 2017, between 18 and 20 degrees latitude and -153 to -123 degrees longitude (qualitatively: a strip of water from Hawaii to North America).[9]

[9] See the Appendix for more detail on the source of these data.

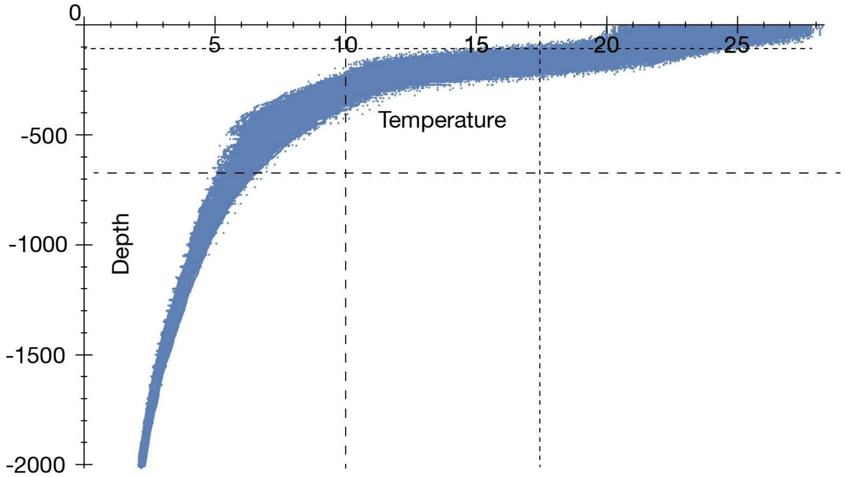

**Fig. 5.** The relationship between depth and temperature in ocean water between Hawaii and North America.

In this case, the depth of the ocean at a location is a cause of water temperature at that location, and not the other way around. But suppose we did not know this, and only knew that either Temperature→Depth or Temperature←Depth. Could we then determine which of these causal hypotheses is correct?

Let's try approaching these data in the same way as our computer-generated data in the last section. We thus begin by assuming that the frequency with which the effect takes on different values when the cause takes on a given value will be Gaussian, but that this is unlikely to hold in reverse. That is, the frequency with which the effect takes on different values—when the cause takes on a given value—will not usually be Gaussian. We can then look at the frequency with which the water takes on various temperatures at a given depth, and the frequency with which the ocean is at different depths when the water is at a given temperature, and see which of these two distributions is more Gaussian.

Fig. 6 shows the distributions of the other variable when we hold each variable fixed at its mean. Fig. 7 shows the distributions of each variable when we fix the other at its mean plus its standard deviation.

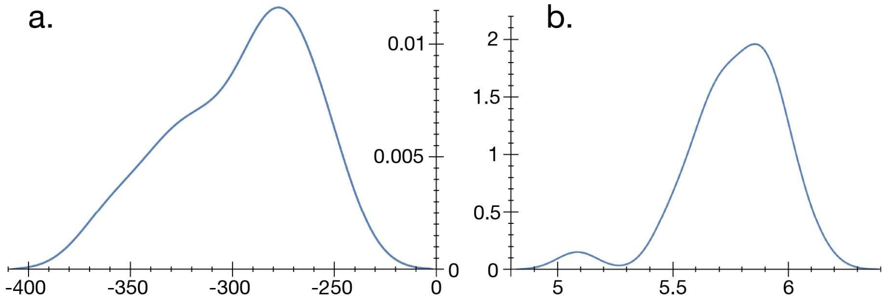

**Fig. 6a.** The recorded distribution of depth values when temperature = 9.96 degrees Celcius.

**6b.** The recorded distribution of temperature values when depth = -679.30 meters.

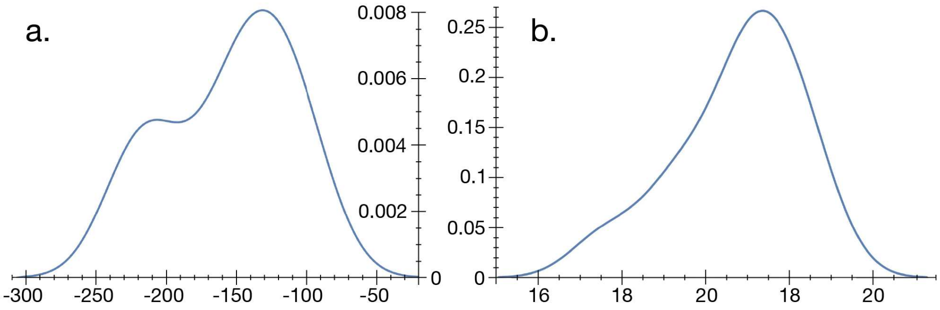

**Fig. 7a.** The recorded distribution of depth values when temperature = 17.33 degrees Celcius.

**7b.** The recorded distribution of temperature values when depth = -104.70 meters.

While the difference is not as clear-cut as in Case 3, visual inspection suggests that, in both cases, the distribution of Temperature given Depth is more Gaussian than the converse—(correctly) indicating that Temperature←Depth, rather than the reverse. This is a real-world example in which the same kind of reasoning applied in our thought experiments in Section 2 seems to deliver the right result.

We stress, though, that the above analysis is primarily illustrative: this kind of visual inspection of individual data points would be a poor (and inefficient) method for drawing causal inferences from most real-world datasets. Statisticians have come up with several procedures that codify the kind of reasoning employed above into something more algorithmic and holistic. One class of these procedures[10] begins with the assumption that the effect is a function of the cause plus Gaussian noise. A test is then run to see whether the data is a better fit to an "additive noise model" on which $y = f(x)$ + N or on which $x = g(y)$ + N, where N is Gaussian noise. The test involves two things. First, a regression is run to express $y$ as a function of $x$, and vice-versa. Second, the resulting deviations from this function—what would be causal noise if this function correctly describes the causal relationship between X and Y—are measured to see how much they deviate from the assumption of Gaussianity. (This is thus a class of NIMs, rather than a single NIM, because it is dependent on both a regression method and a measure of Gaussianity.) The function that comes closer to expressing the one variable as a function of the other plus Gaussian noise is then preferred.

Other NIMs exploit different kinds of asymmetries in patterns of noise, or rely on different assumptions about the noise. For example, Shimizu et al. (2006) describe NIMs that

[10] This kind of NIM was first described in Hoyer et al. 2009. Here we follow the description in Mooij et al. 2016.

exploit assymetries in cases with linear causation but non-Gaussian noise (like Cases 1 and 2 in Section 2).[11]

There are a number of reasons why NIMs might be less accurate when applied to real-world data than when applied to simulated data. The most important reason is that the assumptions of the NIM might be incorrect. For example, there are many reasons why an assumption that causal noise is Gaussian might fail. A smaller number of data points means that random variation is more likely to be lopsided. There may be unobserved causes which systematically bias our observations. Measurement or recording errors may result in anomalous outliers. In addition, most NIMs that have been developed assume that either X→Y or X←Y,[12] and in many real-world cases this assumption is violated.

We take these difficulties to be calls for further work, rather than reasons to doubt the possibility of ever applying NIMs to real-world problems. For example, algorithms that seek to identify outliers in datasets in a principled way might help us remove anomalous data points that

[11] For further discussion of and references for inferring causation in these different kinds of cases, see the review in Mooij et al. 2016.

[12] An exception is Hoyer et al. 2008, which extends Shimizu et al.'s (2006) method for inferring causation from linear causation and non-Gaussian noise to the case where there may be unobserved confounding variables. The strongest result Hoyer et al. argue for in that paper is that "it is possible to estimate, up to a finite set of observationally equivalent models, causal models involving linear relationships between non-Gaussian variables, some of which may be hidden" (Hoyer et al. 2008: 376). In other words, there is a finite set of causal models (specific ways in which X←Z→Y or [X→Y]&[X←Z→Y] could be true) which predict the observed pattern of noise. This is an important result, though it falls short of our being able to use the correlational data to discriminate between these models.

make the distributions of effect given cause less Gaussian.[13] As for common causation, Cases 2 and 4 in the last section show that it is in principle possible to use NIMs to discriminate not only between X→Y and X←Y, but also X←Z→Y. The development of NIMs that can be used in real-world cases involving common causes will require attending to the various ways in which these cases are more complicated than Cases 2 and 4. (For example, in real-world cases, X→Y and X←Z→Y are often not mutually exclusive.)

It is also worth noting that NIMs may still be effective even in cases where their assumptions are violated, and that to a large extent (due to our lack of knowledge of the causal details of real-life cases) the reliability of NIMs when applied to different kinds of datasets is an empirical question. Mooij et al. (2016) study the reliability of different NIMs by applying them to a variety of real-world and simulated datasets in which the direction of causation is already known (either because the dataset is computer-generated, or because we have background knowledge about the real-world phenomenon that already tells us the direction of causation—as in the example of ocean depth and temperature above). In general, they find that NIMs are more accurate when applied to the computer-generated datasets, even when these datasets include perturbations meant to model non-ideal features of real-world datasets. Applied to the real-world

[13] For an overview of outlier detection algorithms, see Hodge and Austin 2004 and Zimek and Schubert 2017. For the purposes of improving the accuracy of NIMs, we would want to use an algorithm that (a) distinguishes outliers and noise, identifying data points that are outside the normal variation resulting from noise in the data (Aggarwal and Yu 2001), and (b) is unsupervised, meaning that it requires no prior knowledge about the data (such as causal directionality) (Hodge and Austin 2004: 88-89). Tests on real-world data sets like those described below could then reveal whether combining NIMs with this algorithm improves their reliability.

datasets, the most effective class of NIMs deliver the right result between 63% and 69% of the time.[14] (By comparison, a "coin flip" algorithm would be expected to deliver the right result 50% of the time.) While these accuracy levels are not yet as high as we might like, this result is already encouraging to the extent that the assumptions of the NIMs (such as the lack of common causes) are often not met in these datasets. We are hopeful that future works on NIMs can lead to even better results, and to their fruitful application to datasets in which the direction of causation is not already known.

**4. Theoretical Implications of NIMs.** The possibility of using NIMs to draw causal inferences has several important theoretical implications. First, it shows that RCTs are not the only way to determine causal relationships: just as certain results are more or less likely in an RCT given X→Y, X←Y, or X←Z→Y, so certain patterns of noise in historical data are more or less likely given these different causal hypotheses. This lends considerable justification to the viability of observational studies as an alternative to RCTs, especially in cases where RCTs are practically or ethically problematic. Our point isn't that RCTs are always bad, or that they are never useful. Nor is it that NIMs necessarily do better than RCTs by way of invariably eliminating confounding causal factors. Rather, our point is to show that RCTs aren't the only game in town when it comes to justifying inference from correlation to causation. Not only can NIMs provide a viable alternative to RCTs when the ethical or practical constraints on RCTs prove too onerous to accept, but there are some cases in which NIMs succeed in determining causal directionality where an RCT would be physically impossible. Our application of NIMs to the correlation

[14] For the details, see pages 34-35 of Mooij et al. 2016.

between ocean temperature and depth is an example: it's not possible for us to randomly assign different locations in the ocean to different depths.

NIMs do not only show that RCTs aren't necessary for determining causal relationships, they show that in some cases no intervention at all is necessary. NIMs involve inference from purly observational data. (In the ocean case, there is no intervention at all. In the Max cases, there was *an* intervention, though it was not yours—namely, Max setting the value on the root computer. But that the value of the root variable is set by Max's intervention is inessential to the cases. Cases 1-4 could be redescribed so that the observed values of X and Y are measurements of some other variables, and the process which determines the distribution of values for the root variable is unknown.)

Not only is no intervention necessary for NIMs, nothing even *like* an intervention is necessary. Scheines (2005), for example, has argued that causal search algorithms that use dependencies and independencies in observed data to derive the causal graphs consistent with those data[15] (given certain assumptions that need not be elaborated here), while they do not require interventions on the network, may require observations of variables that function *like* interventions. In particular, he notes that the Fast Causal Inference algorithm proposed by Spirtes et al. (2000) only allows us to infer that X→Y if we have an observed variable Z that either directly causes X or shares a common cause with X, is not an effect of X, and is not a direct cause of Y. This is similar to the requirement for experimental studies that an intervention I on X should be a direct cause of X and not a direct cause of Y. The observed variable Z has, in effect, taken the place of the intervention variable I on X in allowing us to infer that X→Y.

[15] This is a burgeoning literature, with Pearl 2000 and Spirtes et al. 2000 among the classic texts.

NIMs do not require the presence of any variable like Z. In Case 3 in Section 2, you are able to infer that X→Y even though you only observe two variables: X and Y. You need not observe any variables that either cause X or share a common cause with X. So NIMs allow for causal inference in cases in which Spirtes et al.'s algorithm does not.

Other conditions proposed as necessary for causal inferences are also not needed for NIMs. For example, Clarke et al. (2014), who are critical of the predominance of RCTs in evidence-based medicine, argue that in order to establish X→Y, one needs to establish both that X and Y are correlated and that there is a *mechanism* that can explain the causal influence of X on Y. While some of our examples did rely on background knowledge, we do not necessarily need *mechanistic* knowledge in order to predict the unique noise signatures from different causal hypotheses. The inferences in the Max cases would still be possible even if you had no knowledge of the mechanism of causation. As long as you have some way of observing the values of X and Y, you do not even need to know what these values represent (they could be recordings of measurements of some external variables, for all you know). You could still infer that X←Z→Y in Case 4, for example, provided that you knew the values of X and Y, that either X→Y, X←Y, or X←Z→Y, and that the causal noise was Gaussian.

We have seen that not only are interventions not necessary to draw causal inferences from noise, neither are observed variables that function like interventions, nor are other kinds of background conditions like mechanistic knowledge. One moral of this is that it is a mistake to try to lay down general substantive necessary conditions on when it is possible to draw causal inferences. Causal hypotheses may often have consequences without our immediately being able to see that they have those consequences (or test for them given our current technology and resources). From the fact that right now we can only see one way in which the predictions of two theories differ, we should not conclude that in the future new ways to discriminate between them will not be discovered.

Another major theoretical implication of NIMs concerns how we understand *noise*. Noise is standardly taken to be something unwanted or unexplained. For example, Floridi's (2016) entry in the *Stanford Encyclopedia of Philosophy* defines noise as "data received but unwanted." Pierce (1980: 291) similarly holds noise to be "[a]ny undesired disturbance in a signaling system." And Scales and Snieder (1998: 1123) define noise as "that part of the data that we choose not to explain." The existence of NIMs shows definitions along these lines to be misleading. When employing a NIM, we are using noise to make causal inferences. As Case 1 made clear, in some situations, we can make causal inferences from the data *only if it is noisy*. In such cases, noise is not unwanted, nor do we choose not to explain it. Noise is not invariably what obscures causal signals—it can be a causal signal of its own, providing invaluable clues to the nature of the causal relations.

Finally, the possibility of drawing causal inferences from noise gives us reason to be wary of the goal of striving to reduce the amount of noise in our data. Sometimes the very data that appear to be keeping us from solving our problem can—when looked at in a new way—be used to establish our desired conclusions. Just as we should not close off the possibility of drawing inferences from hitherto unrecognized consequences of different theories, we should be cautious about throwing out evidence whose relevance is not immediately apparent. There may be a considerable amount of data collected by researchers that appears to be useless. The existence of NIMs should, however, give researchers pause when considering the value of their noisy data. Instead of casting such data aside, they should consider whether there may be statistical techniques able to extract causal information from the data. And as data storage is becoming increasingly inexpensive, the prospect of emerging NIMs shedding light on old datasets argues for an increased effort to archive raw data in the hope that they may eventually be reanalyzed with NIMs.

**6. Conclusions.** NIMs call for a revision of the epistemology of causation. Philosophers and scientists often argue that RCTs are needed to demonstrate causation (e.g., Fisher 1947; Papineau 1994), or they simply assume the necessity of RCTs for causal knowledge (e.g., the evidence hierarchies offered by the evidence-based medicine movement). If we are right, RCTs are not the only way of obtaining causal knowledge, and in many cases they may not be the best way of generating this knowledge. Variables that are correlated will be correlated in specific ways. The way that they are correlated may support one or more causal hypothesis. In particular, the noisiness of the causal link between X and Y is a source of causal information that has not been appreciated by the philosophical world.[16] If X and Y are causally linked, and if this link exhibits noise, the noise will tend to be asymmetric. This asymmetry can be a basis of causal inference, allowing us to discriminate between causal hypotheses such as X→Y, X←Y, and X←Z→Y.

NIMs call into question the standard way of understanding the nature and importance of noise. Noise is generally considered a nuisance, something we should strive to eliminate to expose the hidden causal relations. Because NIMs require noise to make causal inferences, while

[16] The word ‘noise’ does not appear in key treatments of the metaphysics and epistemology of causation such as the *Philosophy Compass* article “Introduction to the Epistemology of Causation” (Eberhard 2009), nor does it appear at the time of writing in relevant *Stanford Encyclopedia of Philosophy* entries, such as “The Metaphysics of Causation,” “Causation and Manipulability,” “Probabilistic Causation,” or “Causal Processes.”

some noise may be a nuisance, noise may more often be a blessing—a window into the causal structure of the world.

## Appendix

The graphs for Case 3 (Figures 1-2) were constructed from 250,000 randomly generated data points. We began with a normal distribution of X-values with mean 0 and variance 1. Then we generated Y values according to the function $y = x + x^3$, plus Gaussian noise with mean 0 and variance 1.

To calculate the frequency of different *y*-values when X = 1, we took the 610 data points at which X = 1 to two significant digits, i.e., X = 1.00 +/- .005. To calculate the frequency of different *x*-values when Y = 1, we took the 446 data points at which Y = 1 to two significant digits, i.e., Y = 1.00 +/- .005.

The graphs for Case 4 (Figures 3-4) were constructed from 250,000 randomly generated data points. We began with a normal distribution of Z-values with mean 0 and variance 1. Then we generated X according to the function $x = 1/z$, plus Gaussian noise with mean 0 and variance 1; and Y according to the function $y = 1/z^{1/3}$ + noise, plus (independent) Gaussian noise with mean 0 and variance 1.

To calculate the frequency of different *y*-values when X = 1, we took the 313 data points at which X = 1 to two significant digits, i.e., X = 1.00 +/- .005. To calculate the frequency of different *x*-values when Y = 1, we took the 496 data points at which Y = 1 to two significant digits, i.e., Y = 1.00 +/- .005.

Figures 5-7 were generated from 339,266 data points from Argo (www.argo.ucsd.edu/), a global array of ocean measurement floats. These data points came from the 3067 Argo floats operative from 1998 to 2017 between 18 and 20 degrees latitude and -153 to -123 degrees longitude. The floats measure depth (in meters) by recording pressure (in decibars). We removed data points that were blank, corrupt, or contained obvious recording errors—i.e., numbers several orders of magnitude greater than all the other numbers in the data. These removals did not affect the data used in generating Figures 6-7.

To two significant digits, the mean temperature in our remaining data (in degrees Celsius) was 9.96, and the standard deviation was 7.37. The mean depth (in meters) was -679.32 and the standard deviation was 574.47. Figures 6 and 7 were constructed using the measurements of the uncontrolled variable when the measurements for the controlled variable were closest to their mean (for Figure 6), and mean + standard deviation (for Figure 7). In particular, Figure 6a was constructed from the 30 data points at which recorded temperature was (exactly) 9.95899963378906, Figure 6b was constructed from the 23 data points at which recorded depth was (exactly) -679.299987792969, Figure 7a was constructed from the 10 data points at which recorded temperature was (exactly) 17.3279991149902, and Figure 7b was constructed from the 11 data points at which recorded depth was (exactly) -104.699996948242.

## References

Aggarwal, C. C. and Yu, P. S.: 2001, 'Outlier Detection for High Dimensional Data'. In: Proceedings of the ACM SIGMOD Conference 2001.

Clarke, Brendan, Donald Gillies, Phyllis Illari, Federica Russo, and Jon Williamson 2014. "Mechanisms and the Evidence Hierarchy," *Topoi* 33(2): 339-360.

Eberhardt, F. 2009. "Introduction to the Epistemology of Causation," *Philosophy Compass*, 4: 913–925. doi: 10.1111/j.1747-9991.2009.00243.x

Fisher, R. A. 1947. *The Design of Experiments.* 4th edition. Edinburgh: Oliver and Boyd.

Forster, Malcolm and Elliott Sober 1994. "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions," *The British Journal for the Philosophy of Science* 45(1): 1-35.

Floridi, Luciano, "Semantic Conceptions of Information," *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2016/entries/information-semantic/>.

Hodge, Victoria J. and Jim Austin 2004. "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review* 22: 85-126.

Hoyer, Patrik O, Shohei Shimizu, Antti J. Kerminen, and Markus Palviainen 2008. "Estimation of causal effects using linear non-Gaussian causal models with hidden variables," *International Journal of Approximate Reasoning* 49(2): 362-378.

Hoyer, Patrik O., Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf 2009. "Nonlinear causal discovery with additive noise models," *Advances in neural information processing systems*. 689-696.

Jaynes, E.T. 2003. *Probability Theory: the Logic of Science,* Cambridge: Cambridge University Press.

Mooij, Joris M., Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. "Distinguishing cause from effect using observational data: methods and benchmarks." *Journal of Machine Learning Research* 17(32): 1-102.

Papineau, David 1994. "The Virtues of Randomization." *British Journal for the Philosophy of Science* 45(2):437-450.

Pearl, Judea 2000. *Causality: Models, reasoning, and inference* (Cambridge University Press).

Pierce, J. R., 1980. *An Introduction to Information Theory: Symbols, Signals & Noise*, 2nd edition, New York: Dover Publications.

Savage, Leonard 1962. *The Foundations of Statistical Inference,* London: Methuen and Co.

Scales, John and Roel Snieder 1998. “What is noise?” *Geophysics* 63: 1122-24.

Scheines, Richard. 2005. “The Similarity of Causal Inference in Experimental and Non-experimental Studies.” *Philosophy of Science* 72: 927-940.

Shimizu, Shohei, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen 2006. “A linear non-Gaussian acyclic model for causal discovery.” *The Journal of Machine Learning Research* 7: 2003-2030.

Sober, Elliott 2001. “Venetian sea levels, British bread prices, and the principle of the common cause,” *British Journal for the Philosophy of Science* 52: 331-346.

Spirtes, P., C. Glymour, and R. Scheines 2000, *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Urbach, Peter 1985. “Randomization and the Design of Experiments.” *Philosophy of Science* 52: 256-273

Worrall, John 2007. “Why There’s No Cause to Randomize.” *British Journal for the Philosophy of Science* 58: 451-488.

Zimek, Arthur and Erich Schubert 2017. “Outlier Detection,” in *Encyclopedia of Database Systems*, ed. L. Liu and M.T. Ozsu, Springer.