Data, Phenomena, Signal, and Noise

James Woodward[†‡]

This essay attempts to provide additional motivation for the data/phenomena framework advocated in Bogen and Woodward, "Saving the Phenomena" (1988).

Twenty years ago, Jim Bogen and I published a paper, "Saving the Phenomena" (Bogen and Woodward 1988), in which we suggested that theory structure in science often could be usefully described in terms of a three-part distinction between data, phenomena, and explanatory theory.[1] The ideas in that paper have been embraced by some (e.g., Kaiser 1991) and, as is the practice in philosophy, criticized by others (McAllister 1997; Glymour 2000; Schindler 2007). In what follows I try to provide additional motivation for the framework advocated in our original paper. I also respond to some criticisms.

I begin with some schematic remarks on the epistemic problematic faced by a researcher who wishes to use data as evidence, since much of the motivation for the data/phenomenon distinction derives from our attempt to capture features of this problematic.

1. Data and Phenomena. Data are the individual outcomes of a measurement or detection process, which may involve instruments or unaided human perception. By extension, records or reports of such outcomes may also be regarded as data. Examples include the results of using a thermometer to make repeated measurements of the melting point of a sample of lead, the photographs recording apparent stellar positions that Ed-

†To contact the author, please write to: History and Philosopohy of Science, University of Pittsburgh, Pittsburgh, PA 15260; e-mail: jfw@pitt.edu.

‡Thanks to fellow symposiasts Jim Bogen, Katherine Brading, James McAllister, and Paul Teller for helpful comments and discussion.

1. This initial paper was followed by others (Woodward 1989, 1998; Bogen and Woodward 1992, 2005).

Philosophy of Science, 77 (December 2010) pp. 792–803. 0031-8248/2010/7705-0009$10.00

dington took using several telescopes during his solar eclipse expedition of 1919, and the outcomes of Robert Millikan's measurements of *e*, the charge of the electron, reported in his paper of 1913. In Millikan's case, his data represented (among other things) the time taken for small drops of oil to fall between scratch marks in the focal plane of his telescope, both when an applied electrical field was present and when it was absent.

In serious scientific research, data production is not usually an end in itself.[2] Usually data is produced in order to serve as evidence for something else—for features of phenomena. Examples of phenomena for which the above data might serve as evidence are, respectively, the melting point of lead, the deflection of starlight by the gravitational influence of the sun, and the charge of the electron.

2. Noise. Investigators try to arrange data production processes so that the data are causally influenced by the phenomena they are trying to learn about (and because of this convey information about those phenomena), but commonly data will also reflect the causal influence of many other more local and idiosyncratic factors having to do with properties of instruments, measurement procedures, and the environment. The influence of these additional factors on the data is often conceptualized as "error" or "noise." In the case of data consisting of thermometer readings, these "other" causal factors will have to do with, for example, variations in the temperature of the surrounding air and in the perceptual processing and judgment of the observer who records the measurement results. In Millikan's case, his data regarding time of fall reflected, in addition to the charge on the drops, the influence of small variations in air currents in his apparatus; lack of uniformity in the electric field; as well as the fact that the drops themselves, besides falling vertically, might also drift horizontally in a way that was not independently detectable by Millikan.

In these and other cases, many of these "additional" factors influencing data will vary on different occasions of measurement. In consequence, the data themselves will often exhibit variability or scatter, even when a measurement of the same phenomenon is repeated with the same instrument or procedure. In addition, the values taken by the additional factors influencing the data on particular occasions may be difficult to measure, and the specific details of the ways in which they affect particular measurement outcomes may be unknown.

Millikan's laboratory notebooks provide an illustration of these points. As is well known, Millikan excluded measurements taken on various drops in his final calculations for *e* because he thought that various sources of

2. If it were, it would be unclear, among other things, why researchers ever discard data.

error were present that made the results unusable. On some occasions, he has some idea, although conjectural, about such sources—thus, he attributes one measurement result to the fact that he has mistakenly measured a dust particle rather than an oil drop and another to nonuniformities in the field. But on other occasions, there are remarks like "[calculated charge] very low something wrong" or "something the matter," indicating that he lacks detailed knowledge of the factors producing his data but thinks it justifiable to exclude it. Whatever one thinks of this decision, Millikan's remarks illustrate the epistemic situation of a researcher engaged in data to phenomena inference.

3. An Epistemic Problem. These features create a generic epistemic problem: investigators must somehow use data that is noisy and, where the details of the process producing this noise are not fully understood, learn about the phenomena of interest. The characterizations that Bogen and I provided in early papers of data, phenomena, and explanatory theory and their interrelations, and, in particular, our idea that we are often not in a position to provide systematic explanations of data and that the direction of inference is typically from data to phenomena and not from phenomena to data, were meant to reflect or capture this epistemic problematic. I stress this because critics often assume, without argument, that researchers engaged in data to phenomena reasoning are in an epistemically very different situation in which, for example, detailed understanding of the processes involved in the production of individual data points is available "in principle."

In Bogen and Woodward (1988) and subsequent papers, we allowed all sorts of things, including events, processes, relationships, and instances of property possession, to qualify as instances of phenomena. Despite this ontological diversity, we argued that a distinguishing feature of most phenomena of scientific interest is that, in contrast to data, they have stable, repeatable characteristics. In principle, even if not in fact, phenomena can reoccur with the same features in different circumstances and contexts. For example, the value of e that Millikan attempted to measure is a property of all electrons, even if measurements of this quantity on different occasions are likely to produce somewhat different data.

The features of data and phenomena just described suggest the following rough schematic: Instances of the phenomenon may be regarded as repeated instances of some fixed feature or quantity P. If the measurement procedure is working properly, individual data points d_i observed on different occasions of measurement of P will be some function f of P and of additional causal factors ("error" or "noise") u_i, which will take different values on different occasions of measurement: $d_i = f(P, u_i)$. Although the researcher will be able to observe the d_i, she often will not be

able to observe or measure the values of all of the additional causal factors u_i. She is thus faced with the problem of obtaining information about P from the d_i, despite the fact that the d_i confounds the influence of P and u_i. As a general rule, accomplishing this requires additional empirical assumptions A_i besides whatever information is in the d_i; it is only by combining such A_i with the d_i that reliable inferences about P are possible. However, as illustrated below, it is an important fact about these A_i that in order to secure the reliability of data to phenomena reasoning, they do not need to be assumptions that can be used to provide systematic explanations (or predictions) of individual data points.

Several other features of the epistemic situation just described deserve particular emphasis. First, note that the problem the researcher faces is not one of describing the d_i or of providing derivations or predictions of the d_i. Instead, her goal is to use the d_i to make an inference about something else: P.

Second, the idea that P is a stable, repeatable type detectable from different bodies of data constrains this inference in important ways. In particular, it points to a problem that the researcher must somehow deal with: the possibility of overfitting or, more generally, the possibility that data produced by any particular set of measurements may (because of the influence of particular u_i present on that occasion) exhibit special features that may be misleading about P. Note that such misleading features may be representative of or genuinely present in the data produced on some occasion; the problem is that they are not representative of the phenomenon. One mark of such overfitting is that features of the phenomenon inferred from one data set are not reproduced in the inferences made from other data sets designed to measure the same phenomenon. The original data set fails to predict what we find in new data sets, where "fails to predict" means that features of the phenomenon inferred from the old data are not supported by the new data. (That the two data sets themselves will not be identical goes without saying.)

As a simple statistical illustration, consider a sample of N data points drawn from a Gaussan distribution of a variable X with mean μ and variance σ^2. It is easy to show that the sample variance s_x^2 is a biased estimate of the variance of this distribution. Intuitively, this is because there is likely to be less variability in any particular sample than in the distribution from which the sample is drawn. If we use s_x^2 as an estimator, we overfit the population variance, underestimating it by a factor of $N - 1/N$. Other examples involve fitting a high-order polynomial to a data set that is, in fact, generated by a linear relationship plus noise—because it has many free parameters, the high-order curve may fit the original data well but fit other data generated by the same linear relationship very poorly.

4. Phenomena and Patterns in Data Sets. The way of thinking about the relationship between data and phenomena just described may be contrasted with the picture advocated by James McAllister (e.g., McAllister 1997). According to McAllister, investigators are interested in "patterns" found in data sets; they proceed by "decomposing" data into noise or error (according to McAllister, determined by "stipulation" of an "error level") and a remaining "pattern." A phenomenon is anything that corresponds to such a pattern. An indefinitely large number of such stipulations and decompositions are legitimate. Thus, there are indefinitely many patterns in any data set and, corresponding to these, indefinitely many phenomena.

Assuming this is intended as a description of data to phenomena reasoning, in my opinion it goes wrong at the outset. The discernment of "patterns" in particular data sets (or, more broadly, the description of particular data sets) is not in itself a matter of much scientific interest. If it were, the notion of overfitting a particular data set (and that there is something wrong with doing this) would make no sense—the only issue would be whether the data set is accurately described, whether by signal/noise decomposition or in some other way. The idea of overfitting, or that certain uses of data can be misleading (e.g., that a certain data-based estimator can be biased), only makes sense if one views the data as a potential source of information about something that is not just a pattern in that particular data set but, rather, has some independent existence. For example, if one's object is just to describe (find a pattern in) a particular data set, there is no reason not to use the sample variance s_x^2 of this set—indeed, this will furnish a more accurate description of this data than any "corrected," unbiased estimate, although a less accurate description of the underlying distribution. The general point is that the problem of describing features of a particular data set with no implications for what one might expect in other data sets is a different problem from using that particular data set to infer to some stable feature of the world, where it is assumed that other data sets also carry information about this stable feature. We go wrong at the start if we fail to distinguish these problems. McAllister seems to assume that reasoning with data has to do only with the first problem, whereas I suggest that it is typically the latter problem which is of interest.

5. The Role of Theory. The remaining element in the original Bogen-Woodward framework was what we called explanatory or "general," "high-level" theory. We related this to the data-phenomena distinction in the following way: claims about phenomena but not data were, we contended, potential objects of systematic explanation and prediction by general theory—claims about phenomena rather than data were what such

theories attempted to "capture" or "account for." For example, general relativity (GR) provides, in conjunction with other assumptions, a systematic explanation of the deflection of starlight by the sun but not of the data from Eddington's eclipse expedition, which served as evidence for the starlight deflection.[3]

None of this was meant to deny that empirical assumptions of many sorts are "involved in " or "relied on" in reasoning from data to phenomena—our original papers provided many examples of this. If all empirical assumptions are regarded as part of a "theory" (just in virtue of being empirical), then data to phenomena reasoning is always "theory laden." Our claims about theory explaining phenomena rather than data were not intended to deny this but, rather, were attempts to express several interrelated ideas. First, we thought the empirical assumptions figuring in data to phenomena reasoning were often rather different in character from core explanatory assumptions figuring in theories like general relativity. Relatedly, we thought that the assumptions employed in data to phenomena reasoning often played a different role in reasoning than the role, say, the field equations of GR play in connection with the deflection of starlight by the sun. It was these differences—the different ways that different sorts of "theory" are "involved" in, on the one hand, the data-phenomena relation and, on the other, the general theory-phenomena relation that we were attempting to capture with our talk of high-level theory explaining phenomena, not data.

As a simple illustration, consider the role of assumptions about the distribution of the error in data to phenomena reasoning—for example, the assumptions that the error is normally distributed or that it is uncorrelated with other variables representing the phenomenon of interest. Such assumptions are certainly empirical—indeed, they often turn out to be false in situations in which they are unthinkingly assumed. Nonetheless, their role is not plausibly regarded as one of providing systematic explanations of particular items of data. To explain individual data one would presumably need to know which particular causal factors represented by the error term occurred on particular occasions and how these contributed to individual measurement outcomes. The assumption that the error term is distributed in a certain way does not provide such information. More

3. The relevant notion of systematic explanation is characterized in our original papers—it involves detailed derivations from principles of considerable generality. As noted in those papers, since (if things are working properly) data will reflect the causal influence of the phenomenon detected, there is also a sense in which this phenomenon figures in an explanation of the data. However, this sort of explanation is different in character from the systematic explanations provided in general theory—it has the structure of a singular causal explanation.

importantly, it does not need to provide such information for it to play its usual role in data to phenomena reasoning. For that purpose, very general assumptions about the distribution of the error will often suffice.

6. A More Detailed Example. I turn now to a more detailed look at another statistical example. This will both illustrate some of the features of data to phenomena reasoning described above and will also allow for a more detailed assessment of some of criticisms of that distinction. I emphasize that my use of this example is not meant to imply that all data to phenomena reasoning is just a matter of statistical inference. However, statistical inference is one important component of such reasoning, and, furthermore, it has the virtue that because a formal framework is employed, certain aspects of that reasoning process stand out particularly clearly.

People unfamiliar with basic ideas of statistical inference may find it natural to suppose talk such as McAllister's about decomposing data into patterns and noise simply describes in an abstract but uncontroversial way what is involved in standard statistical inference. This is not the case. Instead, McAllister's framework is radically different than the standard one for posing problems of statistical inference. In particular, it does not allow one to even raise issues that are at the heart of statistical inference, standardly conceived: for example, issues about the properties of estimators and how these properties are related to the structure of the error distribution. I will also suggest (see n. 5 below) that McAllister's framework conflates choices that are genuinely conventional or stipulative (such as choice of an "error level" in the sense of a significance level) with considerations that are not a matter of stipulation at all, such as issues about how errors in the sense of causal factors in addition to the phenomenon of interest are distributed.

As a point of departure, I distinguish *descriptive* from *inferential* uses of statistics. One use of statistics is simply to describe a body of data in a summary, compact way. Given a set of numbers, I may calculate their mean, their standard deviation, fit a linear relationship to them (guided by some conception of what constitutes a best fit, such as least squares), and so on. This is just a matter of carrying out certain mathematical operations on the data—the mean, standard deviation, and so forth, are, mathematically speaking, functions of the data alone, and their calculation does not require any additional assumptions that "go beyond" the data. In my view insofar as our interests are just in describing some body of data, there is no single right way to do this—it might be reasonable to adopt any one of a number of different descriptions, depending on the data itself and the purposes for which we may want to use the description. Thus, we might decide to summarize the data by giving its mode or median rather than its mean, or we might describe various qualitative facts about

the distribution—that it is normal, or bimodal, or skewed. Obviously, such different descriptions, if true, should not be regarded as competing with each other.

It is also characteristic of such data description that it does not involve any substantive or nontrivial notion of "error" (i.e., other than misdescription of the data itself). Of course, one may, like McAllister, "decompose" some particular d_i into two different quantities and label one of them "error" ($d_i = x_i + u_i$). One may also do this with a collection of data points and ask how the resulting "errors" are distributed. However, without further constraints, there is no right or wrong way to carry out such a decomposition and no basis for interpreting the resulting "error" as having to do with additional causal factors besides the phenomenon of interest that influences the data. The decomposition is just a matter of carrying out certain mathematical operations on the data.

I mention this because some of what McAllister says about decomposing data into signal and noise seems correct if one is engaged in purely descriptive statistics—here, talk of a plurality of different decompositions of data into signal and noise, none more legitimate than the others, seems entirely appropriate.

By contrast, data to phenomena inference involves the use of statistics for inference, and this is fundamentally different from descriptive statistics. Here we are interested not (or not just) in describing a given body of data but in using the data to infer to something else. Issues about the conditions under which (or the additional assumptions that must hold in order that) the data may reliably used for this purpose loom large. Now the notion of error has a substantive significance (in terms of, e.g., additional causal factors influencing the data), and it is a mistake to think that error is just the result of performing some arbitrary mathematical operation on the data.

As an illustration, suppose one is interested in estimating a linear regression equation with just one independent variable X and dependent variable Y. The data consist of particular values of X and Y, x_i and y_i for each of n measurement pairs. One possible use for such an equation is simply to describe this body of data in terms of a "best-fitting" line of form $y_i = ax_i + u_i$, with no implication about whether a similar relationship holds in any other body of data and no assumptions or implications about how these data points have been produced or generated. There are many possible criteria for "best fitting," but a widely accepted criterion involves least squares: one chooses the line that minimizes the sum of the squared vertical distances of the data points x_i, y_i from that line. In other words, one chooses that value a^* for the coefficient a that minimizes the sum of the squared differences:

$$Q = \sum (y_i - a^* x_i)^2. \tag{1}$$

This can be accomplished by choosing $a^* = Sxy/S_x^2$. Here Sxy is the sample covariance of X and Y, and S_x^2 is the sample variance of X. Note that on this interpretation, the term u_i is simply whatever results from subtracting a^*x_i when a^* is determined in the way described above—u_i has no more substantive significance (i.e., no interpretation in terms of "other" causal variables). Moreover, the mathematical operation of choosing a^* in accord with (1) can be carried out on the data independently of any further assumptions about the u_i or anything else. Since we are not using a^* as an estimate of anything outside this particular data set, there is no issue about whether or under what conditions this estimate is reliable or predictive. Insofar as our purposes are to describe this data set, we might equally appropriately have carried out some different set of mathematical operations on the data—we might have used another criterion of best fit (e.g., minimizing the sum of the absolute values $|\, y_i - ax_i \,|$) and/or fitted a higher-order curve to the data.

The descriptive project just described should be distinguished from a second project that does involve inferential statistics. Here it is assumed that the data have been generated by some in-principle repeatable process or data generating mechanism (DGM) that generates the data in accord with the linear relationship

$$Y = aX + U. \tag{2}$$

Our goal is to use these data to estimate the value of a in (2). Here we are *not* thinking of (2) just as a description of a pattern that happens to hold in the particular data set at hand but that we have no reason to think will hold outside of this data set. Instead, (2) is regarded as characterizing a stable, repeatable (even if only locally stable) relationship (a "mechanism") that generates values of Y for given values of X but where Y is also influenced by other causal factors represented by U. (The values of U vary but are assumed to follow a stable probability distribution—another example of an empirical assumption made in some data to phenomena inference.) Thus, (2) is the sort of relationship that might be used to predict what we might find in other samples of data generated by the same DGM. Because this is our goal, we face a new problem—we have to worry about the possibility that the way we use the sample data may be misleading about the characteristics of this underlying process. Given some procedure for using the sample data to infer to features of this underlying process, it now makes sense to ask about the conditions under which this procedure will be reliable or not (according to some specified conception of reliability such as unbiasedness, etc.; see below) and also to ask whether there is some way of bounding the magnitude of the error that the procedure introduces. Neither question would make sense if our interest were just in describing the data.

Under what conditions might we use information in the sample data to estimate a value a^* for a in (2) when this coefficient is conceived as characterizing some underlying process generating the sample data? A standard result in statistics is that if the error U satisfies certain distributional assumptions (A), then the estimator given by (1) will have desirable characteristics—in particular, it will be a minimum variance, unbiased estimator.[4]

The distributional assumptions A are as follows:

(i) $E(U_i) = 0$,
(ii) $\text{Var}(U_i) = \sigma^2$, where σ is constant,
(iii) Absence of autocorrelation in U,
(iv) X and U are uncorrelated.

Note that expression (1) is the same as in the purely descriptive task, but now the interpretation and goal have shifted. In particular, a^* is now conceptualized as a random variable that is an estimator for something else, the fixed parameter a, which characterizes the DGM. The goal is now to use the data in the sample to infer to a feature of the process that generates the data, rather than to describe a pattern in the data. (Again, the patterns "in" the data will vary from sample to sample, but a is a parameter that is taken to be stable across different samples generated by the same DGM.) It is this conceptual separation between, on the one hand, the data and statistics that are defined just from the data, such as the sample covariance s_{xy}, the variance s_x^2, or the estimator a^*, and, on the other hand, the parameter a that allows one to even raise the question of whether and in what circumstances a^* is appropriate to use in order to estimate a.

Now consider the assumptions A concerning the error/noise, the satisfaction of which determines whether a^* is a good estimator to use. Although McAllister characterizes data to phenomena reasoning in terms of choosing a "noise level," the assumptions A are not naturally regarded as about the level of the noise. Rather, they have to do with the distribution of U and about its relationship to X. Moreover, in contrast to McAllister's treatment, it is an empirical question (once error/noise is interpreted in terms of additional causal factors) and not a matter that can be settled by "stipulation" how the noise is distributed. I do not see how to make sense of the idea that if, for example, U is correlated with X, then a^*

4. A random variable x^* is an unbiased estimator for parameter X, if $E(x^*) = X$.

should not be used as an estimator of a if it is just a matter of stipulation (and there is no fact of the matter) about whether U is so correlated.[5]

7. Data, Phenomena, Sample, and Population. I conclude with some brief comments on another related issue. In his article (2000), Bruce Glymour suggests that all cases of reasoning from data to phenomena are just cases of reasoning, via some process of statistical inference, from a sample to a population parameter. It follows, according to Glymour, that the data-phenomenon framework is superfluous—it reproduces, in other language, a distinction already adequately described by the sample/population distinction.

Unsurprisingly, I disagree with this claim. Although statistical inference is an important component in much data to phenomena reasoning, it is not involved in all such reasoning. Even when it is, there is typically more to data to phenomena reasoning than statistical inference. The problem Millikan faced in finding a second-order correction to Stokes's law that he could use in the analysis of his data was not a statistical problem.

Even when statistical inference plays a prominent role, it is still not obvious that the phenomena one infers should be regarded as a parameter in some specific population. Suppose that repeated measurements of the melting point of lead M are made with some measurement procedure R, and, on the basis of these, a claim is made about M. What is the population P of which M is a parameter? One suggestion might be that P is a hypothetical population: something like the collection of results one would get by repeating R indefinitely in the same circumstances. Call K the population mean of this population and think of the sample mean from the measurements as an estimate of K. Of course, one hopes K coincides with (or at least is "close" to) the true melting point M. But it seems wrong, as a conceptual matter, to identify M with the mean of any particular population. For one thing, if there is any systematic error in R, K will not even coincide with M. More generally, the true melting point M is supposed to be a feature of all samples of pure lead (for a fixed pressure, etc.), and it can be measured in a variety of different ways, with different thermometers and so forth, all of which will yield different data. It seems farfetched, to say the least, to think of all possible data that might result from different possible measurements of the melting point of lead as samples drawn from a single population.

5. An additional potential source of confusion associated with the notion of a "noise or error level" is that it fails to distinguish "error level" in the sense of choice of significance level (which is a matter of stipulation) from the "error" represented by U.

REFERENCES

Bogen, J., and J. Woodward. 1988. "Saving the Phenomena." *Philosophical Review* 97:303–52.
———. 1992. "Observations, Theories, and the Evolution of the Human Spirit." *Philosophy of Science* 59:590–611.
———. 2005. "Evading the IRS." In *Correcting the Model: Idealization and Abstraction in Science*, ed. Martin Jones and Nancy Carwright, 233–67. Poznan Studies. Netherlands: Rodopi.
Glymour, B. 2000. "Data and Phenomena: A Distinction Reconsidered." *Erkenntnis* 52:29–37.
Kaiser, M. 1991. "From Rocks to Graphs—the Shaping of Phenomena." *Synthese* 89:111–33.
McAllister, J. 1997. "Phenomena and Patterns in Data Sets." *Erkenntnis* 47:217–28.
Schindler, S. 2007. "Rehabilitating Theory: Refusal of the 'Bottom-Up' Construction of Scientific Phenomena." *Studies in History and Philosophy of Science* 38:160–84.
Woodward, J. 1989. "Data and Phenomena." *Synthese* 79:393–472.
———. 1998. "Data, Phenomena, and Reliabilit." *Philosophy of Science* 67 (Proceedings): S163–S179.