Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Questions
- HW: Peer review of data cleaning

Asking and answering questions

Whither big data?

- http://www.slate.com/articles/technology/technology/2017/10/what_happened_to_big_data.html

Whither big data?

- Five years ago—in February 2012—an article in the New York Times' Sunday Review heralded the arrival of a new epoch in human affairs: "The Age of Big Data."

- Society was embarking on a revolution, the article informed us, one in which the collection and analysis of enormous quantities of data would transform almost every facet of life.

- No longer would data analysis be confined to spreadsheets and regressions:

Whither big data?

- Sophisticated data analysis software could help identify utterly unexpected correlations, such as a relationship between a loan recipient's use of all caps and his likelihood of defaulting. This would surely yield novel insights that would change how we think about, well, just about everything.

Whither big data?

- The haste to implement and apply big data, via what’s often called “data-driven decision-making,” resulted in grievous mistakes.

- Some were blatant:

  - There was the time Target sent coupons for baby items to the family of a teenage girl who hadn’t told anyone she was pregnant.

  - Or the time Pinterest congratulated single women on their impending marriages.

  - Or the Google Photos snafu, in which the company’s vaunted A.I. mistook black people for gorillas due to a lack of diversity in the data it was trained on. (It’s worth pointing out that, in this case at least, the “big data” wasn’t quite big enough.)

Whither big data?

- Others were more subtle, and perhaps more insidious. Among these are the types of opaque, data-powered institutional models O’Neil chronicles in her important book:

  - the ones that (arguably) encoded racial bias in recidivism models used by courts to sentence criminals, or

  - fired beloved schoolteachers based on questionable test-score data.

  - the Facebook algorithms that evidently helped Russian agents sow division in the American electorate via well-targeted, hyperpartisan fake news.

Whither big data?

- By its nature, big data is hard to interpret. When you’re collecting billions of data points—clicks or cursor positions on a website; turns of a turnstile in a large public space; hourly wind speed observations from around the world; tweets—the provenance of any given data point is obscured.

- This in turn means that seemingly high-level trends might turn out to be artifacts of problems in the data or methodology at the most granular level possible.

- But perhaps the bigger problem is that the data you have are usually only a proxy for what you really want to know. Big data doesn’t solve that problem—it magnifies it.

Whither big data?

- By its nature, big data is hard to interpret. When you’re collecting billions of data points—clicks or cursor positions on a website; turns of a turnstile in a large public space; hourly wind speed observations from around the world; tweets—the provenance of any given data point is obscured.

- This in turn means that seemingly high-level trends might turn out to be artifacts of problems in the data or methodology at the most granular level possible.

- But perhaps the bigger problem is that the data you have are usually only a proxy for what you really want to know. Big data doesn’t solve that problem—it magnifies it.

Whither big data?

- For instance, public opinion polling is widely used as a proxy for how people will vote in an election.

- But as surprise elections throughout the decades have reminded us—from Tom Bradley’s 1982 loss in the California gubernatorial race on through to Brexit and Trump—there is not always a perfect correspondence between the two.

- Facebook used to measure users’ interest in a given post mainly by whether they hit the “like” button on it. But as the algorithmically optimized news feed began to be overrun by clickbait, like-bait, and endless baby photos—causing user satisfaction to plunge—the company’s higher-ups gradually realized that “liking” something is not quite the same as actually liking it.

Whither big data?

- The wider the gap between the proxy and the thing you’re actually trying to measure, the more dangerous it is to place too much weight on it.

- Take the aforementioned example from early in O’Neil’s book: school districts’ use of mathematical models that tie teacher evaluations to student test scores. Student test scores are a function of numerous important factors outside of a teacher’s control.

Whither big data?

- Aside from swearing off data and reverting to anecdote and intuition, there are at least two viable ways to deal with the problems that arise from the imperfect relationship between a data set and the real-world outcome you're trying to measure or predict

Better data

- One is, in short: moar data.

- This has long been Facebook’s approach. When it became apparent that users’ “likes” were a flawed proxy for what they actually wanted to see more of in their feeds, the company responded by adding more and more proxies to its model.

- It began measuring other things, like the amount of time they spent looking at a post in their feed, the amount of time they spent reading a story they had clicked on, and whether they hit “like” before or after they had read the piece.

- When Facebook’s engineers had gone as far as they could in weighting and optimizing those metrics, they found that users were still unsatisfied in important ways.

- So the company added yet more metrics to the sauce: It started running huge user-survey panels, added new reaction emojis by which users could convey more nuanced sentiments, and started using A.I. to detect clickbait-y language in posts by pages and publishers.

Better data

- One is, in short: moar data.
- One downside of the moar data approach is that it’s hard and expensive. Another is that the more variables are added to your model, the more complex, opaque, and unintelligible its methodology becomes

Better data

- One is, in short: moar data.

- One downside of the moar data approach is that it’s hard and expensive. Another is that the more variables are added to your model, the more complex, opaque, and unintelligible its methodology becomes

Better data

- Another possible response to the problems that arise from biases in big data sets is what some have taken to calling “small data.” Small data refers to data sets that are simple enough to be analyzed and interpreted directly by humans

Better data

- …Big data’s power to make systems more efficient, but its potential to obscure the causes and mechanisms of specific problems that aren’t being effectively measured in the aggregate.

- A safeguard, when making decisions based on things you know how to measure, is to make sure there are also mechanisms by which you can be made aware of the things you don’t know how to measure.

- “The question is always, what data don’t you collect?” O’Neil said in a phone interview. “What’s the data you don’t see?”

Better data

- Yet the threats posed by the misuse of big data haven’t gone away just because we no longer speak that particular term in reverent tones.

- Glance at the very peak of Gartner’s 2017 hype cycle and you’ll find the terms machine learning and deep learning, alongside related terms such as autonomous vehicles and virtual assistants that represent real-world applications of these computing techniques.

- These are new layers of scaffolding built on the same foundation as big data, and they all rely on it. They’re already leading to real breakthroughs—but we can rest assured that they’re also leading to huge mistakes.

What is data science?

Why use the term rather than “statistics” or “data mining” or “analytics”?

Asking and answering questions

- Some suggest that there are X types of questions data science can answer.
    - “It might surprise you, but there are only five questions that data science answers:
        - Is this A or B?
        - Is this weird?
        - How much – or – How many?
        - How is this organized?
        - What should I do next?”
- These people are wrong. More importantly, they are answering the wrong question.

Asking and answering questions

- Is this A or B? Is this weird? How much – or – How many? How is this organized? What should I do next?”

- These people are wrong. More importantly, they are answering the wrong question.

- This assumes that the only thing a data scientist can do is apply algorithms. (A very impoverished set, as well.)

- In reality, good data science should assist organizations in collecting data that answer the questions that they need to answer.

Types of questions

- Causal questions:
  - Why did this happen?
  - What causes what?
- Correlational questions:
  - What happens together?
- Predictive questions:
  - These are typically correlational questions (with some rare exceptions for temporal data)

- **HW: Pick one data set, write notebook that downloads and cleans the data (for general purpose analyzing)**
- Netflix data
    - https://www.kaggle.com/netflix-inc/netflix-prize-data/data
- Yahoo finance
    - https://pypi.python.org/pypi/yahoo-finance
- IMF data
    - https://brianc ne-reading-imf-data-data-retrieval-with-python/

Peer review!

- NYC open data
    - https://opendata.cityofnewyork.us/data/#datasetscategory
    - Examples:
        - http://blog.nycdatascience.com/student-works/r-shiny/noise-coming-case-study-nycs-311-noise-complaints/
        - http://blog.nycdatascience.com/student-works/new-york-city/