Primer on Technical Malpractice in Machine Learning

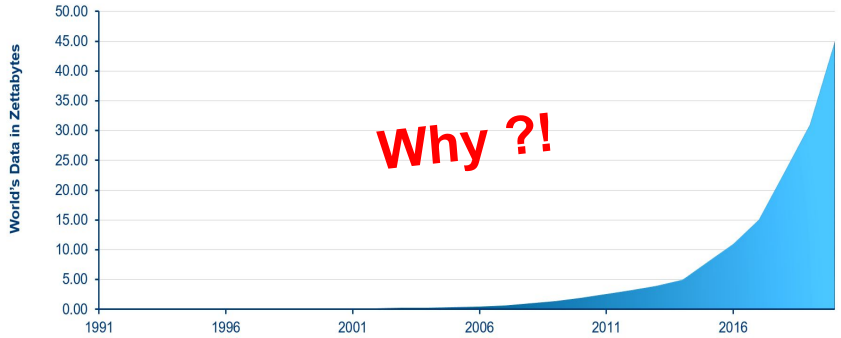
Patrick Hall
George Washington University

All data scientists are bad. Some are useful.

-- Nihlist Data Scientist

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

We've Been Lied to About Data



- Data is often inaccurate, incomplete, obsolete, or just plain wrong.
- Data often encodes human biases, sociological and otherwise.
- Machine learning (ML) algorithms learn only from data. Hence, ML models are often inaccurate and biased.
- We can only make good data-driven decisions if we are very careful.

The Scientific Method:

1. Develop a credible hunch
   (e.g., based on prior experiments or literature review)
2. Record a hypothesis
3. Collect data
   (e.g., using design of experiment)
4. Test hypothesis (e.g., using a double blind random construct)
5. Reproduce

The Data-scientific Method:

1. Assume we'll make millions of dollars
2. Install GPU, download Python
3. Collect wrong, biased data from the internet or the exhaust of some business process
4. Surrender to confirmation bias:
   - 4.1. Study collected data to form a hypothesis, i.e., which **X**, **y** and ML algorithm to use
   - 4.2. Use the same data from hypothesis generation to test our hypothesis
   - 4.3. Test our hypothesis with a high-capacity learning algorithm that can fit almost any set of loosely correlated **X** and **y** well
   - 4.4. Change our hypothesis until our results are "good"
5. Don't worry about reproducing, we're all good bruh

The Unique Properties of Data

Data itself is a quantity with unique properties:

- **Data Generating Processes:** most data is sampled from some known distribution
- **The Law of Large Numbers:** results based on small samples exhibit high variance
- **Randomness:** many real-world events are random, unconnected, and unpredictable
- **Regression to the Mean:** most quantities are dominated by some expected value (mean)

These properties often drive results in data-driven experiments, decisions and products, especially when the data scientific method is employed.

Smart people have been making fun of bad data science before most of us were even born.

Psychological Bulletin
1971, Vol. 76, No. 2, 105–110

BELIEF IN THE LAW OF SMALL NUMBERS

AMOS TVERSKY AND DANIEL KAHNEMAN[1]
Hebrew University of Jerusalem

People have erroneous intuitions about the laws of chance. In particular, they regard a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. The prevalence of the belief and its unfortunate consequences for psychological research are illustrated by the responses of professional psychologists to a questionnaire concerning research decisions.

"Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?"

If you feel that the probability is somewhere around .85, you may be pleased to know that you belong to a majority group. Indeed, that was the median answer of two small groups who were kind enough to respond to a questionnaire distributed at meetings of the Mathematical Psychology Group and of the American Psychological Association.

On the other hand, if you feel that the probability is around .48, you belong to a minority. Only 9 of our 84 respondents gave answers between .40 and .60. However, .48 happens to be a much more reasonable estimate than .85.[2]

Apparently, most psychologists have an exaggerated belief in the likelihood of successfully replicating an obtained finding. The sources of such beliefs, and their consequences for the conduct of scientific inquiry, are what this paper is about. Our thesis is that people have strong intuitions about random sampling; that these intuitions are wrong in fundamental respects; that these intuitions are shared by naive subjects and by trained scientists; and that they are applied with unfortunate consequences in the course of scientific inquiry.

We submit that people view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. Consequently, they expect any two samples drawn from a particular population to be more similar to one another and to the population than sampling theory predicts, at least for small samples.

The tendency to regard a sample as a representation is manifest in a wide variety of situations. When subjects are instructed to generate a random sequence of hypothetical tosses of a fair coin, for example, they produce sequences where the proportion of

[1] The ordering of authors is random. We wish to thank the many friends and colleagues who commented on an earlier version, and in particular we are indebted to Maya Bar-Hillel, Jack Block, Jacob Cohen, Louis L. Guttman, John W. Tukey, Ester Samuel, and Gideon Shwarz.

Requests for reprints should be sent to Amos Tversky, Center for Advanced Study in the Behavioral Sciences, 202 Junipero Serra Boulevard, Stanford, California 94305.

[2] The required estimate can be interpreted in several ways. One possible approach is to follow common research practice, where a value obtained in one study is taken to define a plausible alternative to the null hypothesis. The probability requested in the question can then be interpreted as the power of the second test (i.e., the probability of obtaining a significant result in the second sample) against the

with known variance, one would compute the power of the test against the hypothesis that the population mean equals the mean of the first sample. Since the size of the second sample is half that of the first, the computed probability of obtaining $z \geq 1.645$ is only .473. A theoretically more justifiable approach is to interpret the requested probability within a Bayesian framework and compute it relative to some appropriately selected prior distribution. Assuming a uniform prior, the desired posterior probability is .478.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Computers Do Not Understand.

- Computers don’t understand their tasks.
- They don’t understand the place of their tasks in the world.
- Nor that there is a world.
- Is this intelligence?
- We must account for the context in which an ML system operates, because computers cannot.

Map apps like Waze have been accused of directing California drivers into wildfires, due to decreased traffic during emergency conditions.

Source: https://www.usatoday.com/story/tech/news/2017/12/07/california-fires-navigation-apps-like-waze-sent-commuters-into-flames-drivers/930904001/.

Commentary based on: https://www.lrb.co.uk/the-paper/v43/n02/paul-taylor/insanely-complicated-hopelessly-inadequate.

How to Be Better

- Choose a specific, targeted application for ML tasks.
- Know where your training data comes from and thoroughly verify its quality.
- Select a reasonable number of uncorrelated, well-understood inputs with clear relationships to the modeling target.
- Avoid black-box models, unless absolutely necessary.
- Match models to the structure of their training data and understand how models work.
- Test real-world model performance -- good test data scores are not enough!
- Ensure model performance is monitored once deployed.