Ethics, string manipulation and conclusions

Overview

Very quick review of PCA and clustering

Quick discussion of ethics in Statistics/Data Science

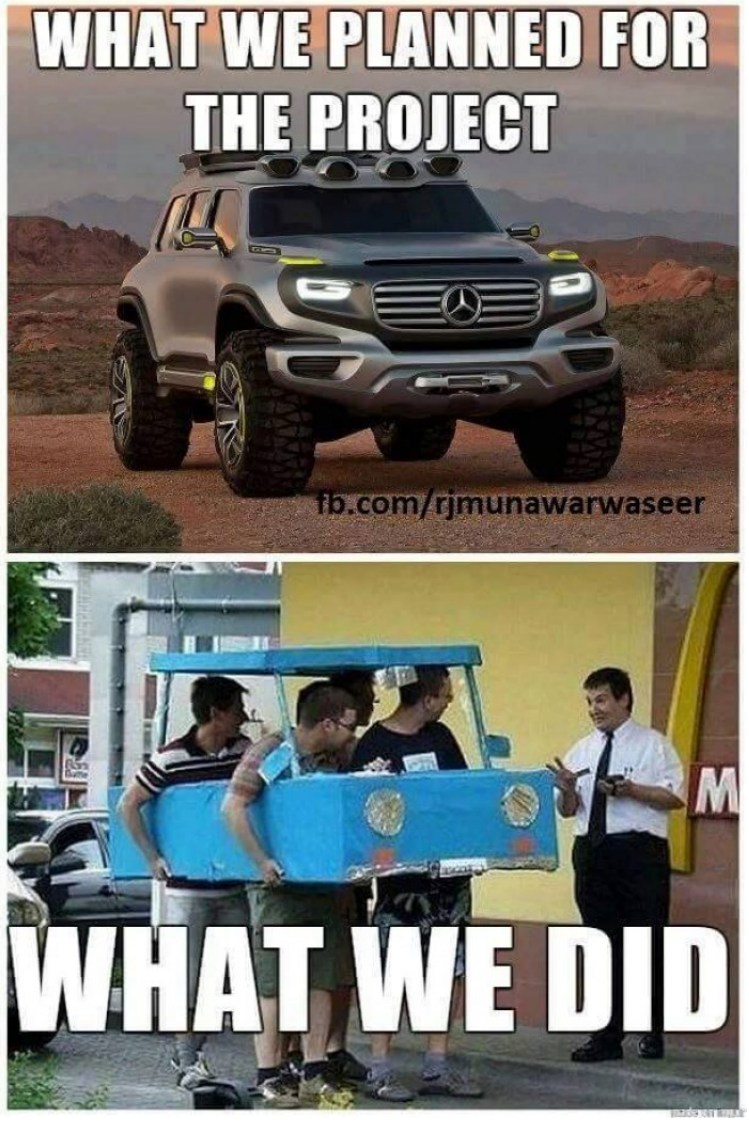
Text manipulation (if there is time)

Wrap up and conclusions

Announcement

Extra office hours tomorrow from 11-12:30pm

How are the final projects going?

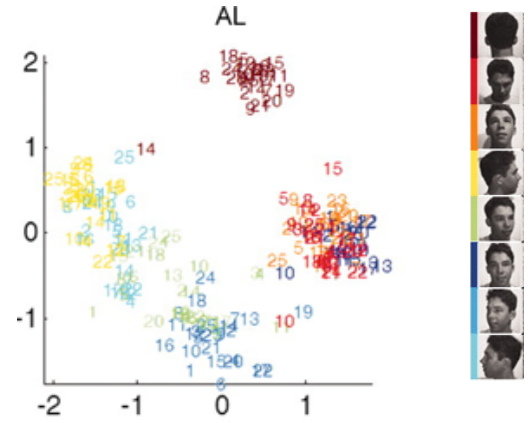


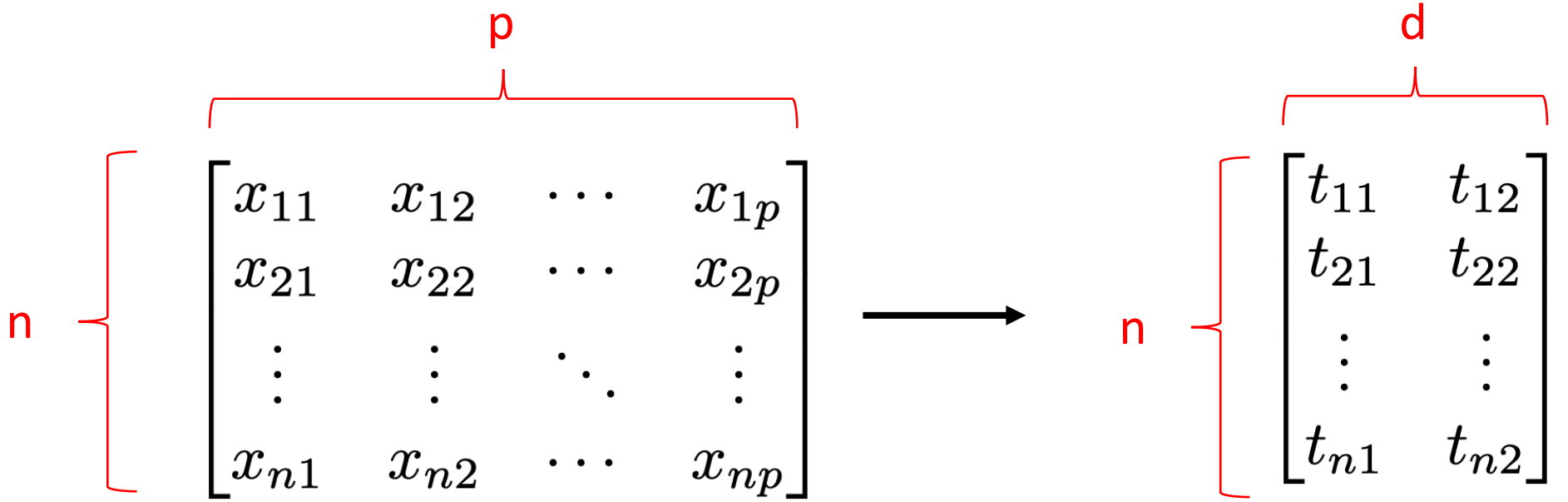
Quick PCA review

Principal Component Analysis

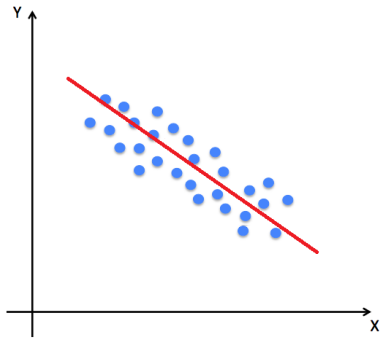
Principal Component Analysis is a linear dimensionality reduction method that tries to capture the dimensions with the highest variability in the original data.

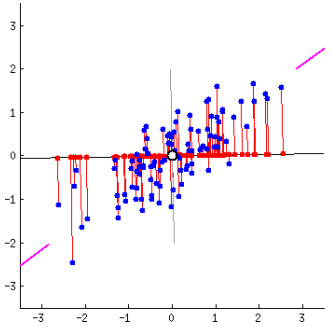
- $x_1, x_2, \ldots, x_p \longrightarrow t_1, t_2, \ldots, t_d$ where $d << p$
- This can be useful for visualization if d is 2 or 3





Principal Component Analysis





Principal component **scores** t_i's are linear combinations of the original variables x_{ij}'s:

$$t_{i1} = \alpha_{11}x_{i1} + \alpha_{21}x_{i2} + \ldots + \alpha_{p1}x_{ip}$$

$$\begin{bmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \alpha_{11} \\ \alpha_{21} \\ \vdots \\ \alpha_{p1} \end{bmatrix}$$

α_{j1} are the **loadings** for the first principal component

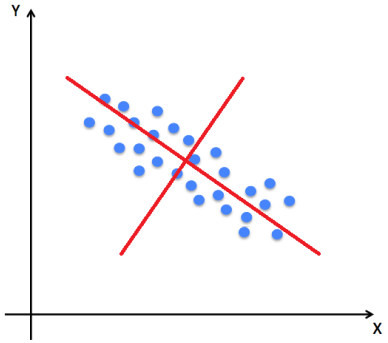
- The "norm" of the loadings is 1: $\sum_{j=1}^{p} \alpha_{j1}^2 = 1$

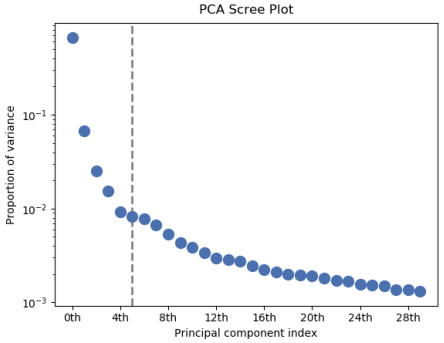
We can do this for each case in our data set we get values: $t_{11}, \ldots, t_{n1}$

We find the loadings α_{j1} such that the variability across the principal component **scores t_{i1}'s is as large as possible**

$$\frac{1}{n-1}\sum_{i=1}^{n} t_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\alpha_{11}x_{i1} + \alpha_{21}x_{i2} + \ldots + \alpha_{p1}x_{ip})^2$$

Higher Principal Components





The second principal component scores t_{i2} is the linear combination of the $x_1, x_2, \ldots, x_p$ that has maximal variance and is **uncorrelated** with the first principal component scores t_{i1}

- $t_{i2} = \alpha_{12}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{p2}x_p$
- $cor(T_1, T_2) = 0$

First and second principal components

$$\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \\ \vdots & \vdots \\ \alpha_{p1} & \alpha_{p2} \end{bmatrix}$$

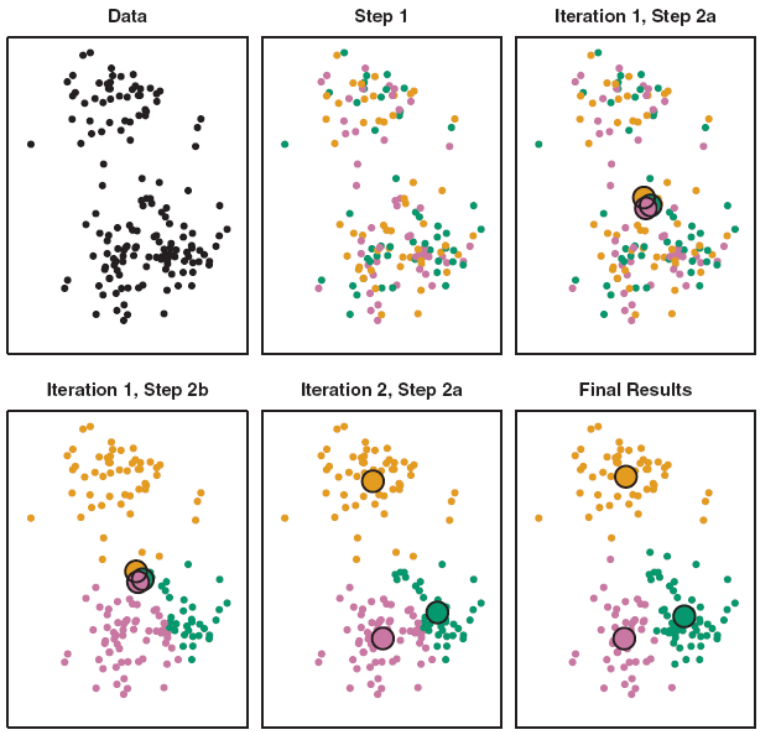
We continue this process until we find all the principal component scores, $T_1, T_2, \ldots T_d$

- We can use a scree plot to decide on how many PCs to keep (the value of ***d***)

All principal components

$$\begin{bmatrix} t_{11} & t_{12} & \ldots & t_{1d} \\ t_{21} & t_{22} & \ldots & t_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \ldots & t_{nd} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1d} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2d} \\ \vdots & \vdots & & \\ \alpha_{p1} & \alpha_{p2} & \ldots & \alpha_{pd} \end{bmatrix}$$

Clustering

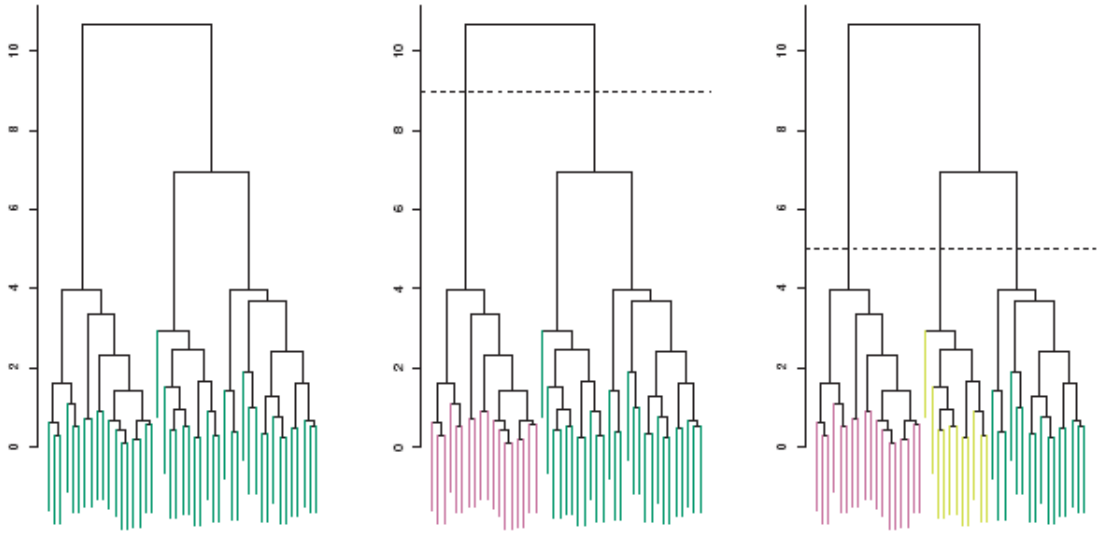


Clustering divides data points x_i's into subgroups

- Items in the same group are similar/homogeneous
- Items in different groups are different from each other

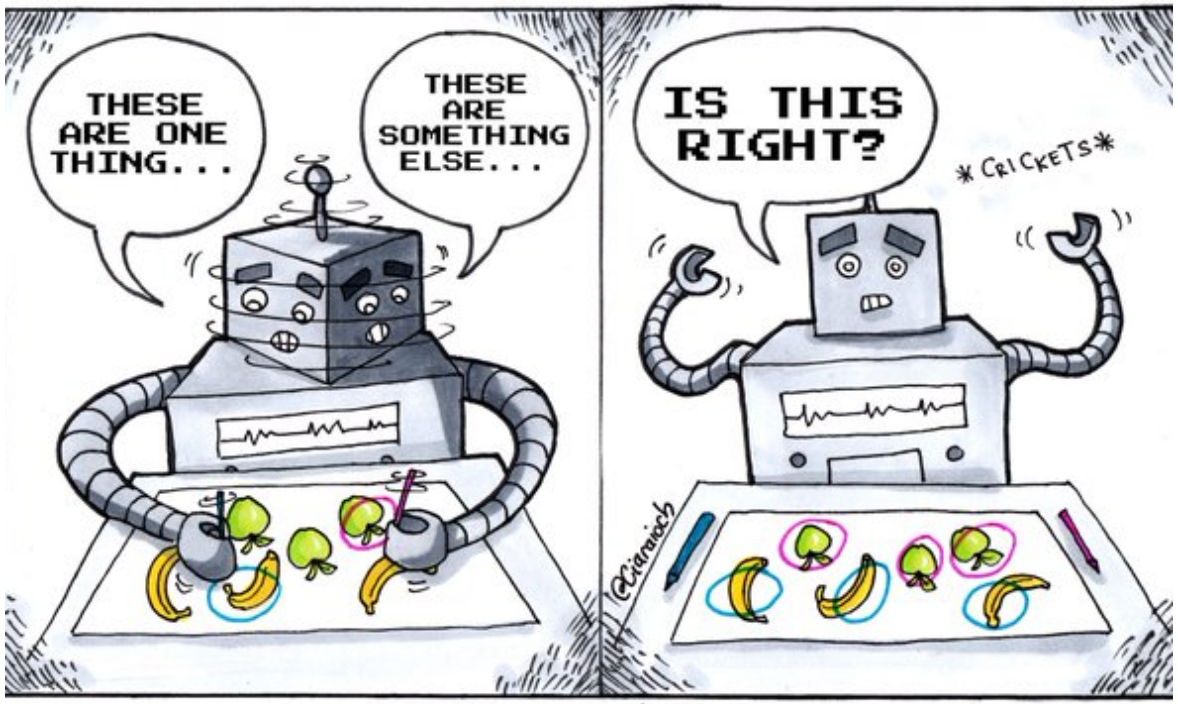
Clustering can be:

- **Flat**: k-means clustering
- **Hierarchical**: visualized via a dendrogram

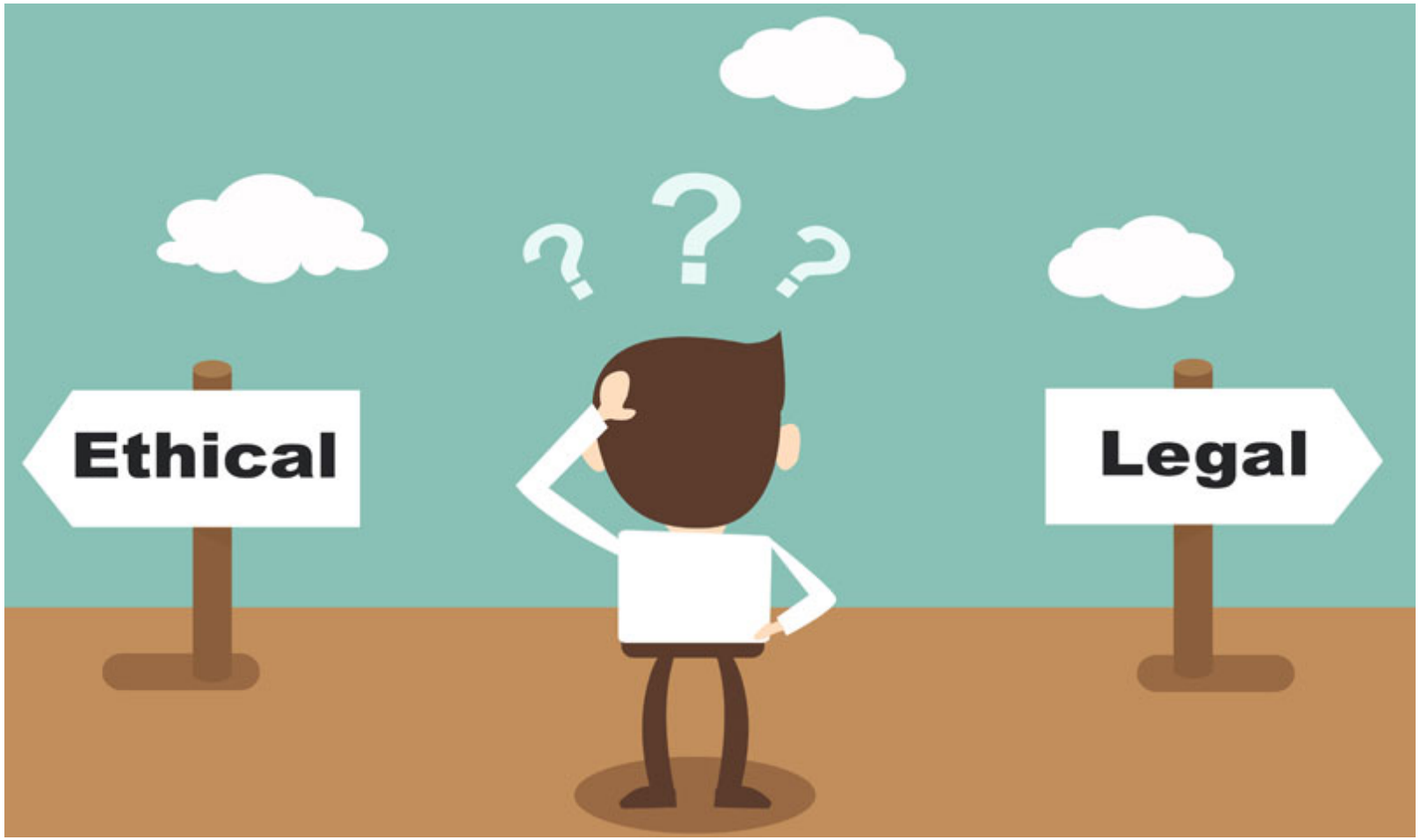


[K-means Shiny app]

Questions?


THESE ARE ONE THING...
THESE ARE SOMETHING ELSE...
IS THIS RIGHT?
CRICKETS
@Ciaraoch
Unsupervised Learning

Ethics in Statistics and Data Science



Ethics in Data Science

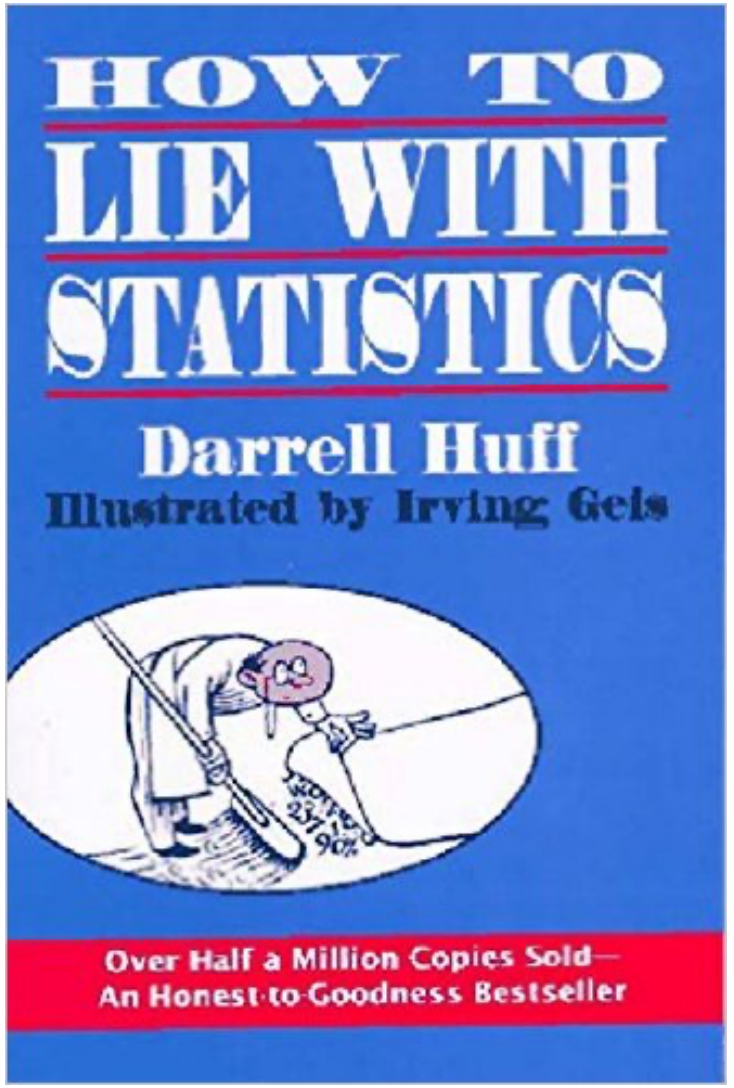
Ethics of:

1. Data presentation
2. Using valid data
3. Data scraping TOS and privacy
4. Reproducibility
5. Citations/peer review
6. Disclosure
7. Ethics in Statistical analyses
8. Ethics of creating powerful tools

1. Ethics of data presentation

Data should be displayed in an honest way that gives an accurate picture of trends

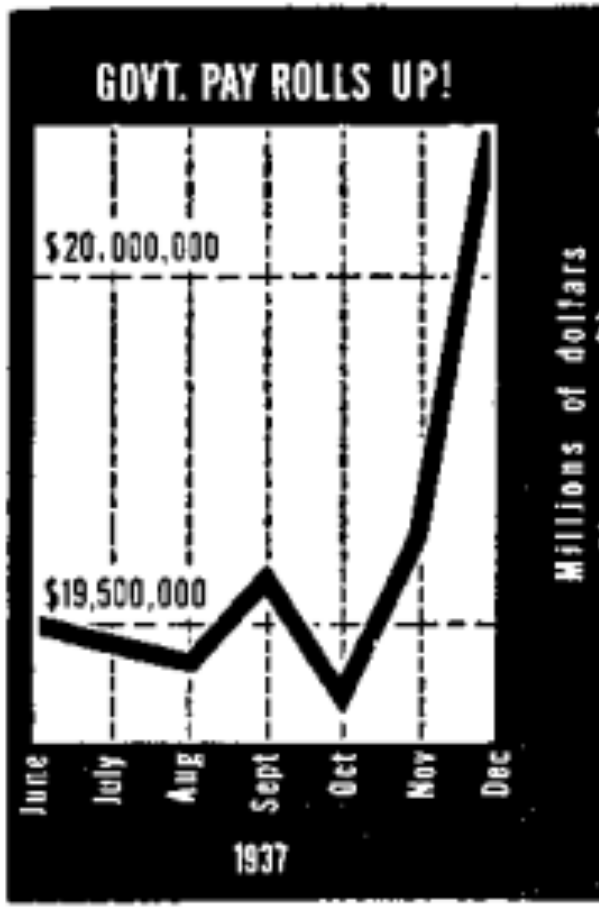
Darrell Huff wrote a classic book in the 1950's pointing out ways that people lie with statistics



Ethics of data presentation

What is potentially misleading with this figure?

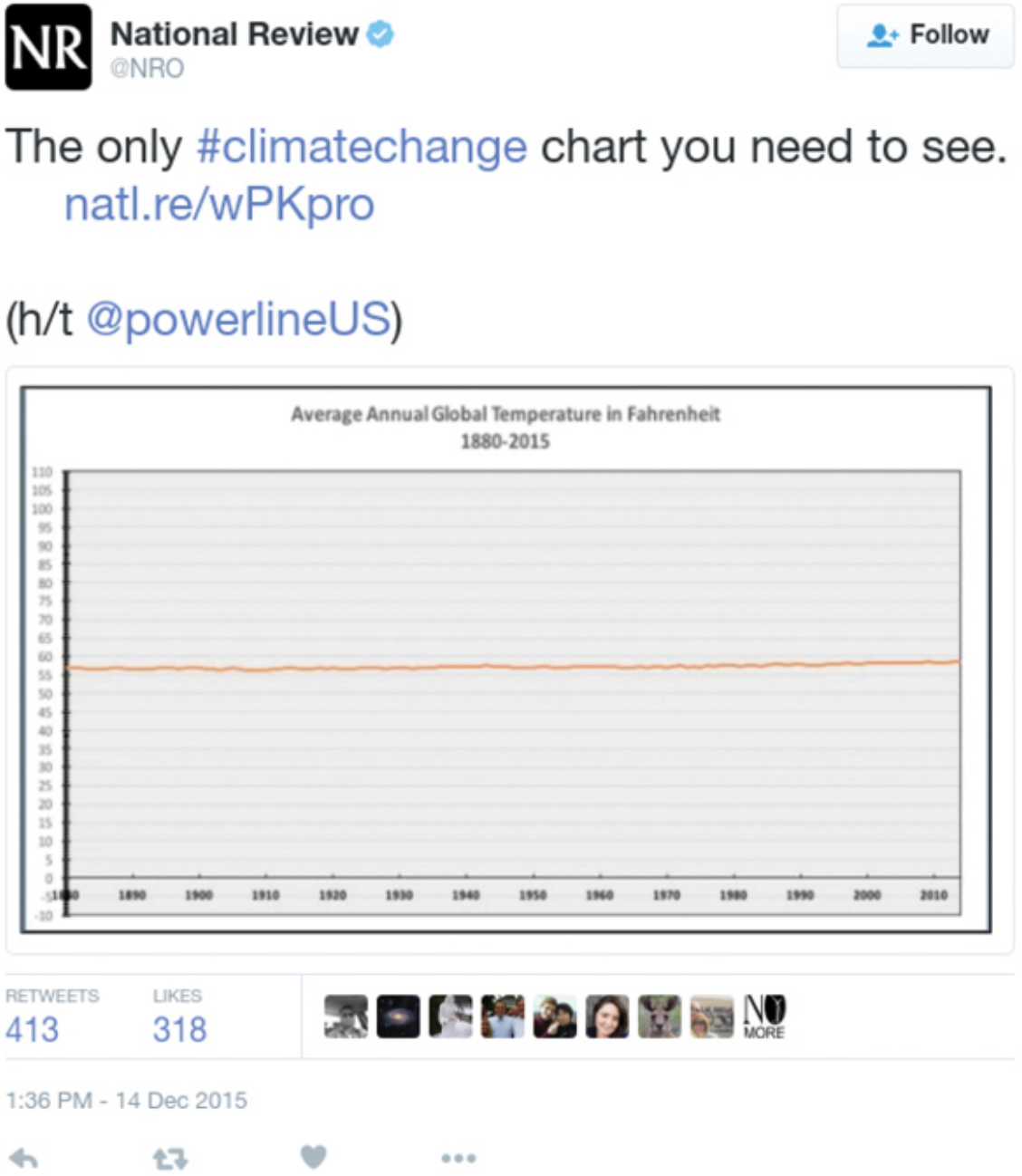
Only a 4% increase in payroll



From a 1938 article in Dun's Review titled 'GOVERNMENT PAY ROLLS UP!'

How much has the climate changed?

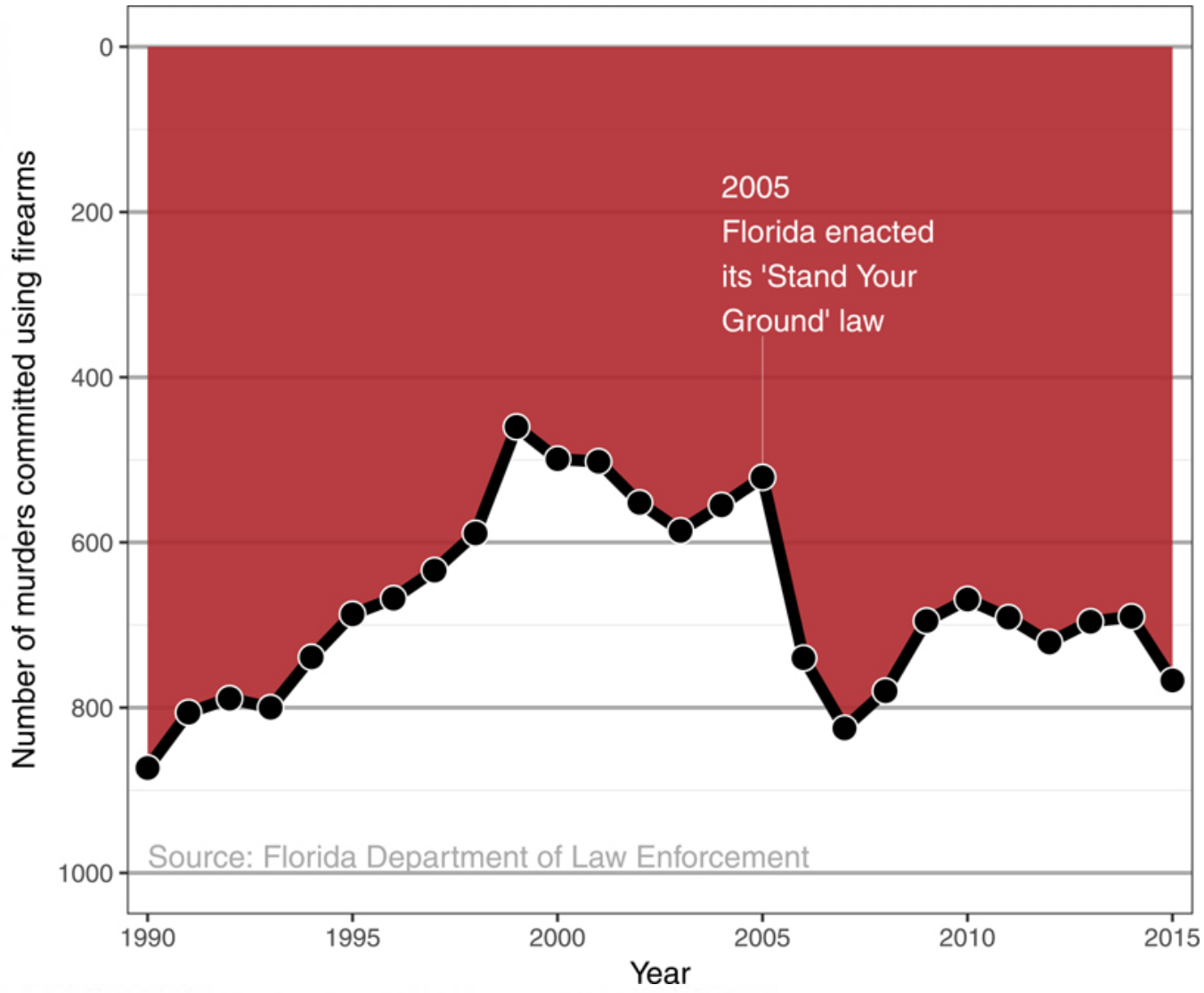
The axes go from 110 to -10 degrees which is not reasonable for the average planet temperature



Did ‘Stand Your Ground’ decrease murder by firearms?

What is misleading with this figure?

The axes are going in the wrong direction



2. Using valid data

Is almost everyone satisfied with Hampshire College?

Did only happy alums response?
(response bias)

Alumni Survey Results

As part of a strategic-planning process, in spring 2013 Hampshire College launched a survey of alums. Via email, the College invited 8,160 alums to fill out an online questionnaire administered by the campus's Alumni and Family Relations and Institutional Research offices. A total of 1,920 surveys were completed, yielding a response rate of 24%.

Note: The percentages in the data (below) are based on the number of responses received for each question.

To what extent do you agree with the following statements?

Strongly Agree or Agree

Please rate your student experience at Hampshire.

2. Using valid data

In a signed affidavit, Miller claimed to show that more than 89,000 ballots requested by Pennsylvania Republicans either were not counted by the state or requested by someone other than the registered Republican.

He used data provided by former Donald Trump campaign staffer Matt Braynard.

The Berkshire Eagle

Williams prof disavows own finding of mishandled GOP ballots

By Francesca Paris, The Berkshire Eagle Nov 24, 2020

Steven Miller, a professor of mathematics at Williams College, issued a statement Monday in which he apologized for a "lack of clarity and due diligence" after his statistical analysis of Pennsylvania mail-in votes was used by conservative lawmakers to push unsubstantiated claims of voter fraud.

- 95% confidence interval for the probability: $p - 1.96\sqrt{\frac{p(1-p)}{n}}$ to $p + 1.96\sqrt{\frac{p(1-p)}{n}}$

We now apply this to our problem. For the first question, we had either 556 out of 1706 who said they did not request a ballot but we know one was requested in their name, or (including the 544 who said they voted in person) we have 556 out of 2250.

We can use this to estimate the number of ballots requested by someone other than the registered Republican:

- 95% confidence interval for such ballots: [50,380, 57,755] or [37927, 43823],

2. Using valid data

The analysis by Steven Miller, a professor of mathematics at Williams, has drawn criticism from statisticians for failing to meet basic standards for a statistical analysis. Academic peers called the analysis “irresponsible” and “naive” for ignoring the shortcomings of the underlying data.

The Berkshire Eagle

Williams prof disavows own finding of mishandled GOP ballots

By Francesca Paris, The Berkshire Eagle Nov 24, 2020

Donald J. Trump ✓
@realDonaldTrump

Thank you!

Elizabeth Harrington ✓ @LizRNC · Nov 21
FEC Chairman: Trump Campaign Bringing 'Legitimate Accusations' of Election Fraud to Court

justthenews.com/politics-polic...

7:25 AM · Nov 22, 2020 · Twitter for iPhone

26.2K Retweets **1K** Quote Tweets **128.4K** Likes

…Pachter also expressed concern that Miller’s academic credentials are giving validity to claims about election misconduct that Trump has pushed for weeks. “He has a degree from Yale, he immediately looks like an expert, so I think this is a very dangerous piece.”

2. Using valid data

Miller told The Eagle that he made a mistake separating his analysis of the data from questions about the reliability of the data itself.

“Especially as the consequences are so important, I should have made a greater effort to go deeply into and share how the data was collected and not treat this solely as an independent calculation,” he wrote in a statement Monday night.

The Berkshire Eagle

Williams prof disavows own finding of mishandled GOP ballots

By Francesca Paris, The Berkshire Eagle Nov 24, 2020

Steven Miller, a professor of mathematics at Williams College, issued a statement Monday in which he apologized for a “lack of clarity and due diligence” after his statistical analysis of Pennsylvania mail-in votes was used by conservative lawmakers to push unsubstantiated claims of voter fraud.

photo provided by

The ethical statistician:

- **Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis. When reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms.**
- **Reports the limitations of statistical inference and possible sources of error.**
- In publications, reports, or testimony, identifies who is responsible for the statistical work if it would not otherwise be apparent.
- **Reports the sources and assessed adequacy of the data, accounts for all data considered in a study, and explains the sample(s) actually used.**
- **Clearly and fully reports the steps taken to preserve data integrity and valid results.**
- **Where appropriate, addresses potential confounding variables not included in the study.**
- In publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics.
- **In publications or testimony, identifies the ultimate financial sponsor of the study, the stated purpose, and the intended use of the study results.**
- **When reporting analyses of volunteer data or other data that may not be representative of a defined population, includes appropriate disclaimers and, if used, appropriate weighting.**
- To aid peer review and replication, shares the data used in the analyses whenever possible/allowable and exercises due caution to protect proprietary and confidential data, including all data that might inappropriately reveal.

3. Data scraping, terms of service and privacy

Scraping publicly available data is fine (e.g., Wikipedia) but what about scraping data if:

- It violates a website's Terms of Service?
- User privacy?

Kirkegaard and Bjerrekaer scraped okcupid and data on 68,371 users publicly available including usernames, dating preferences, etc.

- Is this ok?

Submitted: 8th of May 2016
Published: 3rd of November 2016

The OKCupid dataset: A very large public dataset of dating site users

Emil O. W. Kirkegaard* Julius D. Bjerrekær†

Open Differential Psychology

4. Reproducibility

Do scientists have an ethical obligation to make sure their research is reproducible?

Reproducibility

Do scientists have an obligation to share data/code?

- What if it could hurt your career?
  - Others could prove you wrong, make new findings on your own data, etc.

What should you do if you find one of your papers is wrong?

- You need to retract the paper!

5. Citations

If you got an idea from someone else you should always cite their work

- What is the term for failing to do this?

You should also cite other background work that is relevant

What about citing someone because they will be a reviewer of your paper?

- How do you deal with someone else's questionable behavior?

6. Disclosure of conflicts of interest

If you have a conflict of interest you should always disclose it

- Even if you think it doesn't affect your judgement it might

No Disclosure? No Problem. Sean Hannity Gets a Pass at Fox News.

By MICHAEL M. GRYNBAUM and JOHN KOBLIN APRIL 17, 2018

On his Fox News show, Sean Hannity regularly spoke with and about Michael Cohen, President Trump's personal lawyer. Mr. Hannity was recently named in court as one of Mr. Cohen's clients, which he denies. By SARAH STEIN KERR, ROBIN LINDSAY and NEETI UPADHYE on April 16, 2018. Photo by Carolyn Kaster/Associated Press. Watch in Times Video »

Embed

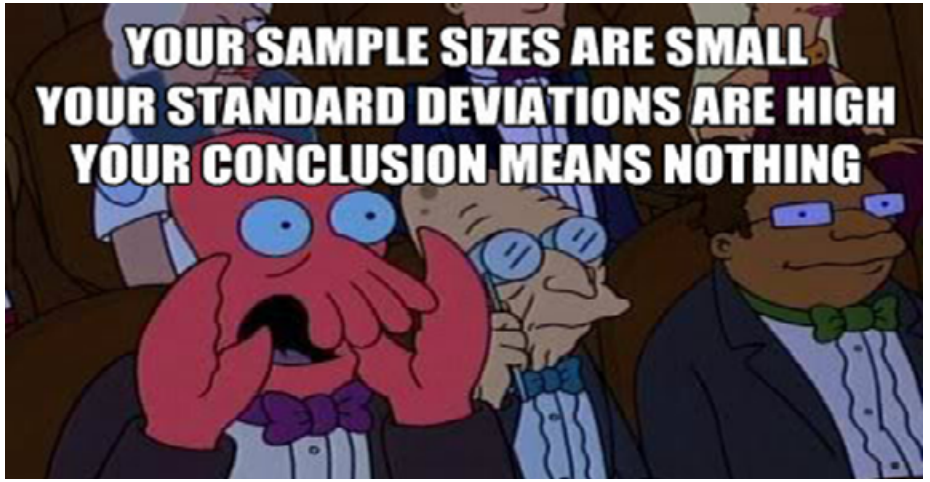
7. Ethics in statistical analyses

P-hacking (data dreging):

Keep trying different hypothesis tests on a data set until you reach ‘statistical significance’ ($p < 0.05$)

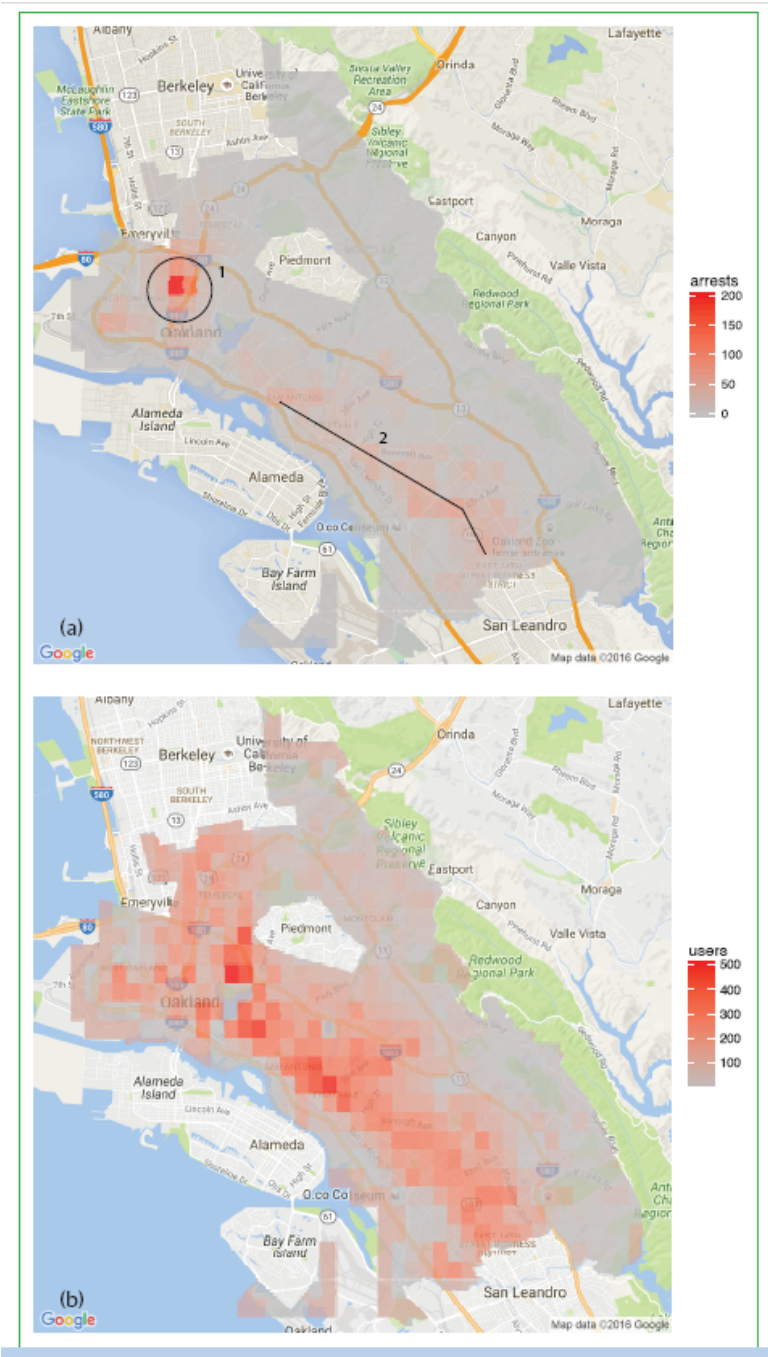
File drawer effect:

- Try a million studies until one is significant



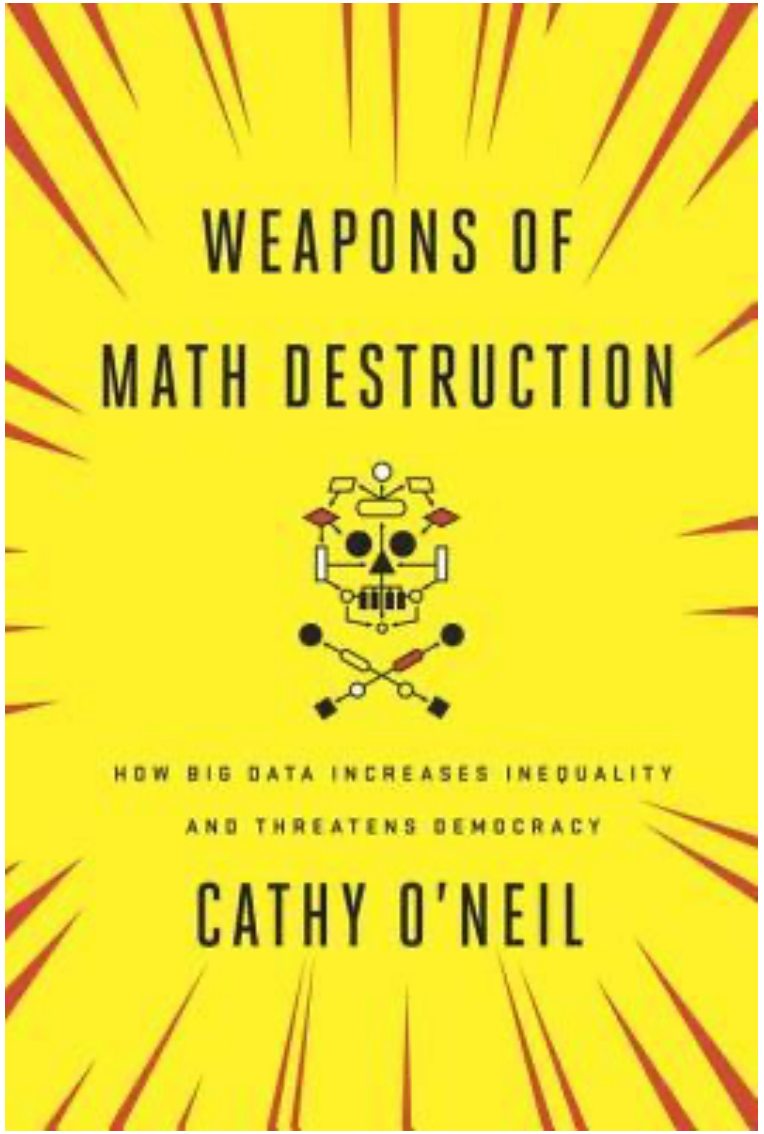
8. Ethics in machine learning

Care must be taken when interpreting the results from machine learning algorithms



Additional reading

https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end



Wrap up and conclusions

Topics we will cover

R and descriptive statistics/plots: Base R, fundamental concepts in Statistics

Review confidence intervals: Sampling and bootstrap distributions

Review of hypothesis tests: Permutation and parametric tests, theories of testing

Data wrangling: filtering and summarizing data, joining data sets, reshaping data

Data visualization: grammar of graphics, mapping

Regression: simple/multiple, non-linear terms, logistic regression

ANOVA: one-way/factorial, interactions

Statistical learning: cross-validation, logistic regression, PCA, clustering

Course objectives

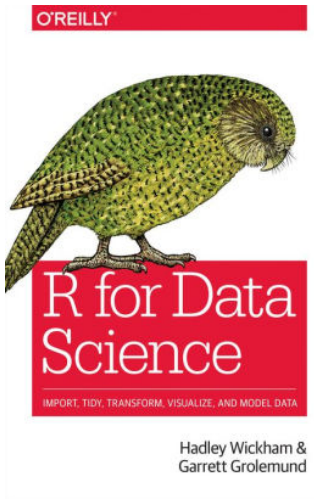
- ✓ Gain experience extracting insights from real data
- ✓ Learn how to use the R programming language to analyze, visualize and wrangle data
- ✓ Extend methods learned in intro stats
  - Non-parametric tests, multiple regression, etc.
- ✓ Solidify understanding of statistical concepts
  - Focus on insights why methods work rather than proofs
- ✓ **Learn how to find patterns in a large noisy data sets and convincingly convey the results to others!**

Next steps

Take more advanced Statistics and Data Science classes offered at Yale!

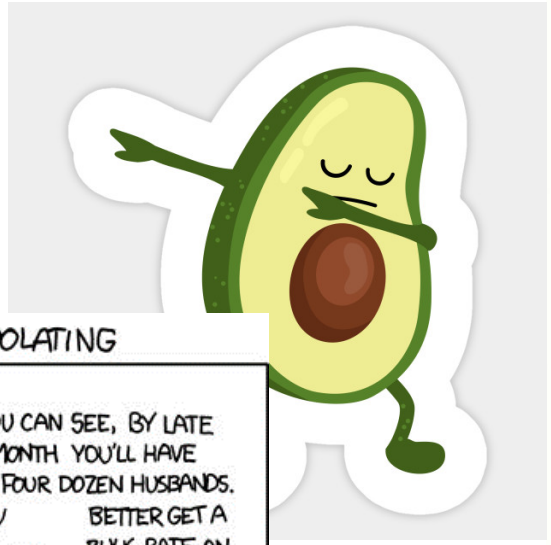
There are many good online resources to learn more R

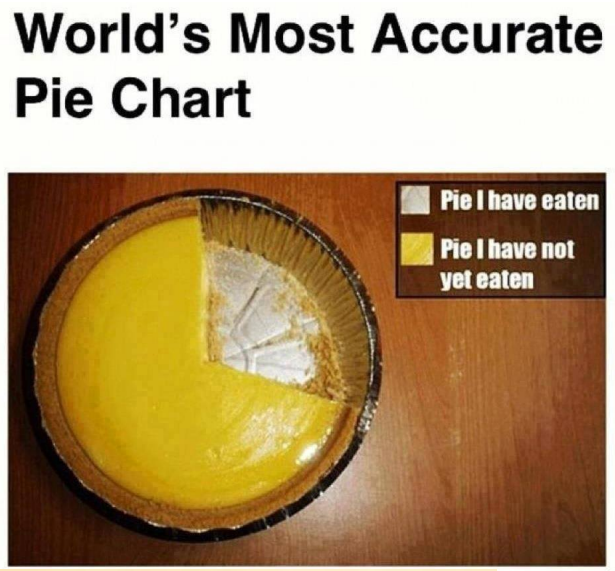
DataCamp

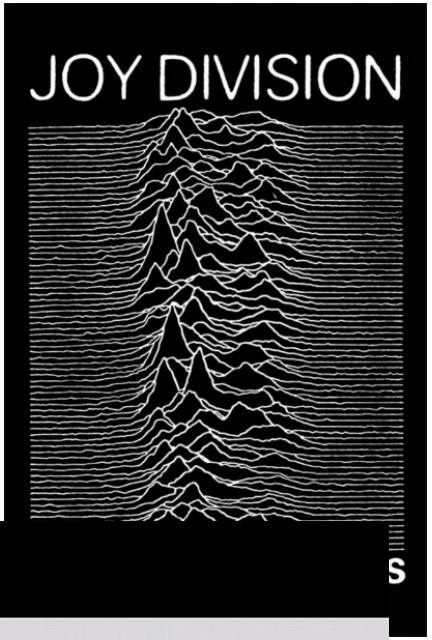


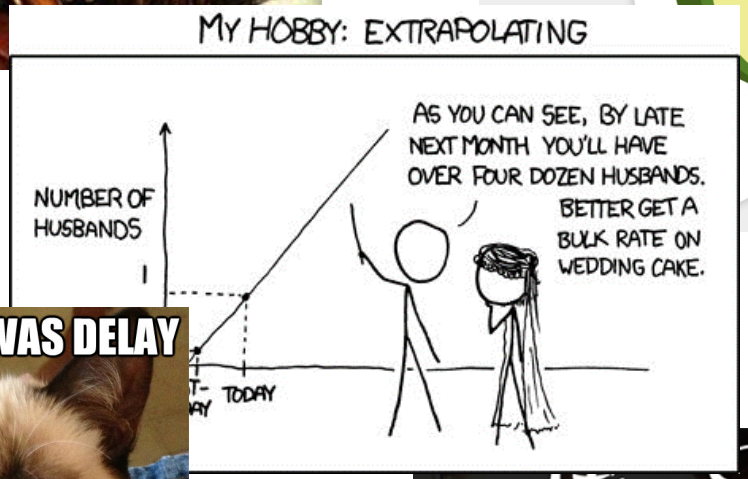
R bloggers

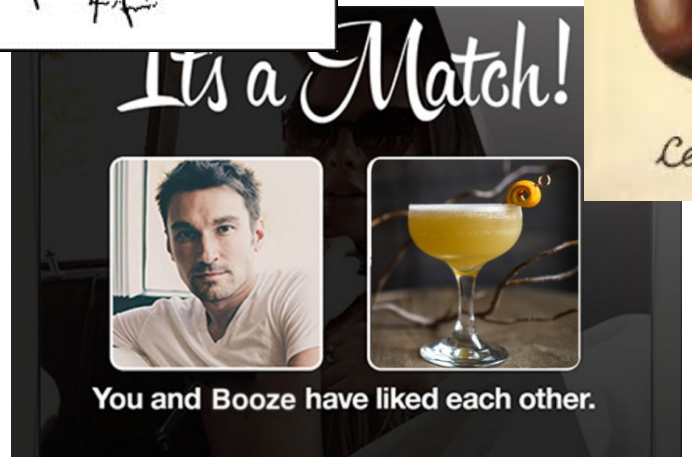
Last question: what was the worst joke of the semester?

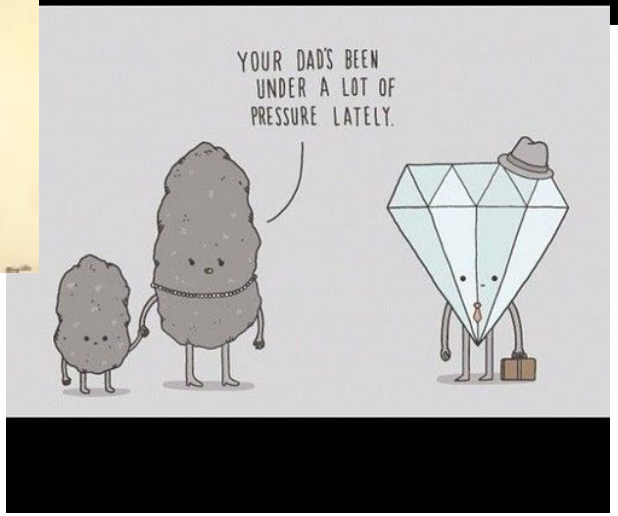












Thanks to the teaching assistants!!!

Teaching Fellows (TF)

- Hayon Michelle Choi: hayonmichelle.choi@yale.edu
- Akshay Surendra: akshay.surendra@yale.edu
- Sam Konstantinov (course manger): Sam.konstantinov@yale.edu

Undergraduate Learning Assistants (ULA)

- Lu Zheng: lu.zheng@yale.edu
- Derek Chen: derek.chen@yale.edu
- Maria (Duda) Eduarda Santana: mariaeduarda.santana@yale.edu
- Stephan Billingslea: stephan.billingslea@yale.edu

Undergraduate Technology Assistant (UTA)

- João Goncalves Cardoso: joao.cardoso@yale.edu
  - Please contact João if you have any questions about how to use Zoom

Good luck with the end of the semester!

Final project is due at 11:59pm on Sunday December 6th

Final exam is on Wednesday December 16 at 9am