PROJECT MUSE®

Reasoning Using Data: Two Old Ways and One New

Michael Baiocchi, Jordan Rodu

Observational Studies, Volume 7, Issue 1, 2021, pp. 3-12 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2021.0016>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/799741>

Reasoning using data: two old ways and one new

Michael Baiocchi

baiocchi@stanford.edu

*Department of Epidemiology and Population Health
Stanford University
Stanford, California 94062, USA*

Jordan Rodu

jsr6q@virginia.edu

*Department of Statistics
University of Virginia
Charlottesville, VA 22901, USA*

Abstract

Instead of two cultures, the story of the last couple decades of data science is about the interplay between three different types of reasoning using data. Two of these types of reasoning were well known when Breiman wrote his *Two Cultures* paper – warranted reasoning (e.g., randomized trials and sampling) and model reasoning (e.g., linear models). Breiman, though he does not appear to have realized it fully, was in fact describing the dynamics arising in a data community that was making progress using the newest, third type of reasoning – outcome reasoning. In this commentary we clarify this dynamic a bit, and suggest some useful language for identifying and differentiating types of problems better suited for outcome reasoning.

Keywords: warranted reasoning, model reasoning, outcome reasoning, common task framework, MARA

1. Introduction

Two decades ago, Leo Breiman wrote one of the few anthropological papers in the field of statistics. He describes how he had gone out of the Statistical Ivory Tower, explored other data cultures, returned to tell us of other thriving cultures, and to – quite forthrightly – warn us that our sleepy tower may be overrun. We, the authors of this particular commentary, were graduate students in a statistics department only a handful of years after Breiman wrote his *Two Cultures* paper. We remember reading his paper for the first time – with absolutely riveting commentaries and a rejoinder providing back-and-forth from some of the most wonderful thinkers, and guardians, of traditional statistical reasoning. At the time, a confluence of technological and methodological advances made it clear there was a shift in the pace of innovation for predictive algorithms. This shift (re-)launched buzzwords like “machine learning” and “artificial intelligence” that seemed to tug at the imagination of thinkers in both data oriented and non-data oriented fields alike. We heard urgent debates about where statistics stood in relationship to this other data culture. Should we jump on their band-wagon, or distance ourselves? Is this a temporary fad that will lead to another AI winter? Or is there a chance we could get left behind as people look elsewhere for data expertise? Though we weren’t around for it, perhaps this is how the Bayesian-

Frequentist “wars” felt at their peak. We point to that cultural schism, and its resolution, to help explain our perspective on Breiman’s two cultures: though there are recognizable differences in how the “algorithmic” and “modelling” cultures approach the specification of their problems, and these specifications can lead to divergent conclusions, these are merely types of reasoning using data that are more, and less, appropriate to certain classes of problems.

Though at some point it felt right to frame this dynamic in terms like “cultures clashing” (which is exciting! high-energy! dynamic!), twenty years later it is clear that what Breiman was describing was not fundamentally a new culture so much as a new form of reasoning using data. This new form of reasoning – which we call “outcome reasoning” – joins the other two types of reasoning that had been well-articulated at the time of Breiman’s writing: “warranted reasoning” and “model reasoning.” And outcome reasoning, like other forms of reasoning (i) has its advantages and its disadvantages and (ii) can be adopted or discarded by an analyst depending on the features of the problem at hand.

In this commentary, we have four goals:

- First, we correct a (very reasonable) error Breiman makes in his assertions about the intellectual foundation of the algorithmic community. We do this to orient the reader. Breiman’s assertion about the foundation, though not correct, seemed so at the time. But now, two decades later, a simple observation shows his assertion was mistaken.
- Second, we will describe “outcome reasoning,” the novel foundation of the algorithmic community’s form of reasoning. And we will discuss how it is currently deployed, as a version of the Common Task Framework (CTF). We suspect other commentators in this issue will also point to the CTF, so we will focus on clarifying its assumptions, as well as how it leverages many of the features present in the current epoch of data analysis (e.g., fast computing). We will also describe limitations to the CTF (for example: given its assumptions it is best used during the algorithmic development phase rather than during the deployment phase).
- Third, we describe an extension to the CTF which, for certain types of problems, allows for outcome reasoning in the deployment phase.
- Fourth, having worked with non-technical colleagues and taught students to use the various forms of data reasoning, we realize there is a lack of rigor and much confusion around using outcome reasoning. Part of this comes from our discipline not having settled on the correct way to describe the assumptions used in outcome reasoning, and therefore not having useful terminology to help communicate the appropriateness and violations of those assumptions. We spend some time here describing how we can teach students and collaborators to identify if a problem is appropriate for outcome reasoning, though we leave the bulk of the development and justification of our framework to work we have done elsewhere (Rodu and Baiocchi (2020)).

A discussion of labels: though Breiman used the term “algorithmic community” we recognize that it’s probably more common nowadays to call this group of data scientists “machine learners.” We’ll use those terms interchangeably, but if you disagree please feel

free to replace and/or nuance; our arguments hold either way. We use the terms statistician and modeler interchangeably (though this doesn't do a good job capturing folks who focus on experiments and sampling) – and recognize that “statistician” (used in this sense) includes people who might call themselves something different like “biostatistician,” “econometrician,” or “epidemiologists.” We use the term “data scientist” and “analyst” to be the most inclusive term, including all data cultures.

2. Correcting an error: reasoning using data

Breiman places the foundation for the new culture of algorithms in a form of reasoning already familiar to the statistics community: warranted reasoning. In Section 7.2 of *Two Cultures*, he makes his most direct attempt at articulating this idea: “The theory in this field shifts focus from data models to the properties of algorithms.” That is a direct rejection of reasoning derived from models. He follows this with, “The one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution.” Anchoring the success of algorithmic culture on warranted reasoning is powerful because it provides cover for the obscuring that accompanied the rapid innovation in machine learning (ML) – the most extreme example being the infamous “black box” algorithms that have emerged. Before we unpack this a bit, we pause to ask a simple question:

Is it true that ML has made progress by using i.i.d. sampling?

No. At this point in its development, essentially none of the progress in ML has come from using data sets generated from carefully crafted sampling schemes.

This misalignment is concerning. Breiman argued that the new culture was a direct rejection of modeling. But in practice, sampling is not the bedrock of progress in machine learning. It's not randomization either. So what is at the core of ML's progress? It is a new form of reasoning – a form of reasoning that Breiman mistook for warranted reasoning – that has come into its own over the last couple of decades: outcome reasoning.

To help understand outcome reasoning we first discuss warranted reasoning. Warranted reasoning is statisticians' foundational type of reasoning, despite Breiman's assertions otherwise. In warranted reasoning, statisticians reason about the data that should come into existence in order to best answer the question, and carefully craft randomization schedules and sampling schemes that give rise to efficient data generating functions. These processes yield data sets that have well understood properties, and have true-by-construction inference – hence, providing a warrant for inference (Draper et al. (1993); Cook (1991)). Warranted reasoning requires direct interaction with the real-world. The challenges of reaching conclusions given the data are lessened at the price of designing the data generation and carefully implementing its collection. Warranted reasoning is a fundamental version of reasoning using data, and we see this style of reasoning at the foundation of modern statistics – in urtexts by R.A. Fisher and Jerzy Neyman. In Fisher's *Design of Experiments*, he describes randomization as the “reasoned basis for inference in experiments” (Fisher (1949)) and similar warrants are provided by sampling designs.

Breiman was mistaken in asserting warranted reasoning (in particular, sampling) as the foundation for the machine learning community. Instead, he was one of the first thinkers to contrast model reasoning and outcome reasoning. The rest of this commentary will continue with the project Breiman set out on, by comparing and contrasting these two forms

of reasoning. With the benefit of an additional two decades, we arrive at an alternative conclusion from Breiman. We are not so excited about there being “two cultures” as we are about there being a new form of reasoning using data.

Let’s say a little bit more about what we mean when we say “reasoning using data” because we’ll invoke that idea a lot. In practice, incorporating data into one’s reasoning boils down to examining how real-world data disagree with what we anticipated. For example, in hypothesis testing we build a null distribution based on beliefs about the data generating function then, using real data, we compute a test statistic and compare it to that null distribution. After that comparison, the next important step in reasoning using data is determining how we adapt our thinking when we see divergence between the data and what our beliefs told us was likely to occur. In classic hypothesis testing, a researcher might see that the observed test statistic appears quite unlikely under the null hypothesis and reject the beliefs that gave rise to the null hypothesis. That is, the researcher considered a hypothesis about the world, and reasoned about what that hypothesis would imply about data generated from a known process. Real data were then generated using such a process and were examined.

In RCTs and sampling studies, our beliefs about the data generating functions are true by construction, because we intervened in the real world. But in a modeling approach to reasoning about data, where we must posit how covariates enter a model as well as the statistical behavior of error terms, our beliefs are contingent and – at best – approximations to the data generating function(s) that gave rise to the data being used for inference. This kind of reasoning using data – what Breiman called the modelling approach and identified as being the dominant form of reasoning inside academic statistics – can be made more or less true by careful design (a.k.a., “data pipelines,” “nonparametric preprocessing”). But, while models have shaky connections to the real-world, they have expansive theory developed to reason about their behaviors and reach conclusions in settings where real-world interventions are infeasible.

Interestingly, while the first two forms of reasoning focused on inference (e.g., using beliefs about the data generating function to describe future variability of test statistics which guides statistical inference about the parameters of the data generating function) the new form of data reasoning bypasses discussions of parameters and instead focuses on using data to assess the quality of predictions. In the next section we discuss this new, third form of reasoning using data: outcome reasoning.

3. Outcome reasoning and the Common Task Framework

In this section we describe outcome reasoning in a couple different ways. First, we examine the dominant vehicle for outcome reasoning: the Common Task Framework (CTF). We then leverage characteristics of the CTF to discuss properties of outcome reasoning and contrast it with both model reasoning and warranted reasoning.

Many readers are likely familiar with the CTF even if the name is unfamiliar; the Netflix Prize (Bennett et al. (2007)) and Kaggle competitions are excellent examples of this framework. Following Donoho (2017), we identify the key features of a problem that exists in the CTF:

- curated data that have been placed in a repository;

- static data (all analysts have access to the same data);
- a well defined task (e.g., predict y given a vector of inputs x for previously unobserved units of observation);
- consensus on the evaluation metric (e.g., the mean squared error of the predictions from the algorithm on a set of observations); and
- an evaluation data set with observations which is not accessible to the analysts.

Today, the CTF exists in many forms and is used in many ways— from competitions run by large companies to the research programs of machine learners in academia. In practice, some of the features of the CTF are relaxed. In particular, outside of competitions, the last feature is self-policed by the analyst who has direct access to (and in most cases is the creator of) the evaluation data set. When performed correctly, the CTF gives us a way to justify a claim that “Algorithm A performs better than Algorithm B on these data sets.” The CTF provides a fast, low-barriers-to-entry means for researchers to settle debates about the relative utility of competing algorithms. This is in contrast to the traditional use of mathematical descriptions of the behavior of an algorithm, or simulations of the algorithm’s ability to recover parameters of a data generating function. In model reasoning, we debate about algorithms and their behavior given data by using careful mathematics or sprawling simulations.

In the CTF, fast computation takes the place of proving theorems, and performance is assessed using held-out data. The consequences of a poorly performing prediction algorithm in the CTF are minimal – e.g., after a failure the analyst tweaks the algorithm and tries again.

Because there is a specific performance metric, there is little ambiguity in the relative ordering of the algorithms conditional on a particular dataset. The ordering gives rise to a ranking of the algorithms, and the public display of these rankings are called “leaderboards.” Leaderboards are cited as using competition to motivate analysts to reach high levels of performance (Boutros et al. (2014); Costello and Stolovitzky (2013); Meyer et al. (2011)).

In contrast to outcome reasoning, model reasoning tends to involve discussions of unobservable parameters and the algorithm’s ability to faithfully recover the values of these parameters. Model reasoning emphasizes thinking carefully about how variation in the co-variates should be linked to variation in the outcome. In stark contrast, outcome reasoning allows the analyst to essentially ignore the input space when justifying their algorithm. Instead, outcome reasoning operates purely in the space of the outcome. In outcome reasoning the debates center on if the task is useful, or if the metrics of assessment are informative, and if the data used to form the performance metrics were adequate. This is where you can see how Breiman made his mistake – he saw that one of the most rigorous ways to select the held out data used to evaluate the performance metric is to use sampling. But sampling is not a necessary condition in most ML settings. ML has used outcome reasoning for decades to make progress, with hardly any rigorous use of sampling from a target population.

The benefit of developing an algorithm using the CTF is that the process is faster than model reasoning and allows for the assessment of extraordinarily complex functions. There are several drawbacks, most of which can be framed as challenges due to extrapolation. The brittleness of these algorithms – surprising failures on seemingly innocuous data – has led

to concerns that we do not understand how these algorithms work and has started to shake confidence. It might feel like sampling is an easy way out of this challenge; give new data to the algorithm and retrain. But there’s a fundamental challenge here: with a black box algorithm it is not clear how the algorithm is using the covariates to form predictions (e.g., two points that are close in euclidean space may produce radically different predictions) so efficient sampling is hard to work out. Defining a target population can be challenging in the ML settings, and so establishing even a simple sampling scheme might not be possible.

It’s easy to get a couple different statistical techniques tangled up when discussing outcome reasoning. The most confusion here is about how k-fold cross validation works in this context. K-fold is a useful statistical tool that is often used in outcome reasoning particularly when K-fold is used to generate estimates of data-set specific performance metrics of the algorithm. Because k-fold uses sampling perhaps it’s a form warranted reasoning? Used in the way described here it isn’t – there is no target estimand from a super population, rather the performance metric is purely in the outcome space and used to rank the algorithm relative to others on a particular data set. In the CTF it may not be obvious how limited k-fold cross validation is relative to the real-world guarantees offered by sampling, but that will become clearer in the next section as we go beyond the purview of the CTF (i.e., development and innovation of algorithms) to using these algorithms in the real-world (i.e., using outcome reasoning to assess an algorithm during its in deployment).

4. Using outcome reasoning in the real-world: MARA(s)

As discussed in Rodu and Baiocchi (2020), it appears the CTF was created to guide breakthroughs in the development of algorithms used in complex settings (e.g., machine translation). That is, the CTF was meant for algorithmic development, not for reasoning about how algorithms will-perform/are-performing in deployment (e.g., producing predictions for *de novo* data). The CTF has serious failures in its logic when it is used to justify an algorithm for deployment.

To address the limitations of the CTF, and to help identify cases in which outcome reasoning is appropriate in deployment, we reworked the assumptions of the CTF and proposed an extension that we refer to using the mnemonic “MARA.” In MARA we classify prediction problems using four features (“problem-features”):

- **[measurement]** ability to measure a function of individual predictions and actual outcomes on future data,
- **[adaptability]** ability to adapt the algorithm on a useful timescale,
- **[resilience]** tolerance for accumulated error in predictions, and
- **[agnosis]** tolerance for potential incompatibility with stakeholder beliefs.

In order to determine if it’s appropriate to use outcome reasoning for a given problem, we classify a problem as either “satisfying” or “not satisfying” each problem-feature individually. If a problem satisfies the MARA problem-features then the problem is suitable for outcome-reasoning – that is, the powerful and easy to implement form of reasoning that is the foundation for the CTF. If the problem fails to satisfy even one of the features then the

problem requires a more complex form of reasoning to justify the algorithm’s deployment – e.g., model-reasoning.

You will see us using $MARA(s)$ a lot, instead of just $MARA$. This is to emphasize that there are real people who need to think about and assess the features of the problem being considered, as well as real people who are impacted by their decisions. The “s” refers to stakeholders, and we reference function notation to help make it clear that changing stakeholders may lead to different assessments of the $MARA$ criteria – that is, if $s_1 \neq s_2$ then it is likely that $MARA(s_1) \neq MARA(s_2)$.

This brief section invokes ideas introduced and developed in Rodu and Baiocchi (2020). In that full paper, we develop the $MARA(s)$ framework in detail, work through several real-world examples of outcome reasoning, and trace some of the history of the CTF’s development. One of our goals in Rodu and Baiocchi (2020) was to provide a paper to help non-technical collaborators get an understanding of these ideas as well. We wanted a paper we could give to collaborators to help them parse the data communities and to help our collaborators craft their own research questions better. In the next section we work through an example using $MARA(s)$ to assess a prediction problem’s suitability for outcome reasoning.

5. Teaching and talking about outcome reasoning

The example here is based on a real example we encountered being deployed, but the details are changed quite a bit to allow anonymity. (We also recognize this example reeks of oversimplified scientific management and Taylorism.)

Suppose a large corporation has identified hiring employees as an absolutely critical area. As part of this recognition, they decide to implement an algorithm that takes in several streams of information about candidates (e.g., details from their CV, interview scores, NLP’ed letters of recommendation) and then returns scores meant to predict the candidates’ prospects for “successful” employment.

This is a prediction problem, and one way to assess the $MARA$ features plays out like: Though probably quite hard, we can imagine **measurements** that could be used to assess the success of an employee such as number of widgets produced, sales number, billable hours, managers’ assessments, number of patents or publications, or some summarizing function of several important metrics. Given this is a large company, it is in a stable period in its growth (unlike say a startup which may pivot its business model rapidly) so it may seem reasonable that there is high potential for the algorithm to be **adapted** on a useful time scale (e.g., the corporation might change what kinds of employees it defines as successful – due to changes in the business cycle, change in market needs, or change in current workforce skills – but not so rapidly that the algorithm cannot be adapted). And, though a mis-hire may be significant, the corporation as a whole is likely rather **resilient** to a mis-hire because there are many other employees involved in carrying out the corporation’s mission and, if observed performance is below a given threshold, a manager may be able limit the downside by removing the employee. From the corporation’s viewpoint, does it really matter why an employee was hired if they’re a great employee? Probably not. We may have incorrect assumptions, biases, about what kinds of people make good employees so it may even feel like having a data-driven algorithm making decisions about employment is to our benefit.

Being **agnostic** about why an employee was hired may be reasonable given that we believe we’re going to get more successful employees. Under this set of assessments, it may feel like this prediction problem is ripe for using an algorithm that can be assessed using outcome reasoning. That is, we deploy the algorithm and monitor how it is doing selecting new employees. But this is just one way to assess the MARA features of a prediction problem, and this assessment was largely (implicitly) derived from management’s perspective on hiring.

If we look at this same prediction problem but factor in new stakeholders we get different results than we did in the previous paragraph. If the set of stakeholders is expanded to include current employees as well as future employees you might imagine critiques of both **measurement** (e.g., “I do so much for this company that isn’t related to the number of widgets I produce!!” as well as “You didn’t hire me because I scored below some cutoff point, how do you know I would be less successful if you have no ability to see me – or people like me – perform?”) and **resilience** (e.g., “This algorithm rejected my application which is a huge loss to me and to the dynamics of your team that I would have been part of.”). Further, imagine if a Department of Labor and unions are at the table, considerations about why an algorithm is making a determination will be centrally important. Being **agnostic** about how the algorithm reached its conclusions (even if it seems to be improving the metrics of “success”) is not an option here: if a candidate was rejected using an algorithm that can be shown to be unethical (e.g., using information about a protected class) then there will likely be serious consequences. Outcome reasoning is not a great option under this assessment. If we’re going to move forward on this prediction problem we’ll need a different way of figuring out if we trust the algorithm for deployment. There are many other criticisms that can be lodged here, but let’s stop at this point because the dynamic is hopefully clear: changing stakeholders changes the kinds of data-reasoning that are appropriate.

Two last points to make in this section: first, the two paragraphs above demonstrate that MARA directs attention to the assumptions that support outcome reasoning, but they do not dictate a particular result for a given prediction problem. Stakeholders are needed to make the assessment of the MARA features. This shouldn’t be a surprise because we see this in experimental design: given the same scientific hypothesis there are usually a handful of ways to design a useful experiment and the final decision about which study design to deploy will depend on assessments of the tradeoffs – and those assessments are made by real people, in anticipation of challenges and with particular goals and beliefs in mind. The second point to make here is that MARA can be used to redesign a prediction problem. For example, in the problem above, one of the criticisms from a potential employee was “My score was below a threshold and so you never hire people like me, how do you know we wouldn’t be successful?” That’s a great critique because it points out a pretty bad violation of the measurement feature. This violation means that there’s no new, de novo data coming in on the “hired” side of things in those regions of the covariate space so we cannot use outcome reasoning to monitor the algorithm’s performance with those kinds of employees (we might be getting “false negatives” for example). Hearing this critique, as well as the argument from management that they are rather resilient against having some potentially less-successful hires, the data scientist might suggest removing the hard threshold and instituting some amount of randomization (e.g., if the algorithm scores a candidate as less than some threshold, call it C , then we hire this person with a probability

of p where p is a rather small probability determined by corporation’s resilience to a mis-hire). This change in the way the prediction problem will work makes a tradeoff between two of the problem features: it reduces a serious violation of the measurement feature by adding in some measures in all parts of the covariate space, at the price of increasing the “hit” to the resilience feature.

6. Discussion: this was good and what comes next will be better

We’ve spent most of this paper “defusing” the feeling that there are two, irreconcilable cultures in data science and providing language to identify and discuss these differences. We point to the origins of these differences as being divergent ways of using data to correct, or guide thinking about the real world. Looking into the future, we anticipate a merging of these “cultures” – because these ways of reasoning using data are just intellectual tools useful in some settings and less useful in others. A non-doctrinaire analyst should be able to identify which forms of reasoning are best for a problem. A sophisticated analyst should be able to work with stakeholders to (re-)design a problem so it is better aligned with a style of reasoning, leveraging that form of reasoning’s strengths and minimizing its weaknesses.

The above is about the future, and we want to take a moment to discuss the past. Though we would reject a “two, *irreconcilable* cultures” framing of the difference between machine learners and statisticians, we do think Breiman (and others) did an extraordinarily important service to data science as a whole by stressing the difference between these two communities, and locating a place for outcome reasoning to grow free from criticisms from model reasoners and experimentalists/interventionalists. When Breiman was writing, model reasoning and warranted reasoning had been undergoing development and advancement for several decades. It is quite possible that there needed to be a speciation in our communities at that point, to allow the necessary intellectual freedom for outcome reasoning to grow, establish extraordinary wins, and to experience failures. When we were graduate students in statistics, it was common for statisticians to dismiss the “sloppy” work done by MLers. That assessment was wrong, and came from a blindness to how the MLers were making progress. We have no proof, but we suspect that those kinds of criticisms would have stifled MLers’ explosive creativity if they had been forced to only operate within traditional groups of statisticians and statistics departments, those of us who had been trained to reason using modeling or warranted reasoning. It is very hard to believe that traditional statistical modes of thinking would have achieved the level of breakthroughs in machine translation and deep learning that the MLers have generated.

The resulting diversity of thought that came from our two cultures’ developments have led to an enriched community of data scientists, capable of tackling a broader range of problems and accomplishing truly extraordinary feats of engineering and science.

References

- James Bennett, Stan Lanning, and Others. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007.
- Paul C Boutros, Adam A Margolin, Joshua M Stuart, Andrea Califano, and Gustavo Stolovitzky. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome biology*, 15(9):1–10, 2014.
- Thomas D Cook. *Clarifying the warrant for generalized causal inferences in quasi-experimentation*. 1991.
- JC Costello and G Stolovitzky. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clinical Pharmacology & Therapeutics*, 93(5):396–398, 2013.
- David Donoho. 50 years of data science. *J. Comput. Graph. Stat.*, 26(4):745–766, October 2017.
- David Draper, James S Hodges, Colin L Mallows, and Daryl Pregibon. Exchangeability and data analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 156(1):9–28, 1993.
- Ronald A Fisher. The design of experiments. 1949.
- Pablo Meyer, Leonidas G Alexopoulos, Thomas Bonk, Andrea Califano, Carolyn R Cho, Alberto De La Fuente, David De Graaf, Alexander J Hartemink, Julia Hoeng, Nikolai V Ivanov, et al. Verification of systems biology research in the age of collaborative competition. *Nature biotechnology*, 29(9):811–815, 2011.
- Jordan Rodu and Michael Baiocchi. When black box algorithms are (not) appropriate: a principled prediction-problem ontology. *arXiv e-prints*, pages arXiv–2001, 2020.