March 25th

Monday, March 25, 2019 2:48 PM

Data Quality

News:

- Facebook stored millions of passwords in plaintext. Poor security for crash-logs
- Data Science has become about leading false credibility to decisions we've already made
    - Data Science is less about good research and fining answers, we data-shop to find the answers we want

Data Validity:

- "Crap-in, Crap out"
- Only 3% of companies data meets basic quality standards
    - Friday Afternoon Measurement method: assemble 1015 critical data attributes
- The data quality benchmark report
    - 1200 companies
    - Sectors: finance, public sector retail manufacturing
    - Asked: "How much of your data do you think is correct"
        - 78% lack sophisticated data quality strategy
        - 88% have some sort of data quality solution in place
    - On avg., US organizations their data is 32% incorrect
    - #1 cause of error: human error/ lack of automation
    - These errors propagate out to other parts of organizational operation.
    -

Case Study: Google Flu Trends

- Flu affects 5-20% of population each year
- Flu Trends identified flu early to enable quick response
- Monitored millions of users'; health by tracking search queries
- Compared trends to historic baseline and CDC data to validate
- Impact: Was able to predict regional flu 10 days before the CDC knew
- Accuracy: it was about 90% accurate compared to CDC in beginning, but declined over time. It heavily over-estimated the flu
- Problems: Lack of domain knowledge
    - Not trained to flu-related symptoms are not flu
- Positive feedback loop: more people search for flu terms as the reported likelihood increases and it gets reported on the news
- Spurious correlations: NBA tournaments and flus both more prevalent in winter
- Now: Google Flu Trends has stopped, no longer publishing estimates. The still do, however, keep running that data and it can be used for research

Improper data can result in Bias

- Data can be correct but, because we live in a society with lots of segmentations, the data may not be representative
- Takeaway: make sure to collect data on right population
- If segmenting study to population groups, think carefully

HCI Study

- Running an HCI Experiment in Multiple Parallel Universes
- Tested what happens when you add a haptic intent on a slider. Tested with difficult and easy task
- A set of data was correct, but not good because it was not represented of the general

population, but was going to be applied to general population.

Group Discussion

- Data quality assurance as an industry: outsourcing this so that smaller companies are able to have better data
- Organizations are largely disconnected as far as data validation goes. Each company, each dept. within that company, and seemingly each employee have different strategy.
- Data managements cannot be a distinct department amount other ones, it needs to be "higher up" in order to provide oversight
- Many companies say that they "plan to invest" in data quality, but it seems dubious to me personally
- Its not flashy or attractive. It just needs to be institutionalized
- What is good data: Harvard Business review says it has to be 97% accurate

Other Groups

- Better survey methods - What is the definition of an "error" and is considering simple mistakes as "errors" may result in underestimated data accuracy
- What are the quality standards? What standards are companies being held to?
    - Data cleaned correctly
    - Collected from representative group?
    - Reported correctly?
- Where do we draw the line for an "error". If it says Zes instead of Yes then they obviously meant Yes so the data isn't lost. However if it says "HDIS" then we really have no idea what it could be and the data is destroyed
- Scale qualitative analysis of data using technical approach
- How do you find and error?
    - Errors will happen no matter what we do - its just a fact of life
    - Its what we do afterwards that matters. Aka prevention is harder and less practical than correction
- Centralized approach is useful b/c not everyone can be an expert in data validation. We can't really be trusted to each regulate our own data - it needs to come from above, a regulatory body