Data Models and the Acquisition and Manipulation of Data

Todd Harris[†]

This paper offers an account of data manipulation in scientific experiments. It will be shown that in many cases raw, unprocessed data is not produced, but rather a form of processed data that will be referred to as a data model. The language of data models will be used to provide a framework within which to understand a recent debate about the status of data and data manipulation. It will be seen that a description in terms of data models allows one to understand cases in which data acquisition and data manipulation cannot be separated into two independent activities.

1. Introduction. This paper will offer an account of data manipulation in scientific experiments. It will be shown that many cases of data acquisition and processing in science are poorly described in terms of the treatment of raw, uninterpreted data. Instead, such cases should be described in terms of interpreted data models. The language of data models will be used to provide a framework within which to understand a recent debate about the status of data and data manipulation, and to better place the role of theoretical goals and assumptions in the acquisition and manipulation of data. In particular, it will be seen that a description in terms of data models allows one to understand cases in which data acquisition and data manipulation cannot be separated into two independent activities.

2. Suppes and the Idea of a Data Model. The concept of a data model can be traced back to a paper by Patrick Suppes titled "Models of Data" (1962). In this paper Suppes describes a hierarchy of models that connect data to theory. There are three features of the Suppean hierarchy that are important to this paper. First, there are three levels of models in Suppes's system, models of theory, models of experiment, and models of data. Models of

[†]To contact the author, please write: Department of Philosophy, University of California Davis, Davis, CA 95616; e-mail: trharris@ucdavis.edu.

Philosophy of Science, 70 (December 2003) pp. 1508–1517. 0031-8248/2003/7005-0054$10.00

https://doi.org/10.1086/377426 Published online by Cambridge University Press

theory and models of experiment are associated with theories,[1] while models of data describe data gathered in a particular experiment. For example, a model of theory can be based on a scientific theory such as the ideal gas law. Models of experiment are based on theories of experiment.[2] Models of data describe data taken in a particular experiment. I will not devote time discussing the relationship between Suppean models and their associated theories as Suppes provides little information about this relationship in this article. However, it is important to note that both models of theory and models of experiment are associated with theories.

Second, unlike the theory to be tested and its associated model, the theory of experiment and its associated model contain references to experiment. For instance, a model of theory might include variables specified in the ideal gas law such as the temperature, pressure, and volume of an ideal gas. The theory of experiment might then specify the relation between the temperature and pressure of a Nobel (nonideal) gas when the volume is held constant in a possible experiment. The model of experiment would then contain possible values of the temperature and pressure in this experiment. The model of data would then contain readings of temperature and pressure taken during a particular run of this experiment.

Third, in Suppes's original formulation, not all data sets collected in a given experiment count as a model of data. In order for a data set to count as a model of data, certain conditions specified in the theory of experiment must hold within an acceptable margin of error. There are also statistical tests that one can perform on data sets that determine whether or not a given data set constitutes a model of data.

3. Data Models. Models represent inexactly. For example, a ball-and-stick chemistry set can be used to construct models that represent certain chemical compounds. Notice that the chemical model is said to represent a particular compound because it is similar to that compound in a relevant respect. Inaccuracies are allowed. A chemical compound has many properties lacking in the ball-and-stick model, and the model has many properties, such as color, that the real-world object lacks.

One can find parallels in every science. One example is Mendel's principle of independent assortment. Put in modern terms, this principle states that the genes that encode for certain traits such seed color or seed type (e.g., smooth seed or wrinkled seed) segregate independently of each other. For example, if a smooth yellow parent and a wrinkled green parent are crossed, there are four possibilities for the next generation: smooth

[1] In this paper he does not give a definition of a theory, but from his definition of model we can assume a theory is a collection of sentences.

[2] Suppes does not define or give a clear example of a theory of experiment in this paper.

https://doi.org/10.1086/377426 Published online by Cambridge University Press

yellow, smooth green, wrinkled yellow and wrinkled green, and each of these possibilities have an equal probability of occurring (25 percent). However, when one examines biological systems one finds that this principle does not hold. One exception is that some genes exhibit linkage, which means there is a higher probability those genes will be passed together to the progeny than predicted by Mendel's principle of independent assortment. Returning to the seed example, if the yellow and wrinkled genes are linked, then there is a much higher probability that the next generation will contain yellow wrinkled progeny than the other possibilities. Rather than claim that Mendel's principle has been refuted by this observation, one could claim that it is a model of inheritance of traits that has certain similarities and certain dissimilarities with biological systems depending on the context.

Turning to data, when one has what one would count as raw data, it is usually not in a form usable by scientists. For this reason, data must undergo some degree of processing before it can be analyzed. A first step in data processing is often aimed at eliminating noise in the data set. For instance, if one was plotting the positions of planets as they move through the nighttime sky, one would find that there is a certain amount of error associated with each of the measurements, due to flaws in the telescope, human error, atmospheric conditions, or other factors in the environment. Before one plotted the trajectory of a given planet, one would want to do two things: (1) eliminate certain points that were the result of error (for example, one saw a planet where there wasn't one), and (2) draw a smooth curve through the remaining points, so that the resulting path of the planet did not jump from point to point. In sum, you would process the data, and produce a smooth, continuous path for a planet from a collection of discrete data points. When information counts as raw data, it must almost always go through some such processing step before it is in a usable form.

This final plotted path of the planet can be considered a data model. Many changes have been made in the original data set. Some data points have been thrown out, and the discrete values have been replaced by a smooth, continuous curve. The first step is called data reduction, and the second can be called curve fitting. These two steps are often used when one constructs data plots in chemistry or physics lab classes, such as a temperature versus resistivity plot in a solid-state physics experiment, or a temperature versus pressure plot in a test of the ideal gas law. The resulting curves are similar to the original data set, but also different, being continuous, and usually not coinciding exactly with most of the retained original data points. The curve is now a model of the data, constructed by the scientist to be similar to the original data set in relevant respects.

An account of data that employs data models can address the role of theoretical goals and assumptions in the processing and interpretation of

https://doi.org/10.1086/377426 Published online by Cambridge University Press

data in a number of ways. For example, it is acknowledged that there are different ways to model data and that these different ways of modeling result in data models that have differing known similarities and dissimilarities with the real-world objects they are intended to represent. When a two-dimensional world map is created, there are a number of possible map projections a cartographer can choose, such as an orthomorphic projection, which retains the correct shape but changes the relative areas of the continents, or an equivalent projection, which preserves the correct relative sizes of the continents but distorts their shape. The cartographer chooses a particular projection with a goal in mind. Similarly, when scientists choose to produce a histogram rather that a smooth curve from a particular data set, they know that certain regularities in the data will be highlighted, while other information will be lost. Scientists make decisions regarding the construction and processing of a data model knowing some of the dissimilarities, as well as some of the similarities between the data model and the real-world object it is intended to represent.

As the concept of raw data plays a role in the debate that is presented in the next section, it is important to note that in many cases the data that has traditionally been referred to as raw is in fact a data model. For example, it will become clear that data such as micrographs are not raw data, but instead data models. Because this might raise an immediate objection, it is worthwhile to examine why researchers refer to some data as being raw.

One possibility is that it is believed that certain data has not been processed, or manipulated by scientists. By this standard, and in any plausible way of understanding manipulation and processing, it is often the case that data models do not represent anything that could be thought of as unmanipulated or unprocessed data. For example, an electron micrograph is the product of an astonishingly complex instrument that requires a specimen to undergo a lengthy preparation procedure. Scientists will change many aspects of this specimen preparation procedure as well as settings on the microscope in order to achieve a desired effect in the resulting micrograph. Because of this purposeful manipulation of both the microscope and the specimen, the electron micrograph cannot be said to be unprocessed data.

A second possibility is that it is believed that certain data is raw because it is not influenced by theory. Using the same example, in many types of microscopy, including electron microscopy, scientists make decisions effecting the micrograph based on their assumptions about the real-world system being tested. These will usually be theoretical assumptions, such as lipids are hydrophobic, or glutaraldehyde covalently cross-links proteins.

A third possibility is that certain data is believed to be raw because it is in a unique position between the real-world system and what is considered to be the first-level data model. Michael Lynch has pointed out (Lynch 1990) that there is often a hierarchy of representations between highly

https://doi.org/10.1086/377426 Published online by Cambridge University Press

processed data models and what is commonly called raw data, each level in some way dependent on the previous level. Many situations are more complex, involving a range of different kinds of levels of representation of the data with models. Each of these models will be variously helpful in achieving the goals of the scientist. It is possible that some of these levels are not arranged in a linear hierarchy. In these cases, some might find it tempting to call the lowest level or first-data model the raw data, but this will no longer mean unmanipulated, uninterpreted data, as many intend when they use the term "raw data." Because of this, I think that many cases should be described as not involving raw data of any kind.

Notice that in these cases the process of data acquisition cannot be separated from the process of data manipulation. From the outset the data must be considered to be the product of a certain amount of purposeful manipulation. This can occur even when simple instruments are used. Two examples involving familiar measuring devices provide cases in which instruments do not produce uninterpreted or "raw" data. At first it might seem as if one could directly read the voltage of a circuit off of the dial of a voltmeter, or the temperature of a liquid off of the mercury column of a thermometer. However, both instruments require a certain amount of learned skill to be properly read. A mercury column can be hard to line up with the temperature scale printed on the thermometer due to the meniscus at the top of the column and the refractive effects of the glass or plastic that holds the mercury. Reading an analog voltmeter also presents difficulties. An analog voltmeter's needle does not hold still, so the observer must learn how to determine the average reading of the needle over a period of time. Until one learns the skills necessary to make these readings, neither instrument will yield an acceptable data model.

When one moves from these relatively simple instruments to an instrument such as an electron microscope, it becomes obvious that the instrument is not producing raw data (in the sense of being unprocessed), but data that has been interpreted and manipulated, in short, a data model. The lack of raw data in many scientific experiments complicates a debate that has recently occurred over the acquisition and manipulation of data.

4. A Recent Debate Concerning Raw Data and Data Manipulation. James Bogen and James Woodward have been concerned with the relation between data, theory, and what they term "phenomena." "We expect phenomena to have stable, repeatable characteristics which will be detectable by means of a variety of different procedures, which may yield quite different kinds of data" (Bogen and Woodward 1988, 317). Phenomena include such things as the quantum tunneling, or the breeding patterns of sharks. When scientists study phenomena, data is produced. "Data, which play the role of evidence for the existence of phenomena, for the most part

https://doi.org/10.1086/377426 Published online by Cambridge University Press

can be straightforwardly observed" (Bogen and Woodward 1988, 306). Data provides evidence for claims about phenomena.

In a later article entitled "Data and Phenomena," Woodward makes the point that data is usually processed in some manner by the investigators before it is in an acceptable form:

> Moreover, data must be made to occur in a form which is tractable with respect to the demands of data-reduction and statistical analysis—a consideration which is crucially important in high energy physics (cf. Section VI). Data must also result from processes in which there has been adequate control for various kinds of experimental error (cf. Section VI). Here again, data having these desirable features rarely occurs naturally. Instead, it is typically the product of laborious, elaborate, and carefully planned contrivance on the part of the investigator. (Woodward 1989, 321)

During this process of data manipulation, scientists are looking for features of their observations or experimental outcomes that they can take to be evidence for claims about phenomena.

Bogen and Woodward also discuss how scientists determine if their observations provide reliable evidence for claims about phenomena. They give several methods by which scientists "control confounding factors" in their observations. To give the reader some indication of what control of confounding factors includes, the example of calibration will be mentioned here, which is credited to Allan Franklin (Bogen and Woodward 1988, 327). Calibration is the use of an instrument to detect phenomena of known characteristics in order to gauge the accuracy of that instrument. One of Franklin's examples involves using a spectrometer to determine known spectra such as the Balmer series of hydrogen (Franklin 1990, 104–105). By doing this, scientists can look for errors or deviations from the reported series in the output of the spectrometer. Another example is putting a thermometer in boiling water (which one knows boils at 100°C at 1 atm) to check if the thermometer is accurate.

Edward McAllister criticizes Bogen and Woodward in "Phenomena and Patterns in Data Sets." In this paper McAllister notes that Bogen and Woodward do not explain how scientists choose which features of their observations or experimental outcomes they will accept as being reliable evidence of phenomena:

> Some of these patterns are taken by investigators as corresponding to phenomena, not because they have intrinsic properties that other patterns lack, but because they play a particular role in the investigators' thinking or theorizing. The two accounts [Bogen and Woodward's and McAllister's] differ also in their epistemological and

https://doi.org/10.1086/377426 Published online by Cambridge University Press

> methodological implications. Bogen and Woodward hold that investigators discover from data sets which phenomena there exist; on my account, investigators discover the patterns that are exhibited in data sets, but stipulate that some of these correspond to phenomena. (McAllister 1997, 224)

McAllister gives the example of a hypothetical situation in which Newton and Kepler are examining a data set consisting of the positions of the planets observed at different times (1997, 226). Depending on the amount of noise allowed, this data set could be used to support either Kepler's ellipses, or Newton's more complicated model in which gravitational effects between the planets cause their paths to deviate from perfect ellipses. In McAllister's terms, both patterns exist in the data, and it is the scientist who stipulates which pattern or patterns will be taken to correspond to the phenomena of interest by specifying the acceptable noise level.

McAllister's objection is that Bogen and Woodward have minimized the role of interpretation in their theory of data. It is true that scientists look for a particular kind of pattern when examining their data. Another example of this is X-ray crystallography, where electron density maps are processed by a computer to produce three dimensional pictures of macromolecules such as proteins. A certain amount of information not contained in the electron density map must be added before one gets a clear three-dimensional output from the computer. For instance, until one tells the computer that the proteins are expected to be in an ordered crystal lattice (among other things) the computer will not pick out a peptide chain from the data. As McAllister points out, there are always competing patterns in data. A scientist must choose one pattern over the other competing patterns.

5. Data Models and Theoretical Objectives. An account of data that utilizes data models can accommodate McAllister's observation regarding the role of interpretation in data processing. Bogen and Woodward claim that scientists use a number of procedures to eliminate confounding factors, and that this process of elimination results in reliable data (Bogen and Woodward 1988, 392). As pointed out by McAllister, Bogen and Woodward's formulation assumes that there is only one reliable data model, given the data. This is not the case. Consider, for example, an electron micrograph of a cross section of a kidney cell.[3] This data could be processed to produce a variety of data models. Scientists could create a chart that shows the probability of finding a mitochondria in different regions of cell (e.g., close to the Golgi or near the plasma membrane). Another

[3] In order to keep this example simple, I will treat the electron micrograph as raw data here.

https://doi.org/10.1086/377426 Published online by Cambridge University Press

possible data model would be a histogram displaying the average distance between mitochondria in different regions of the cell. Finally, a line could be drawn through the long axis of each mitochondria to create a vector representing the orientation of that mitochondria relative to a designated axis in the cell, then the average orientation of mitochondria in different regions of the cell could be presented in a graph.

Each of these possible data models would be produced with a theoretical objective in mind. One would not want to claim that one of these data models is the acceptable data model, and the rest are unacceptable. In order to understand why each of these data-manipulation processes results in an acceptable data model, one must examine the theoretical objectives that guided the production of each data model.[4] For example, the first data model might be acceptable if scientists were interested in studying the fission and fusion of mitochondria, while the last data model might be acceptable if they were studying the transport of mitochondria. In general, data models are acceptable relative to a theoretical goal, and cannot be evaluated independently from that goal.

Rather than say that scientists stipulate patterns in data (as McAllister does), one should say that the construction of data models involves interpretation. In many cases what has traditionally been taken to be raw data already embodies theoretical principles, or at least is a manufactured object that contains input from nature, but is also the product of a number of decisions scientists have made. Such cases are better described in terms of data models that involve interpretation from the start, and are to be evaluated for reliability by various standards, depending differentially on the theoretical objectives in question.

Given this account, some will worry about the influence of scientists' interests on their judgments about what counts as relevant similarity between a data model and its target. As shown above, judgments concerning data models involve theoretical assumptions. Using McAllister's terms, the theoretical assumptions and interests of the scientists will influence the patterns they stipulate. Similarly, Mary Morrison has pointed out that not all theoretical models are models of theory in the sense of being derived from theory, nor is every phenomenological model free from theoretical concepts and parameters (Morrison 1999, 44). In particular, as scientists construct data models they use elements of theory to guide the production of data models. Nancy Cartwright has emphasized that many models are used to get beyond theory to better represent physical systems:

[4] I am not ruling our other kinds of objectives, or claiming there is any clear separation between theoretical objectives and other kinds of objectives.

https://doi.org/10.1086/377426 Published online by Cambridge University Press

> I want to argue that the fundamental principles of theories in physics do not represent what happens; rather, the theory gives purely abstract relations between abstract concepts: it tells us the "capacities" or "tendencies" of systems that fall under these concepts. No specific behavior is fixed until those systems are located in very specific kinds of situations. When we want to represent what happens in these situations we need to go beyond theory and build a model. (Cartwright 1999, 242)

Along these lines one can note that a data model can be the lowest level of representation of what happens. It is specific to the experiment at hand. However, it is not simply a copy of what happens. Instead it incorporates elements of theory to create a representation that contains features of interest to the scientist.

A description in terms of data models facilitates an improved understanding of the interplay between theoretical principles, the theoretical interests of scientists, and aspects of the data model that are due to experimental constraints imposed by nature. When one is faced with complex experiments involving multiple instruments, each producing manipulated data, an account of scientific methodology that requires the identification of raw, unprocessed data will be inadequate. As McAllister points out, we will be left to explain how scientists identify patterns in that data, and given the fact that a number of patterns exist in a given data set, this will be challenging, if not impossible. Instead, we should realize that experiments will often contain a number of data models. The task then becomes to identify the theoretical assumptions and objectives that produced these data models. The data-model approach can accommodate the wide range of situations one encounters in laboratory science, including cases in which there is no raw data, in other words, cases in which data acquisition, manipulation, and elements of interpretation are not independent activities.

REFERENCES

Bogen, James, and James Woodward (1988), "Saving the Phenomena", *Philosophical Review* 97: 303–352.

Cartwright, Nancy (1999), "Models and the Limits of Theory: Quantum Hamiltonians and the BCS Models of Superconductivity", in Mary S. Morgan and Margaret Morrison (eds.), *Models as Mediators*. Cambridge: Cambridge University Press, 241–281.

Franklin, Allan (1990), *Experiment, Right or Wrong*. Cambridge: Cambridge University Press.

Lynch, Michael (1990), "The Externalized Retina: Selection and Mathematization in the Visual Documentation of Objects in the Life Sciences", in Steve Woolgar and Michael Lynch (eds.), *Representation in Scientific Practice*. Cambridge: MIT Press, 153–186.

McAllister, James W. (1997), "Phenomena and Patterns in Data Sets", *Erkenntnis* 47: 217–228.

Morrison, Margaret (1999), "Models as Autonomous Agents", in Mary S. Morgan and

https://doi.org/10.1086/377426 Published online by Cambridge University Press

Margaret Morrison (eds.), *Models as Mediators*. Cambridge: Cambridge University Press, 38–65.

Suppes, Patrick (1962), ‘‘Models of Data’’, in Ernest Nagel, Patrick Suppes, and Alfred Tarsi (eds.), *Logic, Methodology, and Philosophy of Science*. Stanford: Stanford University Press, 252–261.

Woodward, Jim (1989), ‘‘Data and Phenomena’’, *Synthese* 79: 393–472.

https://doi.org/10.1086/377426 Published online by Cambridge University Press