Exploring your data

Fraida Fund

Contents

Garbage in, garbage out

If you remember nothing else from this semester, remember this!

If you use “garbage” to train a machine learning model, you will only ever get “garbage” out. Even worse, since you are testing on the same data, you might not even realize it is “garbage” until the model is in production!

Recall: ML as a “leaky pipeline”

Example: a data problem (1)

Data analysis: use PubMed, and identify the year of first publication for the 100,000 most cited authors.

What are our expectations about what this should look like?

Example: a data problem (2)

Example: a data problem (3)

The explanation: in 2002, PubMed started using full first names in authors instead of just initials.

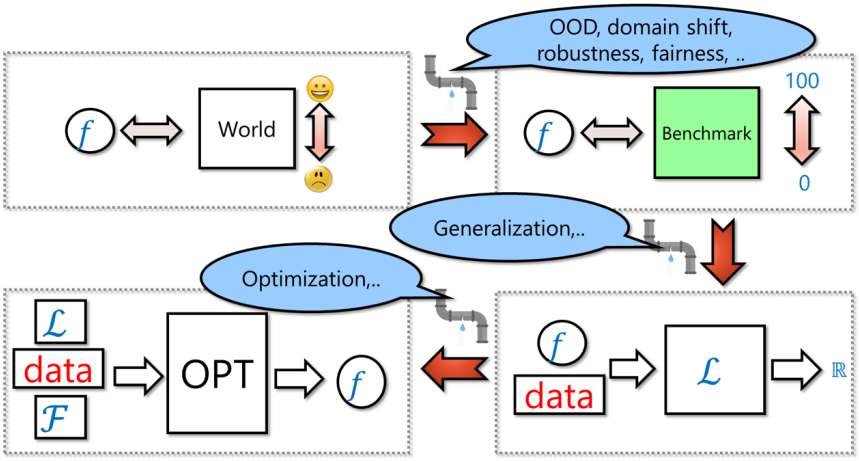


Figure 1: Source: Boaz Barak.

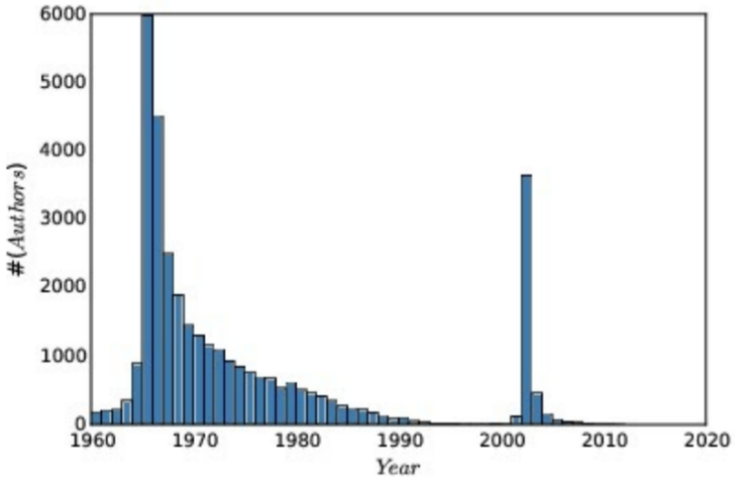


Figure 2: Does this look reasonable?

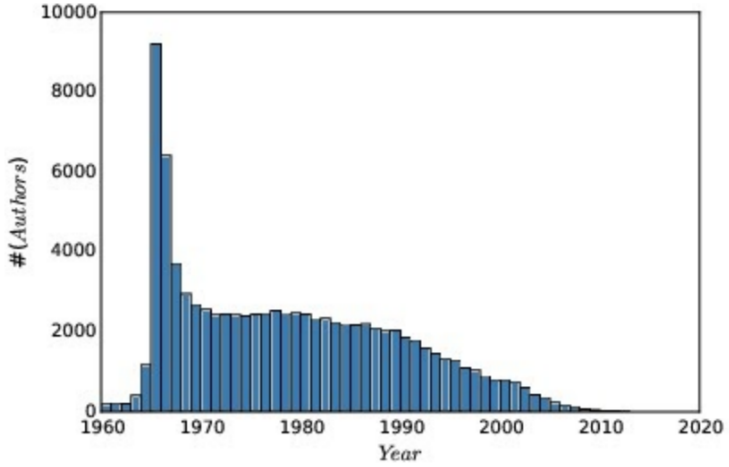


Figure 3: The real distribution. Example via Steven Skiena @ SBU.

```
▼ 187:
  ▼ vote_shares:
      trumpd:          0.566
      bidenj:          0.42
    votes:             2984468
    eevp:              42
    eevp_source:       "edison"
    timestamp:         "2020-11-04T04:07:43Z"
▼ 188:
  ▼ vote_shares:
      trumpd:          0.56
      bidenj:          0.426
    votes:             2984522
    eevp:              42
    eevp_source:       "edison"
    timestamp:         "2020-11-04T04:08:51Z"
```

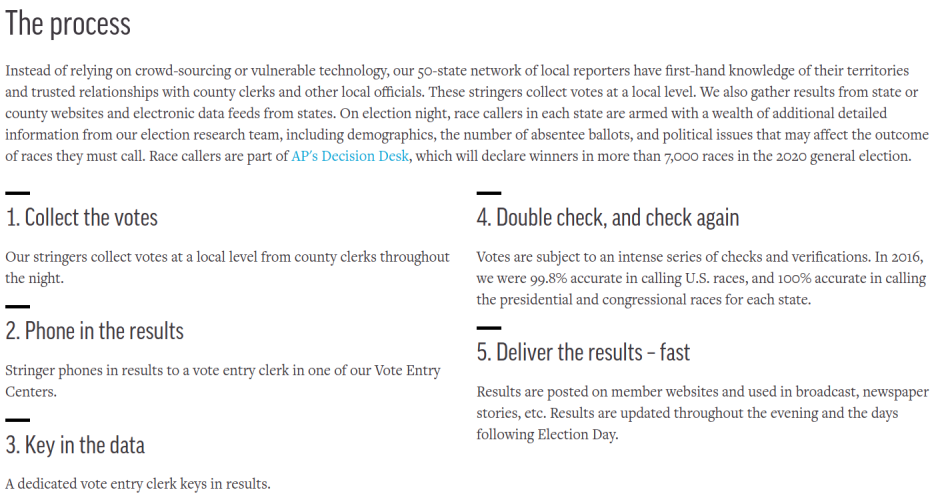

Figure 4: Data like this was widely (wrongly) used as evidence of anomaly in the 2020 U.S. Presidential election.

Example: another data problem (1)

What are our assumptions about this data, and how are they violated here?

What are possible explanations?

Example: another data problem (2)



The process

Instead of relying on crowd-sourcing or vulnerable technology, our 50-state network of local reporters have first-hand knowledge of their territories and trusted relationships with county clerks and other local officials. These stringers collect votes at a local level. We also gather results from state or county websites and electronic data feeds from states. On election night, race callers in each state are armed with a wealth of additional detailed information from our election research team, including demographics, the number of absentee ballots, and political issues that may affect the outcome of races they must call. Race callers are part of AP's Decision Desk, which will declare winners in more than 7,000 races in the 2020 general election.

1. Collect the votes

Our stringers collect votes at a local level from county clerks throughout the night.

2. Phone in the results

Stringer phones in results to a vote entry clerk in one of our Vote Entry Centers.

3. Key in the data

A dedicated vote entry clerk keys in results.

4. Double check, and check again

Votes are subject to an intense series of checks and verifications. In 2016, we were 99.8% accurate in calling U.S. races, and 100% accurate in calling the presidential and congressional races for each state.

5. Deliver the results – fast

Results are posted on member websites and used in broadcast, newspaper stories, etc. Results are updated throughout the evening and the days following Election Day.

Figure 5: Process by which data is collected by Edison and AP.

How Edison/AP collects the data for the data feed used by New York Times and other sites on Election Night:

- There are “reporters at county elections offices who call results” into their phone center
- They use “data feeds provided by some states and counties”
- They have people who “scour state and county websites for results” to enter into the system
- They have people who monitor “results sent from counties, cities, and towns via email or fax”
- They have people (“chasers”) who monitor other news sources for results not yet in the system.

Source: AP, Edison

What kinds of data problems?

What kind of problems might you encounter? (1)

- Rows where some fields are missing data
- Missing data encoded as zero
- Different units, time zones, etc. in different rows
- Same value represented several different ways (e.g. names, dates)
- Unreasonable values

How should you handle little bits of missing data? Depending on the circumstances, it may make sense to:

- omit the row
- fill with mean
- fill back/forward (ordered rows)
- train a model on the rest of the data to "predict" the missing value
- what **not** to do: fill with zero

How should you handle unreasonable values or outliers?

- e.g. suppose in a dataset of voter information, some have impossible year of birth - would make the voter a child, or 120 years old. (Voters with no DOB, who registered before DOB was required, are often encoded with a January 1990 DOB.)

What kind of problems might you encounter? (2)

- Rows that are completely missing
- Data is not sampled evenly
- Data or labels reflect human bias
- Data is not representative of your target situation

Examples:

- Twitter API terms of use don't allow researchers to share tweets directly, only message IDs (except for limited distribution, e.g. by email). To reproduce the dataset, you use the Twitter API to download messages using their IDs. But, posts that have been removed are not available - and posts are not equally likely to be removed! (For example: you might end up with a dataset that has offensive posts but few "obvious" offensive posts.)
- Many social media datasets used for "offensive post" classification are subject to human bias (especially if they were produced without adequate training procedures in place). For example, they may label posts containing African-American dialects of English as "offensive" much more often. Source, User-friendly article
- A dataset of Tweets following Hurricane Sandy makes it looks like Manhattan was the hub of the disaster, because of power blackouts and limited cell service in the most affected areas. Source
- The City of Boston released a smartphone app that uses accelerometer and GPS data to detect potholes and report them automatically. But, low income and older residents are less likely to have smartphones, so this dataset presents a skewed view of where potholes are. Source

What kind of problems might you encounter? (3)

- Data ethics fails
- Data leakage

Some data ethics fails:

- On the anonymity of the Facebook dataset
- 70,000 OkCupid Users Just Had Their Data Published; OkCupid Study Reveals the Perils of Big-Data Science; Ethics, scientific consent and OKCupid
- IBM didn't inform people when it used their Flickr photos for facial recognition training

Data leakage

In machine learning, we train models on a training set of data, then evaluate their performance on a set of data that was not used in training.

Sometimes, information from the training set can "leak" into the evaluation - this is called data leakage.

Some types of data leakage

- Learning from adjacent temporal data
- Learning from duplicate data
- Learning from features that are not available at prediction time
- Learning from a feature that is a proxy for target variable

Example:

- human activity recognition data
- email spam detection data
- credit card approval

COVID-19 chest radiography (1)

- **Problem**: diagnose COVID-19 from chest radiography images
- **Input**: image of chest X-ray (or other radiography)
- **Target variable**: COVID or no COVID

COVID-19 chest radiography (2)

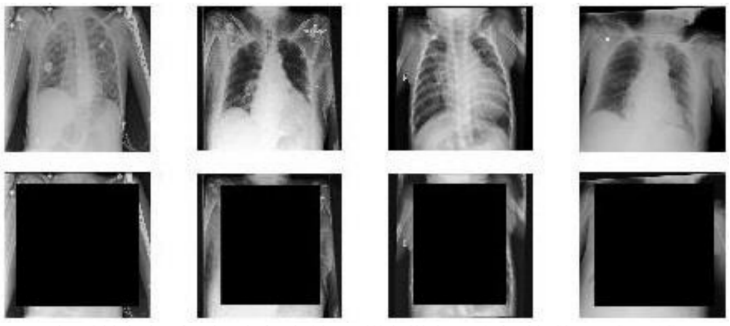


Fig. 5. Original and transformed samples from the 4 datasets, 300 sized black square (Left to right: COV, NIH, CHE, KAG)

Figure 6: Neural networks can classify the source dataset of these chest X-ray images, even *without lungs*! Source

In Spring 2020, many papers were published that claimed to use machine learning to diagnose COVID-19 patients based on chest X-rays or other radiography.

To train these models, people used an emerging COVID-19 chest X-ray dataset, along with one or more existing chest X-ray dataset, for example a pre-existing dataset used to try and classify viral vs. bacterial pneumonia.

The problem is that the chest X-rays for each dataset were so "distinctive" to that dataset, that a neural network could be trained with high accuracy to classify an image into its source dataset, even without the lungs showing!

COVID-19 chest radiography (2)

Findings:

- some non-COVID datasets were pediatric images, COVID images were adult
- there were dataset-level differences in patient positioning
- many COVID images came from screenshots of published papers, which often had text, arrows, or other annotations over the images. (Some non-COVID images did, too.)

COVID-19 chest radiography (3)

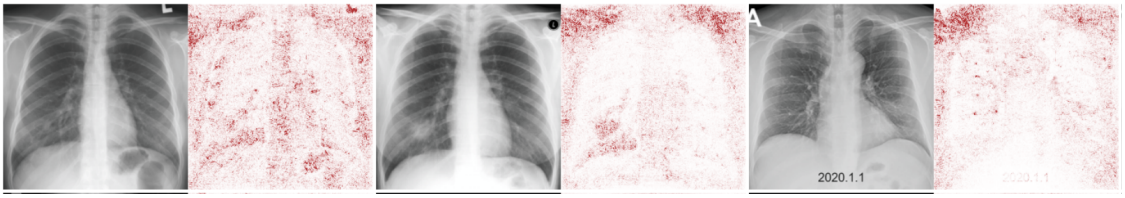


Figure 7: Saliency map showing the "important" pixels for classification. Source

These findings are based on techniques like

- saliency maps, where the model is made to highlight the part of the image (the pixels) that it considered most relevant to its decision.
- using generative models and asking it to take a COVID-negative X-ray and make it positive (or vice versa)

Many of the findings are not interpretable without domain knowledge (e.g. knowing what part of the X-ray *should* be important and what part should not be.) For example: should the diaphragm area be helpful?

Signs of potential data leakage (after training)

- Performance is "too good to be true"
- Unexpected behavior of model (e.g. learns from a feature that shouldn't help)

Detecting data leakage

- Exploratory data analysis
- Study the data before, during, and after you use it!
- Explainable ML methods
- Early testing in production