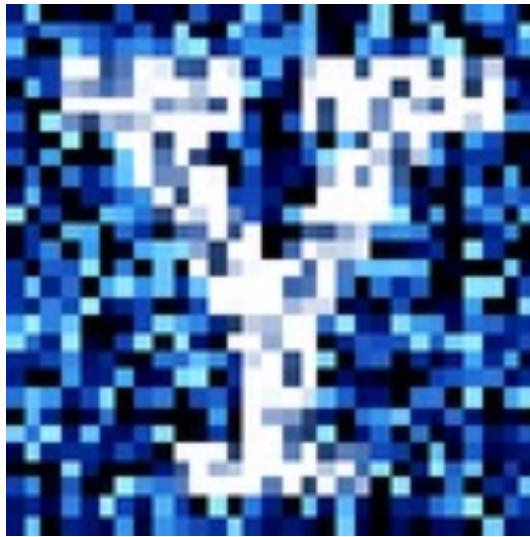


YData: Introduction to Data Science



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

Overview

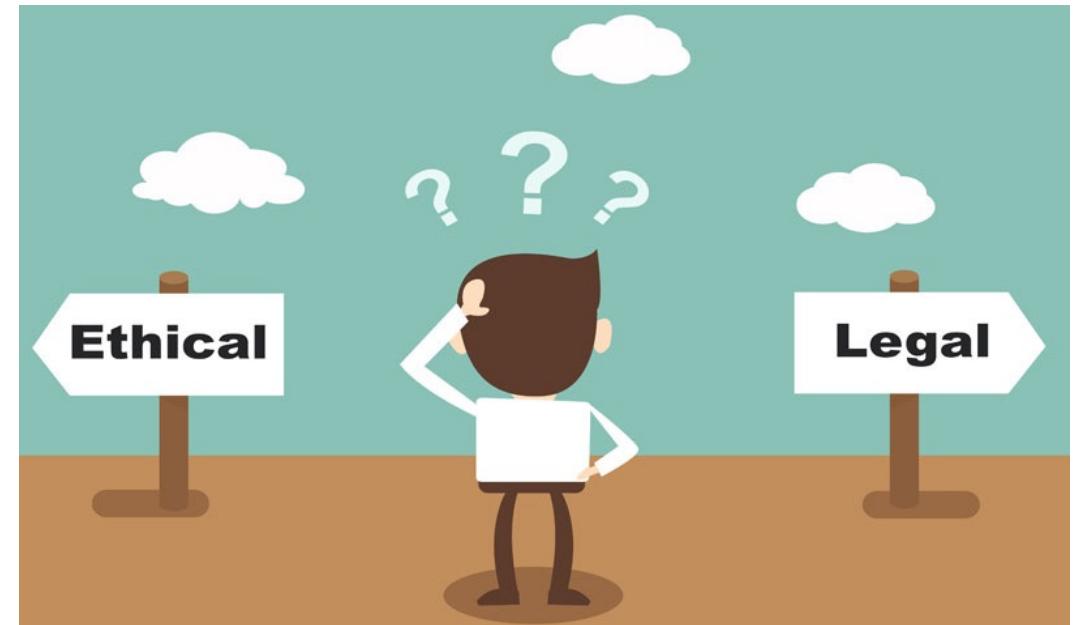
Unsupervised learning/clustering

- Review: k-means cluster
- Hierarchical clustering

Brief demo of running a LLM/chatbot
in Python

Ethics

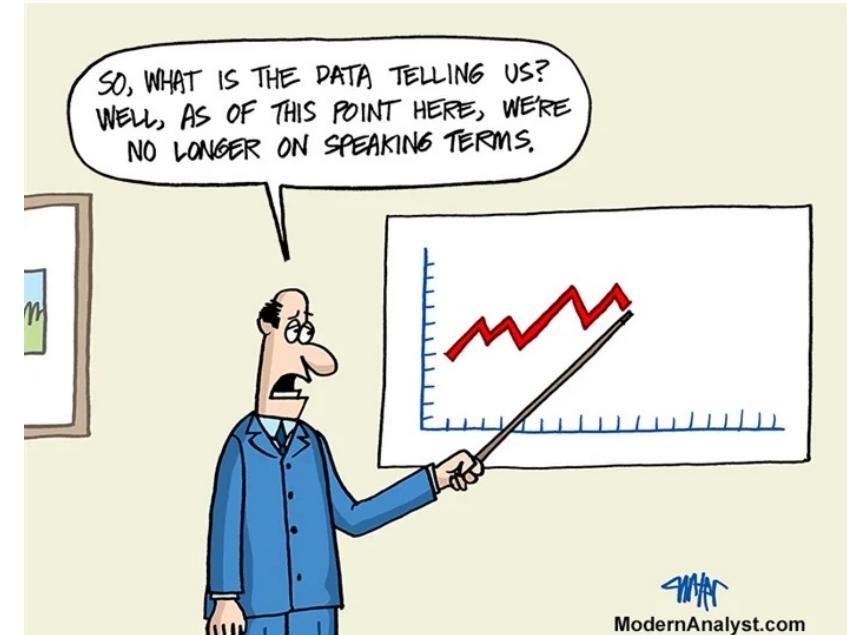
Wrap-up



Project timeline

Sunday, April 28th

- 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

Announcement

Exam review session: Tuesday, April 30th from 4-5pm in DL 220

Final exam: May 5th at 2pm in Davies Auditorium



Unsupervised learning

Supervised learning and unsupervised learning

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

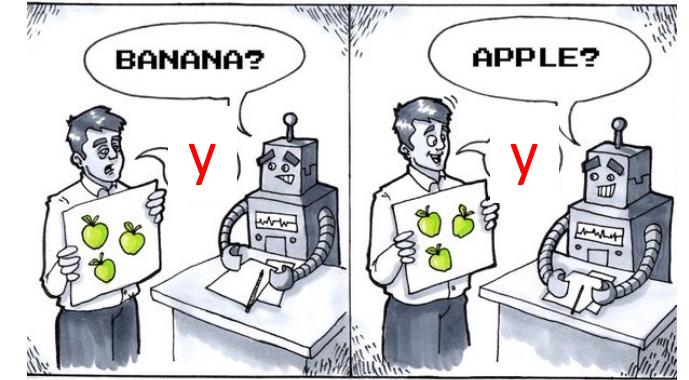
- We use the features X to predict y on new data

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

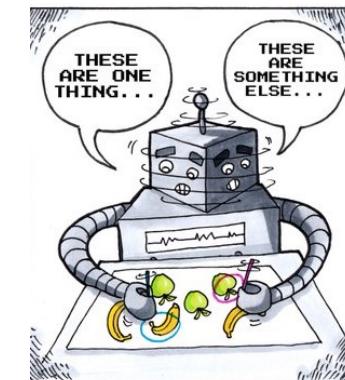
Unsupervised learning can be useful in order to find structure in the data and to visualize patterns

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



Supervised Learning



Unsupervised Learning

Clustering review and continuation



So tell me how
many clusters do
you see?



How many clusters?



Six Clusters



Two Clusters

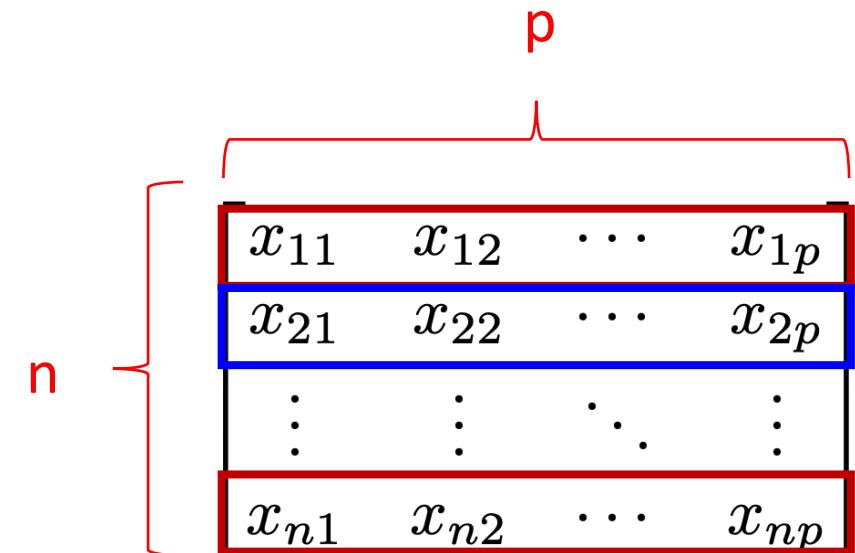


Four Clusters

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



Clustering can be:

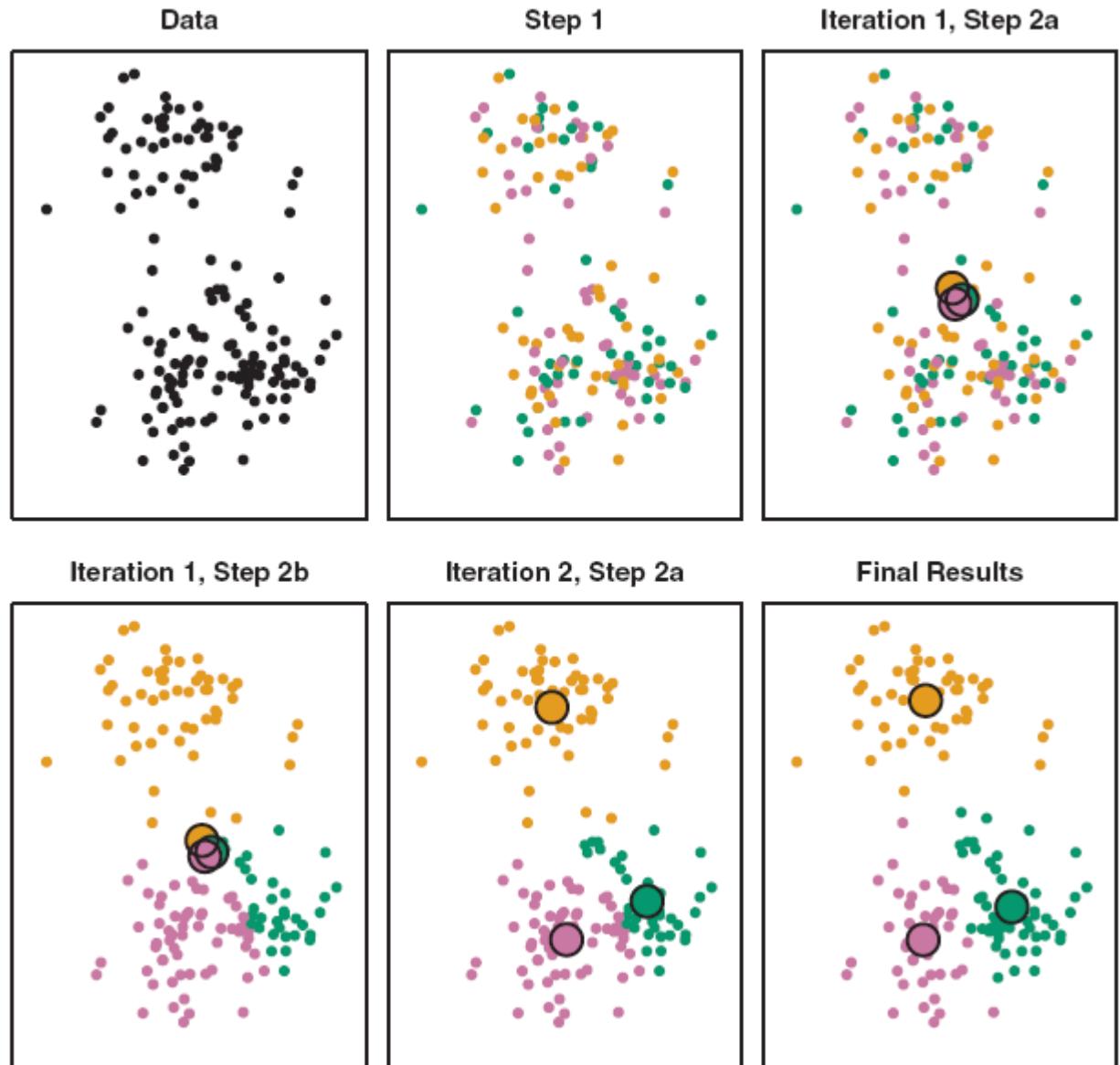
- **Flat:** no structure beyond dividing points into groups
- **Hierarchical:** Population is divided into smaller and smaller groups (tree like structure)

Review: K-means clustering

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

- The number of clusters, k , needs to be specified prior to running the algorithm

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

```
from sklearn.cluster import KMeans
```

```
kmeans = KMeans(n_clusters = 3)
```

```
kmeans.fit(X_penguin_features)
```

```
kmeans.predict(X_penguin_features)
```



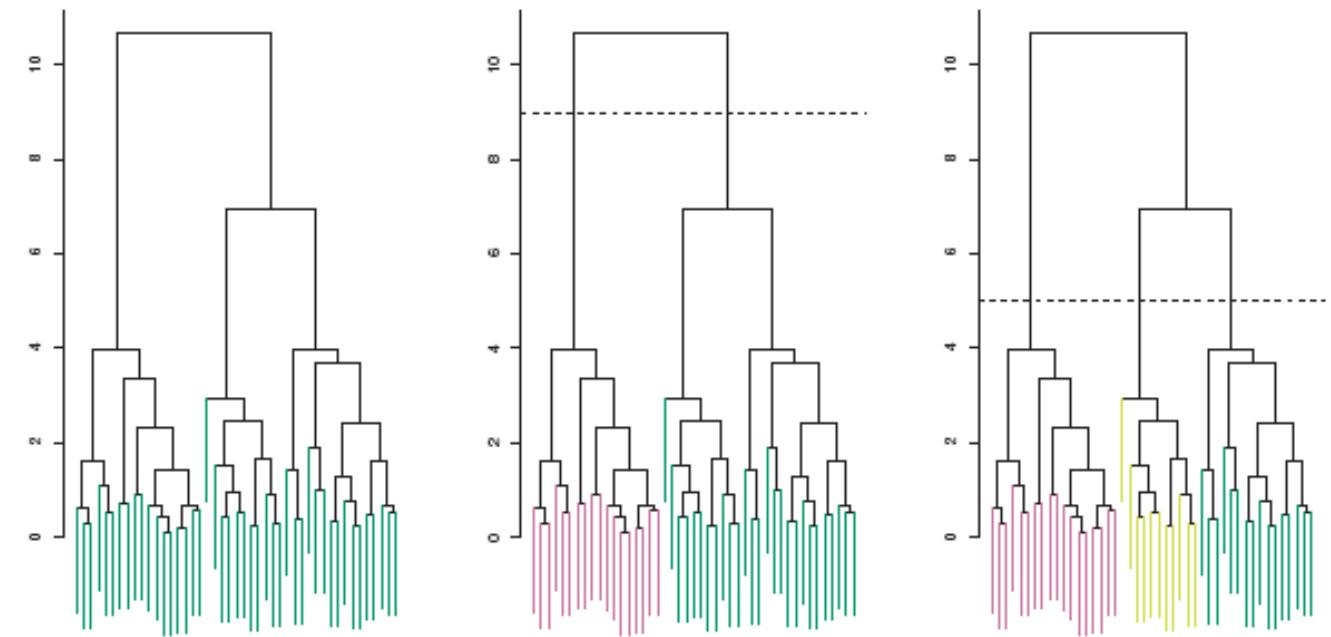
Hierarchical clustering

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

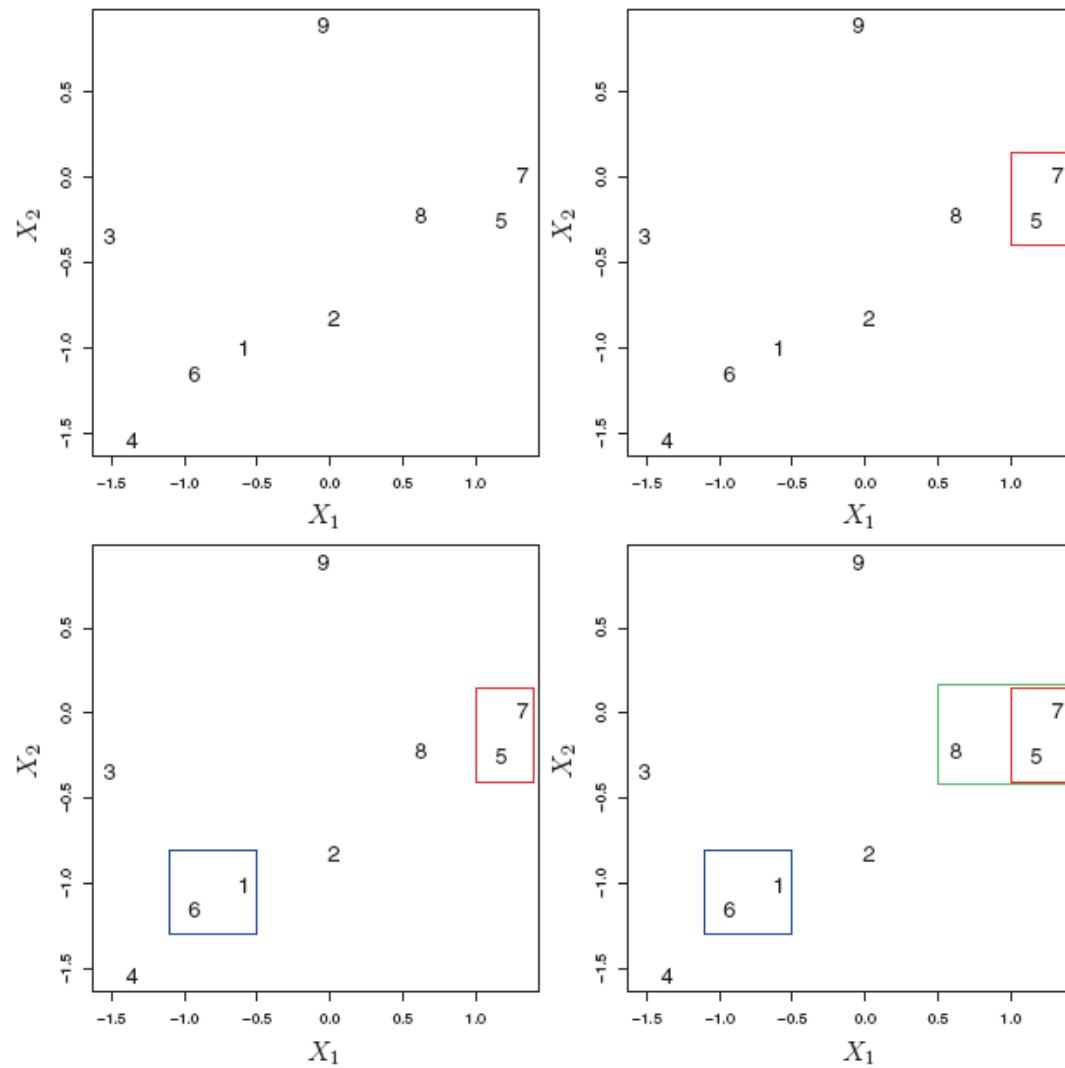
- i.e., don't need to specify the number of clusters, K , beforehand



Hierarchical clustering

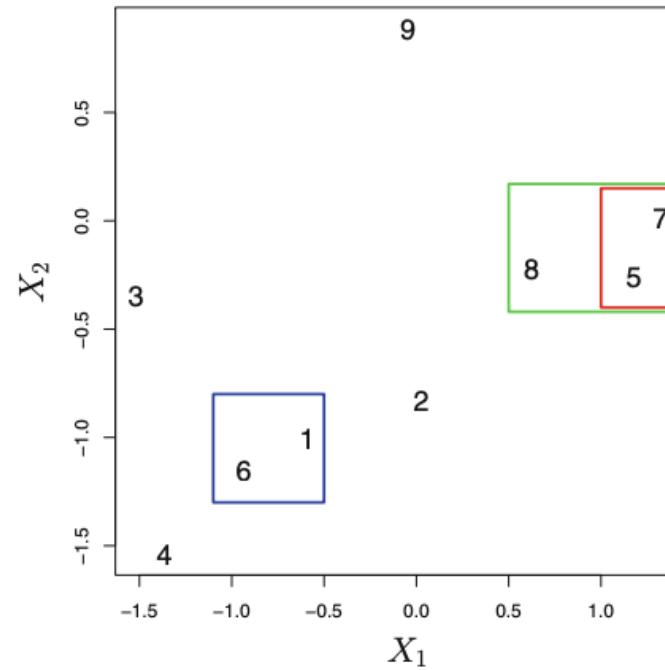
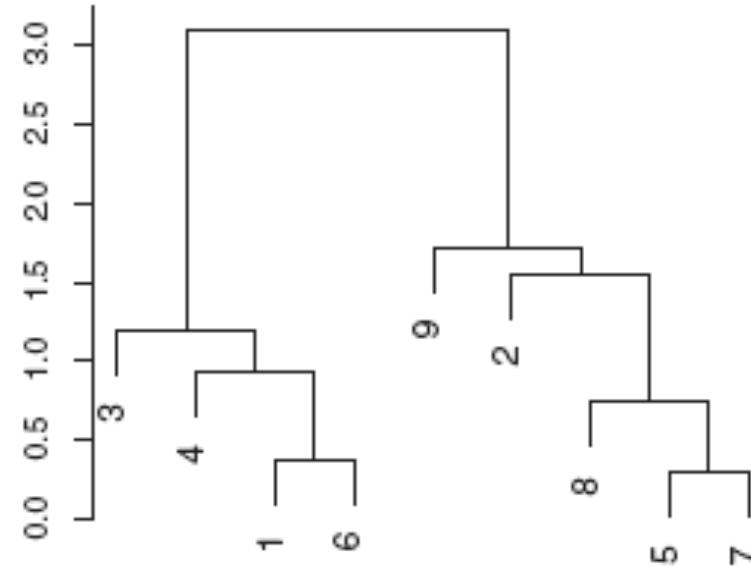
We can create a hierarchical clustering of the data using simple bottom-up agglomerative algorithm:

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

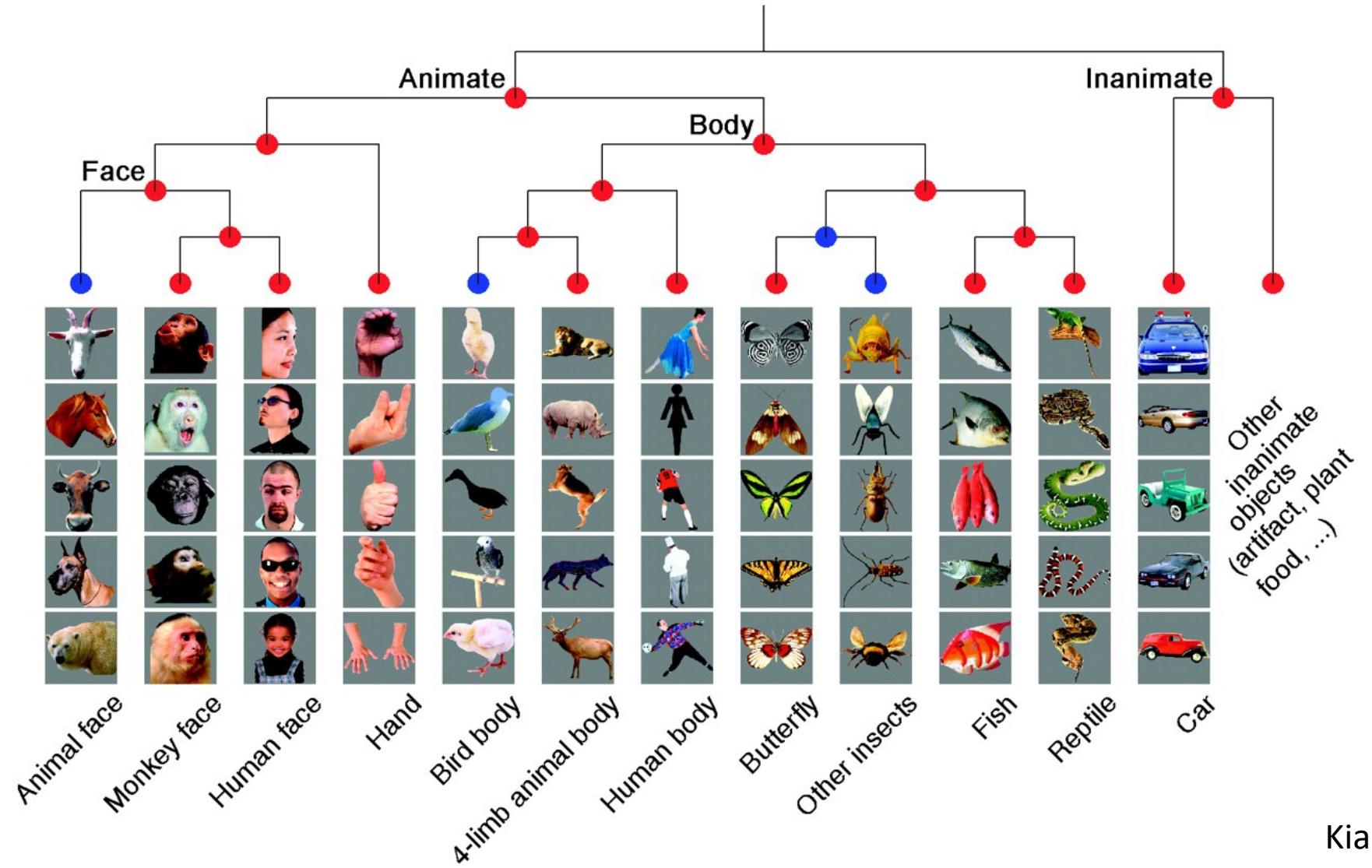
The vertical height that two clusters/points merge show how similar the two *clusters* are



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

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

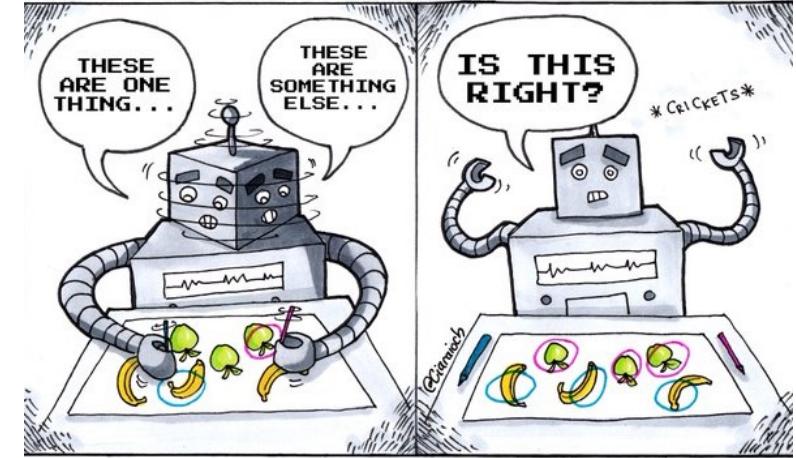
Hierarchical clustering example



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



Unsupervised Learning

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!

Brief discussion of Large Language Models

Brief discussion of Large Language Models

Large language models (LLMs) are taking over the world

- They can write code and even analyze data

One can download free, open source, LLMs through the Hugging Face platform

Let's very briefly look at running a LLM locally on our own computers...

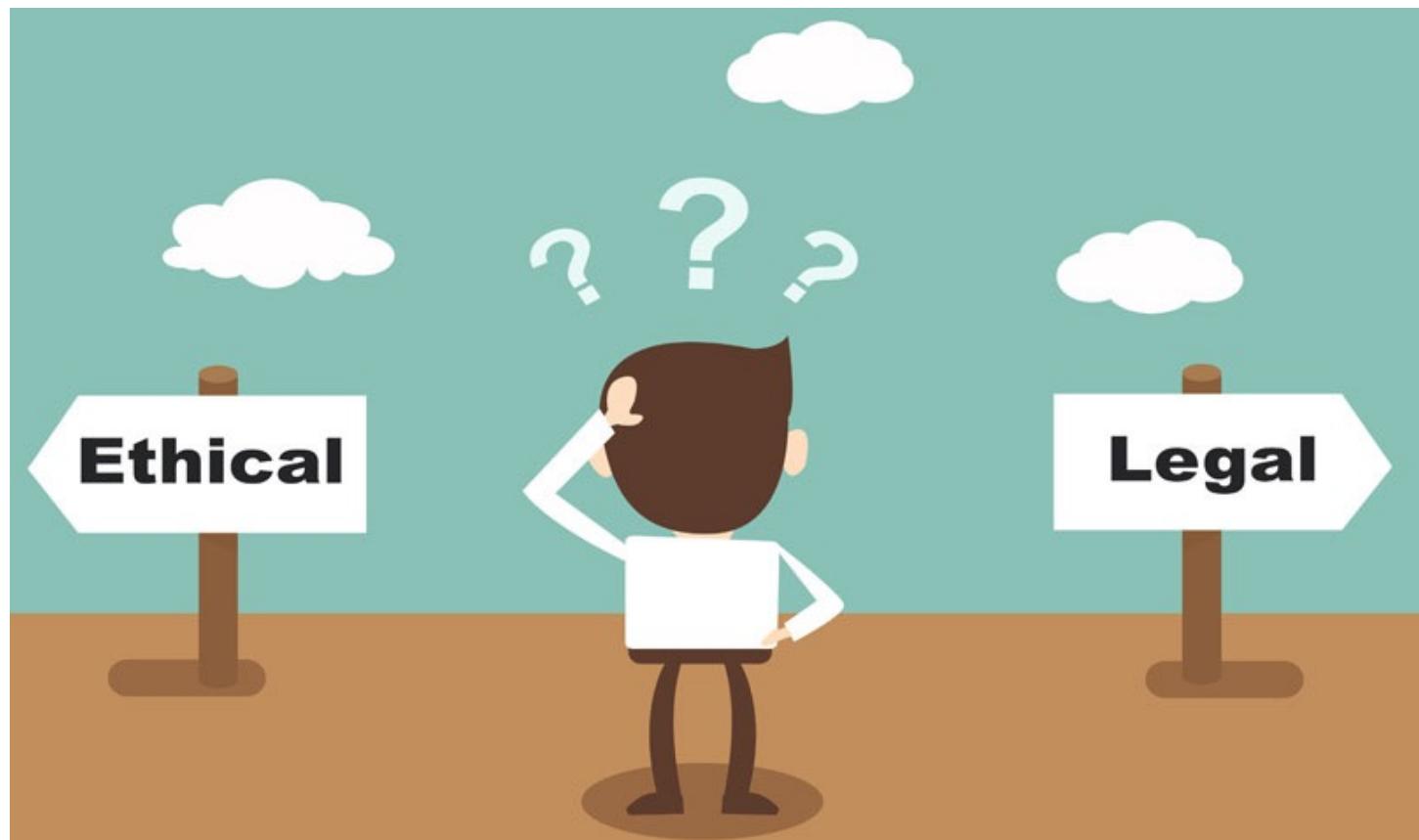


ChatGPT

Gemini



Ethics



Ethics in Data Science

Ethics of:

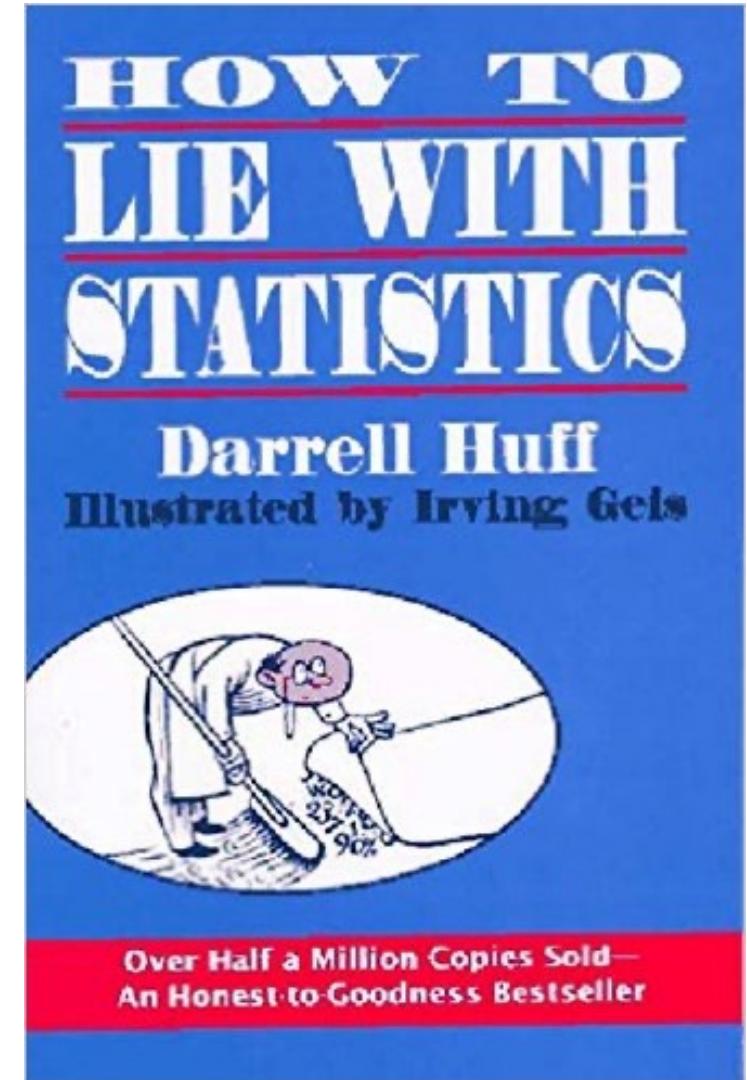
1. Data presentation
2. Data scraping TOS and privacy
3. Reproducibility
4. Citations/peer review
5. Disclosure
6. Ethics in Statistical analyses
7. Ethics of creating powerful tools

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

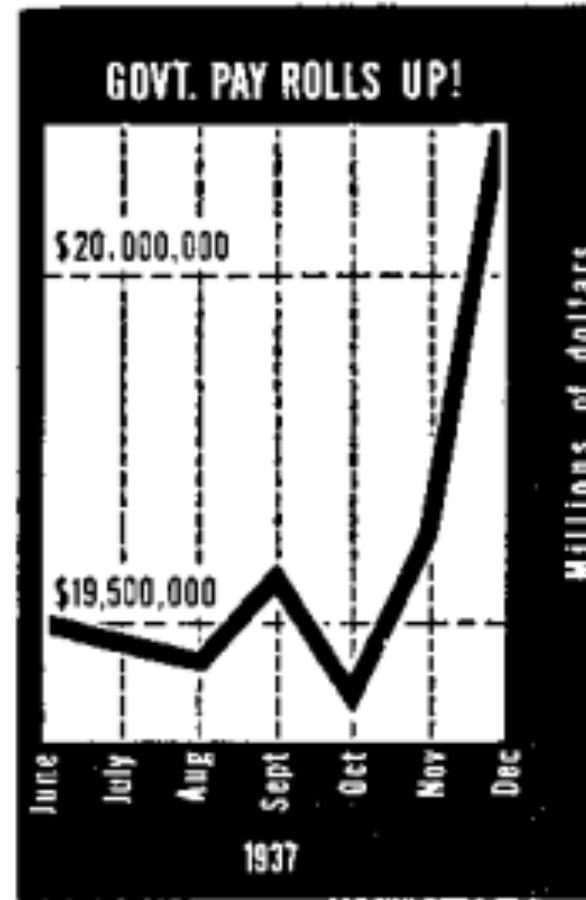
The book was banned as training material at the VA



Ethics of data presentation

What is potentially misleading with this figure?

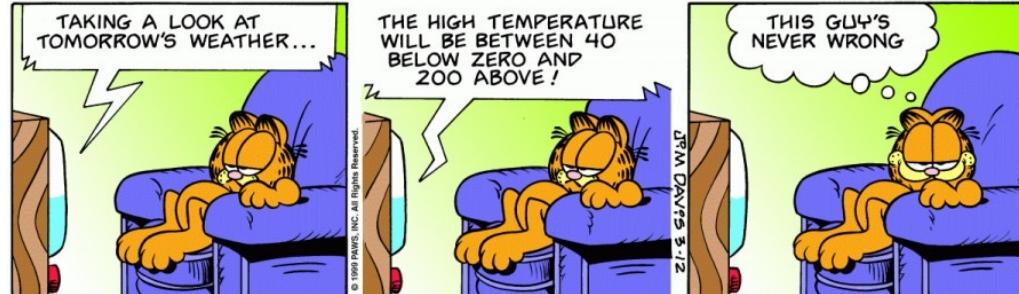
Only a 4% increase in payroll



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



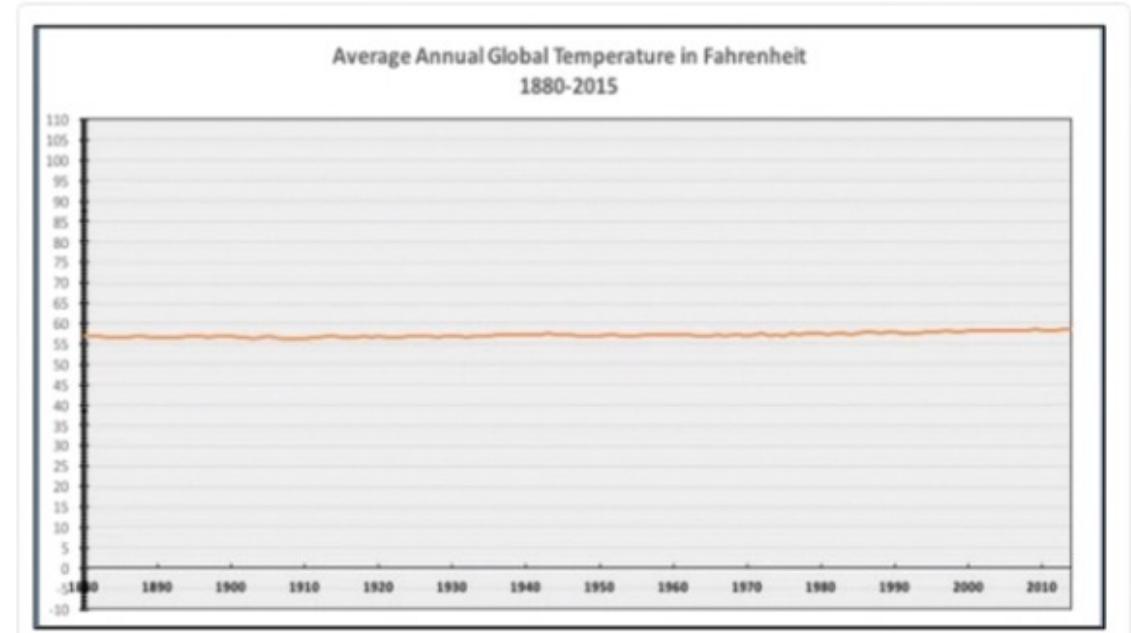
National Review

@NRO

Follow

The only #climatechange chart you need to see.
natl.re/wPKpro

(h/t [@powerlineUS](#))



RETWEETS

413

LIKES

318



NO
MORE

1:36 PM - 14 Dec 2015

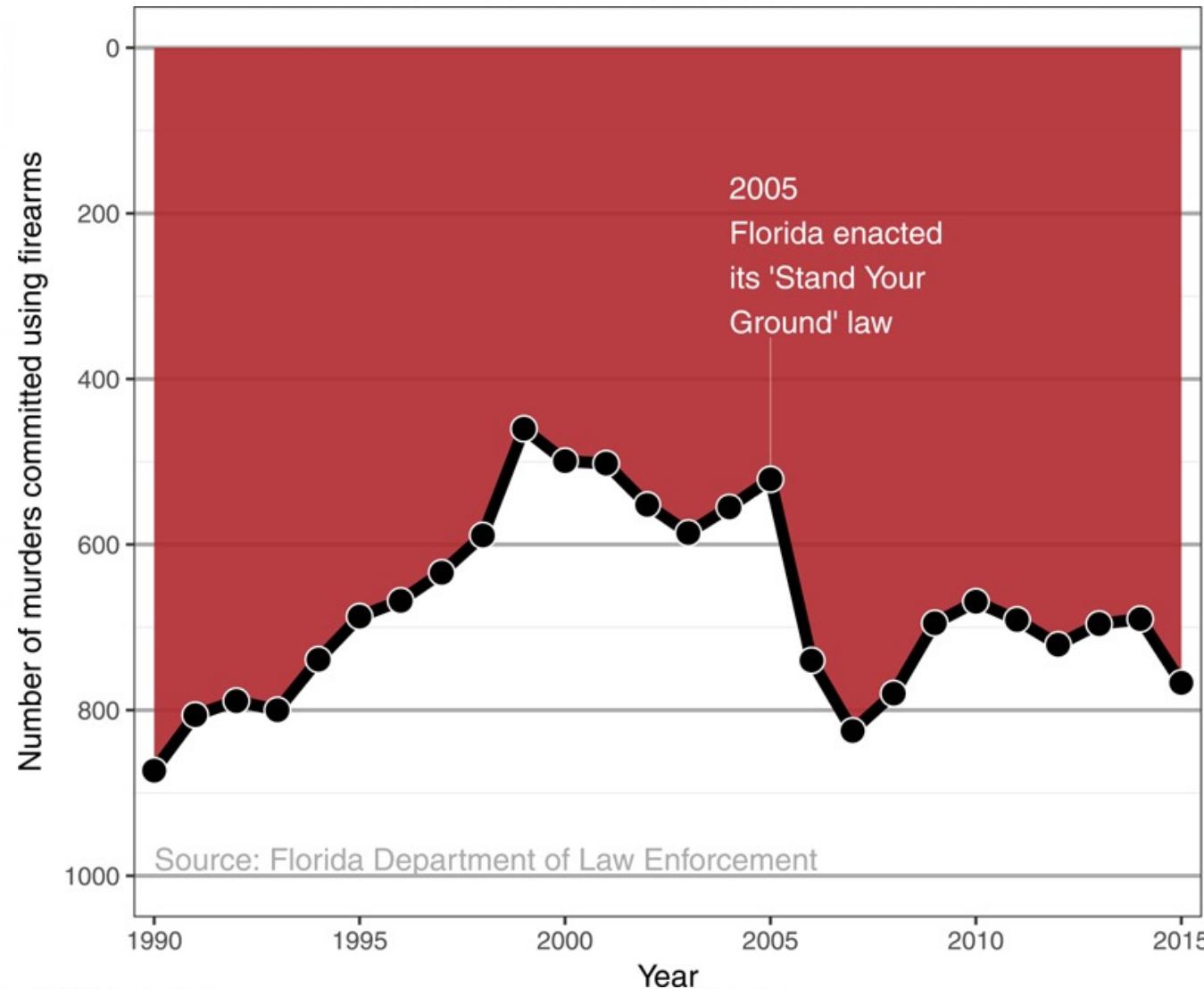


...

Did 'Stand Your Ground' decrease murder by firearms?

What is misleading
with this figure?

The axes are going in
the wrong direction

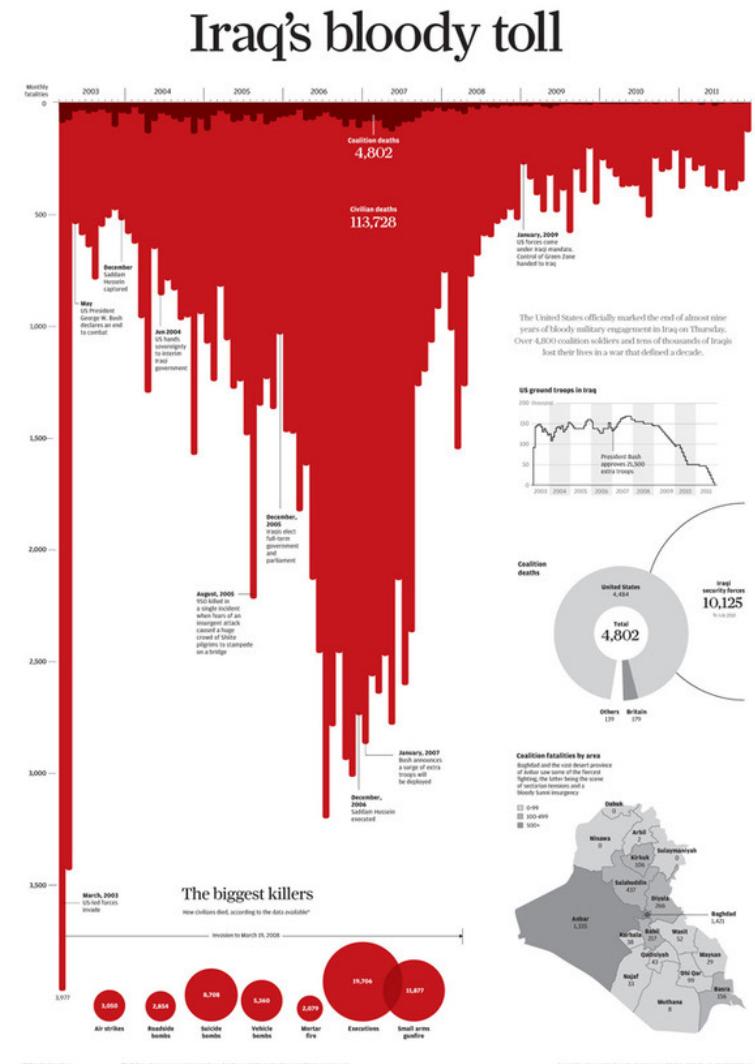
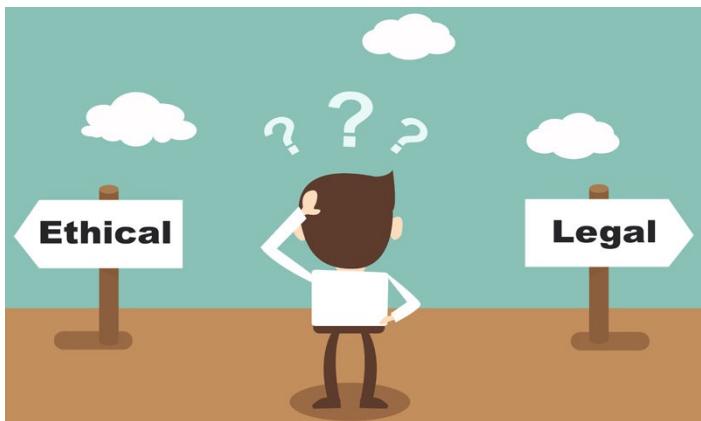


Did 'Stand Your Ground' decrease murder by firearms?

A12 Saturday, December 17, 2011

South China Morning Post

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.

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

- Is this ok?

3. Reproducibility

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

The screenshot shows a research article page from the journal **nature methods**. At the top, the journal logo is displayed in white on a dark blue background. Below the logo, the title of the article is shown in a smaller white font. To the right of the title, there is a note about access provided by Massachusetts Institute of Technology. Below the title, there is a horizontal bar containing several pieces of information: a blue progress bar, the text "Altmetric: 5", the text "Citations: 5", and a link "More detail >".

Access provided by Massachusetts Institute of Technology

Altmetric: 5 Citations: 5 More detail >

Commentary

Ethical reproducibility: towards transparent reporting in biomedical research

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!



[NEWSLETTERS](#) [SIGN IN](#) [NPR SHOP](#)

[NEWS](#) [CULTURE](#) [MUSIC](#) [PODCASTS & SHOWS](#) [SEARCH](#)

EDUCATION

Harvard professor who studies dishonesty is accused of falsifying data

JUNE 26, 2023 · 1:15 PM ET

 Juliana Kim



Francesca Gino has been teaching at Harvard Business School for 13 years.
Maddie Meyer/Getty Images

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



Concerns about Justice Clarence Thomas' disclosures are sent to judicial panel

A pair of Democratic lawmakers asked the U.S. Judicial Conference to weigh in on Thomas' failure to disclose gifts and luxury travel he allegedly accepted from a GOP billionaire.



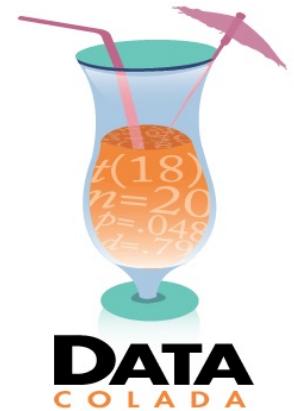
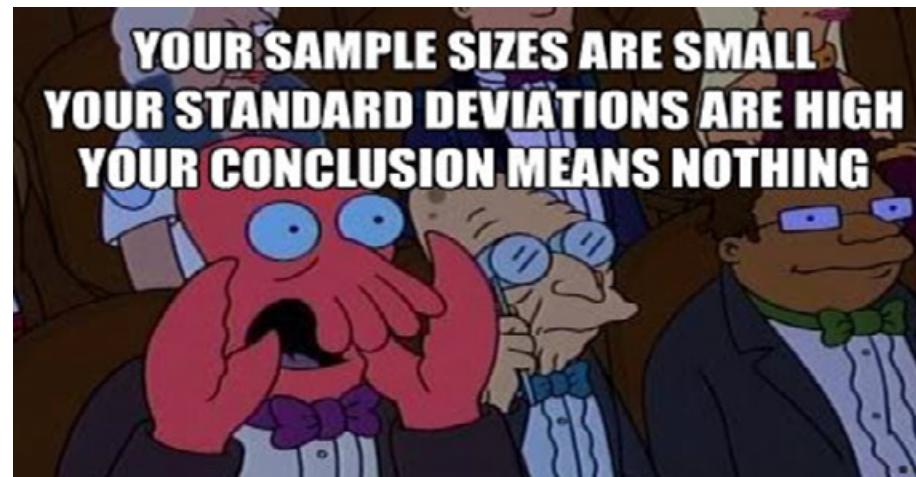
6. Ethics in Statistics

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

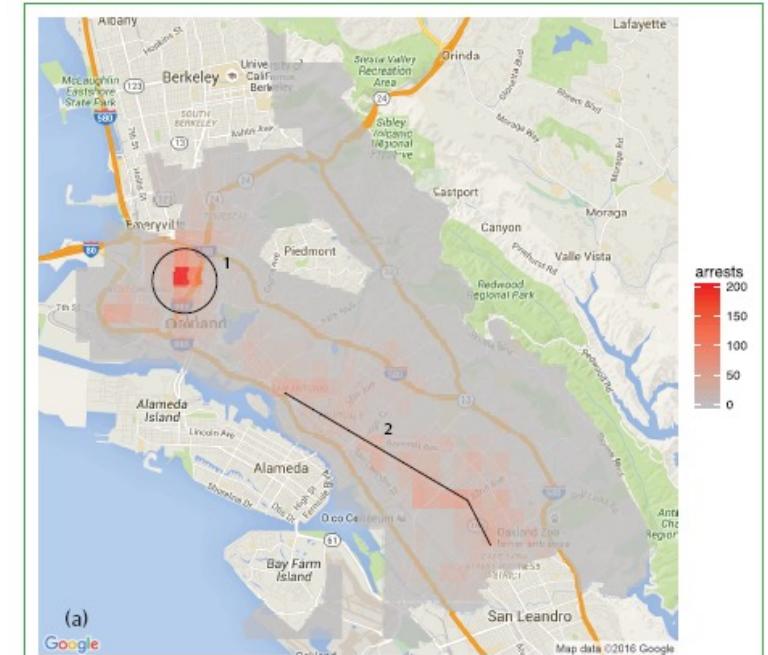
- Is this something we should be concerned with as Data Scientists?



Ethics in machine learning

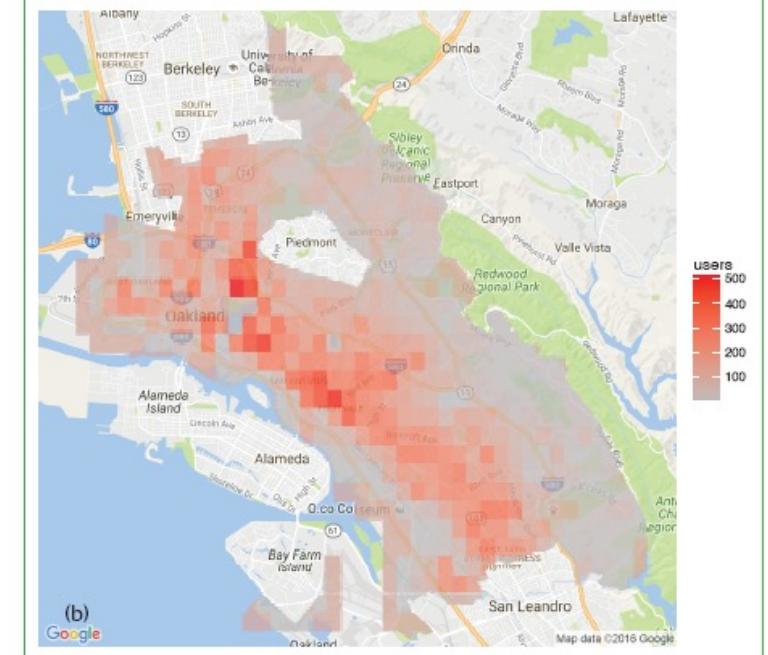
Idea: use ML to police areas with most crimes

- 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

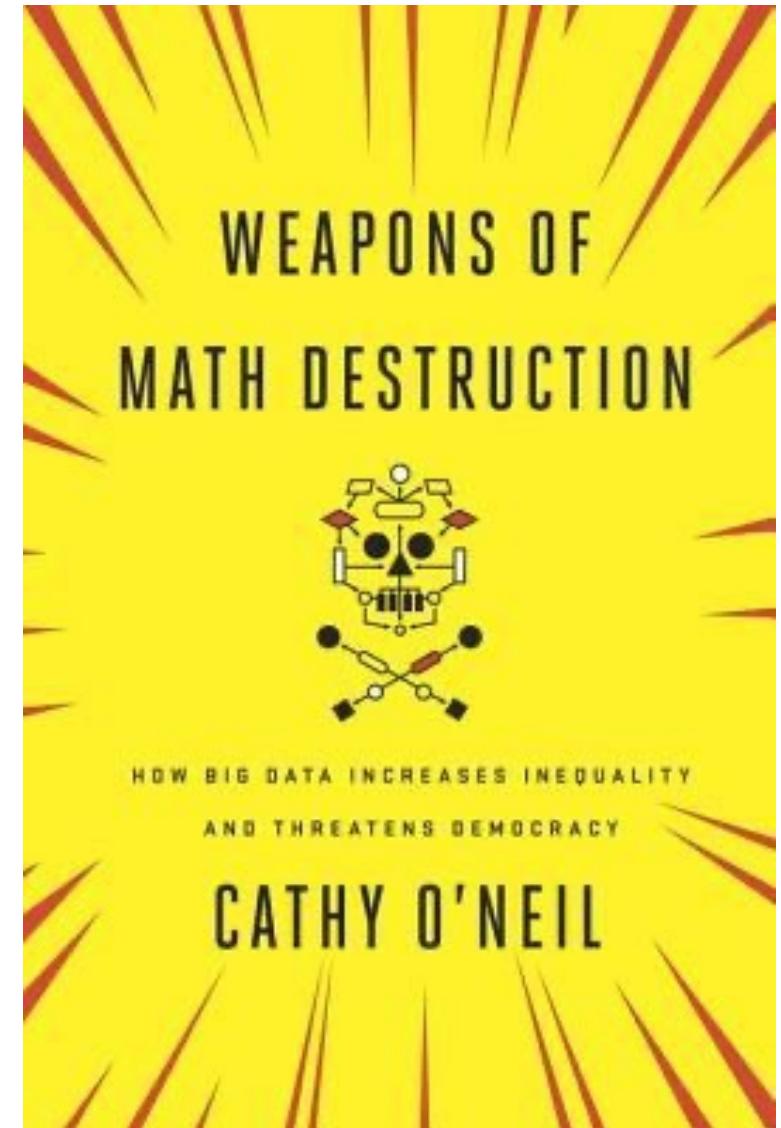


Any potential problems with this?

- Are more arrests 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](https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end)



Wrap up and conclusions

Topics covered

~~What is Data Science~~

~~Python basics~~

~~Array/Matrix computations~~

~~Data manipulation/wrangling~~

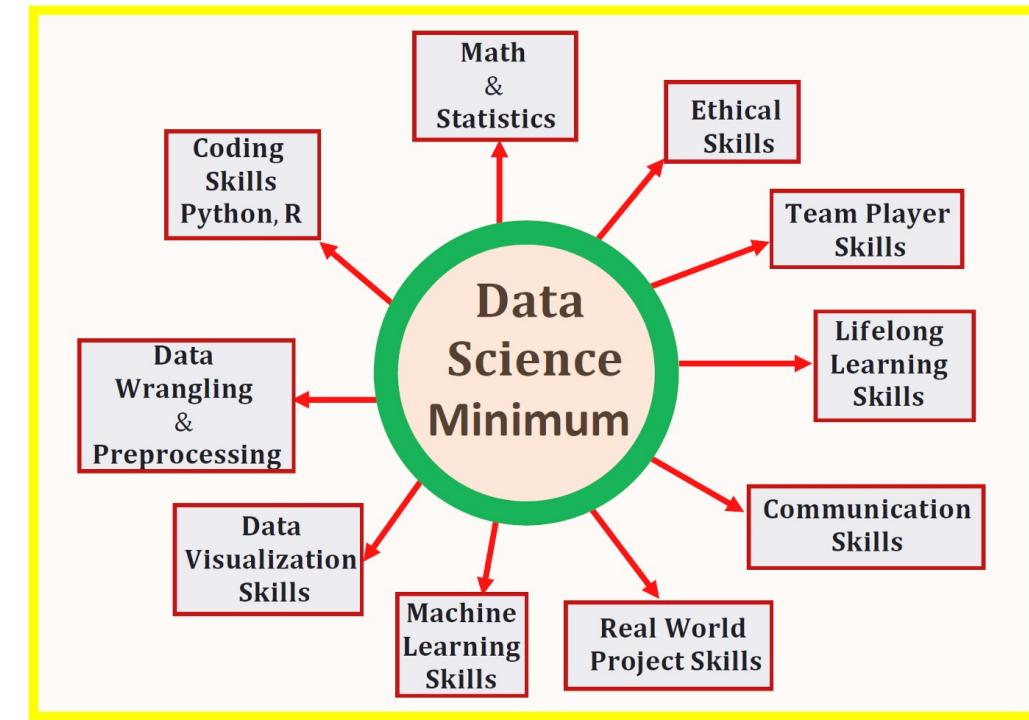
~~Data visualization~~

~~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

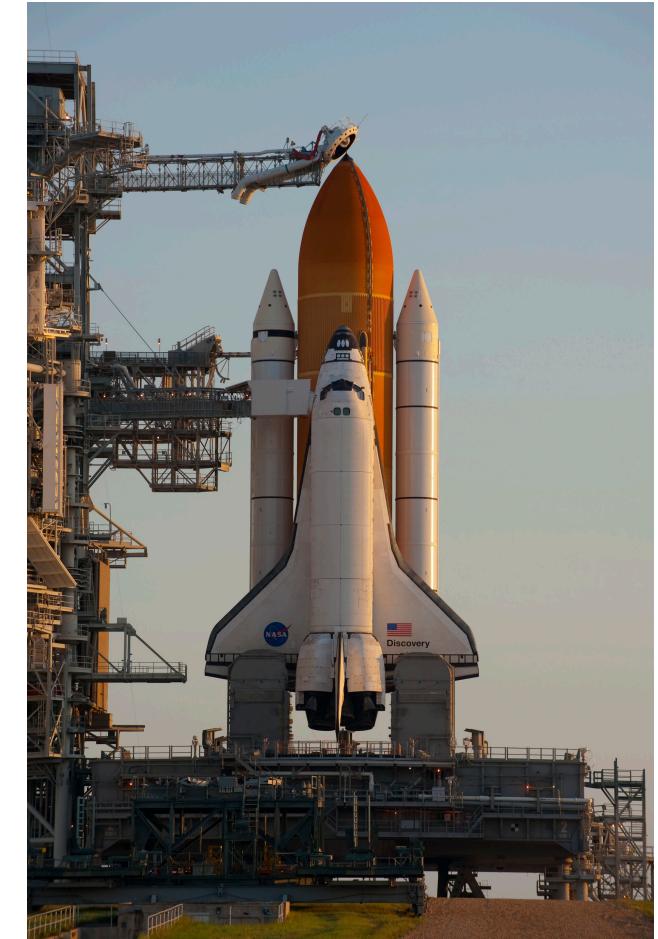
2. Gain practical data science skills applicable to any domain

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

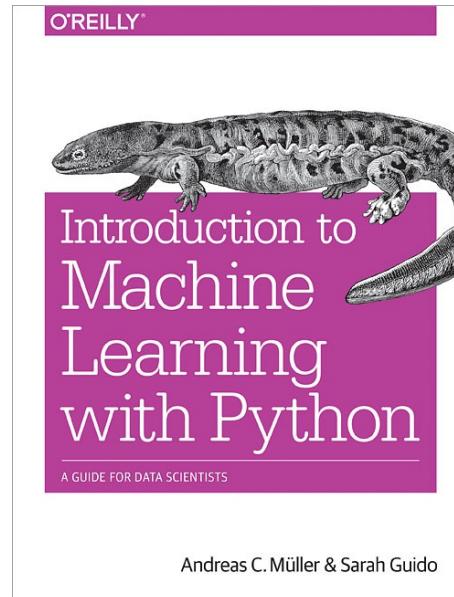
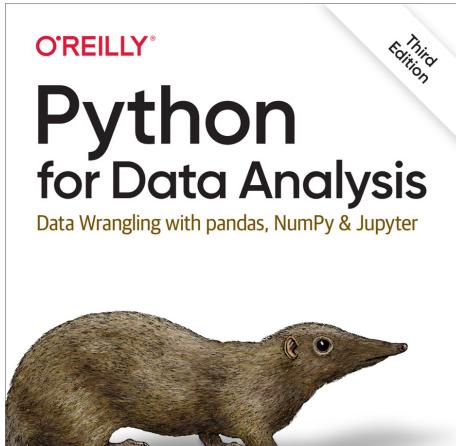
There are no prerequisites for this class

- E.g., no prior knowledge of statistics or programming is required



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!

THE WALL STREET JOURNAL

[Home](#) [World](#) [U.S.](#) [Politics](#) [Economy](#) [Business](#) [Tech](#) [Markets](#) [Opinion](#) [Books & Arts](#) [Real Estate](#)

LIFE & WORK | JOURNAL REPORTS: COLLEGE RANKINGS

Top Colleges for High-Paying Jobs in Data Science

Median salary over first 10 years is \$100,323

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

Teaching Assistants

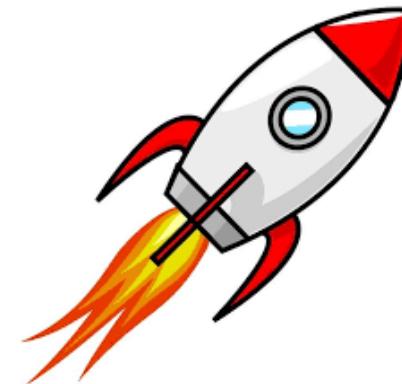
Teaching Fellows

- Ashley Oaks: ashley.oaks@yale.edu



Undergraduate Learning Assistants

- Dani Mekuriaw: dani.mekuriaw@yale.edu
- Corinne Evans: corinne.evans@yale.edu
- Cindy Mei: cindy.mei@yale.edu
- James Poe: james.poe@yale.edu
- Carly Benson: carly.benson@yale.edu
- Christian Baca: christian.baca@yale.edu



If you are interested in being a ULA next year, email Jay Emerson

- john.emerson@yale.edu

Good luck with the end of the semester!

Good luck finishing your final projects!

Review session: Tuesday, April 30th
from 4-5pm in DL 220

[Data Science vs. Statistician](#)

