# YData: Introduction to Data Science

## Lecture 26: LLMs, Widgets, Ethics, and Conclusions

# Overview

Quick review of clustering

Running an LLM/chatbot in Python

Interactive Jupyter notebooks widgets

Ethics

Wrap-up

If there is time:  putting a jupyter notebook on the internet

# Project timeline

Sunday, December 7th

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

Please also fill out the final project reflection on Canvas!

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

# Announcement

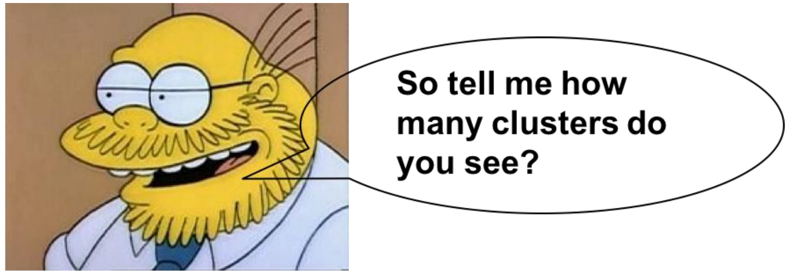
**Exam review session**: Tuesday December 9th from 2:40-3:45pm in this room

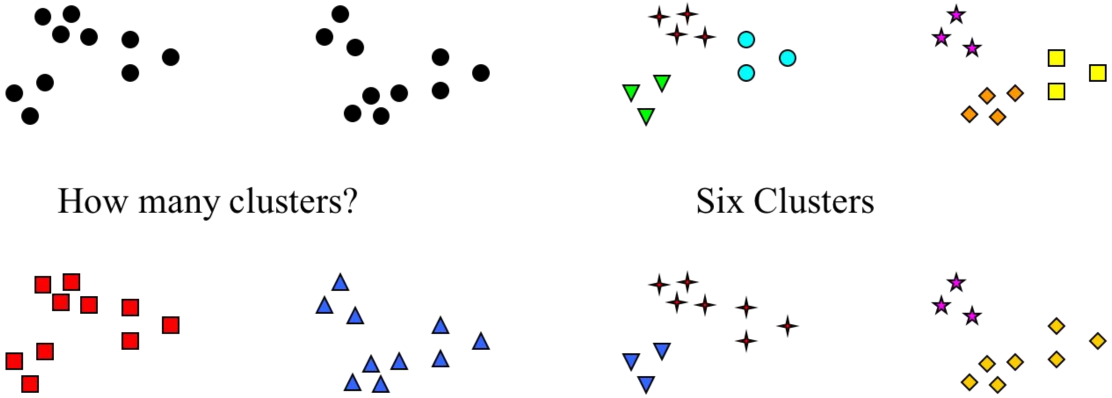
**Final exam**: Monday December 17th at 2pm

- **Location**: Marsh Auditorium

Also, be sure to download any work from the class you want to save from the JupyterHub server since accounts will be deleted toward the beginning of next semester

# Quick review of clustering





How many clusters?

Six Clusters

Two Clusters

Four Clusters

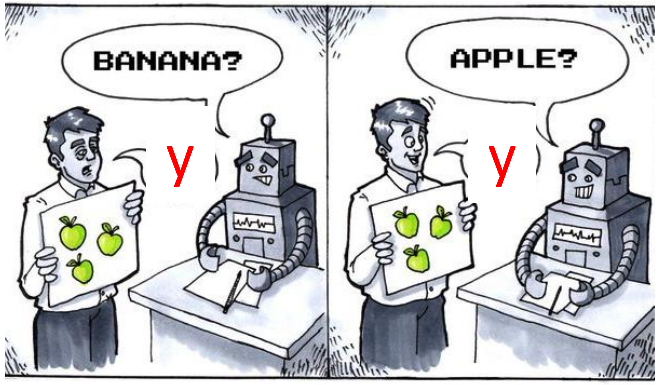
# Supervised learning and unsupervised learning

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

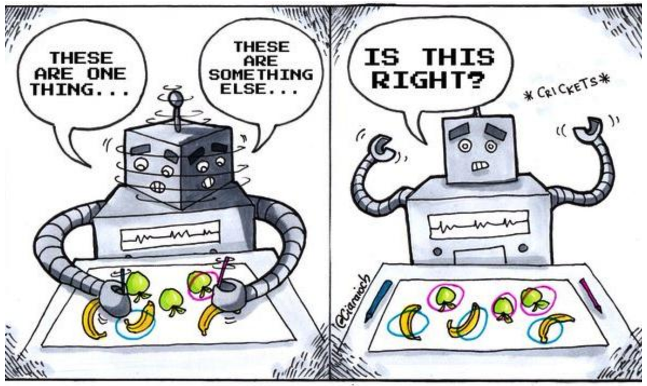
- Unsupervised learning can be useful in order to find structure in the data and to visualize patterns,
- but there is no real ground truth response variable y

1. **Clustering:** we try to group similar data points together

2. **Dimensionality reduction:** we try to find a smaller set of features that captures most of the variability in the data
   - Principal component analysis (PCA)



Supervised Learning



Unsupervised Learning

# Clustering

$$n\left\{\overbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}}^{p}\right.$$

Clustering divides n data points $x_i$‘s into subgroups

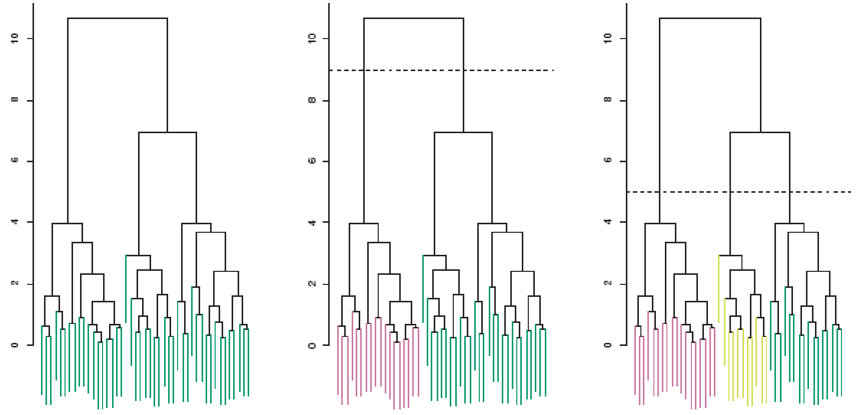
- Data points in the same group are similar/homogeneous
- Data points in different groups are different from each other

**Flat clustering**: k-means

- Specify k clusters in advance is each point is assigned to one cluster

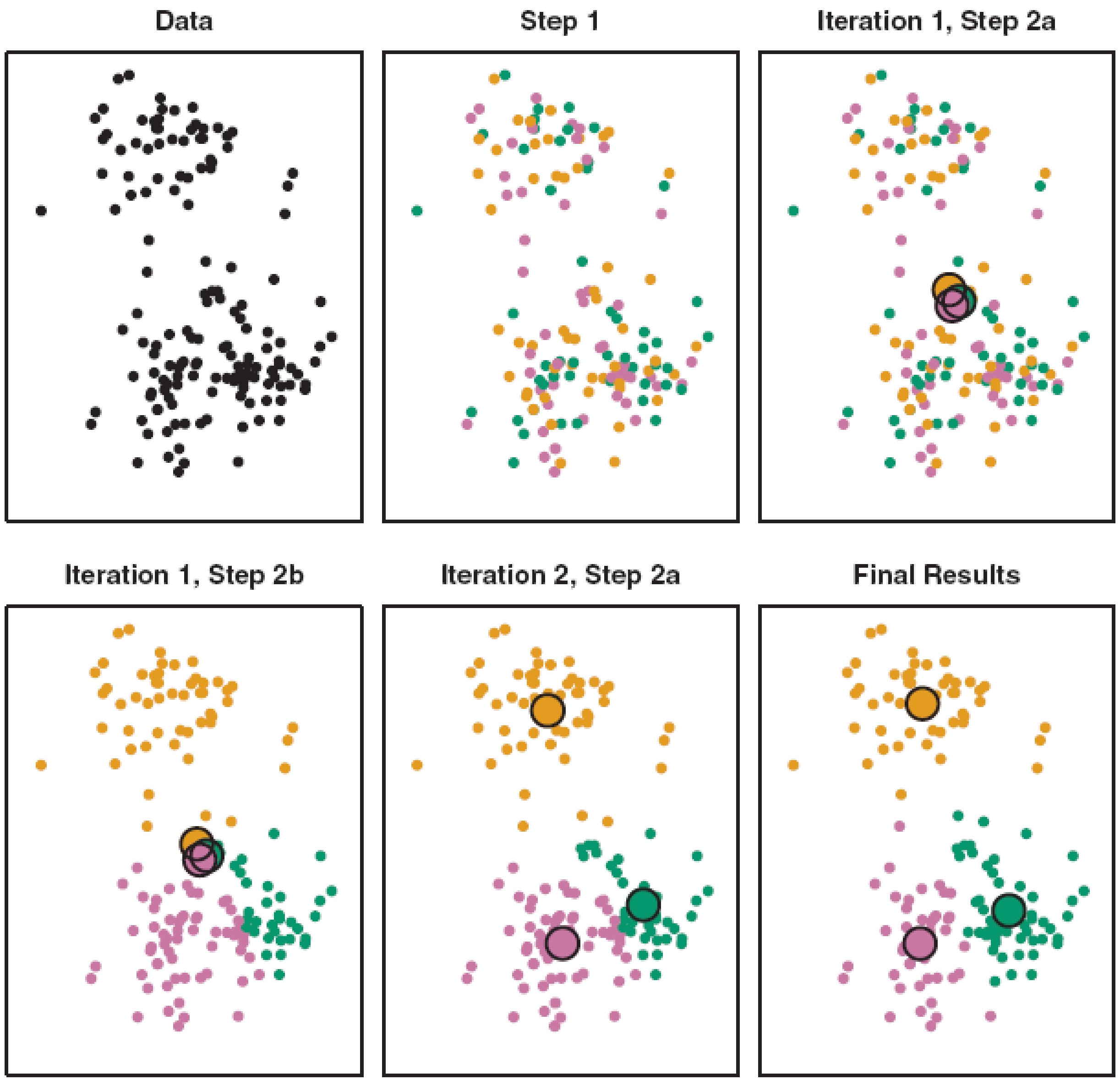
**Hierarchical clustering**

- Tree of nested clusters is created and we “cut” the tree to get a particular number of clusters



# 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



# Hierarchical clustering

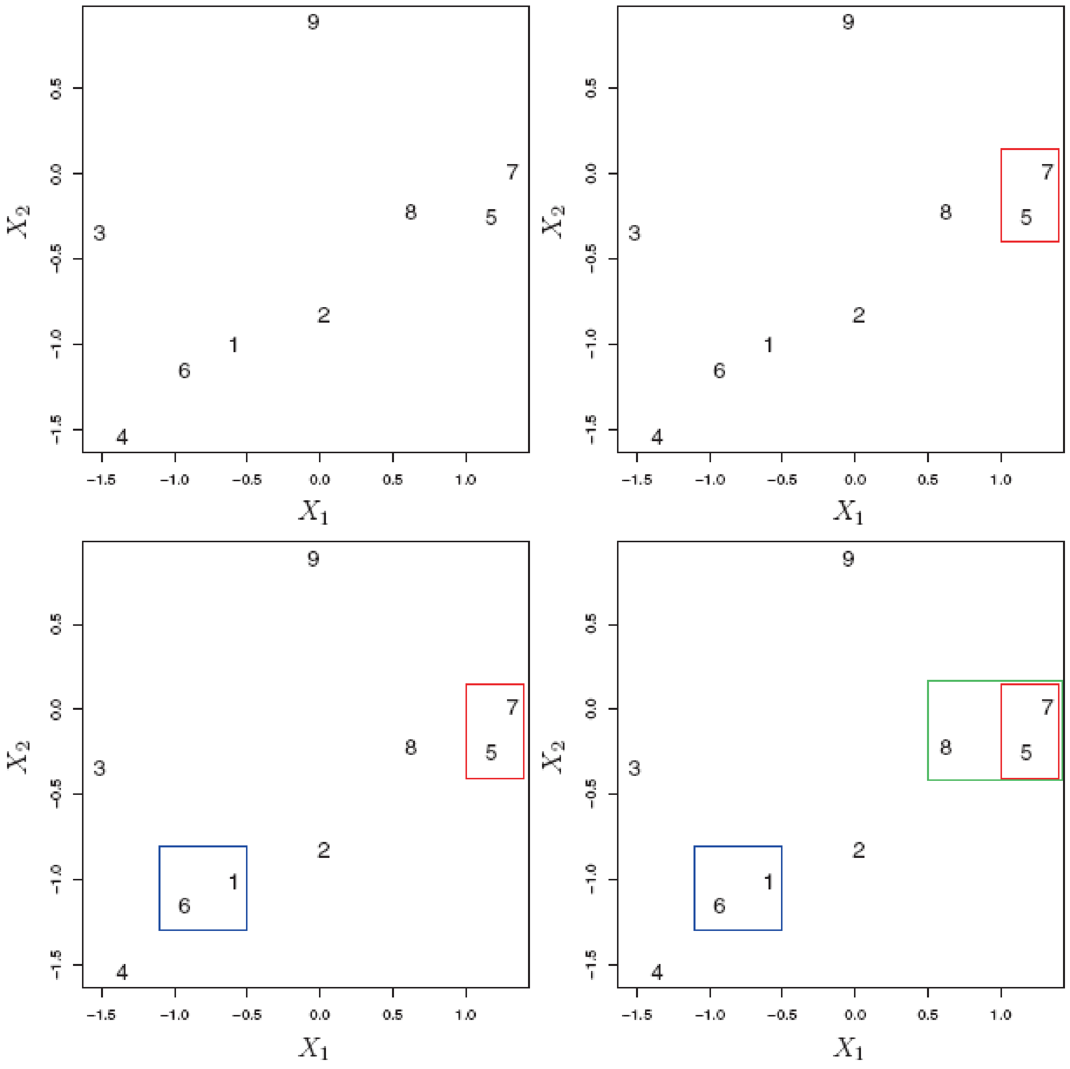
We can create a hierarchical clustering of the data using simple bottom-up agglomerative algorithm:

1. Choosing a (dis)similarity measure
   - E.g., The Euclidean distance

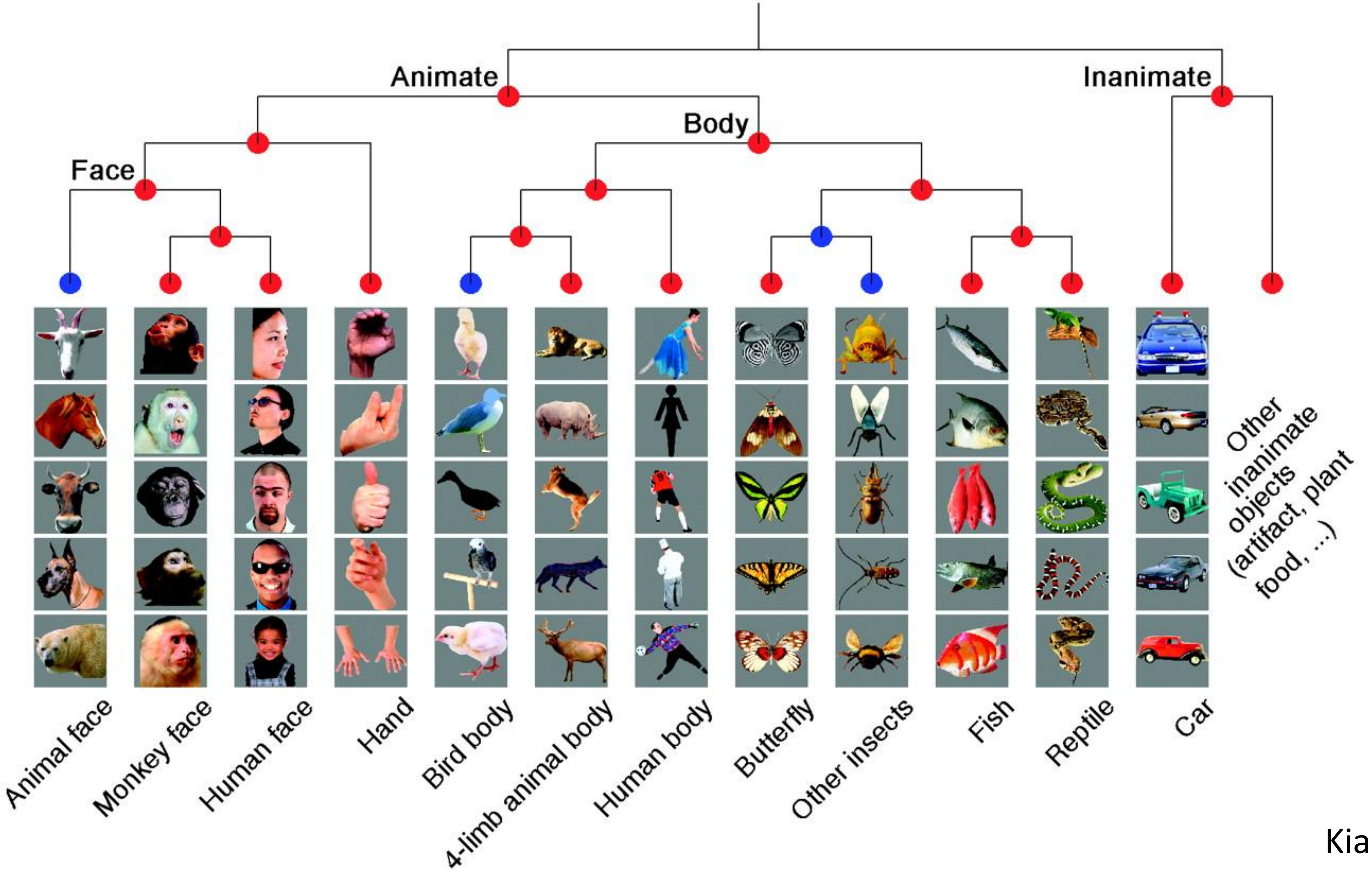
2. Initializing the clustering by treating each point as its own cluster

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 example



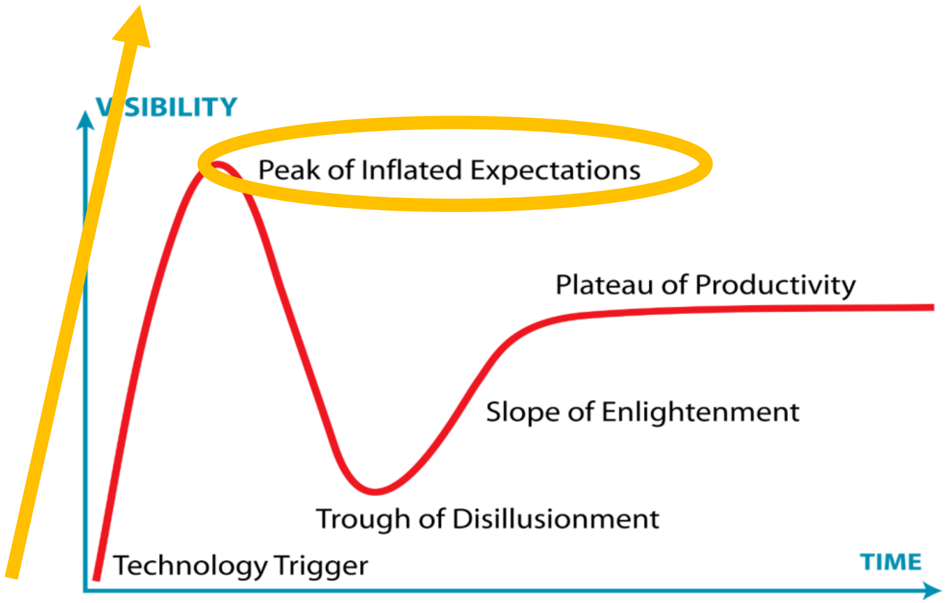
Kiani et al, 2007

# Questions?

# Brief discussion of Large Language Models

# Brief discussion of Large Language Models

Large language models (LLMs) are taking over the world



ChatGPT

Gemini

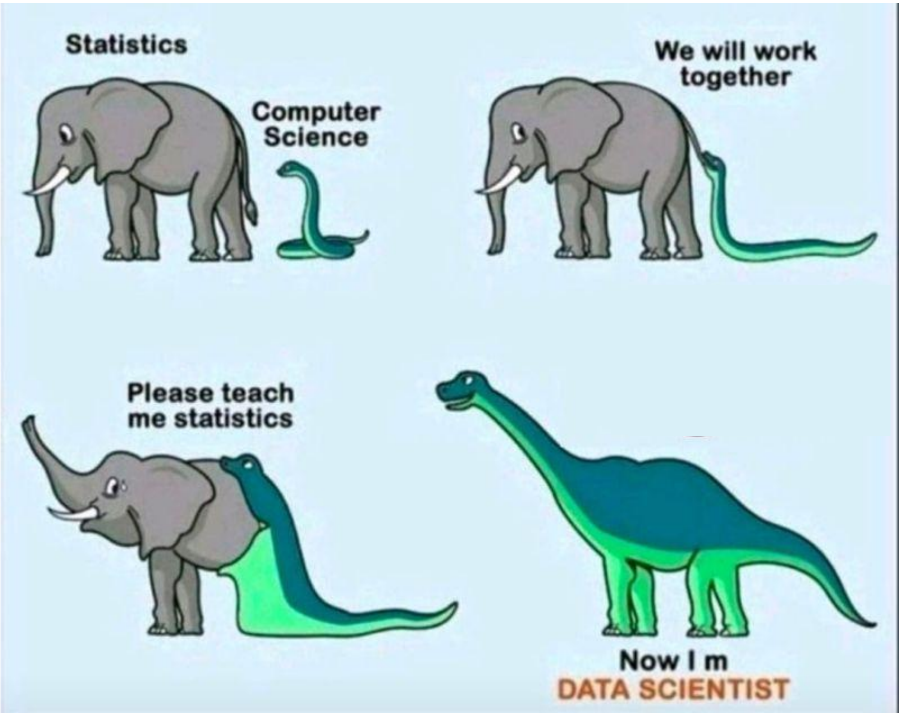
Claude

# Brief discussion of Large Language Models

LLMs can write code and 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 computers…



# Jupyter widgets

# Jupyter widgets

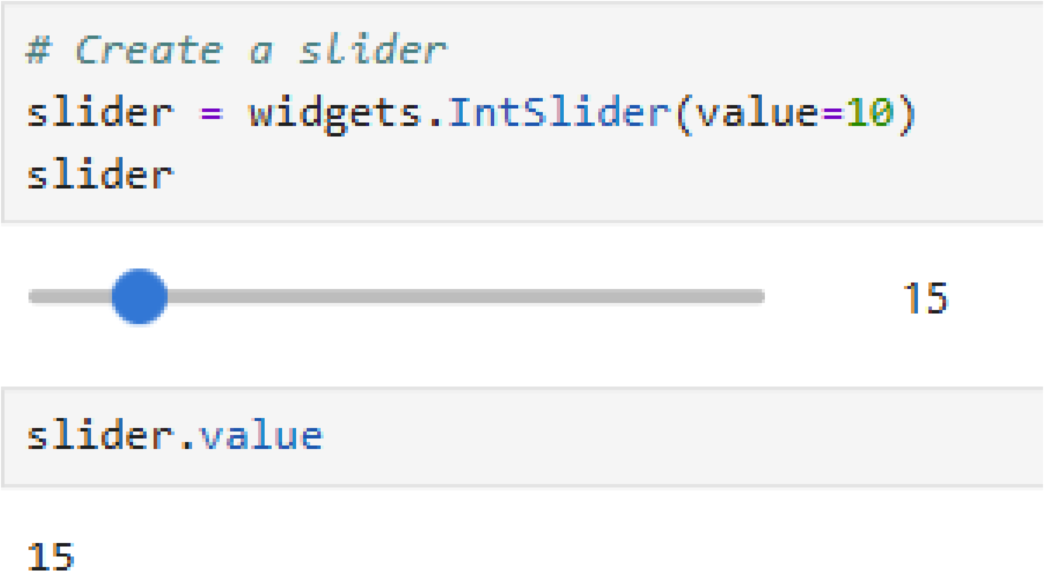
Jupter widgets allow you to add buttons, sliders, etc. to a Jupyter notebook so that you can create interactive visualizations

We can add widgets by importing:

```
import ipywidgets as widgets
```

We can then create a widget using:

```
slider = widgets.IntSlider(value=10)
```



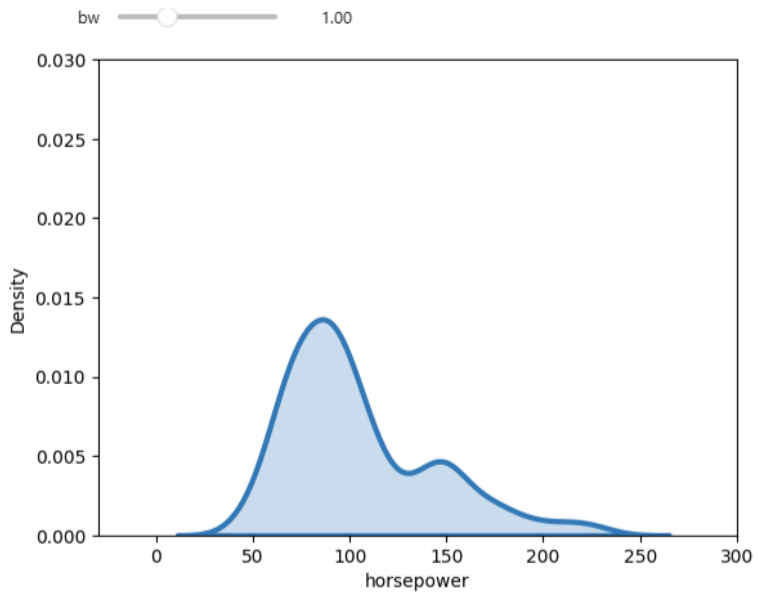
```
# Create a slider
slider = widgets.IntSlider(value=10)
slider
```

```
slider.value
```

```
15
```

# Jupyter widgets

This slider updates the figure



There are several ways to connect widgets to figures

- One way is through widgets.interact()

Boolean argument: checkbox

String argument: textbox

List argument: dropdown menu

Numeric arguments: slider

```
def bandwidth_widget(bw = 1):
        sns.kdeplot(cars.horsepower,
                    bw_adjust=bw)
```

Overrides default arguments

```
widgets.interact(bandwidth_widget, bw = (.1, 3));
```

Let’s try it in Jupyter!

# Ethics

# Ethics in Data Science

Ethics of:

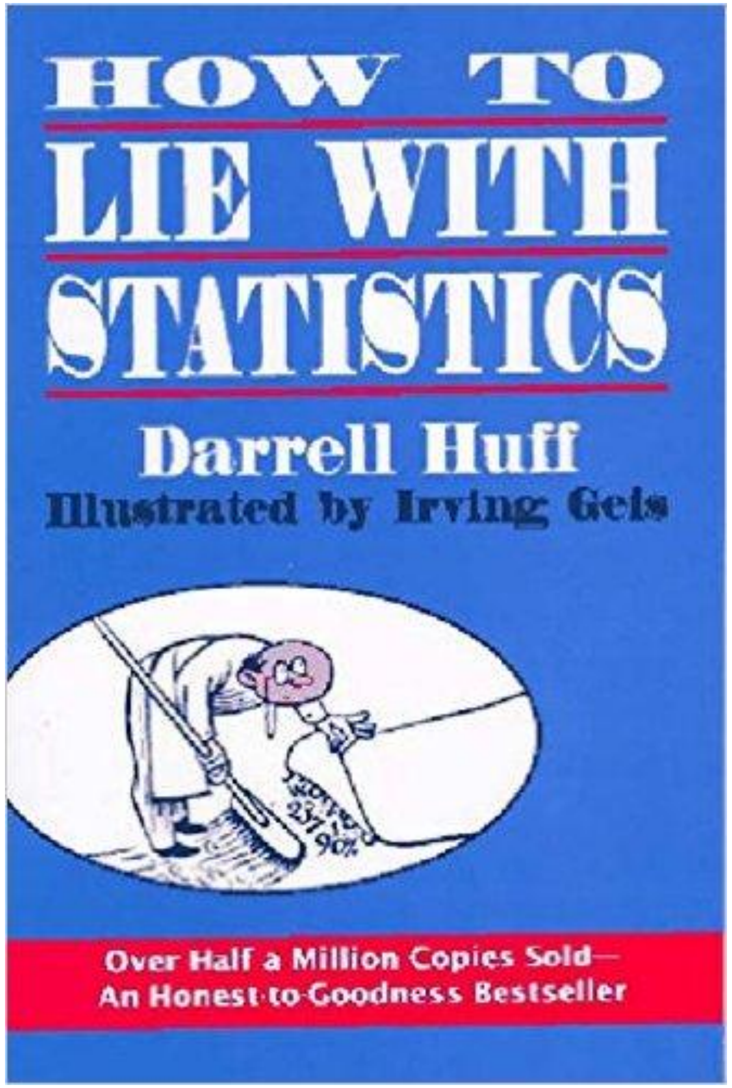
1. Data presentation
2. Using valid data
3. Data scraping TOS and privacy
4. Reproducibility
5. Citations/peer review
6. Disclosure
7. Ethics in Statistical analyses
8. Ethics of creating powerful tools

# 1. Ethics of data presentation

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

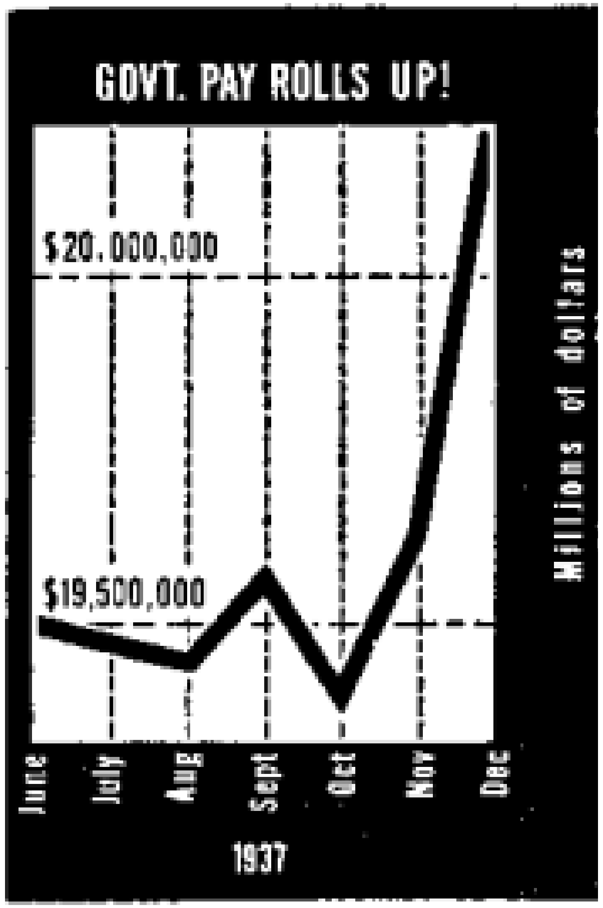
Darrell Huff wrote a classic book in the 1950’s pointing out ways that people lie with statistics

The book was banned as training material at the VA


HOW TO
LIE WITH
STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

# Ethics of data presentation

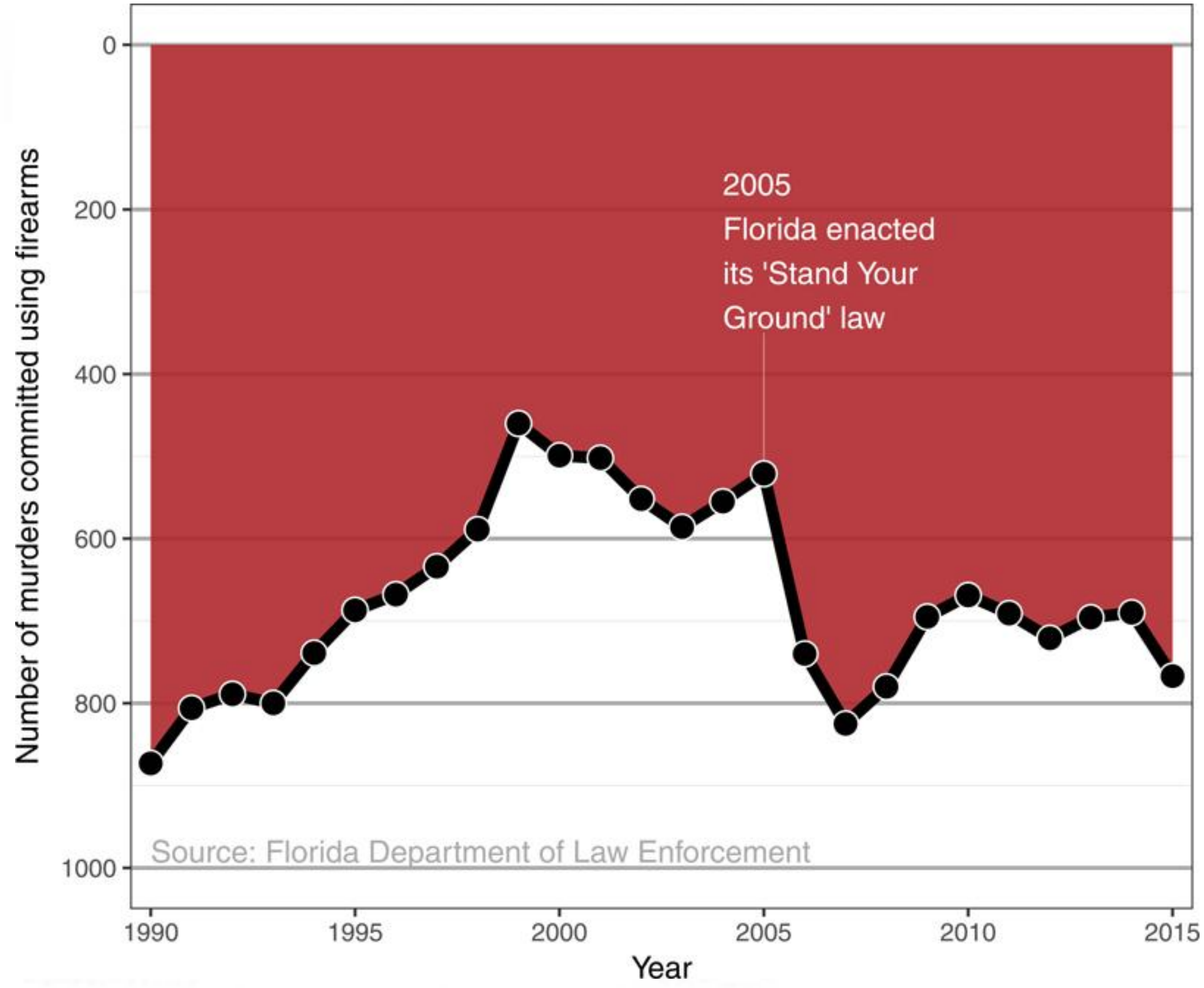
What is potentially misleading with this figure?



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

# Did ‘Stand Your Ground’ decrease murder by firearms?

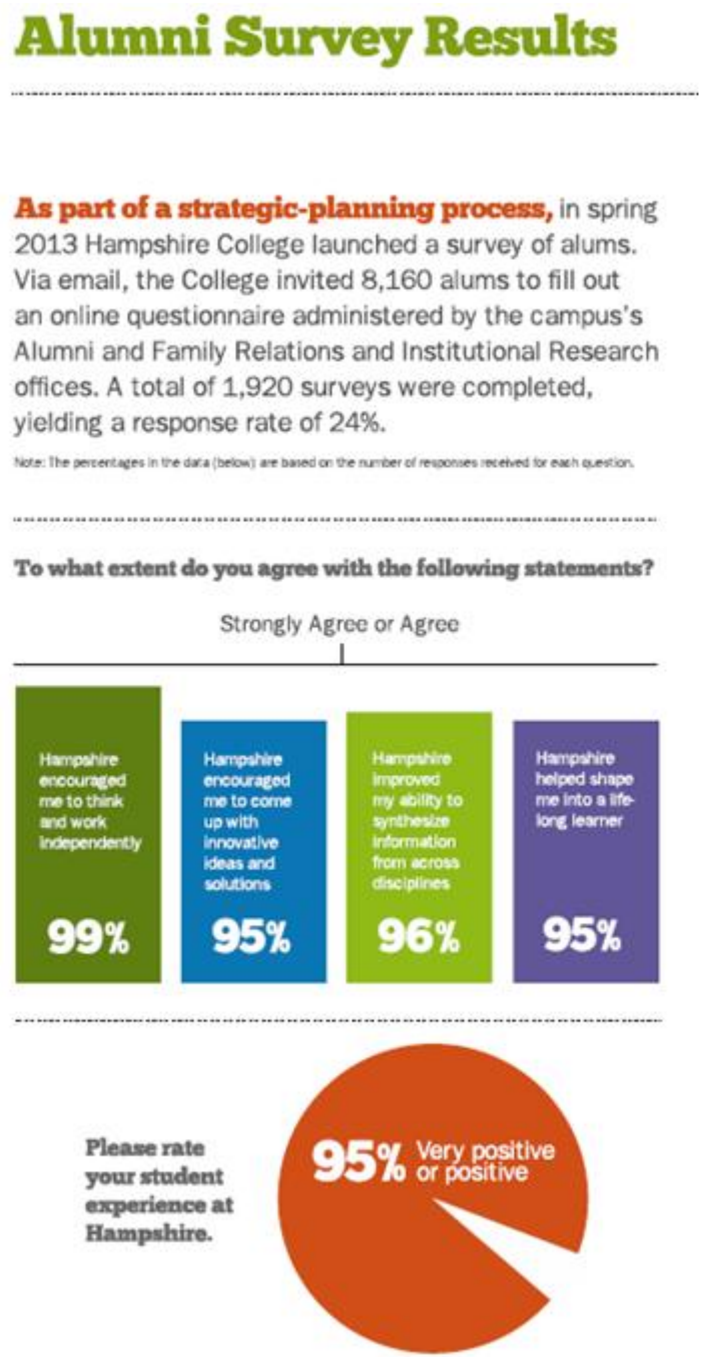
What is misleading with this figure?



Source: Florida Department of Law Enforcement

# 2. Using valid data

Is almost everyone satisfied with Hampshire College?



## 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

| Hampshire encouraged me to think and work independently | Hampshire encouraged me to come up with innovative ideas and solutions | Hampshire improved my ability to synthesize information from across disciplines | Hampshire helped shape me into a life-long learner |
|---|---|---|---|
| 99% | 95% | 96% | 95% |

**Please rate your student experience at Hampshire.**

95% Very positive or positive

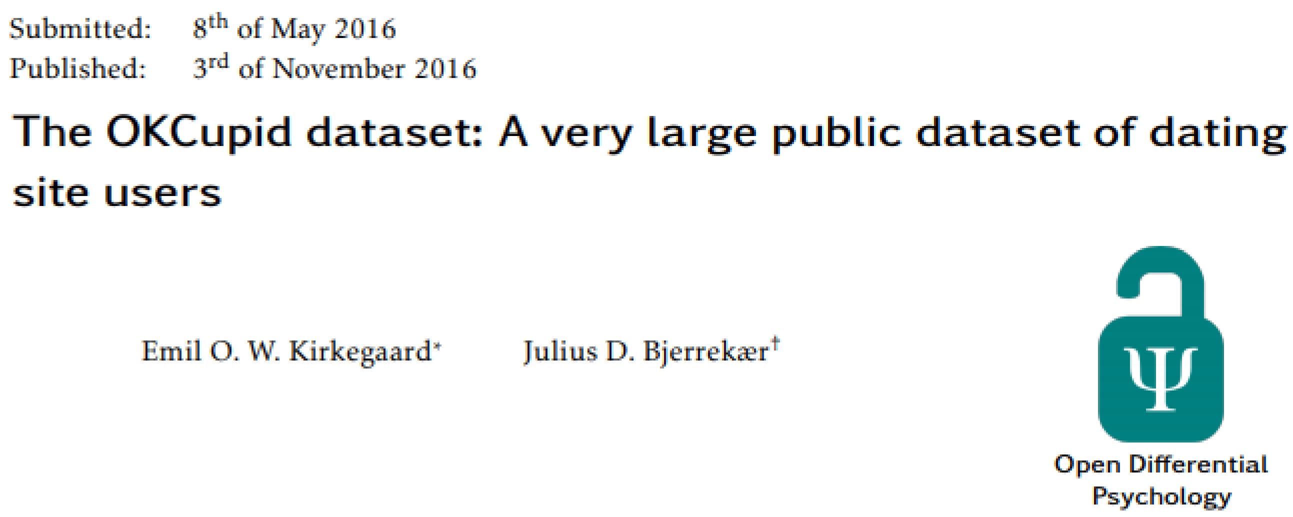
# 3. Data scraping, terms of service and privacy

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

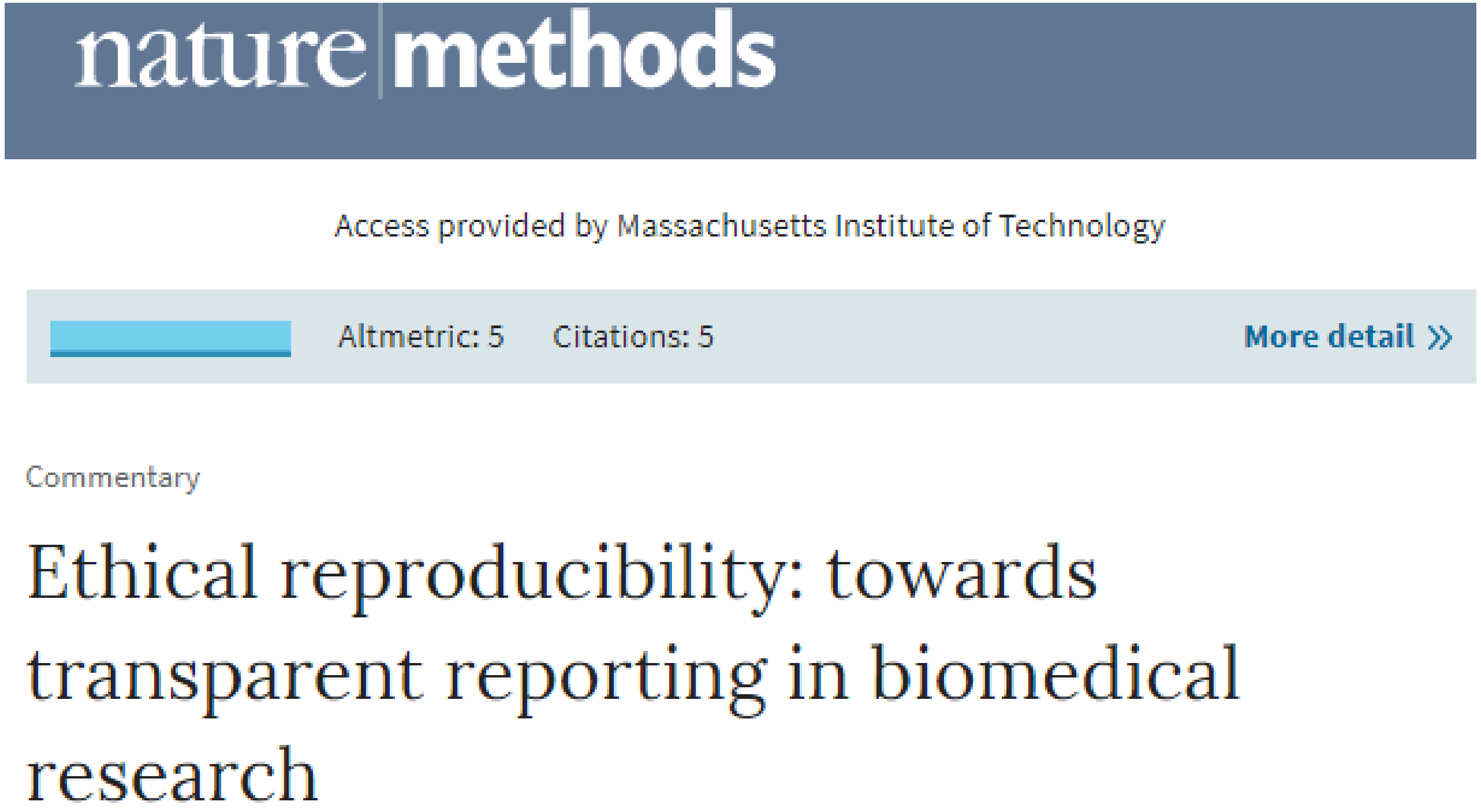
The OKCupid dataset: A very large public dataset of dating site users

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?



nature methods

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?

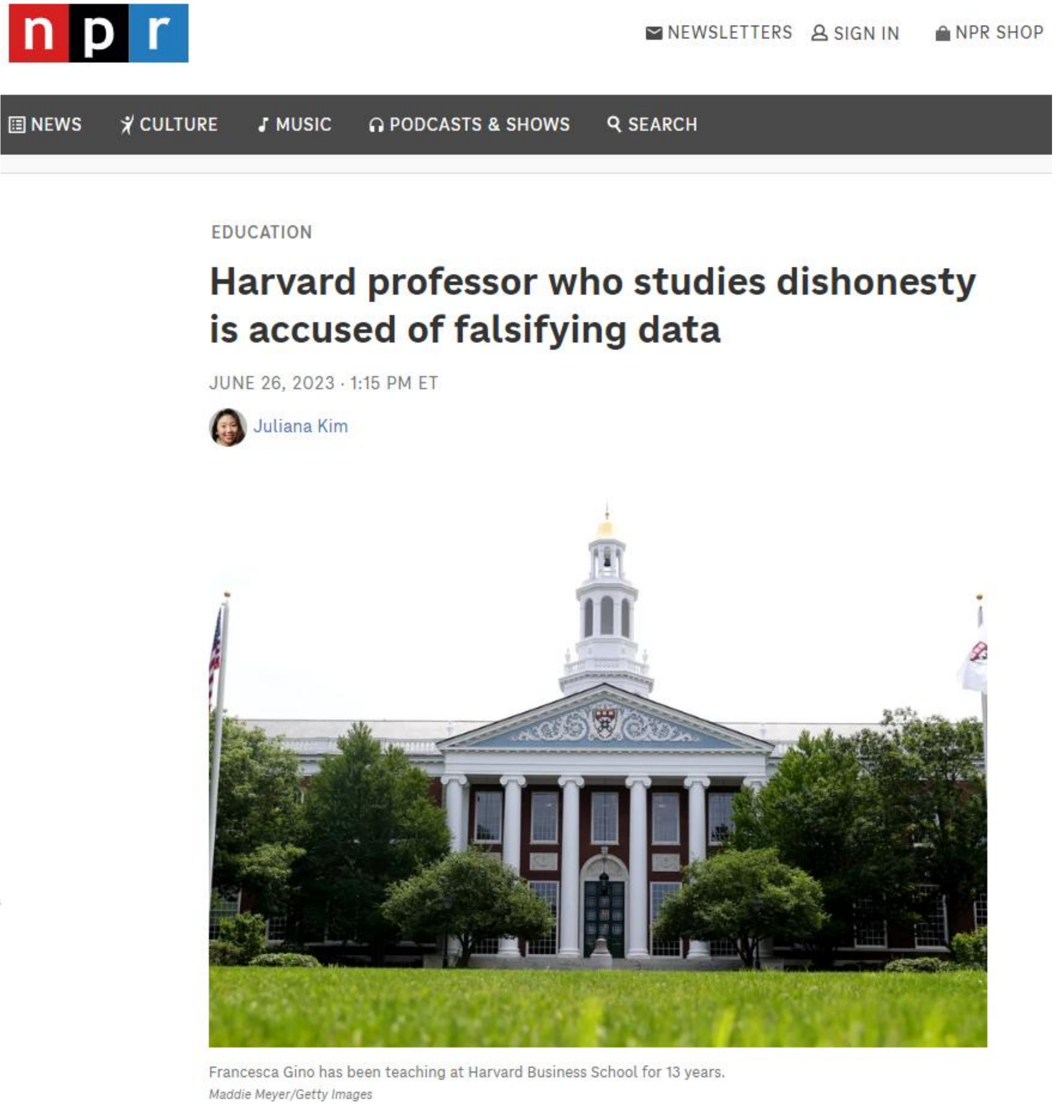
- What if it could hurt your career?
    - Others could prove you wrong, make new findings on your own data, etc.

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

- You need to retract the paper!

Retraction Watch

DATA COLADA



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*

# 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?

# 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

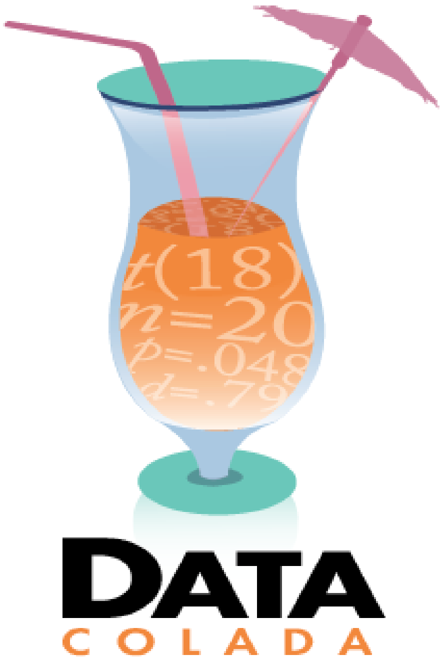
# 7. Ethics in Statistics

## 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



# 8. 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

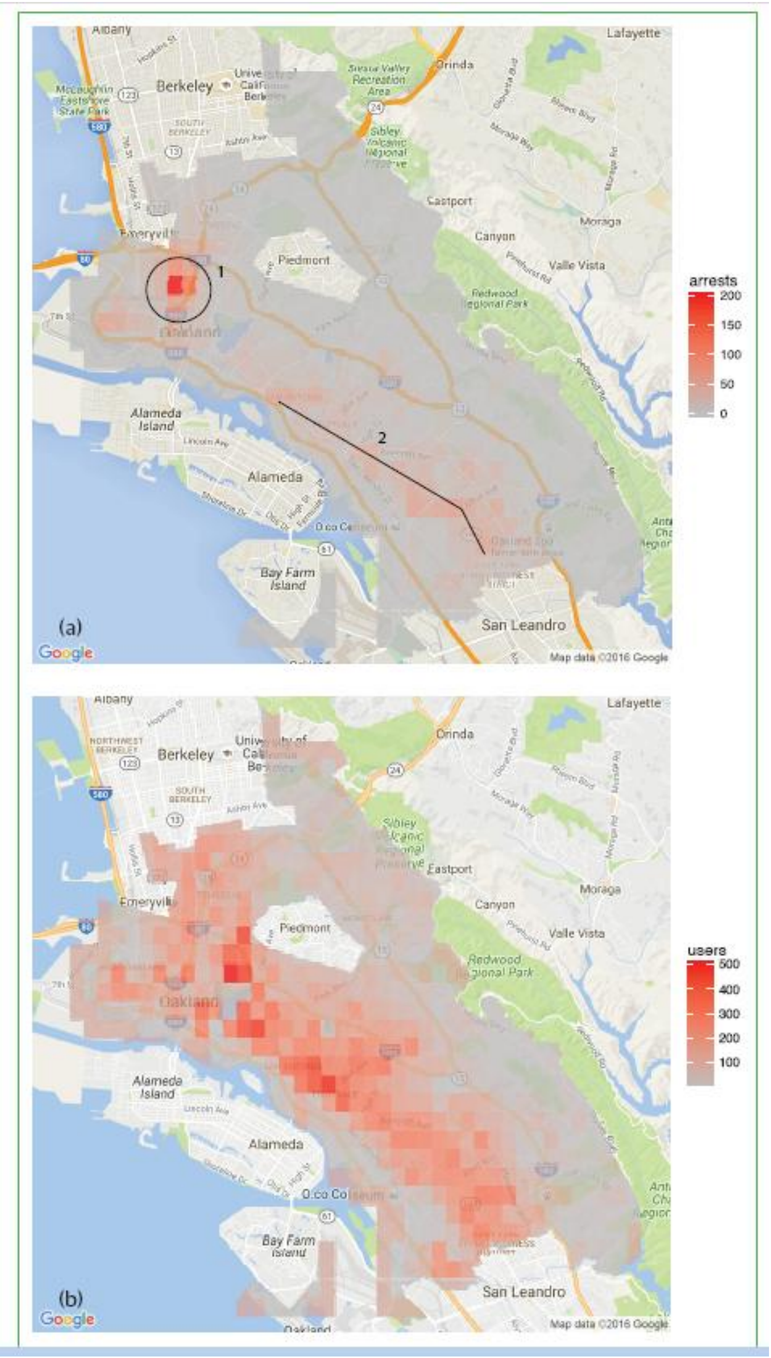
Idea: use ML to police areas with most crimes

- E.g. more police where most drug arrests have made in the past

Possible results

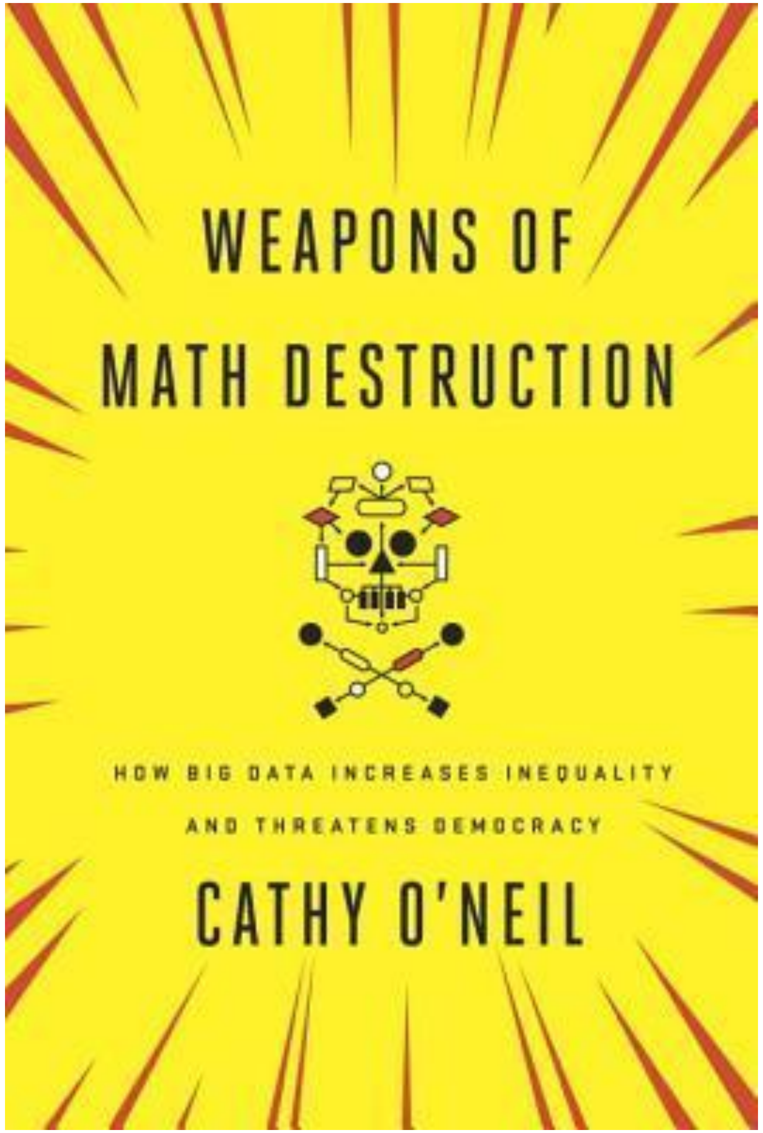
- Continued higher arrest rates in these areas seemingly showing the ML algorithm is working

Any potential problems with this?



# Additional reading

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



WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

# Wrap up and conclusions

# Topics covered

What is Data Science?

Python basics

Descriptive statistics

Array computations

Manipulating data tables

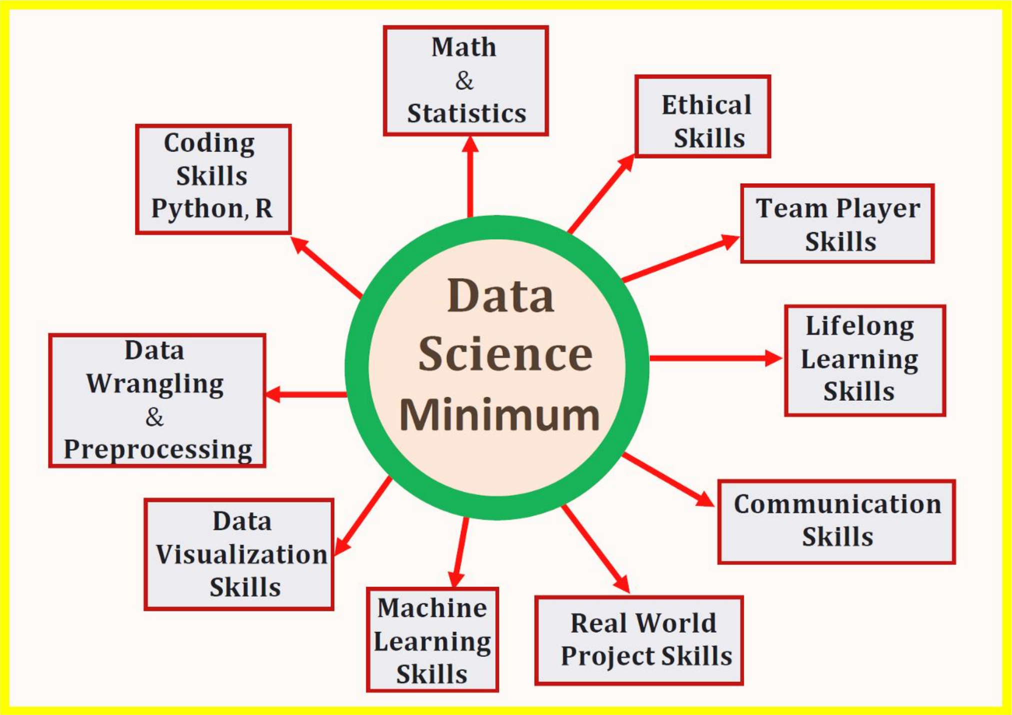
Data visualization

Mapping

Text manipulation and data cleaning

Statistical perspective: hypothesis tests and confidence intervals

Machine learning perspective: supervised and unsupervised learning



# Learning goals

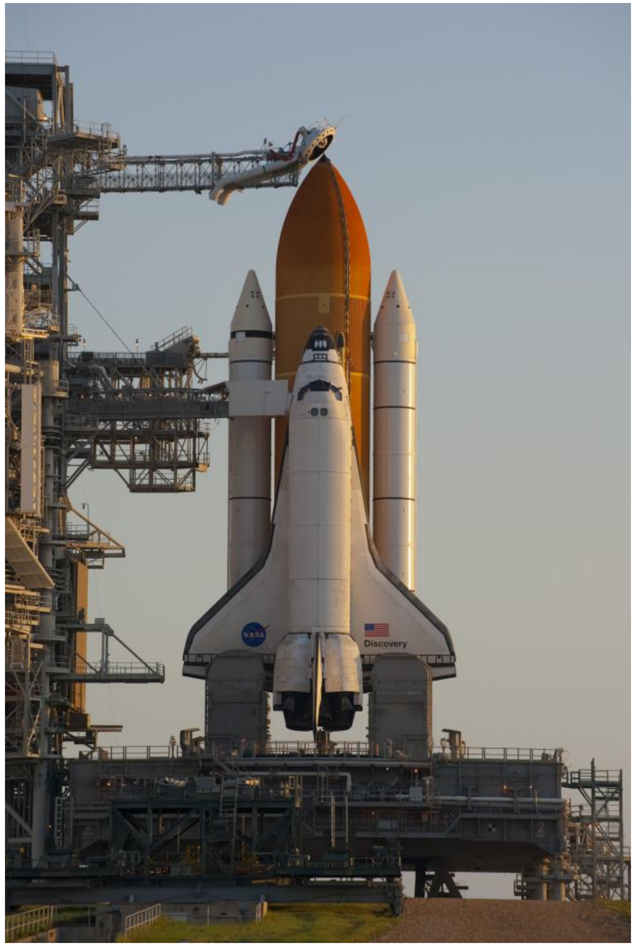
1. Understand concepts in data science
   - Learn basic computational skills for analyzing data
   - Understand concepts in statistics and machine learning

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

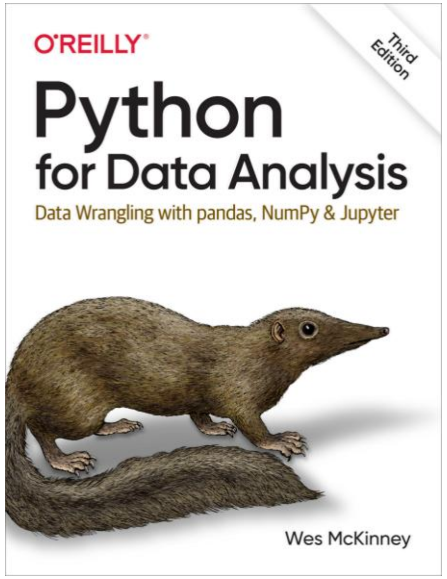
- E.g., no prior knowledge of statistics or programming is required

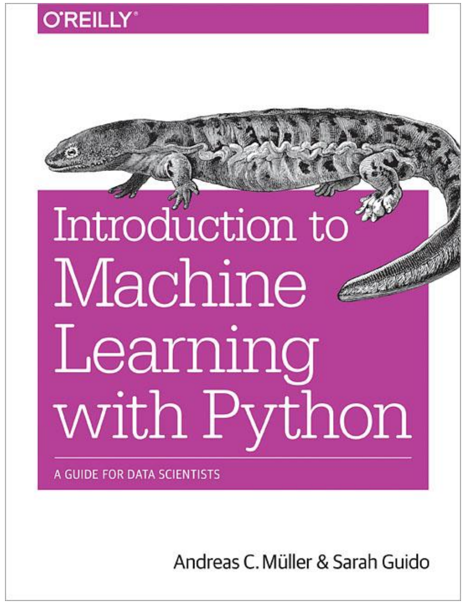


# Next steps

1. Take more advanced Statistics and Data Science classes offered at Yale!
    - S&DS 1000, S&DS 2400, YData connector classes, …

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




O'REILLY
Introduction to Machine Learning with Python
A GUIDE FOR DATA SCIENTISTS
Andreas C. Müller & Sarah Guido

# Teaching Assistants

Preceptor

- Shivam Sharma

Teaching Fellows

- Steve Ward
- Ben Green

Undergraduate Learning Assistants

- Kyle Levesque
- Brunokai Ong
- Sloane Huey
- Christian Baca

# Good luck with the end of the semester!

Good luck finishing your final projects!

**Review session**: Tuesday December 9th from 2:40-3:45pm in this room

# Hosting webpages on GitHub pages

# Webpages we’ve created

GitHub.com is a service that allows people to share code

- Also allows one to host webpages for free!
- So if you save your project as a .html document, you can share it on the web!

Let’s go through the steps now to host a webpage on GitHub