The Golden Age of Empiricism

Tim Savage

March 7, 2017

All opinions expressed here are my own and do not represent the views of CBRE.

My Background

- I started in the pre-digital age with administrative data.

- NLSY79: ~3,000 young men tracked annually for about 20 year.

- Complicated algorithm to evaluate potentially adverse impacts of early labor market “shocks.”

- The data were cheap, but the computation was very expensive.

My Languages

- English
- Basic
- Latin
- Italian
- Fortran
- SAS/Stata
- R
- Python

The Changed Landscape

- If it can be digitized, it can be analyzed. And now everything is digitized.

- Most data now arise as the digital exhaust of human activity. It is both non-experimental and non-IID.

- Statisticians and economists have lost the language wars. It is now data science and machine learning. Nobody cares about whether you divide by n or (n-1) when you have a trillion observations. This is as close to asymptotics as mortals will get.

- But machine learning is just stochastic modeling. We started it, but were very poor marketers. Computer scientists are currently in charge, but they rarely think about DGPs.

The Same Paradigm

- Every algorithm has three components: representation; evaluation; and optimization.

- Algorithms and hypothesis testing are conceptually different animals. An algorithm is an application. A hypothesis is a conjecture.

- We can have one without the other. The point of this effort is the logically-consistent application of the first to the second.

The Revolution Within

- Fisher/Neyman/Pearson(s) were also the product of the pre-computer age: closed-form solutions and asymptotic arguments. It was the obvious extension of mathematics.

- But The Reverend Bayes was conceptually (if incoherently) correct. Laplace corrected the formulation, and we have now Bayes Rule of inverse probability. What is the likelihood of an outcome conditional on what I've seen and subject to my transparent (albeit subjective) beliefs.

- We now are able to massively scale MCMC simulation. For prediction and forecasting, we now simulate posterior marginal distributions to make probabilistic statements about an inherently unknowable outcome.

The Myth of AI

- Humans writing code is not artificial intelligence. (Either that or statisticians and economists in the 1950's invented AI.)

- We still live in the old paradigm: target data; computer code; desired output.

- This unlabeled image from the internet contains a four-legged hairy animal. 84.56% of all similar images on the internet are labeled as cats. Therefore, this unlabeled image should be labeled as a cat. Entirely algorithmic.

- Deep learning (neural networks on steroids) are the most supervised learners in existence.

- True AI is an unsupervised learner.

Our Common Future

- We live in a golden age of empiricism, but the world is upside down (for me, at least).
- Useful data are very expensive, but computation is cheap.
- We need to think deeply about empirical causation in a world of digital exhaust.
- It can be done: https://github.com/thsavage/Causation/blob/master/Poster.pdf
- We had this conversation before and must again: Judea Pearl.
- The statisticians have begun to strike back: Efron and Hastie.

Thank You