YData: Introduction to Data Science

Lecture 26: Unsupervised learning continued, ethics, wrap-up

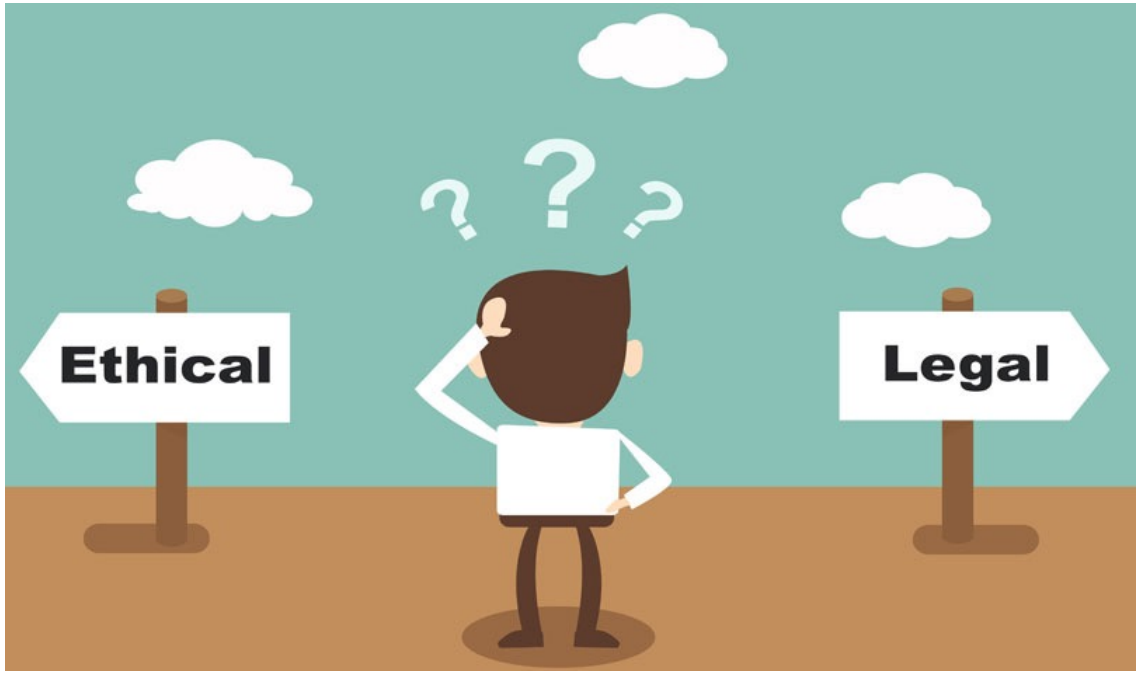
Overview

Unsupervised learning/clustering

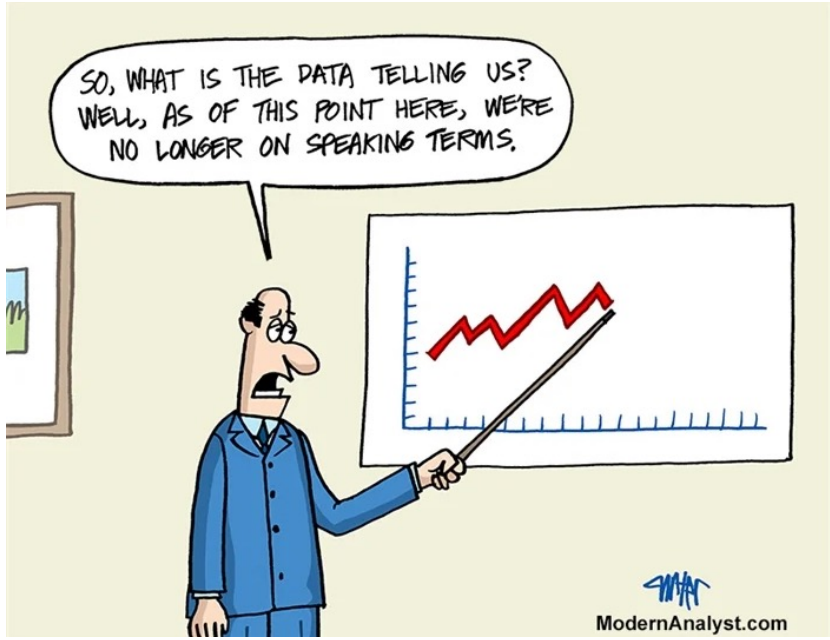
- K-means cluster
- Hierarchical clustering

Ethics

Wrap-up



Project timeline



Sunday, April 30th

- Project is due on Gradescope
  - Add peer reviews to an Appendix of your project

Please also fill out the final project reflection!

- Will be very valuable to have your feedback on how the project and class overall went

Exam review session: Tuesday, May 2nd at 2:30-3:30

- Davies Auditorium

Unsupervised learning

Supervised learning and unsupervised learning

In **supervised learning** we have a set of features X, along with a label y

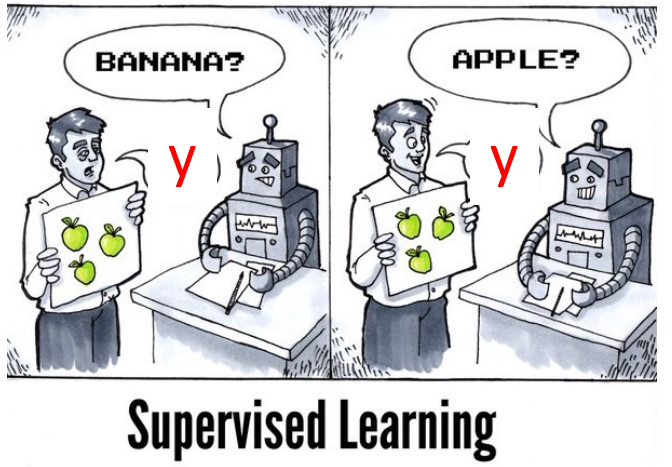
- We use the features X to predict y on new data

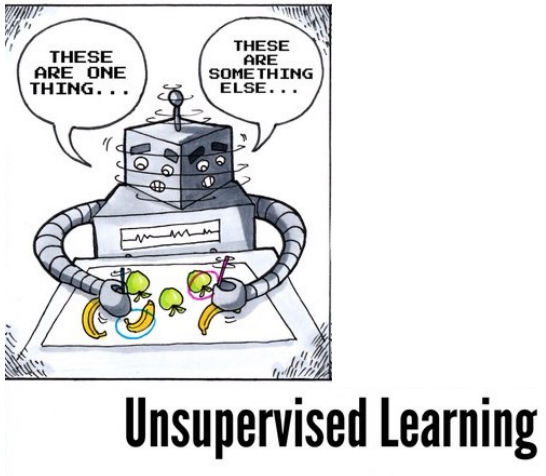
In **unsupervised learning**, we have features X, but **no** response variable y

Unsupervised learning can be useful in order to find structure in the data and to visualize patterns

A key challenge in unsupervised learning is that there is no real ground truth response variable y

- So we don't have measures like the mean prediction accuracy





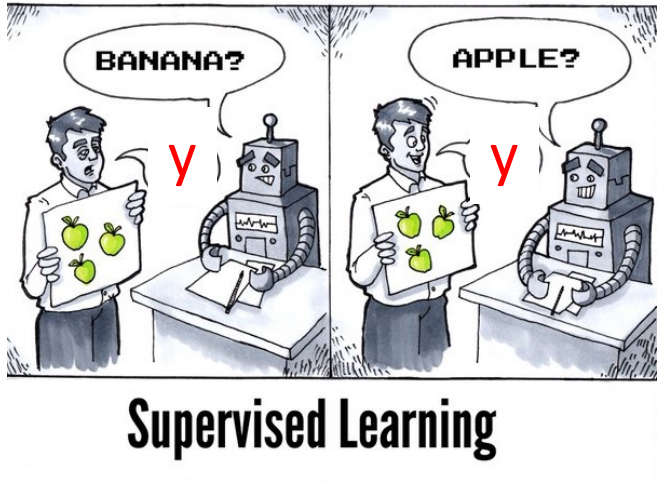
Unsupervised learning

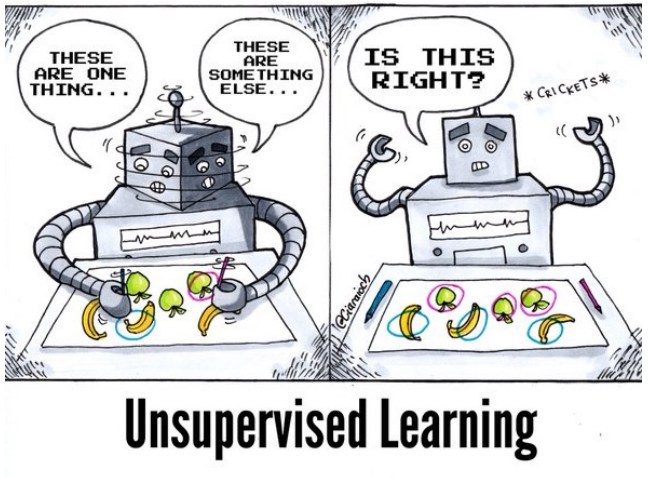
Given we are almost at the end of the semester, we will focus on clustering, which is one type of unsupervised learning:

In **clustering** we try to group similar data points together

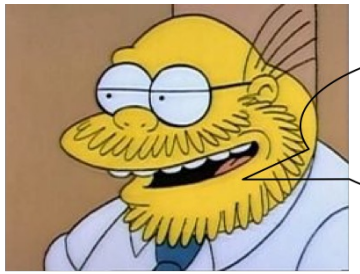
Another type of unsupervised learning:

- **Dimensionality reduction** where we try to find a smaller set of features that captures most of the variability original larger feature set
  - E.g., principal component analysis (PCA)





Clustering



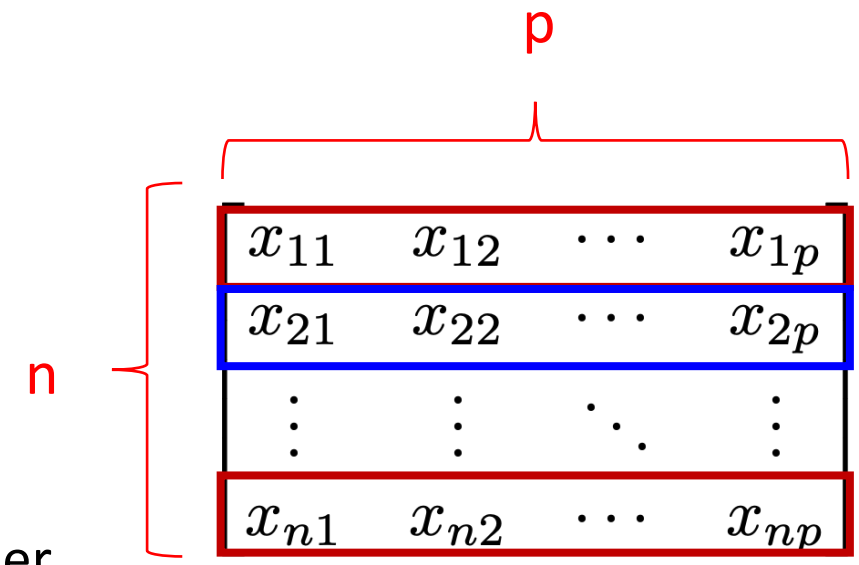
How many clusters?

Six Clusters

Two Clusters

Four Clusters

Clustering



Clustering divides n data points x_i's into subgroups

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

Examples:

- Examining gene expression levels to group cancer types together
- Examining consumer purchasing behavior to perform market segmentation

Clustering can be:

- **Flat**: no structure beyond dividing points into groups
- **Hierarchical**: Population is divided into smaller and smaller groups (tree like structure)

K-means clustering

K-means clustering partitions the data into ***K*** distinct, non-overlapping clusters

- i.e., each data point x_i belongs to exactly one cluster C_k

The number of clusters, ***K***, needs to be specified prior to running the algorithm

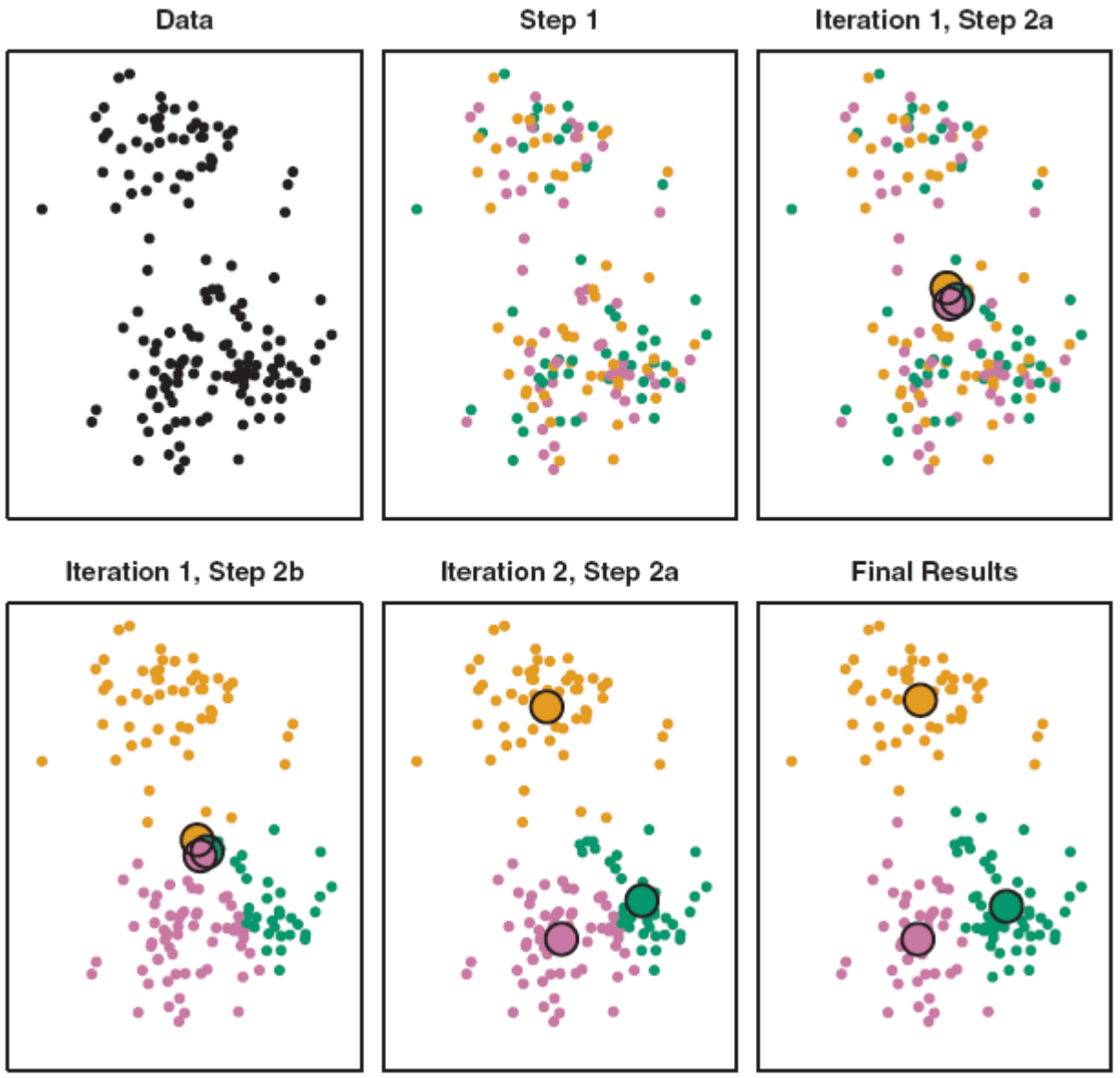
The goal is to minimize the within-cluster variation

- e.g., to make the Euclidean distance for all points within a cluster as small as possible

Finding the exact optimal solution is computationally intractable (there are k^n possible partitions), but a simple algorithm exists to find a local optimum which is often works well in practice.

K-means clustering

1. Randomly assign points to clusters C_k
2. Calculate cluster centers as means of points in each cluster
3. Assign points to the closest cluster center
4. Recalculate cluster center as the mean of points in each cluster
5. Repeat steps 3 and 4 until convergence



K-means clustering

Because only a local minimum is found, different random initializations will lead to different solutions

- One should run the algorithm multiple times to get better solutions

Let's explore this in Jupyter!

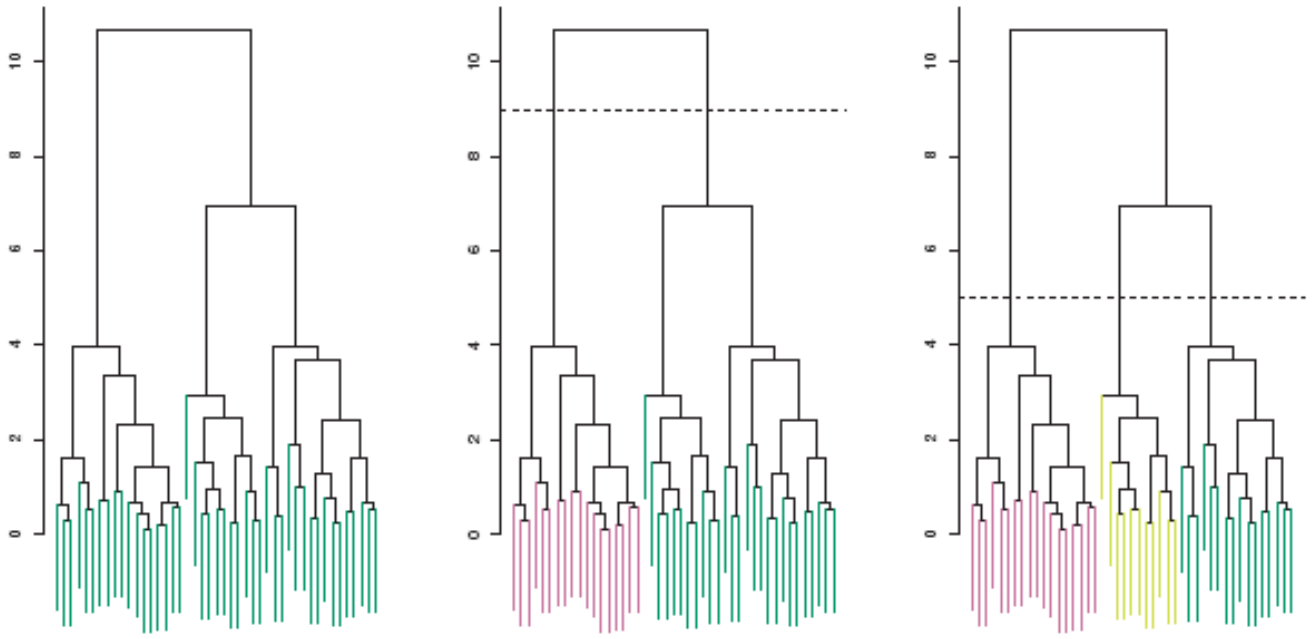
Hierarchical clustering

Hierarchical clustering

In **hierarchical clustering** we create a dendrogram which is a tree-based representation of successively larger clusters.

We can cut the dendrogram at any point to create as many clusters as desired

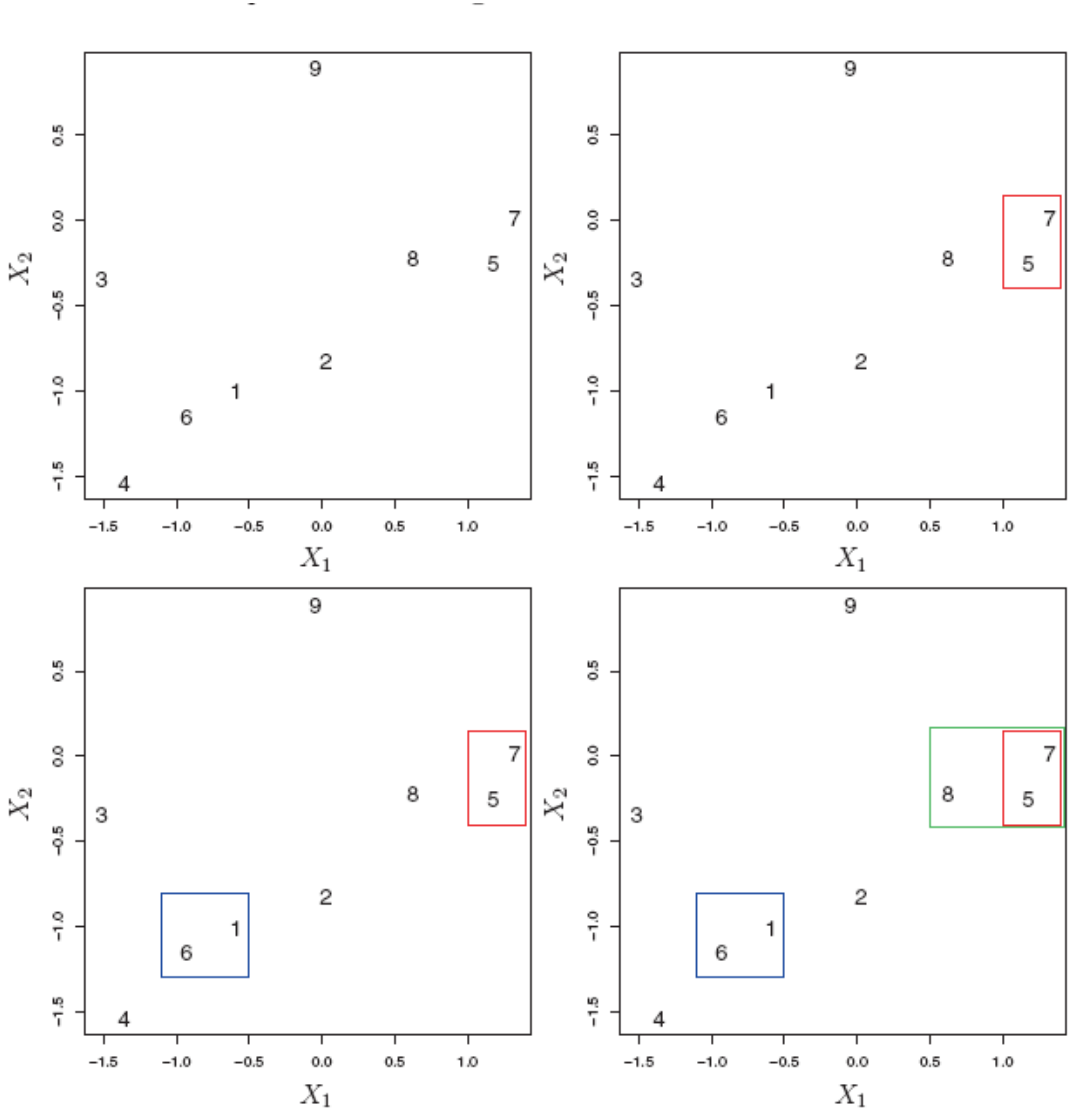
- i.e., don't need to specify the number of clusters, K, beforehand



Hierarchical clustering

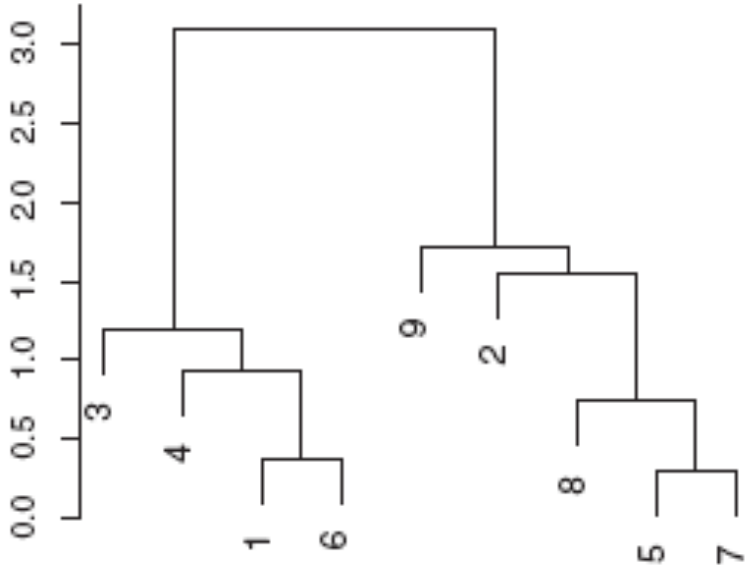
We can create a hierarchical clustering of the data using simple bottom-up agglomerative algorithm:

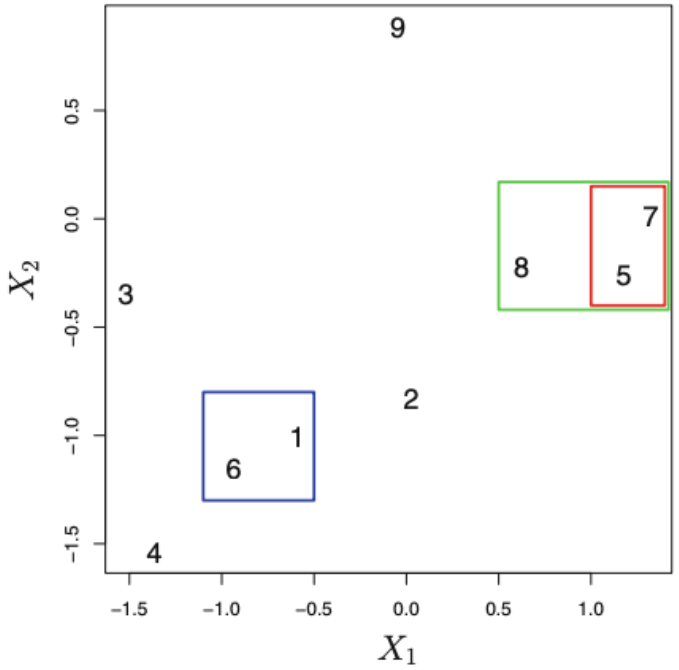
1. Choosing a (dis)similarity measure
   - E.g., The Euclidean distance
2. Initializing the clustering by treating each point as its own cluster
3. Successively merging the pair of clusters that are most similar
   - i.e., calculate the similarity between all pairs of clusters and merging the pair that is most similar
4. Stopping when all points have been merged into a single cluster



Hierarchical clustering

The vertical height that two clusters/points merge show how similar the two ***clusters*** are

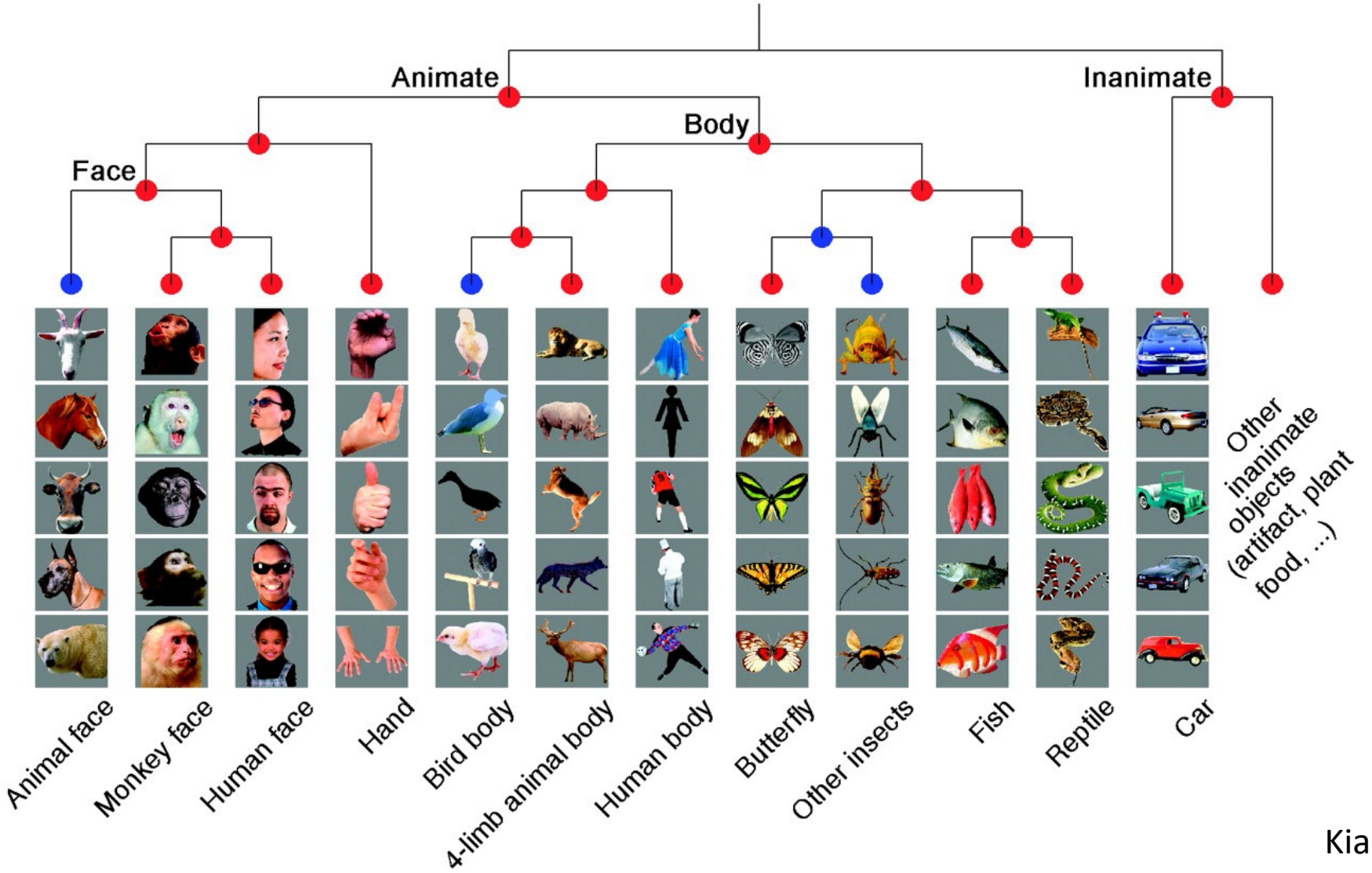




Note: horizontal distance between ***individual points*** is not important:

- point 9 is considered as similar to point 2 as it is to point 7

Hierarchical clustering example

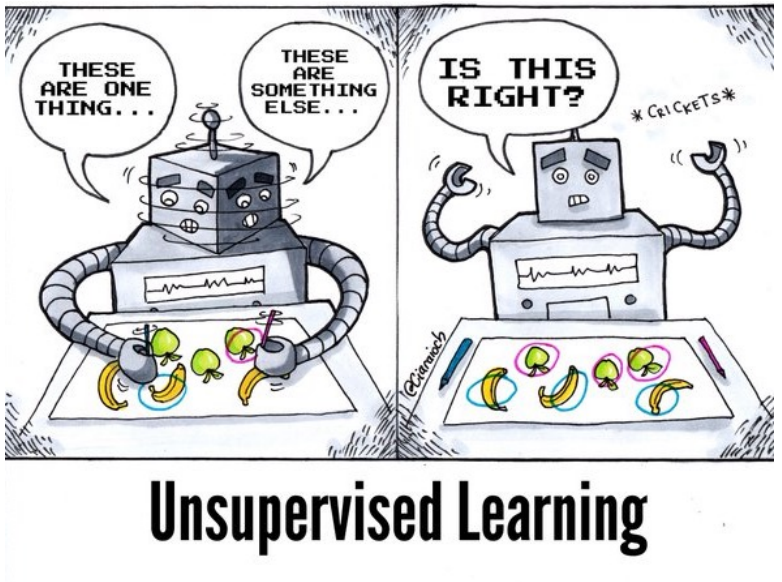


Kiani et al, 2007

Issues with clustering

Choices made can effect the results:

- Feature normalization and/or dissimilarity measure
- K-means: choice of K
- For hierarchical cluster: linkage and cut height

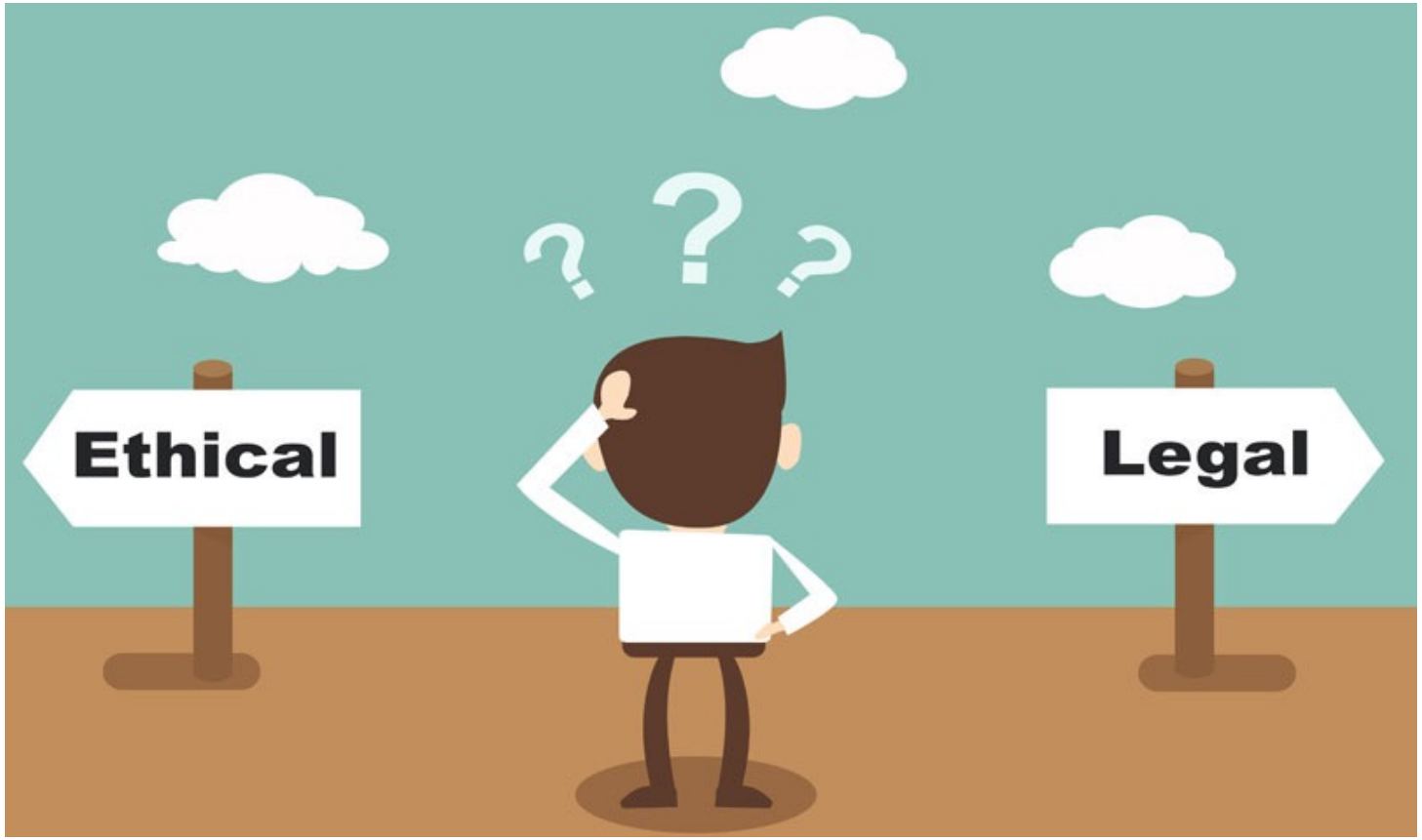


Potential approaches to deal with these issues:

- Try a few methods and see if one gives interesting/useful results
- Validate that you get similar results on a second set of data

Let's explore this in Jupyter!

Ethics



Ethics in Data Science

Ethics of:

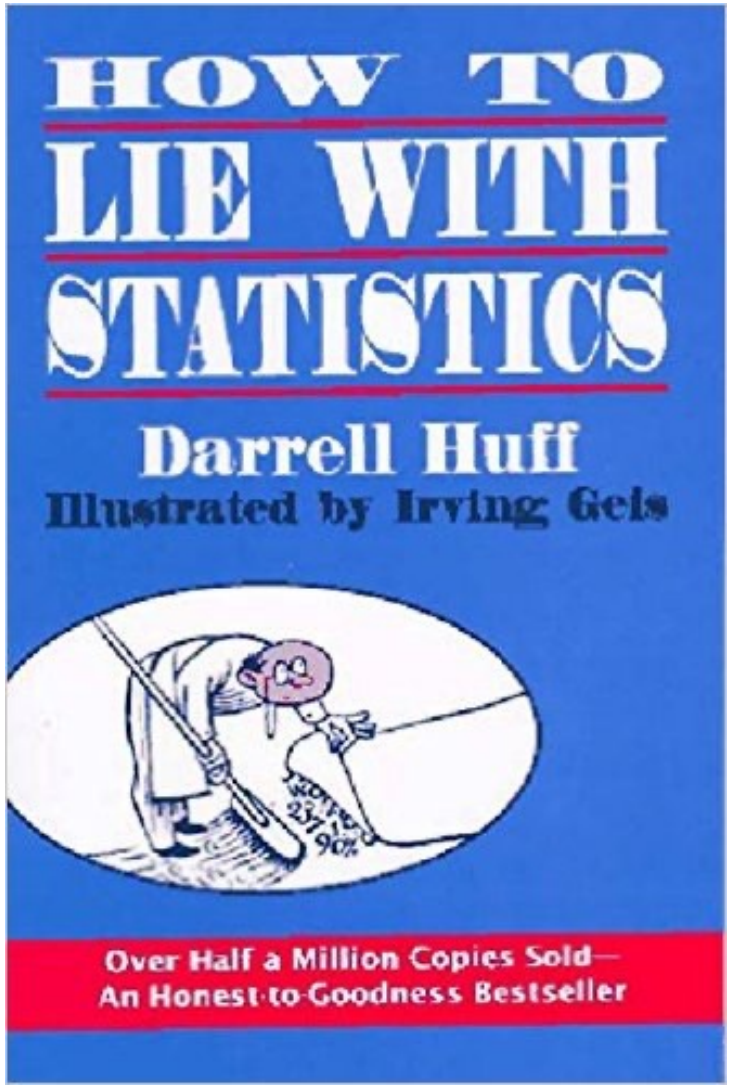
1. Data presentation
2. Data scraping TOS and privacy
3. Reproducibility
4 Citations/peer review
5. Disclosure
6. Ethics in Statistical analyses
7. Ethics of creating powerful tools

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

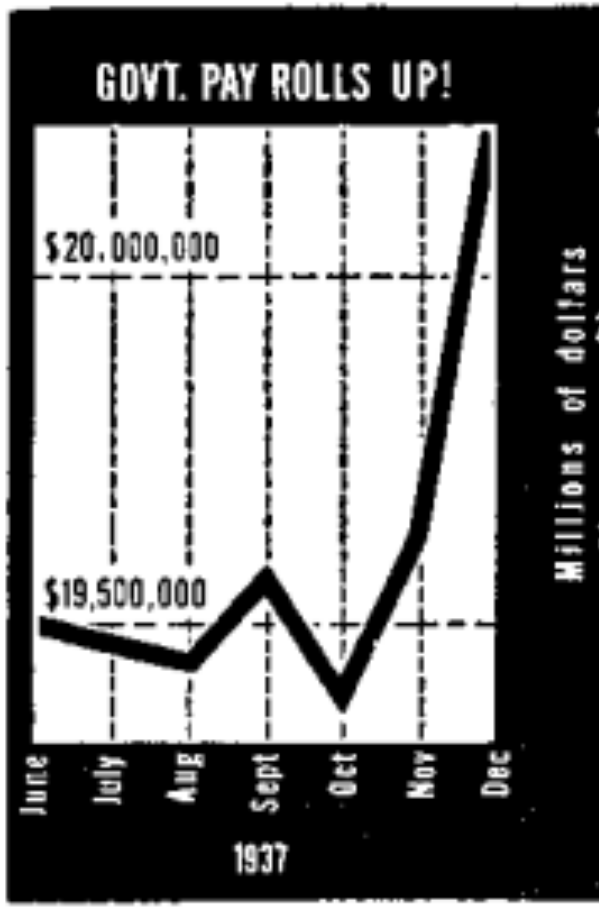
[The book was banned as training material at the VA]



Ethics of data presentation

What is potentially misleading with this figure?

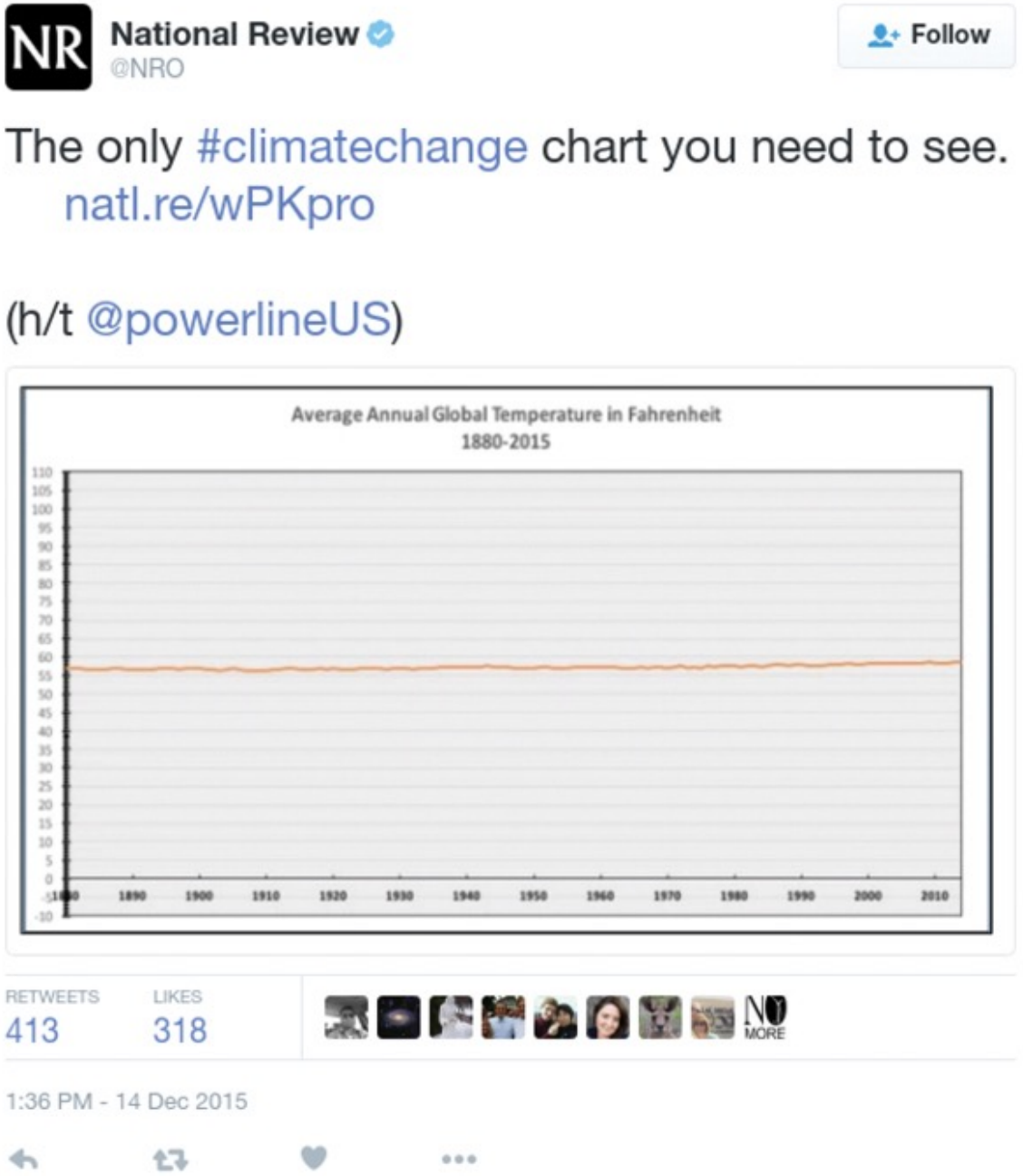
Only a 4% increase in payroll



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

How much has the climate changed?

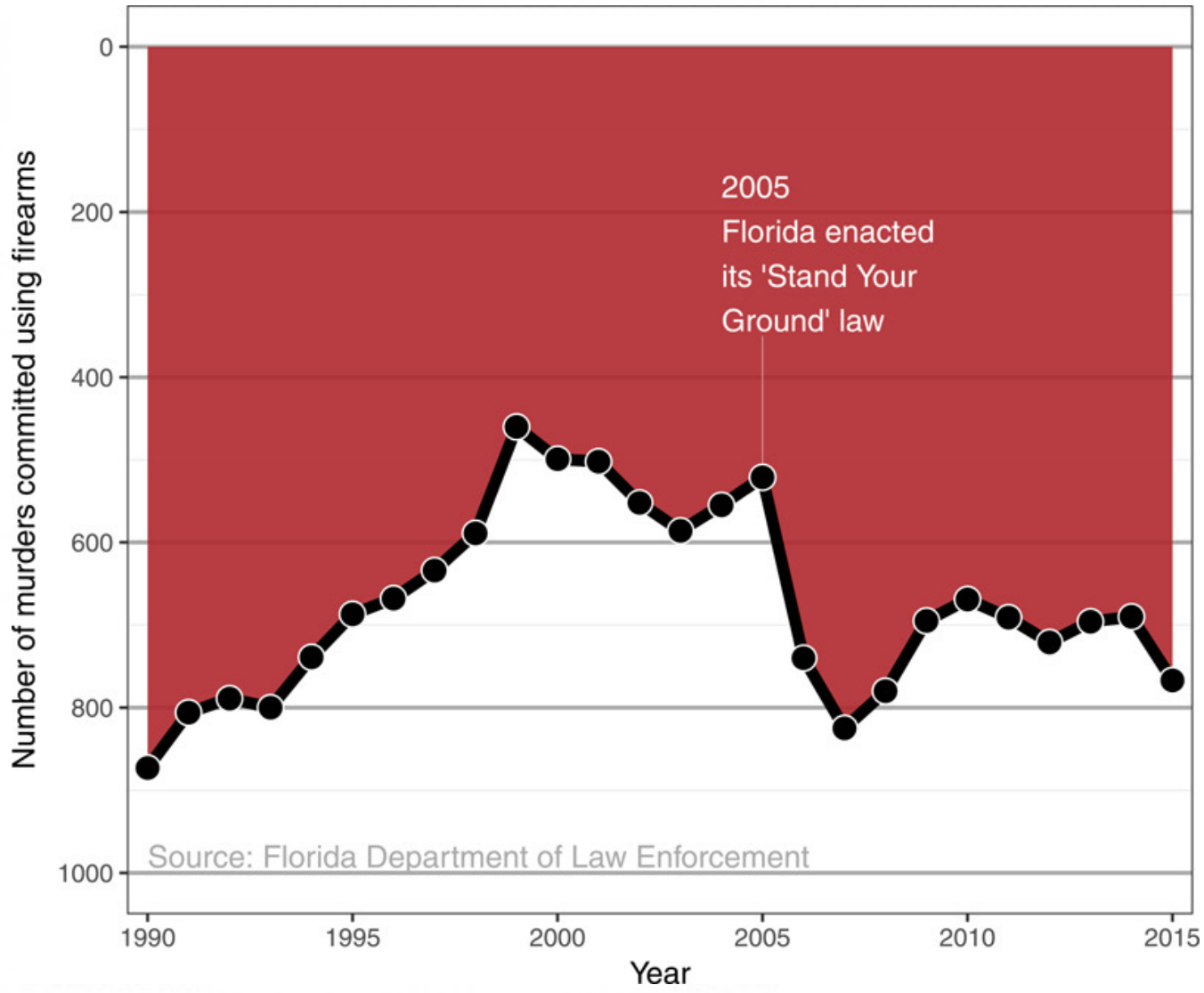
The axes go from 110 to -10 degrees which is not reasonable for the average planet temperature



Did ‘Stand Your Ground’ decrease murder by firearms?

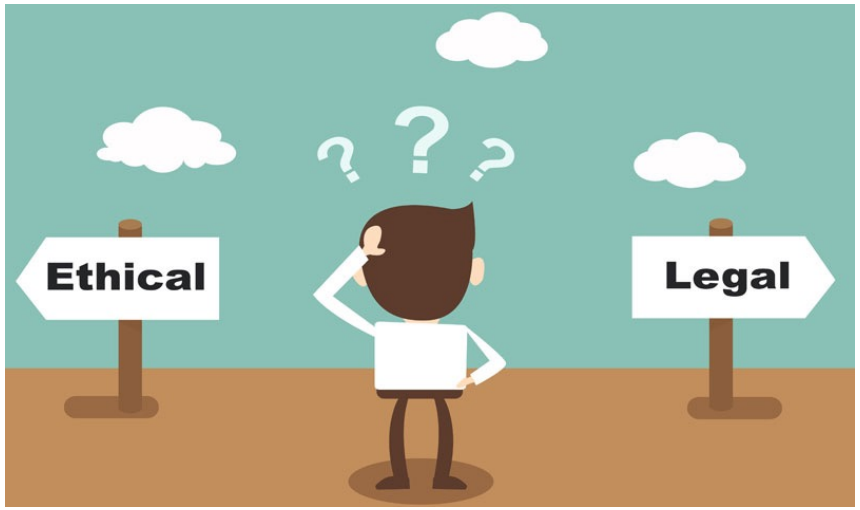
What is misleading with this figure?

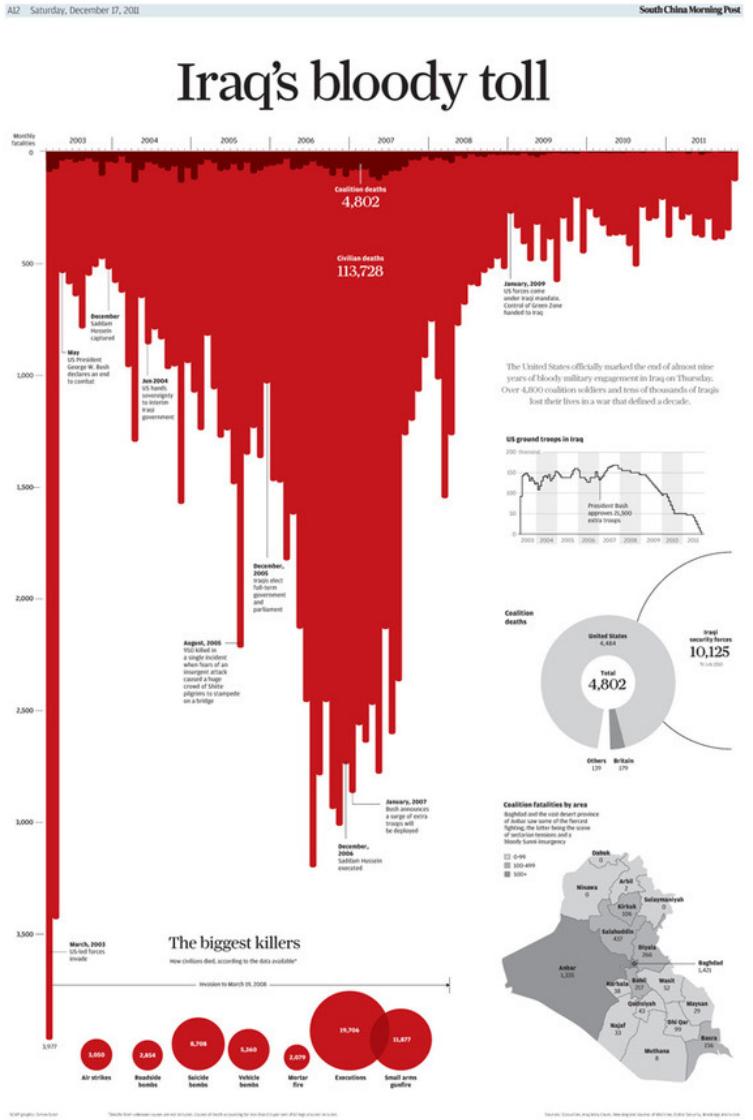
The axes are going in the wrong direction



Did ‘Stand Your Ground’ decrease murder by firearms?

The Authors' claimed they were copying the style of an NY Times infographic…





2. 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

3. 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!

4. 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 if they didn't cite you?

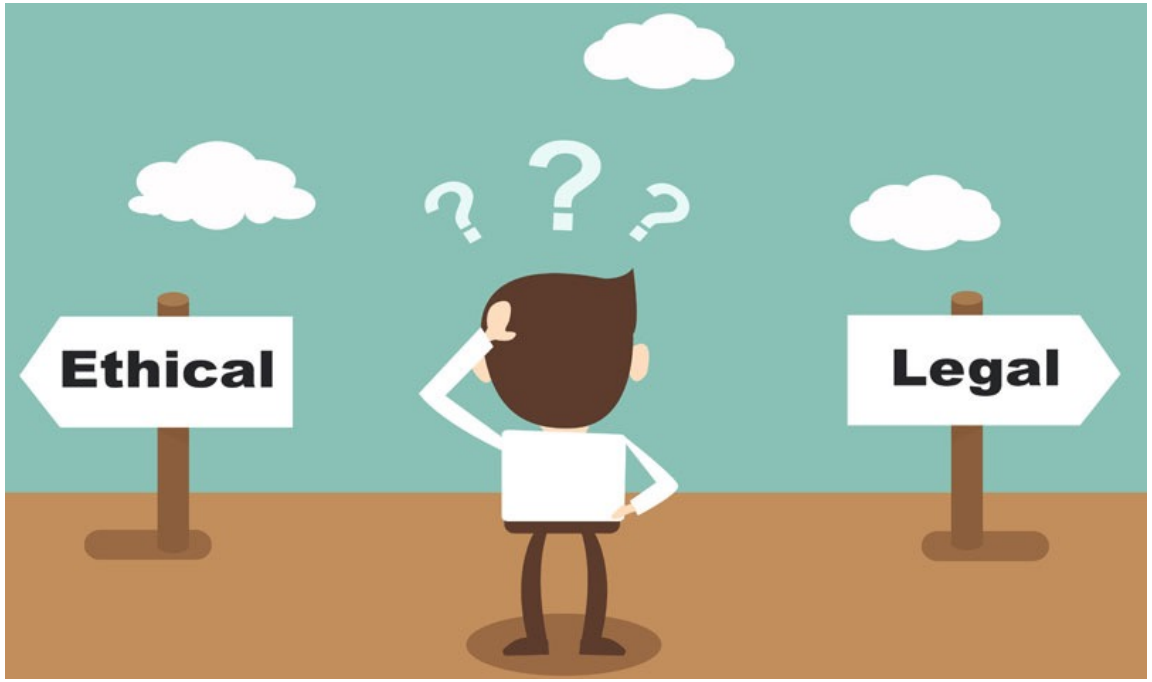
What about citing someone because they will be a reviewer of your paper?

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

5. 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



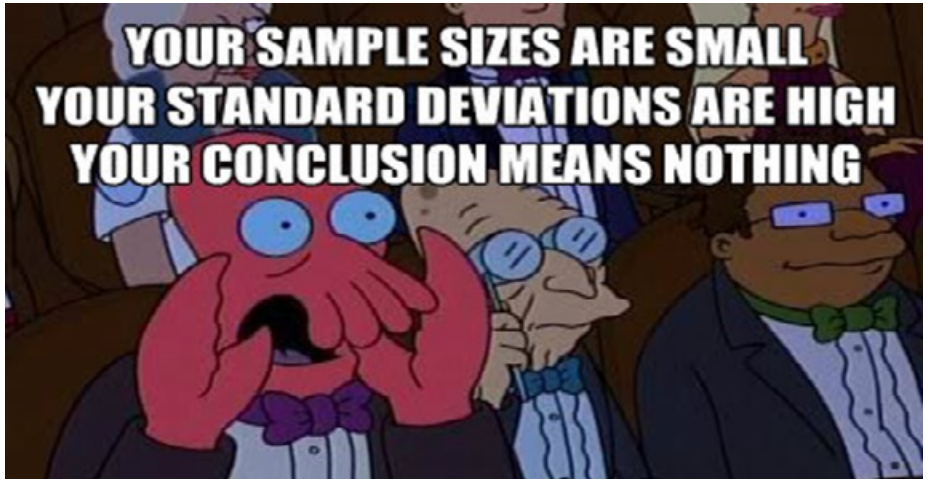
6. Ethics in Statistics

P-hacking (data dredging):

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



7. Ethics of creating powerful tools

Some prominent people are concerned about job loss due to machine learning, or even computers posing an existential threat to humans

- Is this something we should be concerned with as Data Scientists?

Ethics in machine learning

Idea: use ML to police areas with most crimes

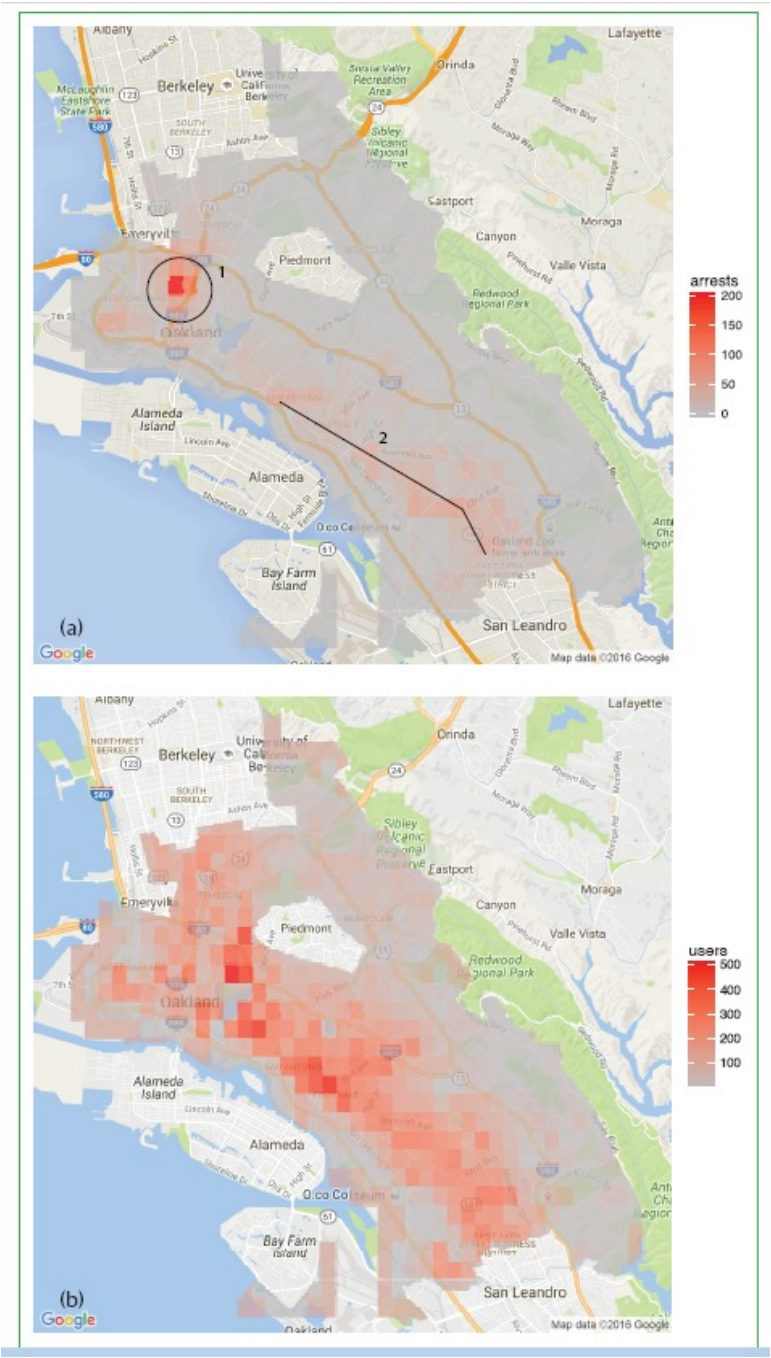
- E.g. more police where most drug arrests were made

Possible results

- Higher arrest rates for drugs found in these areas seemingly showing the ML algorithm is working

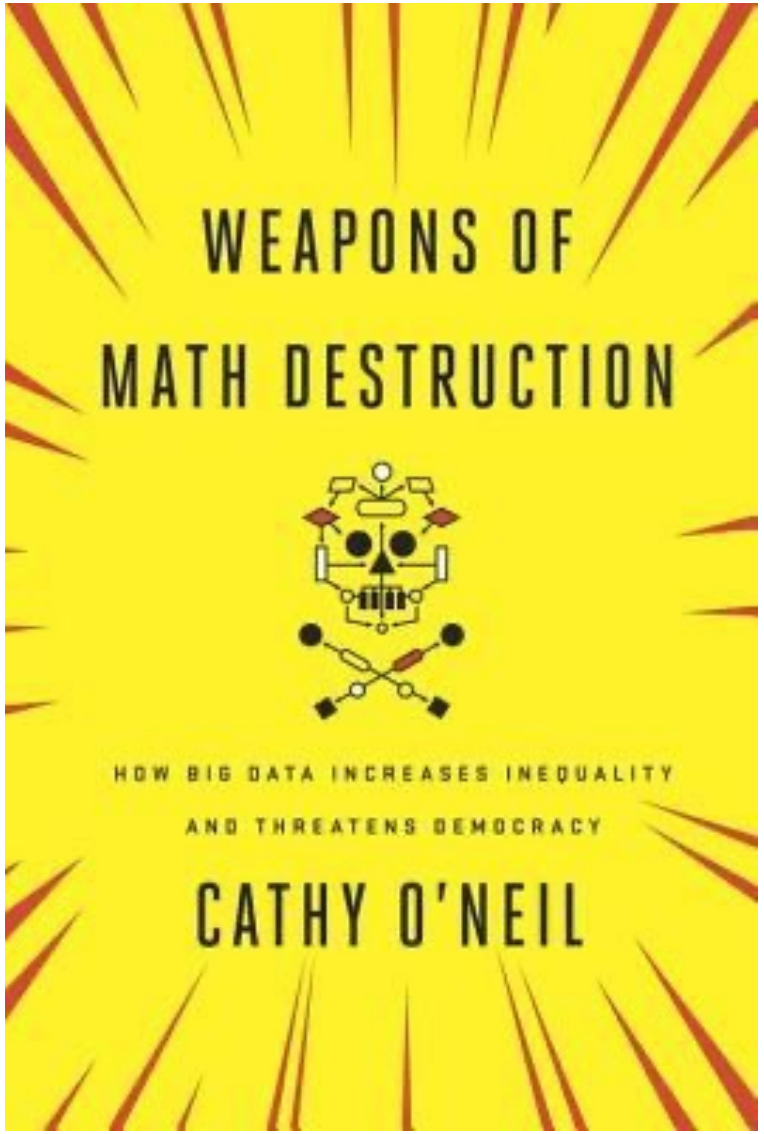
Any potential problems with this?

- Are more arrest made because drug use is higher or because there are more police



Additional reading

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

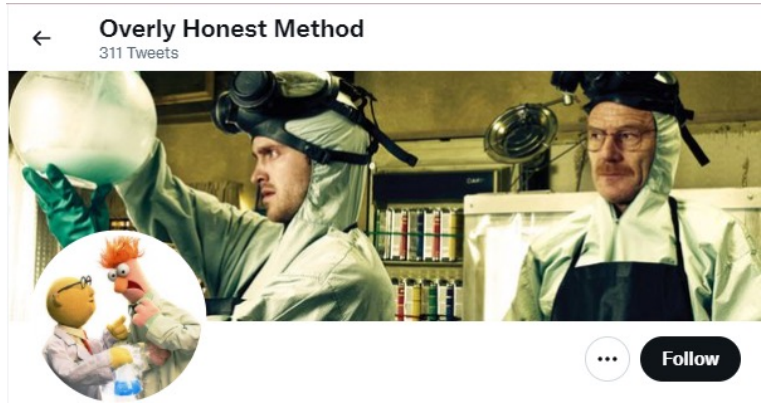


Wrap up and conclusions

Wrap up and conclusions

Congratulations, you survived the semester!

About this class



Complete reorganization of the class from previous years

Focused more on giving you real skills

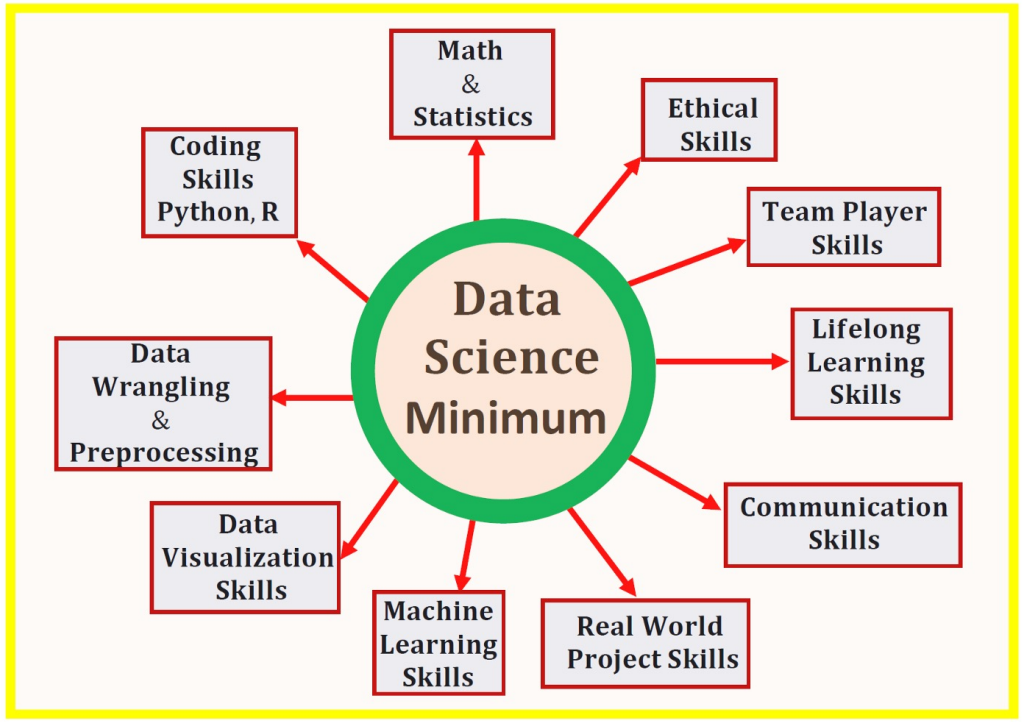
- We will use real Python data science packages instead of Berkeley's datascience package

We will will hit rough spots!

- e.g., some HW might be too easy/hard, etc.
- I will ask for a lot of feedback!

Topics covered

- ~~What is Data Science~~
- ~~Python basics~~
- ~~Array/Matrix computations~~
- ~~Data manipulation/wrangling~~
- ~~Data visualization~~
- ~~Mapping~~
- ~~Text manipulation and data cleaning~~
- ~~Statistical perspective: hypothesis tests and confidence intervals~~
- ~~Machine learning perspective: supervised and unsupervised learning~~



Learning goals

1. Understand concepts in Data Science
    - Learn basic computational skills for analyzing data
    - Understand concepts in Statistics and Machine Learning

2. Gain practical Data Science skills applicable to any domain

3. See how Data Science analyses can be applied to real-world data from a variety of domains
    - There will be ~weekly readings on Data Science related topics

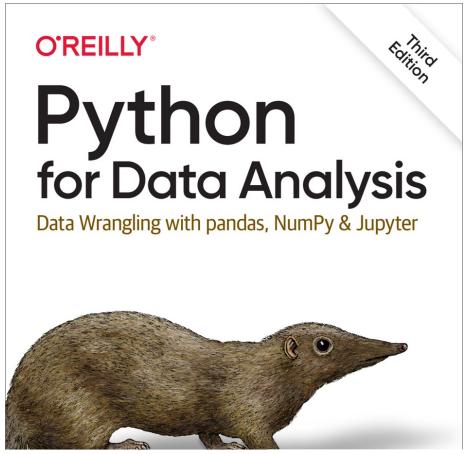
There are no prerequisites for this class
    - E.g., no prior knowledge of Statistics or Programming is required

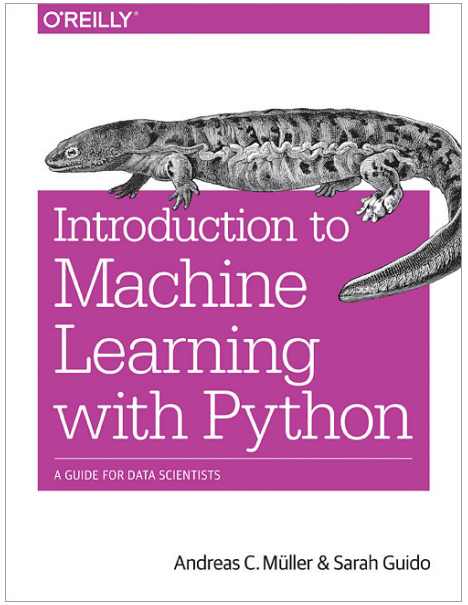
We had fun along the way???

Next steps

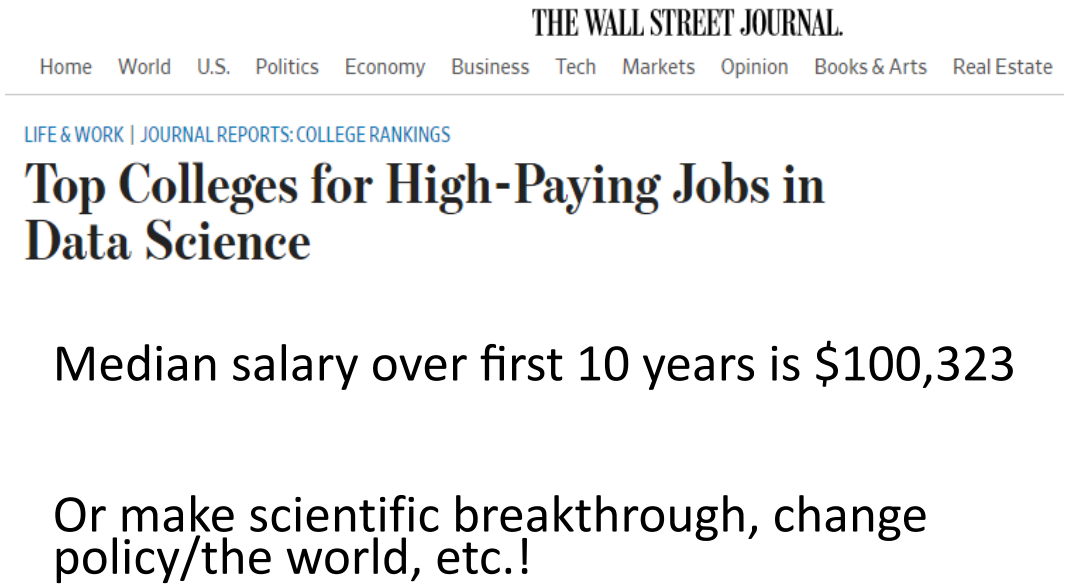
1. Take more advanced Statistics and Data Science classes offered at Yale!

2. There are many good books and online resources to learn more Python





3. Profit!



Median salary over first 10 years is $100,323

Or make scientific breakthrough, change policy/the world, etc.!

Teaching Assistants

Teaching Fellows

- Weiyi Li: weiyi.li@yale.edu

Undergraduate Learning Assistants

- Dani Mekuriaw: dani.mekuriaw@yale.edu
- Irine Juliet Otieno juliet.otieno@yale.edu
- Mark Ayiah mark.ayiah@yale.edu
- Rose Bae rose.bae@yale.edu
- Vivian Vasquez vivian.vasquez@yale.edu

Good luck with the end of the semester!

Good luck finishing your final projects!

Review session May 2nd at 2:30-3:30

Data Science vs. Statistician

Greek national anthem