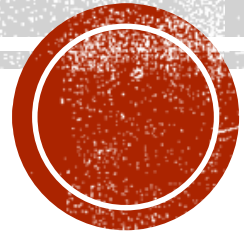
UNIT 2 LESSON 9

Data: Check your Assumptions



FILL OUT THE CLASS DATA TRACKER

- Take this quiz everyday: https://goo.gl/forms/wrVahnYfiGQxTXKX2

CHECK YOUR ASSUMPTIONS

- We must separate the what and why when looking at data
- Analyzing and interpreting data will typically require some assumptions to be made about the accuracy of the data and the cause of the relationships observed within it.
- When decisions are made based on a collection of data, they will often rest just as much on that set of assumptions about the data aa the date itself.
- Identifying and validating (or disproving) assumptions is therefore an important past of data analysis.
- Furthermore, clear communication about how data was interpreted should also include an account of the assumptions made along the way.

DATA: SEPARATE THE WHAT FROM THE WHY

- Video: Google Flu Trends Failure: https://www.youtube.com/watch?v=6l1lnS66Dpk
- What are the potential beneficial effects of using a tool like Google Flu Trends?
  - Incorrect assumptions about a dataset can lead to faulty conclusions

DATA: SEPARATE THE WHAT FROM THE WHY

- Then read one of the articles and be prepared to share:
- Article 1: https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/
- Article 2: http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/?_r=0
- Article 3: http://www.nature.com/news/when-google-got-flu-wrong-1.12413
- Article 4: http://time.com/23782/google-flu-trends-big-data-problems/
- Article 5: https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data

- Why did Google Flu Trends Eventually Fail?
- What assumptions were made that were NOT true?

GOOGLE FLU TRENDS FAIL KEY POINTS

- Google Flu Trends worked well in some instances but often overestimated, underestimated, or entirely missed flu outbreaks. A notable example occurred when Google Flu Trends largely missed the outbreak of the H1N1 flu virus.
- Just because someone is reading about the flu doesn't mean they actually have it.
- Some search terms like "high school basketball" might be good predictors of the flu one year but clearly shouldn't be used to measure whether someone has the flu.
- In general, many terms may have been good predictors of the flu for a while only because, like high school basketball, they are more searched in the winter when more people get the flu.
- Google began recommending searches to users, which skewed what terms people searched for. As a result, the tool was measuring Google generated suggested searches as well, which skewed results.

ACTIVITY GUIDE: THE DIGITAL DIVIDE

- Follow the activity guide and be prepare to share out (10-15 min)

- Complete the second part about Checking Your Assumptions

THE DIGITAL DIVIDE & CHECKING YOUR ASSUMPTIONS

- Key Points about the Digital Divide:
  - Access and use of the Internet differs by income, race, education, age, disability, and geography.
  - As a result, some groups are over or underrepresented when looking at activity online.
  - When we see behavior on the Internet, like search trends, we may be tempted to assume that access to the Internet is universal and so we are taking a representative sample of everyone.
  - In reality, a "digital divide" leads to some groups being over or underrepresented.
  - Some people may not be on the Internet at all.
- Key Assumptions about DATA:
  - The data collected is representative of the population at large (e.g., ignoring the "digital divide").
  - Activity online will lead to activity in the real world (e.g., people expressing interest in a candidate online means they will vote for him or her in real life).
  - Data is being collected in the manner intended (e.g., ratings are generated by actual customers, instead of business owners or robots).

SUMMARY & HOMEWORK

- HOMEWORK: Complete U2L9 Reflection PINK SHEET about data