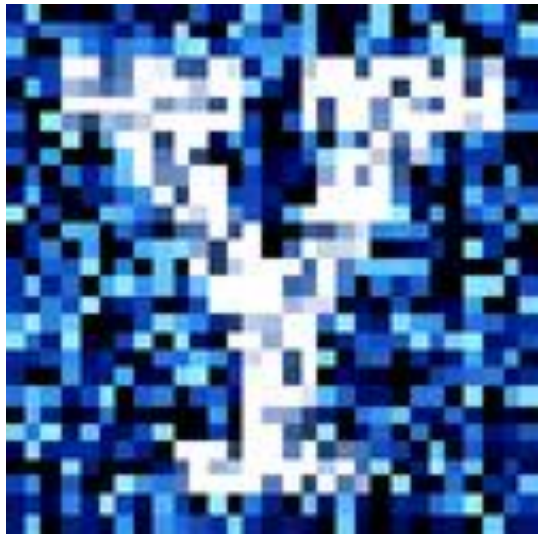
YData: Introduction to Data Science



Lecture 26: Web scraping, LLMs Ethics, and wrap-up

Overview

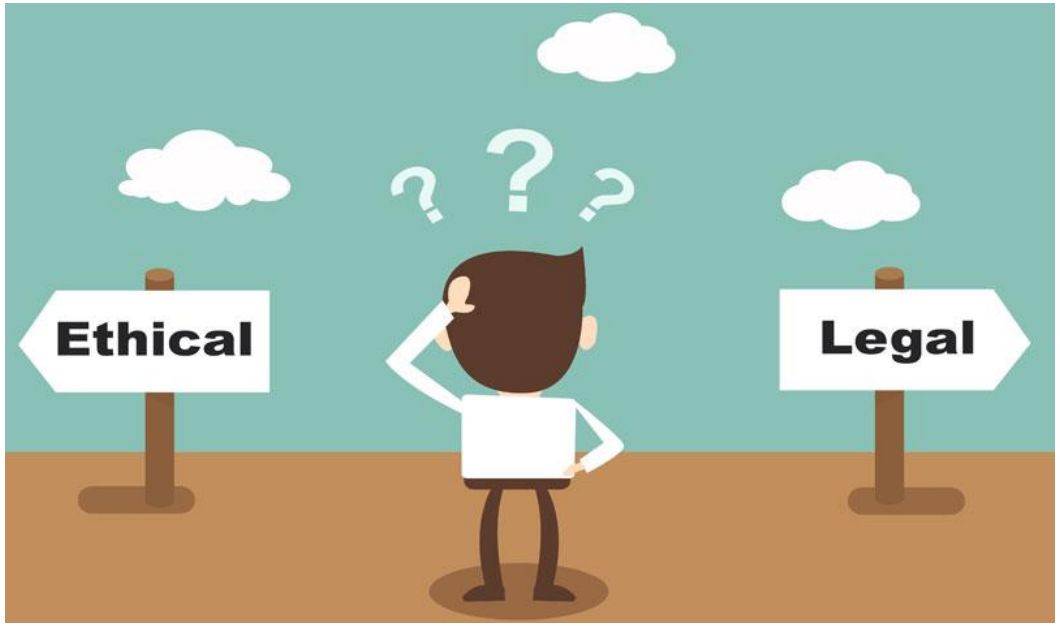
Quick review of web pages

Web scraping

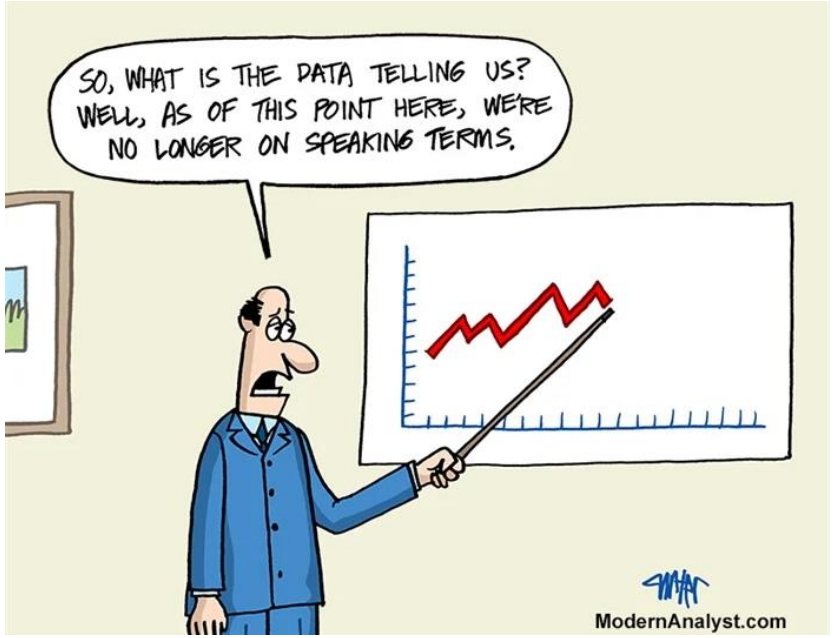
Brief demo of running a LLM/chatbot in Python

Ethics

Wrap-up



Project timeline



Sunday, December 8th

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

Please also fill out the final project reflection on Canvas!

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

Announcement

Exam review session: Monday December 9th from 3-4pm in this room

Final exam: Monday December 16th at 7pm in this classroom!

Quick review of webpages

Review: Webpages

Webpages are written in Hypertext Markup Language (HTML)

- Typically webpages end with the extension .html

Webpage consists of text and tags

- Text is the content displayed on the webpage
- Tags allow one to insert links, images, and other meta-data

Tags have the form <> </>

- For example, we can make text bold using

The word bold is bold

Review: Example of a basic webpage

```
<html>

<title> My cool page </title>

<body>

    This is my <b>cool</b> webpage

    <br><br>

     Here is <a href="https://canvas.yale.edu/">Canvas </a>

</body>

</html>
```

This is my **cool** webpage

This is a link to Canvas

Additional webpage tags

Headers: <h2> Second biggest header </h2>

Inserting images: <img src="ant.jpg">

Lists:

```
<ul>
    <li> Item 1 </li>
    <li> Item 2 </li>
</ul>
```

Additional webpage tags - tables

```
<table>
    <tr>
        <td> 1 </td>
        <td> 2 </td>
    </tr>

    <tr>
        <td> 4 </td>
        <td> 5 </td>
    </tr>

</table>
```

1 2

4 5

Let's take a quick look at some webpages

- A page I created hosted on GitHub
- Department of Statistics and Data Science
- Wikipedia page

Let's look at the source code and inspect elements of these pages using the Chrome web browser

W3School is a good site to learn more about creating webpages

Web scraping

Web scraping

Web scraping is the processing of extracting webpage content through code

In Python, the most widely used package for web scraping is Beautiful Soup which can extract relevant information from webpages

```
import requests
from bs4 import BeautifulSoup
```

Web scraping

We first need to download the webpage using the requests module

```
# the web address
url = "http://blah.com/page.html"

# request the webpage
response  = requests.get(url)

# see if the request was successful
print(response)
```

Common responses

- **200**: success
- **400**: URL was badly formed
  - i.e., bad web address
- **403**: "Forbidden" Server understood the request but refused it
  - Often because it detects you're trying to scrape the site

Web scraping

We can parse a successful response with Beautiful Soap using:

```
soup = BeautifulSoup(response.text, 'html.parser')   # create the object

tables = soup.find_all('table') # get all tables on a webpage
headers = soup.find_all(['h1', 'h2'])   # get all tables on a webpage
# etc.
```

The .find_all() method returns tag elements which can be used to extract the relevant information related to particular tags

Let’s try 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…

Ethics

Ethics in Data Science

Ethics of:

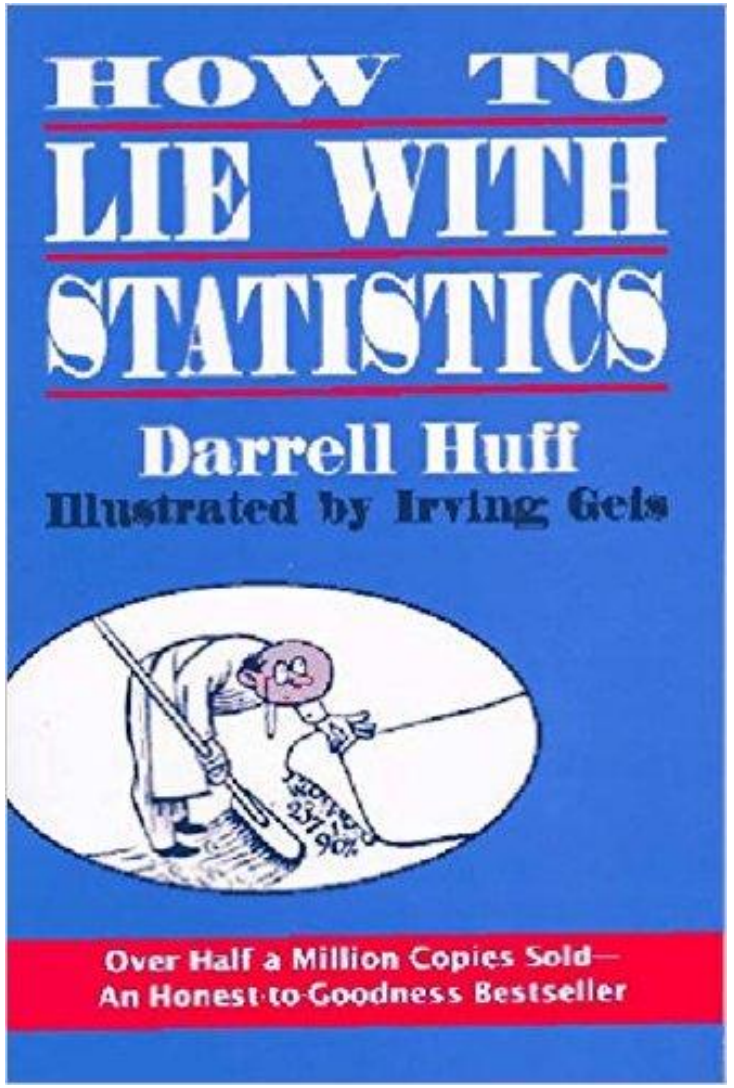
1. Data presentation
2. Using valid data
3. Data scraping TOS and privacy
4. Reproducibility
5. Citations/peer review
6. Disclosure
7. Ethics in Statistical analyses
8. Ethics of creating powerful tools

1. Ethics of data presentation

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

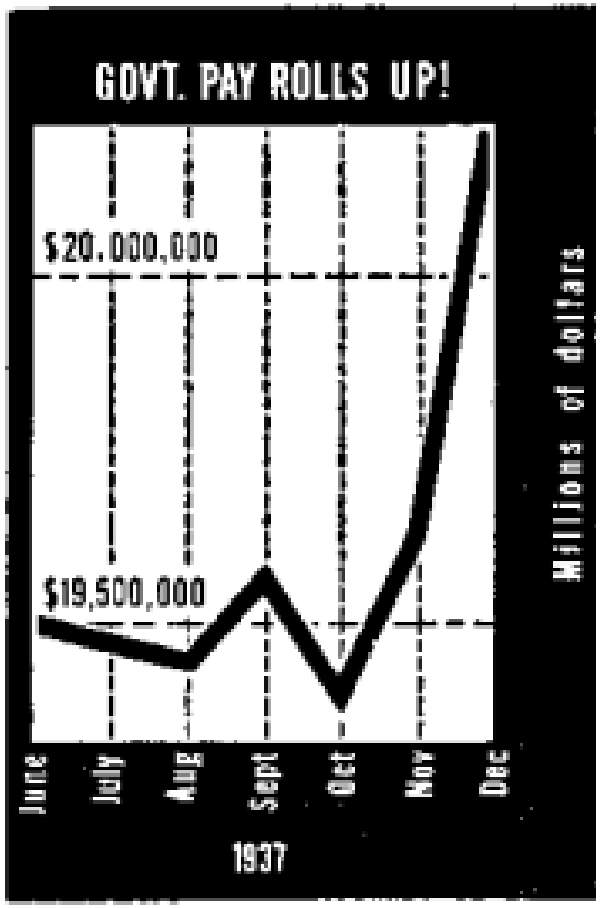
Darrell Huff wrote a classic book in the 1950's pointing out ways that people lie with statistics

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



Ethics of data presentation

What is potentially misleading with this figure?



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

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

Please rate your student experience at Hampshire.

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

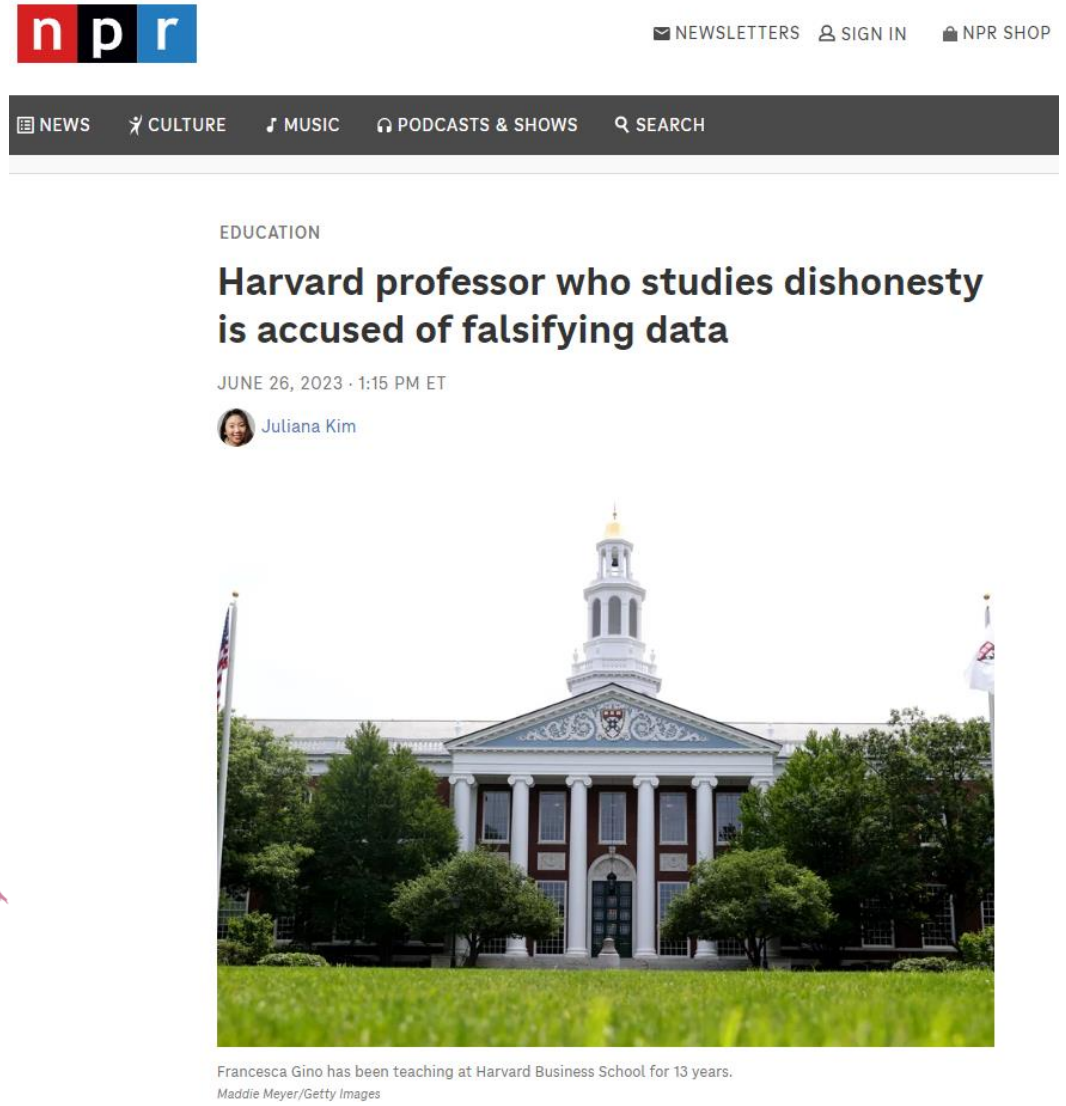
Do scientists have an obligation to share data/code?

- 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!

Retraction Watch



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 if they didn’t cite you?

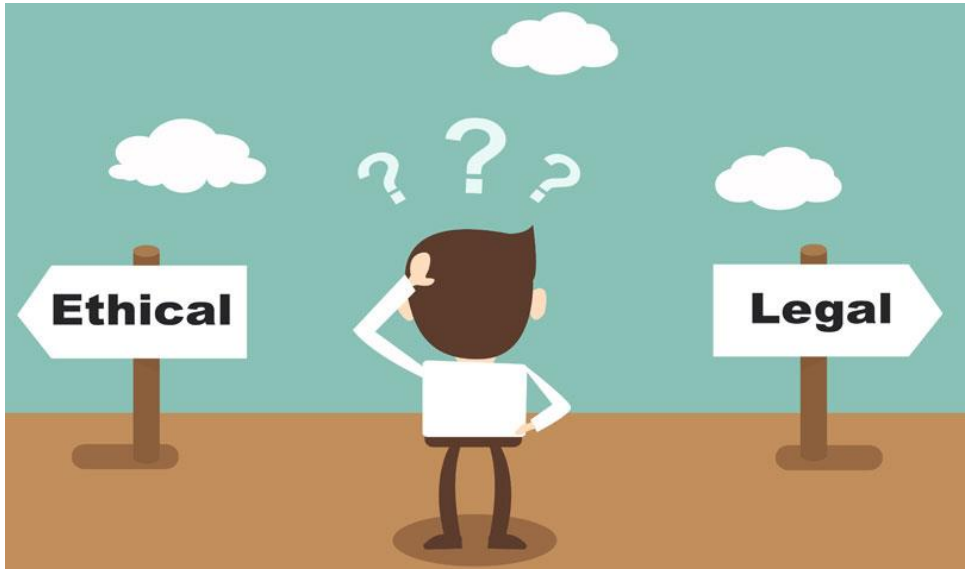
What about citing someone because they will be a reviewer of your paper?

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

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

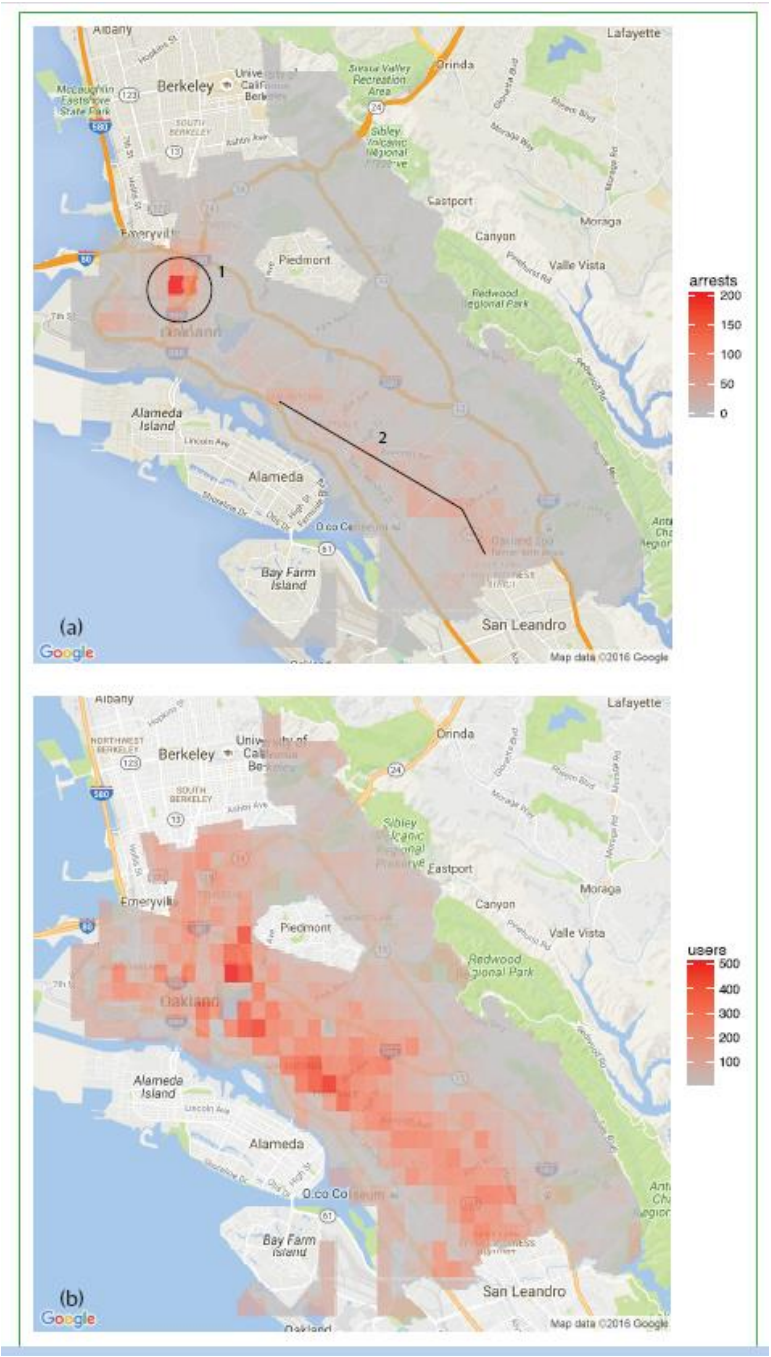
Idea: use ML to police areas with most crimes

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

Possible results

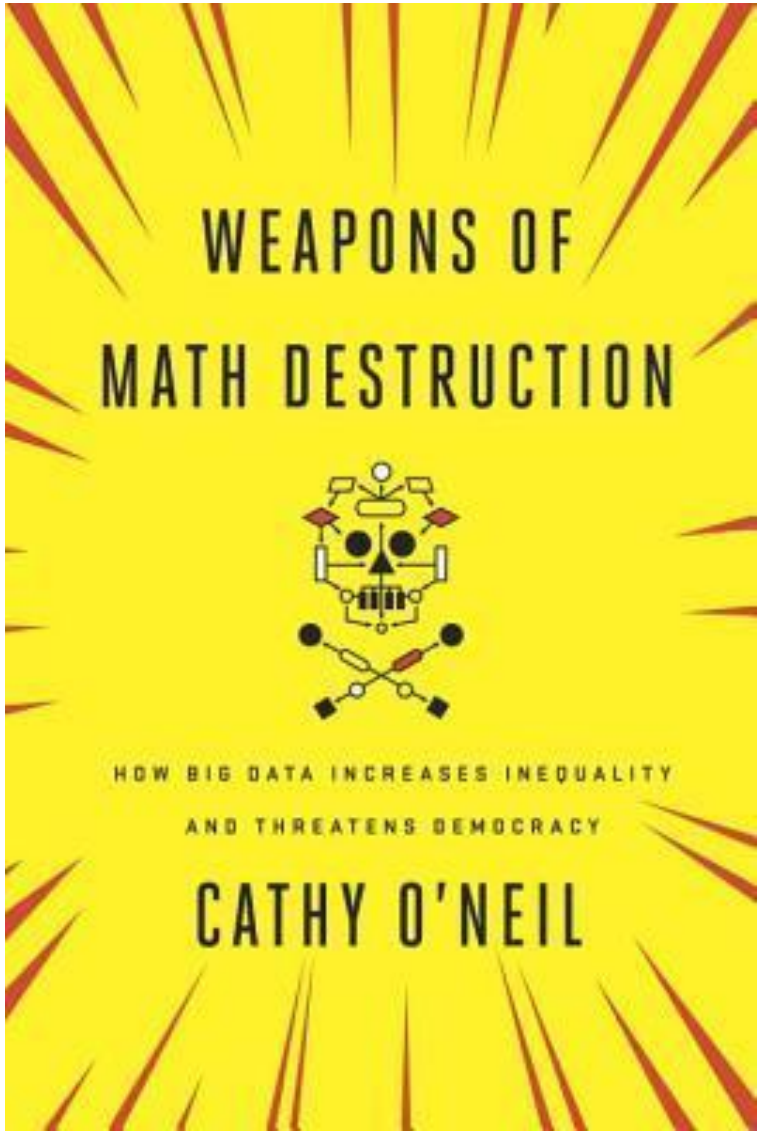
- Higher arrest rates for drugs found in these areas seemingly showing the ML algorithm is working

Any potential problems with this?



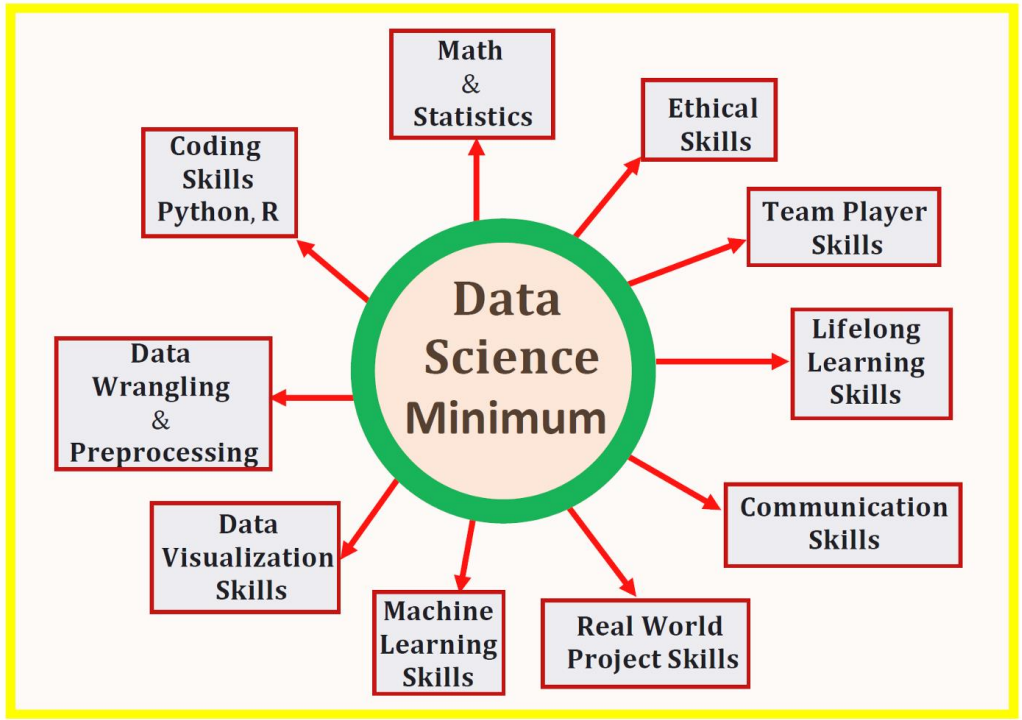
Additional reading

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



Wrap up and conclusions

Topics 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

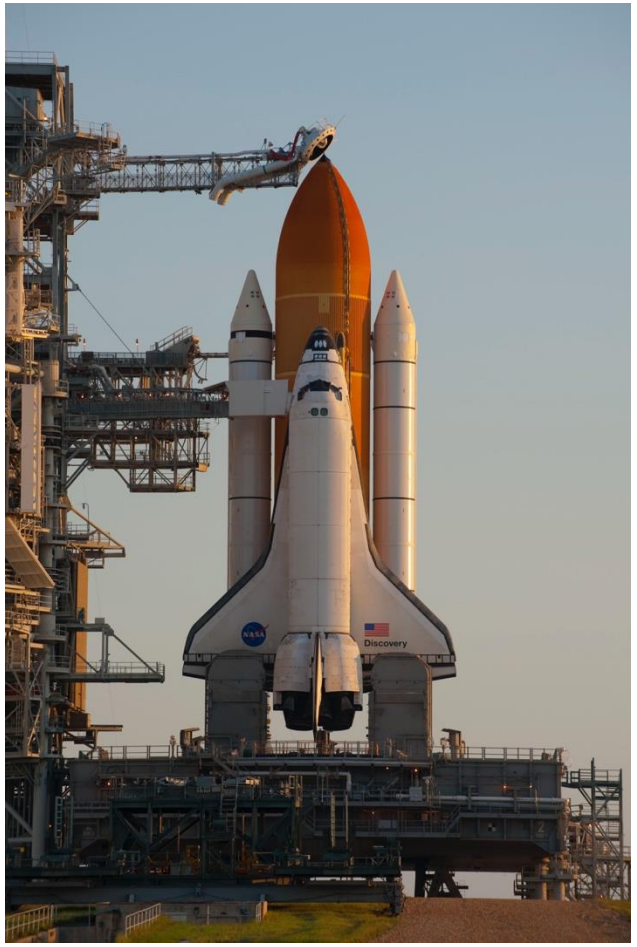
Learning goals

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

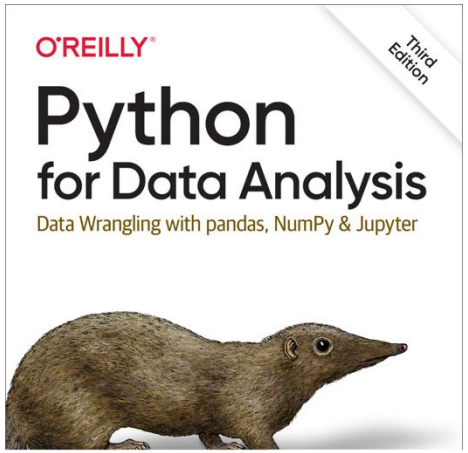
There are no prerequisites for this class
    - E.g., no prior knowledge of statistics or programming is required

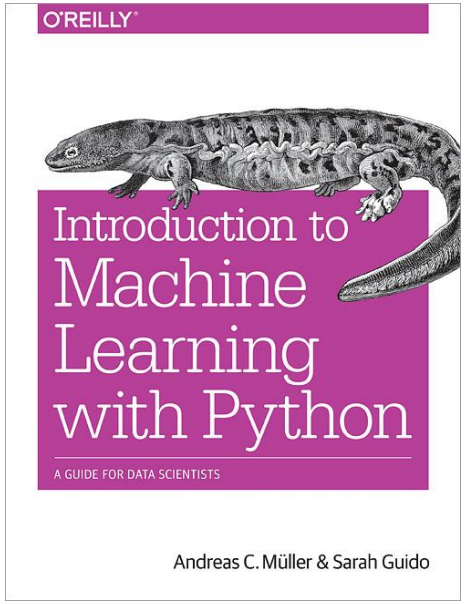


Next steps

1. Take more advanced Statistics and Data Science classes offered at Yale!
   - S&DS 100, S&DS 240, YData connector classes, …

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.!

Good luck with the end of the semester!

Good luck finishing your final projects!

Review session: Monday December 9th from 3-4pm in this room