

# 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.

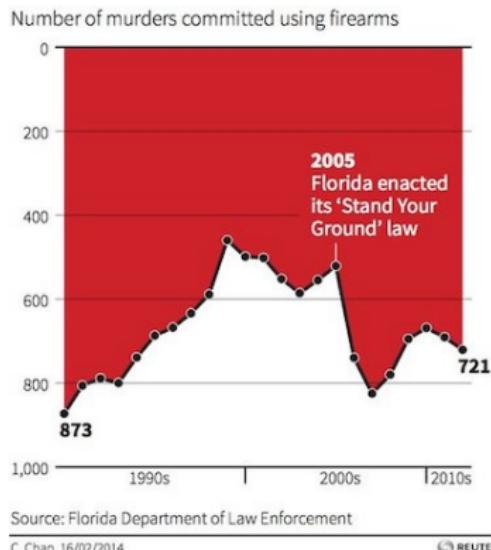
# What's wrong with this graph?



How would you fix it?

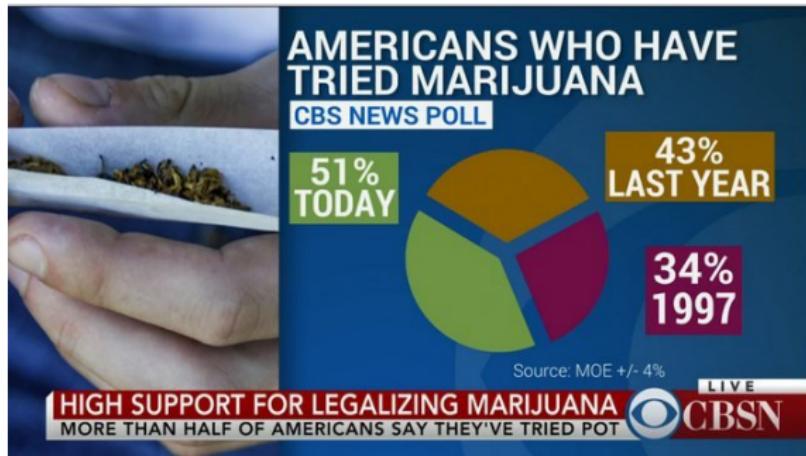
# What's wrong with this graph?

## Gun deaths in Florida



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. | erhu1979/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



**Ethan Jewett** @esjewett · May 11, 2016



Replies to @KirkegaardEmil

@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?



**Emil O W Kirkegaard**

@KirkegaardEmil

@esjewett No. Data is already public.

3 12:30 PM - May 11, 2016



[See Emil O W Kirkegaard's other Tweets](#)



“OkCupid study reveals the perils of big-data science”

# Facebook and Cambridge Analytica

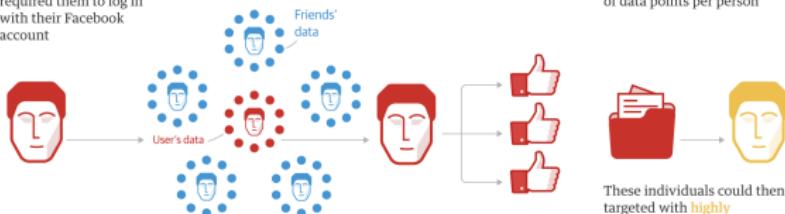
## Cambridge Analytica: how 50m Facebook records were hijacked

**1** Approx. 320,000 US voters ('seeders') were paid \$2-5 to take a detailed personality/political test that required them to log in with their Facebook account

**2** The app also collected data such as likes and personal information from the test-taker's Facebook account ...

**3** The personality quiz results were paired with their Facebook data - such as likes - to seek out psychological patterns

**4** Algorithms combined the data with other sources such as voter records to create a superior set of records (initially 2m people in 11 key states\*), with hundreds of data points per person



These individuals could then be targeted with highly personalised advertising based on their personality data

Guardian graphic. \*Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

“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.”

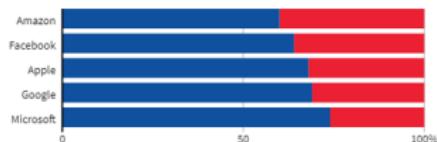
# Amazon's experimental hiring algorithm

## Dominated by men

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

### GLOBAL HEADCOUNT

■ Male ■ Female



### EMPLOYEES IN TECHNICAL ROLES



Note: Amazon does not disclose the gender breakdown of its technical workforce.

Source: Latest data available from the companies, since 2017.

By Han Huang | REUTERS GRAPHICS

“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

