What is the Purpose of Statistical Modelling?

David Hand

Published on: Jun 22, 2019

Updated on: Oct 04, 2019

DOI: [10.1162/99608f92.4a85af74](https://doi.org/10.1162/99608f92.4a85af74)

Chris Anderson, the former editor of *Wired* magazine, famously wrote (2008) that “[t]oday companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.” He went on to say, “We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.” As an example, he cites Craig Venter’s sequencing of genotypes.

The notion that we can manage without models and that sufficient quantities of data—big data—can take the place of models is a seductive one. After all, coming up with realistic models describing the way (natural) processes might work is hard mental effort. If we can instead simply crunch vast data sets, relying on the awesome power of modern computers, so much the better. This notion that having enough data means we do not have to worry about constructing models invokes the saying that “the numbers speak for themselves,” although this adage has something of a history (see Hand 2019).

But Anderson was wrong. He failed to take account of the fact that there are two fundamentally different types of model, and that while “big data” might partly replace one, it will not do so for the other. Some authors add other types of model, or make other distinctions (e.g. Cox 1990; McCullagh 2002; Neyman and Scott 1959), and computer scientists use the term “data model” to describe the relationships between the aspects of the structure of a data set. However, for data science I think the key distinction—at least as far as responding to Anderson’s comment goes—lies between two types of model, which appear under various names.

On the one hand we have *theory-driven*, *theoretical*, *mechanistic*, or *iconic models*, and on the other hand we have *data-driven*, *empirical*, or *interpolatory models*. Theory-driven models encapsulate some kind of understanding (theory, hypothesis, conjecture) about the mechanism underlying the data, such as Newton’s Laws of motion in mechanics, or prospect theory in psychology. In contrast, data-driven models merely seek to summarize or describe the data.

The two types of model need not be exclusive—both can be used in any particular application, and indeed, the division between the two types may not always be sharp. Moreover, models often start out as data-driven and gradually become theory-driven as understanding grows. Certainly, a given statistical technique might be used to fit models of either type.

The distinction between the two types of model comes into focus when we recall an even more famous comment than Anderson’s, which is George Box’s remark to the effect that, while all models are wrong, some models are useful (e.g. in Box 2005). In the context of the two model types, we see that this comment is not quite right. Data-driven models *cannot* be wrong—though they can be poor, or of varying degrees of usefulness for any particular purpose—because they are simply summarizing data and are not describing any purported underlying reality. Theory-driven models, on the other hand, can

be wrong, inadequate, or indeed wildly misleading descriptions of the reality they are being used to represent. Box, of course, was fully aware of this. He gave an example: “In many circumstances even though no theoretical model is available, perfectly good empirical approximations can be obtained by fitting a polynomial or some other flexible graduating function over the region of interest” (2005).

Box’s point that models are approximations applies to theory-driven models as well as empirical models. No model can capture all of the niceties of the real world, so that (theory-driven) models are idealizations and simplifications. This fact was nicely captured in the title of Theodore Micceri’s (1989) paper, “The Unicorn, the Normal Curve, and Other Improbable Creatures.” The normal curve does not occur in nature, but it’s a very useful mathematical model – rather like the points and lines of elementary geometry. As Box (1979) says, “[I]t would be very remarkable if any system existing in the real world could be *exactly* represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations.”

The fact that there is a distinction between these two types of model is not a new observation.

Neyman (1939) distinguishes between models for fitting empirical data distributions and models which provide “an ‘explanation’ of the machinery producing the empirical distributions.” However, the importance of the distinction has suddenly grown because of the data revolution. In particular, the availability of massive data sets along with hardware and algorithms which allow elaborate models to be fitted effortlessly has led to dramatic advances such as autonomous vehicles, machine translation, chess and Go machines, credit scoring, automatic diagnosis, and a host of other applications, many occurring online and in real time. The ultimate models underlying these advances are almost always primarily empirical and data-driven, with virtually no theoretical guarantee or representation of a fundamental ‘mechanism.’ This example underpins Anderson’s observation, quoted at the beginning of this article.

For prediction, data-driven models are ideal—indeed in some sense optimal. Given the model form (e.g. a linear relationship between variables) or a criterion to be optimized (e.g. a sum of squared errors), they can give the best fitting model of this form, and if the criterion is related to predictive accuracy, the result is necessarily good within the model family. In contrast, theory-driven models are required for understanding, although of course they can also be used for prediction. It is sometimes suggested that prediction will be superior if it is based on a model of an underlying mechanism, but this is not necessarily the case. Examples showing this include the elaborate and purely empirical models implicit in deep learning and random forests, and oversmoothed models which increase bias (and so depart from accurately representing underlying reality) but have smaller variance and hence a lower overall total of squared errors.

In a recent discussion contrasting models for prediction with those for understanding, Shmueli (2010) points out that “applied statisticians instinctively sense that predicting and explaining are different,” although she precedes this by saying the distinction is “not explicitly stated in the statistics methodology literature.” That last observation is untrue—see, for example, Cox (1990), Lehmann (1990), Breiman (2001), or Hand (2009). Breiman (2001) used the words prediction and information for the two objectives, with “information” meaning “to extract some information about how nature is associating the response variables to the input variables.” Lehmann (1990) pointed out that the two types of model differ in their basic purpose—empirical models being used as “a guide to *action*, often based on forecasts of what to expect from future observations” (so we are in the realm of prediction), whereas explanatory models “embody the search for the basic mechanism underlying the process being studied; they constitute an effort to achieve *understanding*.” (His italics.) Cox (1990) also introduced a third model type, “indirect” models, where probability models are used to suggest methods of analysis.

In his famous paper, “Statistical Modeling: the Two Cultures,” Breiman extended the theory-driven/data-driven modelling continuum further. Rather than thinking in model terms, he says “[T]he problem is to find an *algorithm* $f(\mathbf{x})$ such that for future $\mathbf{x}$ in a test set, $f(\mathbf{x})$ will be a good predictor of $\mathbf{y}$ ” (2001, p.205, original emphasis). He continued: “The theory in this field shifts the focus from data models to the properties of algorithms.” Algorithms, of course, are the lifeblood of computer science, and the recognition that data-driven models can be looked at from two perspectives—models and algorithms—is a core strength of data science.

Even further along this continuum is the school represented by the Albert Gifi consortium. (e.g. Gifi, 1990). This approach is summarized in their preface: “Models do have a place in Gifi’s philosophy: not as tentative approximations to the truth, but as devices to sharpen and standardize the data analytic techniques. [...] First choose a technique (implemented in a computer program) on the basis of the format of your data, then apply this technique, and study the output” (p.v).

The distinction between data-driven and theory-driven models can be important. Most of the big, attention-grabbing illustrations of data science in action are data-driven. But if theory-driven models can be wrong, data-driven models can be *fragile*. By definition they are based on relationships observed within the data which are currently available, and if those data have been chosen by some unrepresentative process, or if they were collected from a non-stationary world, then their predictions or actions based on the models may go awry.

References

- Anderson C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired*. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
- Box G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L.; Wilkinson, G. N. (Eds.), *Robustness in Statistics* (201-36). New York, NY: Academic Press.
- Box G.E.P. and Hunter W.G. (1965). The experimental study of physical mechanisms. *Technometrics*, 7, 23-42.
- Box G.E.P., Hunter W., and Hunter S. (2005). *Statistics for Experimenters* (2nd ed.). New York, NY: Wiley.
- Breiman L. (2001). Statistical modeling: the two cultures. *Statistical Science*, 16, 199-215.
- Cox D.R. (1990). Role of models in statistical analysis. *Statistical Science*, 5, 169-174.
- Gifi A. (1990). *Nonlinear Multivariate Analysis*. Chichester, England: Wiley.
- Hand D.J. (2009). Modern statistics: the myth and the magic (RSS Presidential Address). *Journal of the Royal Statistical Society, Series A*, 172, 287-306.
- Hand D.J. (2019). Talking data. To appear in *Bulletin of the Institute for Mathematical Statistics*.
- Lehmann E.L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science*, 5, 160-168.
- McCullagh P. (2002). What is a statistical model? *The Annals of Statistics*, 30, 1225-1310.
- Micceri T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Neyman J. (1939). On a new class of ‘contagious’ distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, 10, 35-57.
- Neyman J. and Scott E.L. (1959). Stochastic models of population dynamics. *Science*, 130, 303-308.
- Shmueli G. (2010). To explain or predict? *Statistical Science*, 25, 289-310.

This article is © 2019 by David Hand. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (<https://creativecommons.org/licenses/by/4.0/legalcode>), except where

otherwise indicated with respect to particular material included in the article. The article should be attributed to the author identified above.