Data Science Ethics

STA 198: Introduction to Health Data Science

Yue Jiang

July 22, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

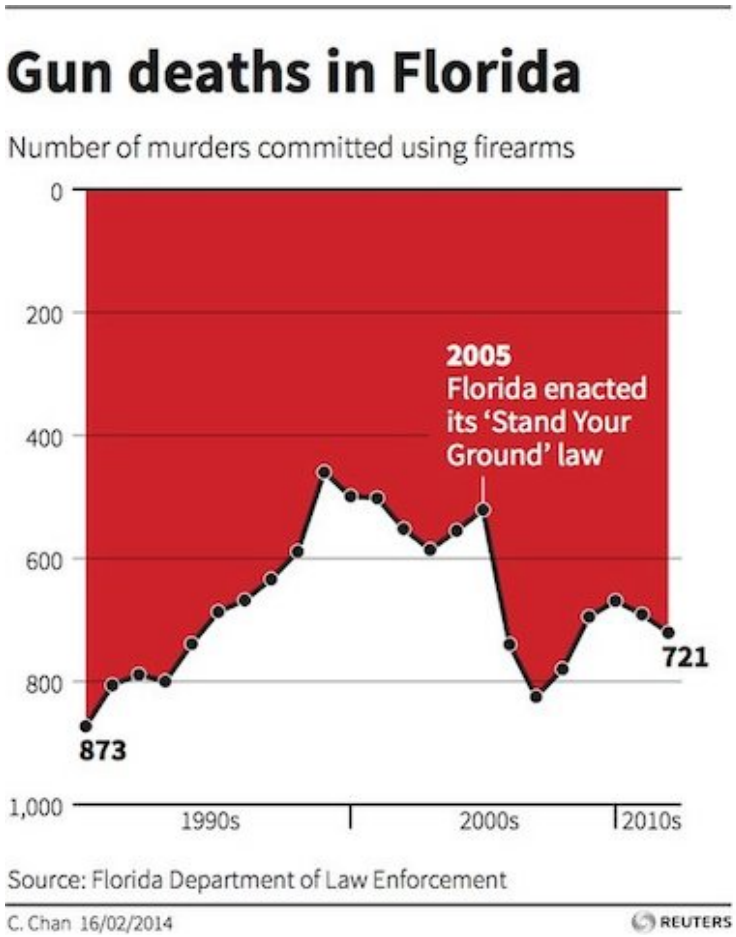
Acknowledgements

These slides are adapted from earlier lectures by Drs. Mine Cetinkaya-Rundel and Maria Tackett for STA 112FS and STA 199 taught during Fall 2018 and Fall 2019, respectively.

What's wrong with this graph?

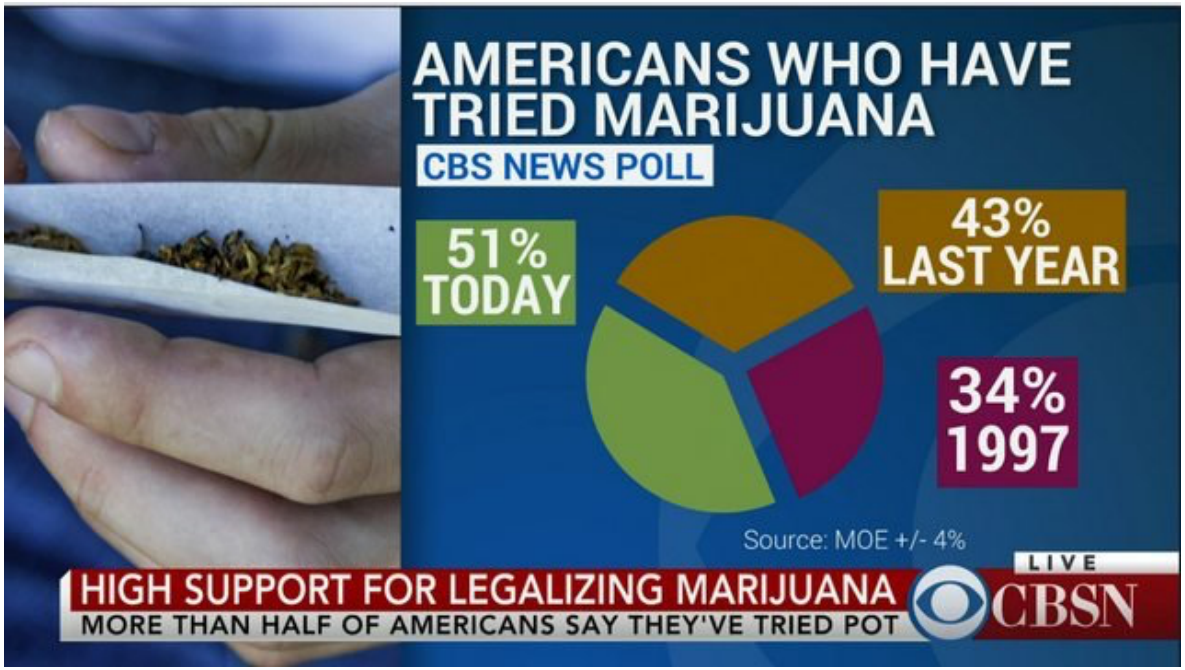
How would you fix it?

What's wrong with this graph?



How would you fix it?

What's wrong with this graph?



How would you fix it?

What is a p-value?

NOV. 24, 2015, AT 12:12 PM

Not Even Scientists Can Easily Explain P-values

By Christie Aschwanden

Filed under Scientific Method

"Not even scientists can easily explain p-values"

(hopefully STA 198 has helped you do better than these professional scientists!)

What is a p-value?

There's a strong case that chasing p-values has led science astray. | erhui1979/Getty Creative Images

"800 scientists say it's time to abandon 'statistical significance'"

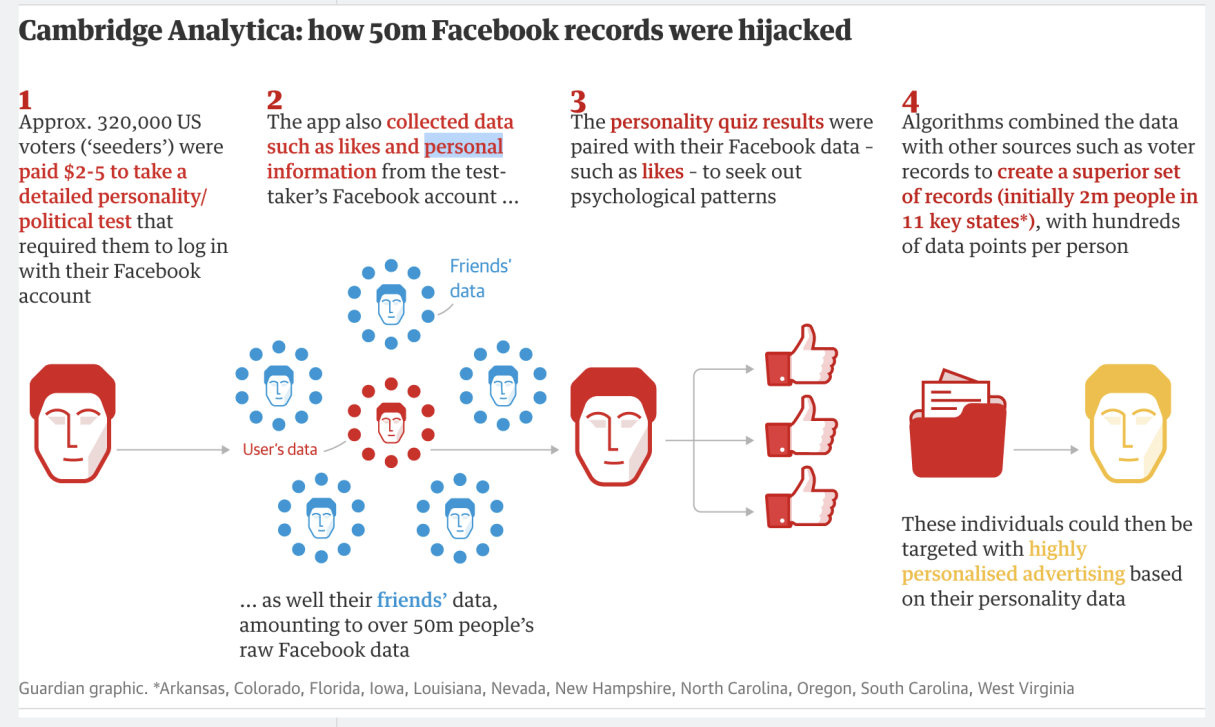
How else might we evaluate evidence?

- p-values don't necessarily tell us if experiments have "worked"
- Statistical significance is different from clinical or scientific significance (effect sizes)
- Ask whether the result is from a novel study or a replication
- Ask whether underlying data is freely available (so anyone can check)
- Use alternative techniques such as Bayesian methods to ask "what is the probability my hypothesis is the best explanation for the results we've found?" instead of "how rare are my results under a certain hypothesis?"

OkCupid Data Breach

"OkCupid study reveals the perils of big-data science"

Facebook and Cambridge Analytica



"How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool"

Algorithmic bias

TECHNOLOGY

Does Anne Hathaway News Drive Berkshire Hathaway's Stock?

ALEXIS C. MADRIGAL MAR 18, 2011

Given the awesome correlating powers of today's stock trading computers, the idea may not be as far-fetched as you think.

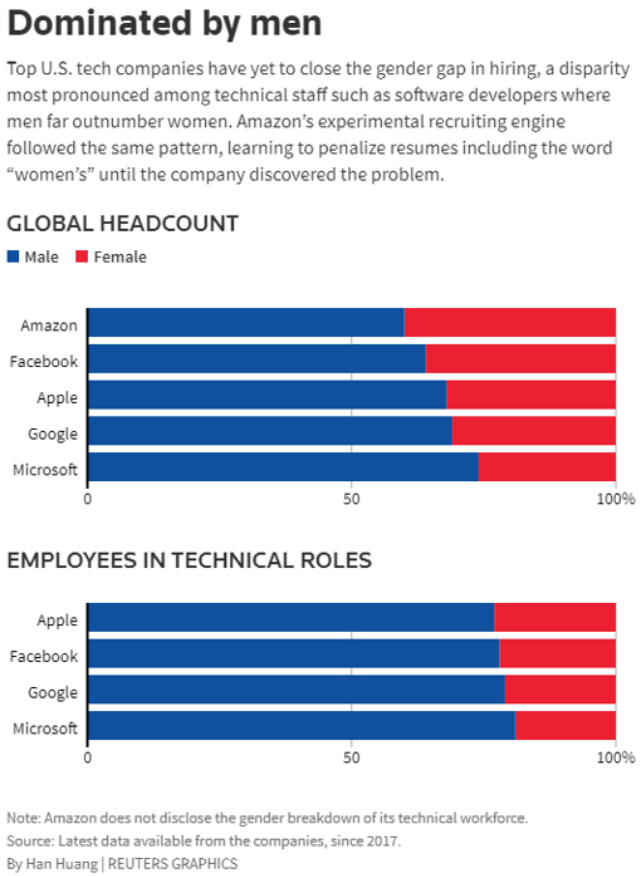
"Does Anne Hathaway news drive Berkshire Hathaway's stock?"

Machine bias in law enforcement

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

"There's software used across the country to predict future criminals. And it's biased against blacks."

Amazon's experimental hiring algorithm



"Amazon scraps secret AI recruiting tool that showed bias against women"

Parting thoughts

At some point during your data science journey you will learn tools that can be used unethically

You might also be tempted to use your knowledge in a way that is ethically questionable either because of business goals or for the pursuit of further knowledge (or because your boss told you to)

How do you train yourself to make the right decisions (or reduce the likelihood of accidentally making the wrong decisions) at those points?

Further reading

- Statistical Inference in the 21st Century: A World Beyond $p < 0.05$ (Free access through Duke library!)
- Ethics and Data Science (free Kindle book download)
- How Charts Lie: Getting Smarter about Visual Information
- Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy