DATA MODELS

Data models, or 'models of data,' are processed versions of data, typically prepared in such a way as to make the data useable as evidence. The term first gained prominence in a 1962 paper by the philosopher of science Patrick Suppes, and after a period of relative neglect has now become an area of active research. This entry will summarize some of the key distinctions and debates, including those regarding the ontology of data, the difference between raw data and data models, the purpose(s) of data modeling, the variety of data processing methods used in the production of data models, and how data models should be evaluated.

All STEM research— indeed all empirical inquiry—is heavily reliant on data collection, processing, and interpretation; hence data models are of foundational importance. Conceptualizing the research process as involving data modeling helps to clarify the relationship between data and the world, data and theories, and the role of modeling in the scientific process.

Ontology of Data & Data Models

One of the most basic theoretical questions one can ask about data concerns what kinds of things data are. Some researchers place no restrictions on which elements of the world can count as data. On this view, laboratory mice could themselves count as data. Other researchers define data as *records* of a process of inquiry involving a physical interaction between the researcher, measuring instrument, and the world: it is only the results of observations or measurements performed on the mice colony that count as data. On this latter view, data necessarily involve a level of abstraction. Data involve replacing the thing in the world with a number, a photograph, or a recorded description. Which view one adopts arguably has epistemic implications, such as whether data are unproblematically "given" or whether they are subject to the sort of questions raised by the philosophy of observation, philosophy of experiment, and philosophy of metrology (measurement).

What kind of thing is a data model? When Suppes first introduced the notion of a data model he conceptualized them as Tarskian or set-theoretic models. Although Tarskian models are prominent in mathematics, the more commonly used notion of model in science today treats models as *representations* of some target system. Some have argued that this representational notion of a data model is important for recognizing that data are *about* the world, and hence for making sense of their ability to serve an evidential role.

Distinguishing Raw Data, Data Models, & Models

Data models are typically contrasted with "raw data," but it is debatable whether and how such a distinction can be meaningfully drawn. Raw data are sometimes defined to be the immediate output of an observation, measuring process, or instrument. Often, however, such immediate outputs are not yet in a form that is useable by scientists. This can occur for a wide variety of reasons: first, the data-collection process might involve some obvious errors resulting in a certain portion of "bad" data that should be excluded from the data set; second, there could be a source of "noise" in the data signal that needs to be filtered out or subtracted; third, the output of the instrument might be of a quantity (measurand) that is closely related, but not identical, to the data quantity of interest, and hence must be converted before being useful; fourth, data are often discrete samples of a continuous quantity, hence need to be smoothed or interpolated; and, fifth, data often need to be organized or ordered before patterns can emerge. The outcome of these various processes of data wrangling—that is, of correcting, converting, smoothing, organizing, etc.—is what we call a *data model*.

The term ‘data model' can refer to any type of data product, including data sets, graphs, equations, images, and even processed artifacts. Statisticians often take ‘data model’ or ‘model of the data’ to refer more narrowly to equations that summarize a data set or the relationships therein. On this construal, models of data are akin to phenomenological models that summarize empirical relationships without representing or positing underlying mechanisms and processes. Phenomenological models, however, are typically distinguished from models of the data by playing a broader interpretive and inferential role. Data models can also be contrasted with data-driven models, which are typically constructed through machine learning and are trained on data sets to predict the value of a dependent variable from a set of independent variables.

On the other side, data models can also be contrasted with raw data. In practice, however, this distinction too can become blurred, with the term ‘raw data’ often taking on a relative rather than absolute meaning. There are frequently many different layers of data processing, depending on the intended use of the data, not all of which are performed at the same time or by the same researchers. Scientists thus often use the term "raw data" to describe the data model at the stage before the data processing they are about to engage in. The raw data/data model distinction also becomes blurred as measuring instruments become more sophisticated, with a lot of the data processing (correction, conversion, etc.) happening within the instrument itself. Given the substantial, and often theory-mediated, processing involved, it is not clear that even the instrument output should be called "raw." While some researchers are comfortable with the idea that it is "data models all (or almost all) the way down," others have tried to recover a principled distinction, such as by taking raw data to be nonrepresentational and only becoming representational when they are processed as a data model.

The Purpose(s) of Data Models

When Suppes first introduced the concept of a model of the data, he conceived of it as part of a hierarchy of models, which included not only models of data, but also models of experiment and models of theory. The purpose of a data model on this view was to provide a bridge for linking theories to data. Being tied specifically to the Tarskian notion of model (which Suppes used for both models of theory and models of data) the bridge was to be effected formally in terms of the presentation or instantiation of shared structure.

Most researchers, however, see the purpose of data modeling as helping data better function as evidence. For those who adopt a representational view of models, this involves drawing out some signal or representational component of the data in order to make it a more accurate or transparent representation of that feature of the world. Here the aim of data modeling is not necessarily to bridge to some theory, but instead to arrive at a better or more useful data product. Some researchers argue that, rather than being part of a synchronic hierarchy to theory, data modeling should instead be viewed as part of a diachronic data lineage. Although one may use a data model to test a theory, this is just one of many possible purposes.

How Data Models are Produced

Many different manipulations and processes can be involved in data modeling. Some involve statistical analysis methods such as data reduction, curve fitting, correlation analysis, regression analysis, statistical significance tests or calculation of confidence intervals, and error estimation. Other data modeling methods go beyond statistical analysis but are just as essential for making data usable as scientific evidence. Which techniques should be used in a particular case will depend on the nature of the data, the aims of the researchers, and the norms of their fields of study.

All data interpretation involves theoretical or conceptual assumptions about the experimental processes by which the data were produced, which are necessary for their proper semantic interpretation. Some data-processing methods, however, can make more substantial use of theoretical knowledge in the production of data models. One widely used example is data conversion, which involves converting a measure of one experimentally accessible quantity into data about another closely related quantity of interest. Examples of this include converting data about the travel time of some acoustic or electromagnetic signal into data about depth or distance, or converting data about the height of a mercury column in a thermometer into data about temperature. Another category of data-processing methods used in the production of data models falls under the rubric of data correction, which involves removing various sources of noise from the data or correcting for other errors, such as in a process of data calibration. This may be particularly important in cases where the data have been collected in the field, rather than in a controlled lab setting. A third process is data interpolation, which involves filling in gaps in the data, especially when the collected data represent some continuous quantity or need to be evenly distributed in space or time for a context of use. These are just a few examples of the many ways raw data are turned into data models, and in many cases several of these data-processing methods will be used together.

For those who think of physical objects themselves (and not just numerical or other records of those objects) as data, the production of associated data models requires methods other than statistical manipulation, such as specimen preparation (e.g., preparing fossils) or image processing techniques for photographs. Data modeling can also involve the preparation or ordering of data for storage and dissemination, including the transfer of data from one substrate or vehicle to another (such as in transferring data from descriptive text in a book to a computer database).

There remain substantive questions about how the lines between data modeling and various other data practices, such as 'data curation' or 'data visualization,' should be drawn. Indeed, standardizing the definitions of these and other terms related to data processing is an ongoing project.

Evaluating Data Models

When thinking about data as the output of a measurement, data are typically evaluated in terms of accuracy, precision, and resolution. These three terms are often explicated using the intuitive analogy of a dartboard: accuracy is how close the dart lands to the bullseye, precision is how closely together a number of darts cluster (anywhere on the dartboard), and resolution is how thin or thick the dart point is. Precision, however, is arguably better thought of as describing the conditions of measurement, rather than the quality of resulting data. More useful is the notion of measurement *uncertainty*, which can arise from many different sources and whose estimate should accompany the data, typically in the form of metadata (i.e., data about data). The quality of data, thus can also depend on the quality of the metadata that accompanies the data model.

There are substantive issues in both philosophy and metrology concerning exactly how these various concepts should be defined and deployed.

In addition to the epistemic evaluation of data models, involving for example how close data values are to some 'true value,' there are socially-driven criteria for evaluating data models, which are often articulated in terms of the so-called FAIR data principles. FAIR is an acronym for Findability, Accessibility, Interoperability, and Reusability. The FAIR principles are taken to apply "not only to 'data' in the conventional sense, but also to the algorithms, tools and workflows that led to that data" (Wilkinson et al. 2016, p.1). These principles are central to the 'open data' movement.

More generally, scholars in both the philosophy of science and metrology communities have argued that data models should be evaluated, not *in abstracto*, but rather contextually in terms of a multi-dimensional problem space that assesses their adequacy or fitness for particular purposes. In the *adequacy-for-purpose* view of data models, the relevant question is not just how closely the data model resembles some feature of the world, but whether or not it is an adequate data model given a certain set of resources, methodologies, and other constraints for accomplishing a particular scientific aim. This view recognizes the essential role that context plays in data evaluation; a data model of a certain accuracy, etc. that is adequate for a scientific purpose in one context, may or may not be adequate in another context.

Yet another dimension along which data models can be evaluated is ethical. Data can be collected, curated, modeled, and disseminated in ways that are biased against—and harmful to—particular communities or social groups, such as women and Black, Indigenous, or other People of Color (BIPOC). Data models can embed or obscure systematic biases, especially in social contexts such as medicine, finance, and criminal justice. For example, the Black Lives Matter movement has called attention to the ways in which racial data collected by law enforcement can be biased along many dimensions, from over-policing certain social groups to biased police reports, leading to what AI Now researchers have called 'dirty data.' These data ethics issues become especially prominent when data models are used for machine learning, artificial intelligence, and other automated decision making. Such data arguably require not just an evaluation of the epistemic risks, but a moral evaluation of the ethical risks as well. The Global Indigenous Data Alliance has argued that the FAIR principles of data evaluation are inadequate in that they ignore power differentials and historical contexts. They argue for a complimentary set of what they call CARE principles for data evaluation, which stands for Collective benefit, Authority to control, Responsibility, and Ethics.

Alisa Bokulich and Aja Watkins

See also Measurement, Accuracy, Data and Phenomena, Big Data, Evidence, Modeling, Philosophy of Science, Statistics

Further Readings

Bokulich, A. (2020). Towards a Taxonomy of the Model-Ladenness of Data. *Philosophy of Science,* 710516. https://doi.org/10.1086/710516

Bokulich, A., & Parker, W. (2021). Data Models, Representation, and Adequacy-for-Purpose. *European Journal for the Philosophy of Science*, https://doi.org/10.1007/s13194-020-00345-2.

Harris, T. (2003). Data Models and the Acquisition and Manipulation of Data. *Philosophy of Science*, *70*(5), 1508–1517. https://doi.org/10.1086/377426

Humphreys, P. (2014). X-Ray Data and Empirical Content. In P. Schroeder-Heister, W. Hodges, G. Heinzmann, and P. Bour (eds.) *Logic, Methodology and Philosophy of Science: Proceedings of the Fourteenth International Congress*, pp. 1-15.

Leonelli, S. (2019). What Distinguishes Data from Models? *European Journal for Philosophy of Science*, *9*(2), 22. https://doi.org/10.1007/s13194-018-0246-0

Leonelli, S. and N. Tempini (Eds.) (2020). *Data Journeys in the Sciences*. Springer Open. https://doi.org/10.1007/978-3-030-37177-7

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

Suppes, P. (1962). Models of Data. In P. Suppes (ed.), *Studies in the Methodology and Foundations of Science: Selected Papers from 1951 to 1969* (pp. 24–35). Springer Netherlands. https://doi.org/10.1007/978-94-017-3173-7_2

Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3:160018. https://doi.org/10.1038/sdata.2016.18