YData: Introduction to Data Science

Class 01: Introduction

Overview

Course overview

- Introductions
- Syllabus and logistics

What is Data Science?

- Brief history of Data Science

Intro to Python

Office hours and contact information

Ethan Meyers (he/him)

Email: ethan.meyers@yale.edu

Office hours:

- Mondays and Wednesdays, 1:30-2:30pm
  - (subject to change)

Office: 24 Hillhouse room 206

- https://yale.zoom.us/j/95197359657

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

TA office hours are on the calendar on Canvas

Introductions

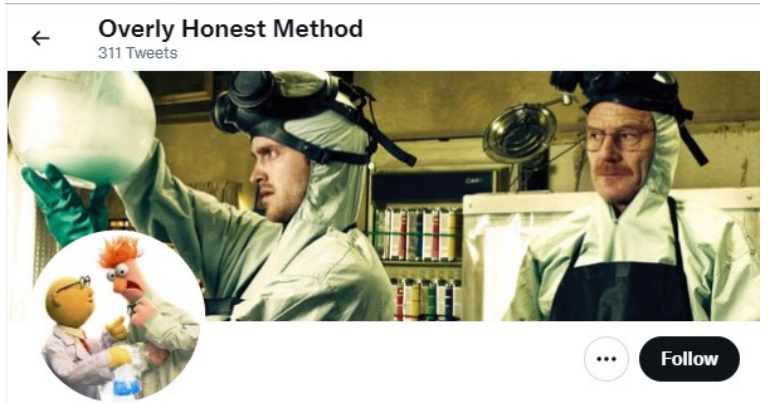
Let’s do some quick introductions

Create groups of ~4 people:

- Your name and preferred gender pronouns
- Your major/grad dept (research area)
- Why you are interested in this class
- Anything else you would like to share with your group

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

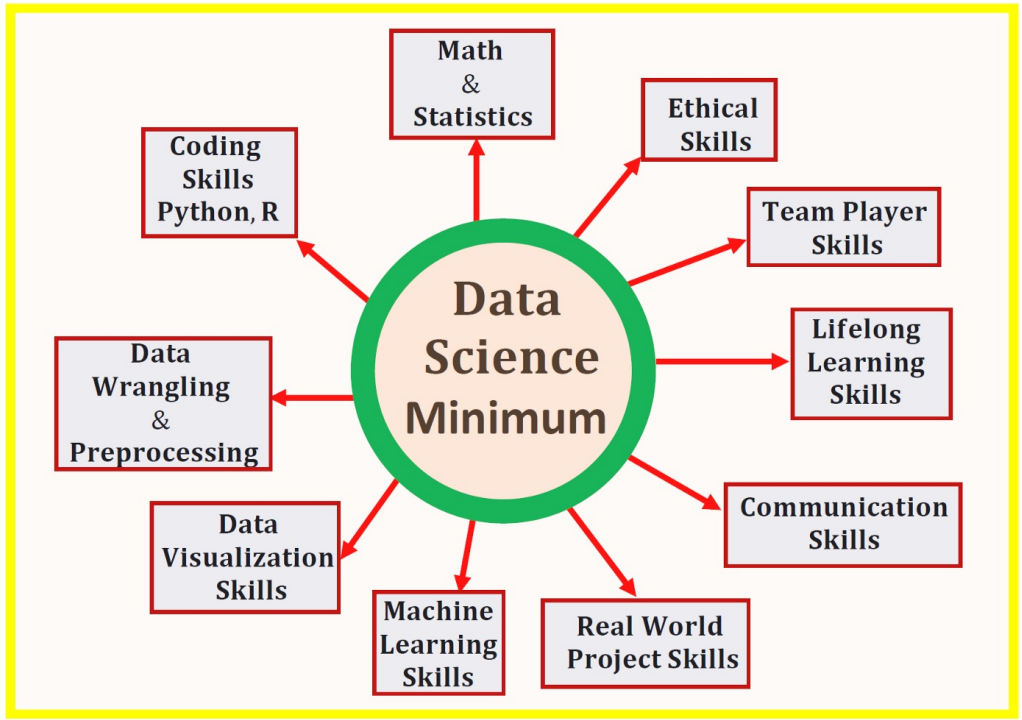
Data visualization

Mapping

Text manipulation and data cleaning

Statistical perspective: hypothesis tests and confidence intervals

Machine learning perspective: supervised and unsupervised learning



Tentative plan for the semester: subject to change!

Week	Date	Topic	HW Assigned	HW Due
1	Jan 17-19	Class overview and intro to Python	0	
2	Jan 24-26	Descriptive statistics and plots	1	29-Jan
3	Jan 31-Feb 2	Data manipulation	2	5-Feb
4	Feb 7-9	Data visualization	3	20-Feb
5	Feb 14-16	Array computations	4	19-Feb
6	Feb 21-23	Mapping	5	26-Feb
7	Feb 28-Mar 2	Text manipulation and data cleaning	6	5-Mar
8	Mar 7	Review		
	Mar 9	Midterm		
		Spring break!		
9	Mar 28-30	Statistics perspective: hypothesis tests	7	2-Apr
10	Apr 4-6	Statistics perspective: confidence intervals	Draft of final project	9-Apr
11	Apr 11-13	Machine Learning: supervised learning	8	17-Apr
12	Apr 18-20	Machine Learning: unsupervised learning	9	23-Apr
13	Apr 25-27	Ethics and review	Final project	30-Apr
Final exam	May 10th, 9am	In person final exam		

Learning goals

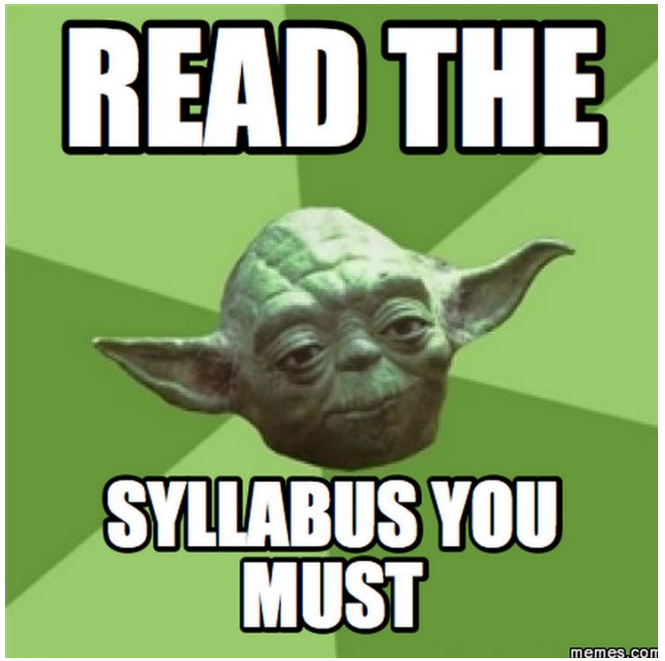
1. Understand concepts in Data Science
    - Learn basic computational skills for analyzing data
    - Understand concepts in Statistics and Machine Learning

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

Course structure



Two lectures per week

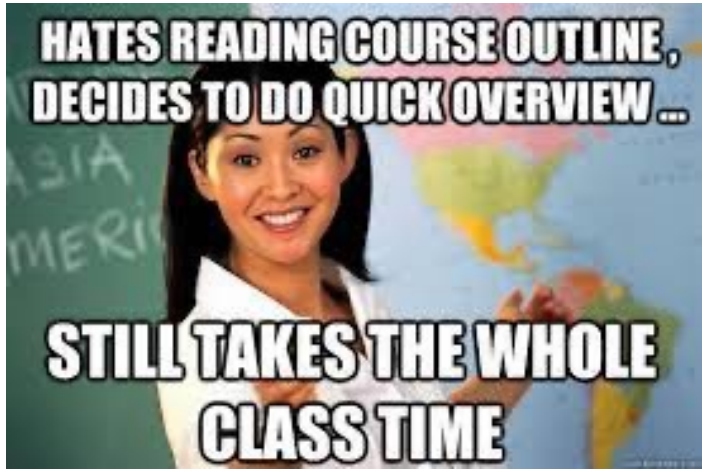
Weekly homework assignments

A class project

Weekly drop-in office hours to get help (see Canvas)

Midterm and final exam

Class readings

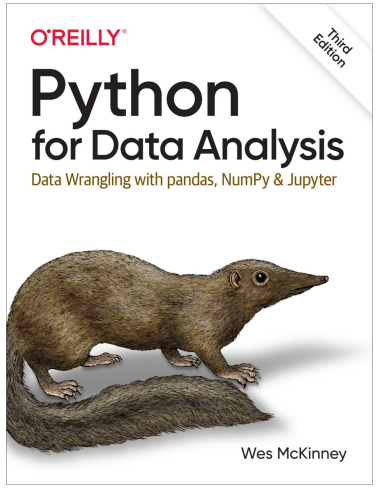


There will be short readings on data science topics approximately every other week

- These will be on Canvas

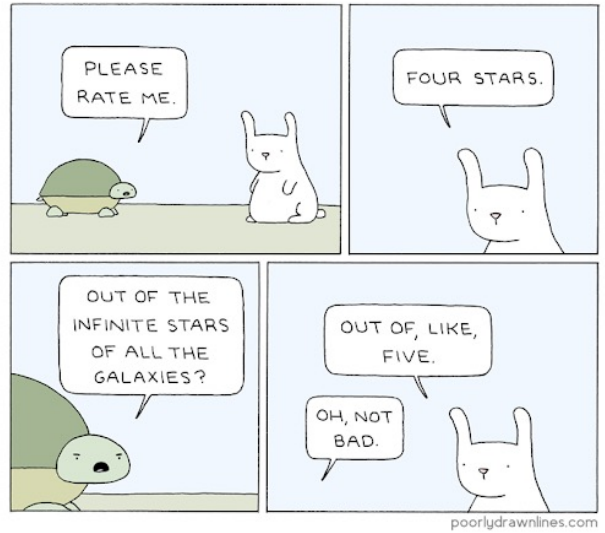
Readings will also be taken from:

- Adhikari and DeNero (2018). Computational and Inferential Thinking
- McKinney (2022). Python for Data Analysis, 3E
- Other sources on the Internet



Resources related to programming will also be posted on Canvas under the appropriate class

Assignments and grades



1. Homework problem sets (48%)
   - Exploring concepts and analyzing data using Python
   - Weekly: 9 in total

Homework policies

- You may discuss questions with other but the work you turn in must be your own!
- Worksheets assigned on Mondays and are due at 11pm on Sundays
  - (with a 59 minute grace period)
- Late worksheets (90%) credit if turned in by 11pm on Monday
  - For any other extensions a Deans Excuse is needed
- Lowest scoring worksheet will be dropped

Assignments and grades

2. Project (10%)
   - A draft of your class project is due 2/3^{rds} of the way through the semester
   - You will give and receive feedback from your peers
   - Final version of the project will be turned in at the end of the semester

3. Exams (40% total)
   - Midterm: March 9th during the regular class time (15%)
   - Final during finals period (25%)

4. Participation (2%)
   - Active asking and answering questions on Ed Discussions

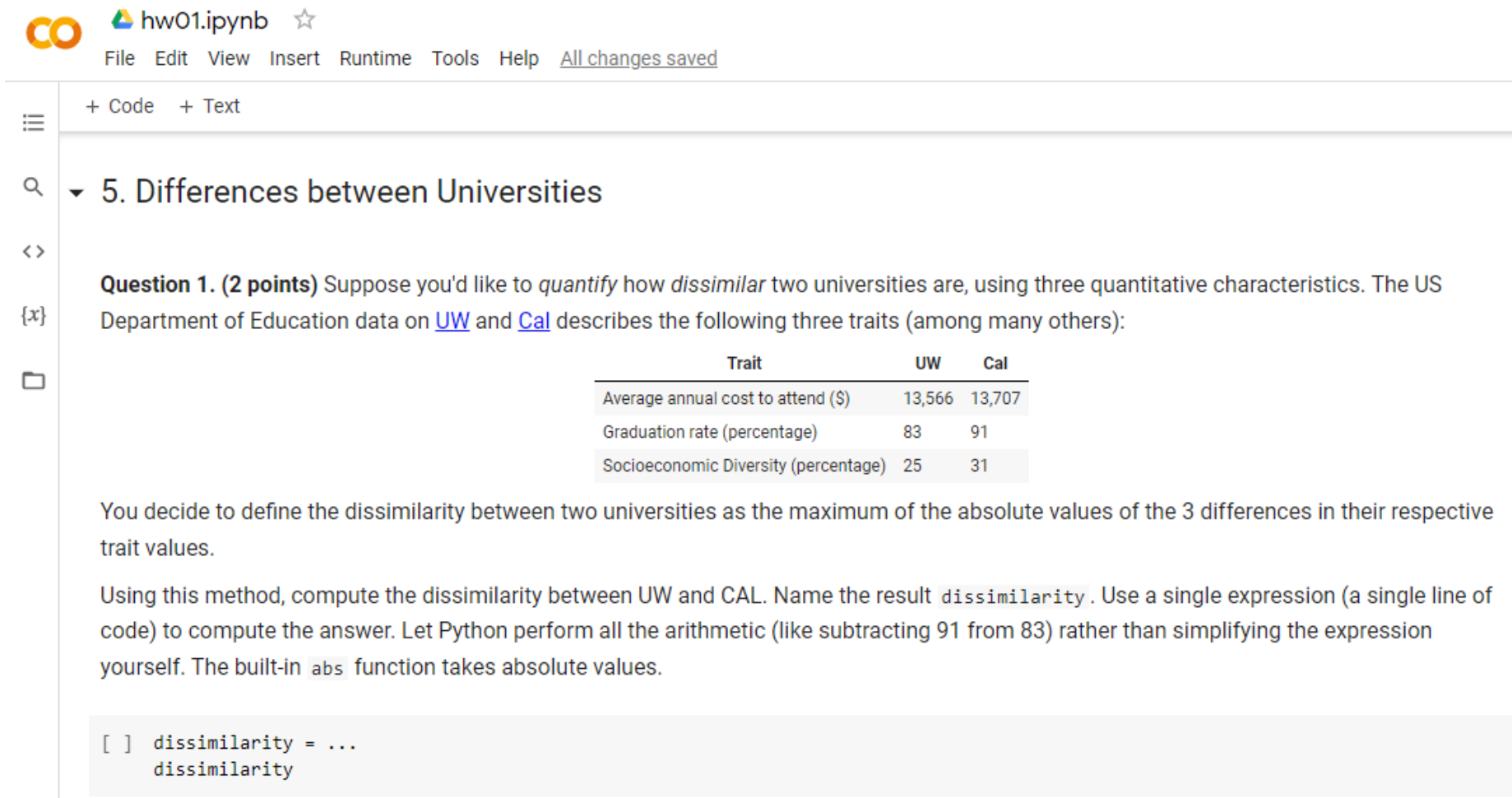
Policies

Accommodation: please let me know if you have accommodations for homework and/or exams

Academic dishonesty: Don't do it!

- You can work with others on the homework but the work you turn in needs to be your own
- Any student who turns in work for credit that is identical, or similar beyond coincidence, to that of another student may face appropriate disciplinary action at the department, college, or university level. *Cheating and/or plagiarism will not be tolerated.*
- If you get ideas or words from a website, journal article, book, another person, etc., cite the source in your work.
- You can't talk with others on exam, etc.

A typical homework assignment



hw01.ipynb ☆

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

5. Differences between Universities

Question 1. (2 points) Suppose you'd like to *quantify* how *dissimilar* two universities are, using three quantitative characteristics. The US Department of Education data on UW and Cal describes the following three traits (among many others):

Trait	UW	Cal
Average annual cost to attend ($)	13,566	13,707
Graduation rate (percentage)	83	91
Socioeconomic Diversity (percentage)	25	31

You decide to define the dissimilarity between two universities as the maximum of the absolute values of the 3 differences in their respective trait values.

Using this method, compute the dissimilarity between UW and CAL. Name the result `dissimilarity`. Use a single expression (a single line of code) to compute the answer. Let Python perform all the arithmetic (like subtracting 91 from 83) rather than simplifying the expression yourself. The built-in `abs` function takes absolute values.

```
[ ]  dissimilarity = ...
     dissimilarity
```

Running Jupyter Notebooks

In order to do the homework, you will need to be able to run Jupyter Notebooks and install the datascience package

There are a few ways to do this:

- Install Anaconda on your own computer
- Use Google Colabs with Google drive

Homework 0 allows you to test that you have a working Jupyter Notebook environment

- Homework 0 is not turned in, but please try it soon
- Ask questions on Ed Discussions or go to office hours to get help

Class survey

In order to get to know you and to adjust the class to everyone's interests, please fill out the class survey on canvas

- Under the Quizzes link on the left

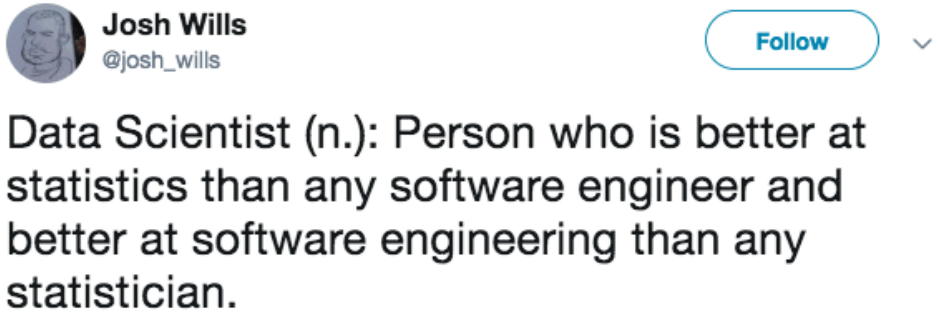
Any questions about the class logistics???

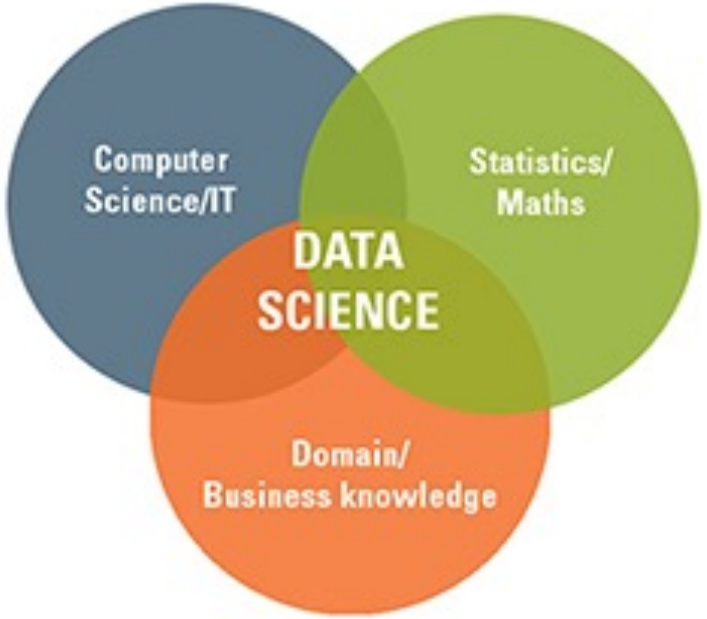
- Ask on Ed Discussions!

What is Data Science?

What is Data Science?

Thoughts?





"A Data Scientist is a Statistician who lives in San Francisco"

Let's see if we can gain an understanding of what Data Science is by looking at some history…

Brief history of Data Science: data

The first data we know of:

- The **Ishango bone** is a bone tool and possible mathematical device discovered at in the Democratic Republic of Congo
- Believed to about 20,000 years old

Cuneiform tablets from Uruk, a Mesopotamian settlement 5,000 years old contained transaction data on commodities

Brief history of Data Science: early computational devices

Some early computational devices include:

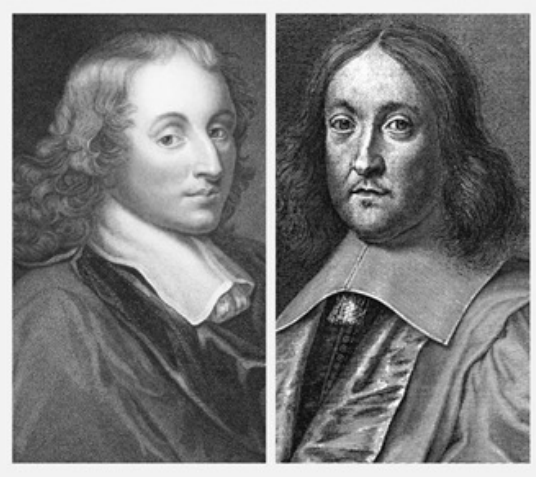
- The abacus comes from Babylon in 2400 BCE
- Antikythera mechanism (~100 BCE) is an ancient Greek hand-powered device describe as the oldest example used to predict astronomical positions and eclipses decades in advance.

Brief history of Data Science: demography and probability

John Graunt (1620-1674) develops statistical census methods that provided a framework for modern demography. He is credited with producing the first life table, giving probabilities of survival to each age.

The mathematics of probability began to be developed in Europe starting in the 17th century

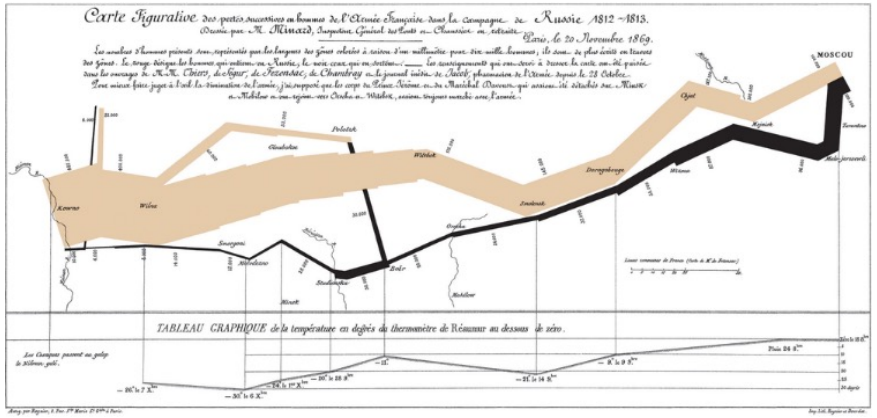
- Fermat and Pascal (1654), Bernoulli (1713), De Moivre (1718), Gauss and Laplace (1812)



Brief history of Data Science: visualization and math

In the second half of the 19th century:

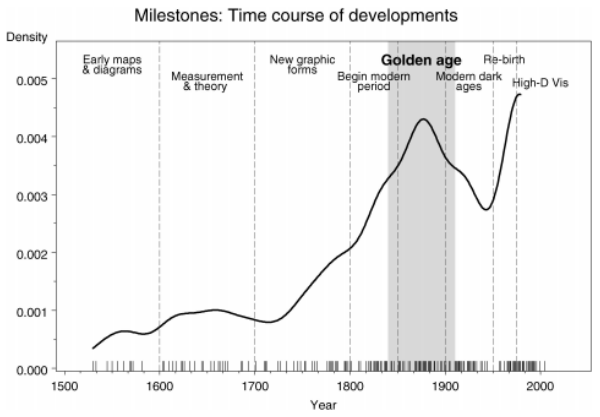
1. The field of Statistics uses probability models to analyze data
   - Galton, Pearson, Fisher, Neyman

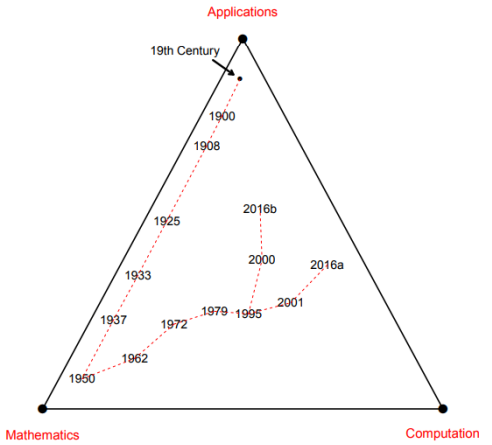


2. Elaborate visualizations of data were published

Probability models dominate Statistics in the first half of the 20th century

Experimental data becomes dominant in the science and medicine in the 2nd half of the 20th century



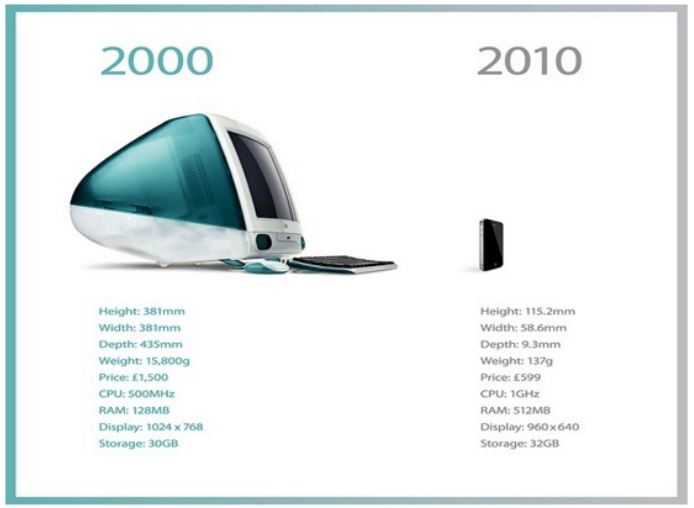


Brief history of Data Science: the rise of computers

Herman Hollerith develops the Hollerith Tabulating Machine for the 1890 census (reduces 10 years of work to 3 months). Creates IBM.

Computer technology develops rapidly over the second half of the 20th century

- Mainframe computers developed in the 1940's
- Relationship database developed in 1970
- Personal computers developed in the 1970's and 1980's
- World Wide Web developed in 1989
- iPhone developed in 2007
- Etc.

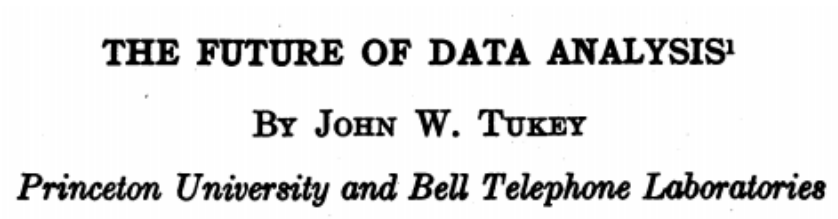


Brief history of Data Science: the rise of Data Science

The rise of powerful computers and plentiful data has given rise to new approaches to analyzing data.

- John Tukey (1962) looks for a broadening of data analysis beyond mathematics
- Breiman (2001) describes a mathematical modeling culture and algorithmic culture
- The term "Data Science" starts being used in the 2000's to describe computational approaches to analyzing data

[Yale renames the department of Statistics to be the Department of Statistics and Data Science]


THE FUTURE OF DATA ANALYSIS[1]

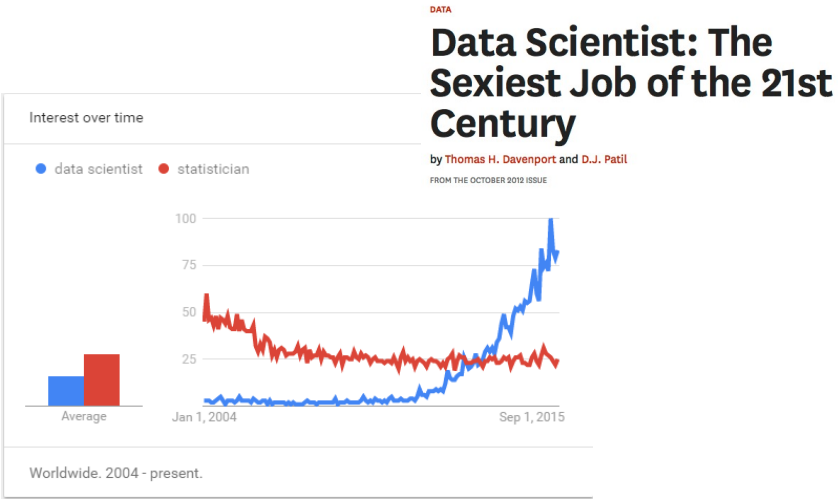
By John W. Tukey

Princeton University and Bell Telephone Laboratories


Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman



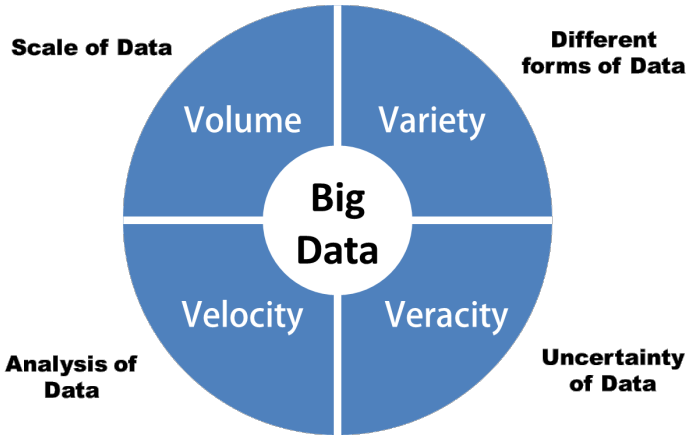
Big Data

New insights:

- Lots of new data from Internet, sensors etc., can be mined to transform our understanding in a range of fields
  - E.g., health, cosmology, social sciences, etc.

New analysis and approaches:

- Hypothesis test pick up on very small (meaningless) effects with very large samples
- Data manipulation and programming are needed to extract insights
- Also, new standards for choosing the best data analysis methods

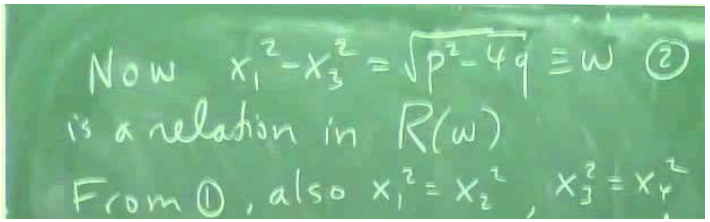


New ways to choose the best methods

Statistics focuses on mathematical models (probability distributions) to analyze data

- Best methods are the ones that have mathematical guarantees (proofs)

The proof is in the math



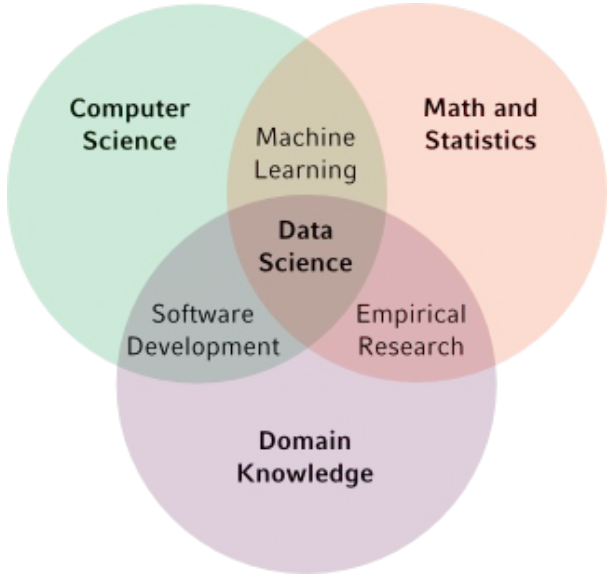
Data Science empirically evaluates data analysis methods

- Best methods are the one that gives the most insight in practice

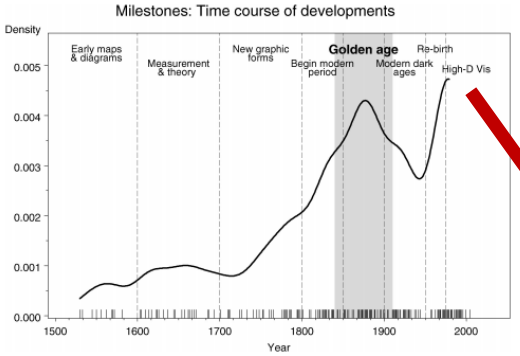
The proof is in the pudding

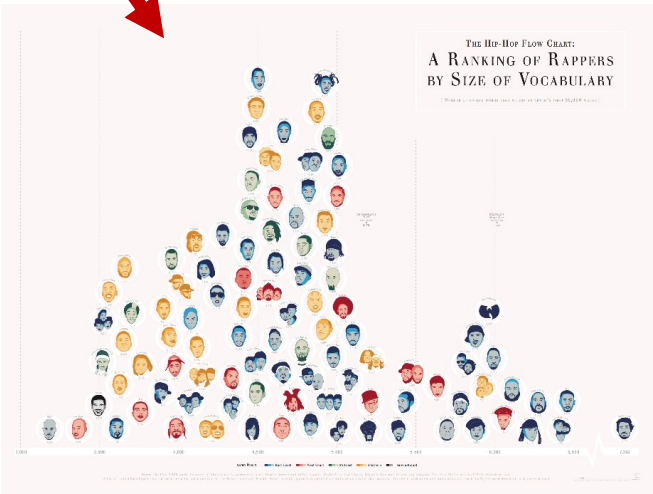
Data Science vs. Statistician video

What is Data Science?



Data Science is a broadening of data analyses beyond what traditional Statistical mathematical/inferential analyses to use more computation





Many other fields impacted by 'Data Science

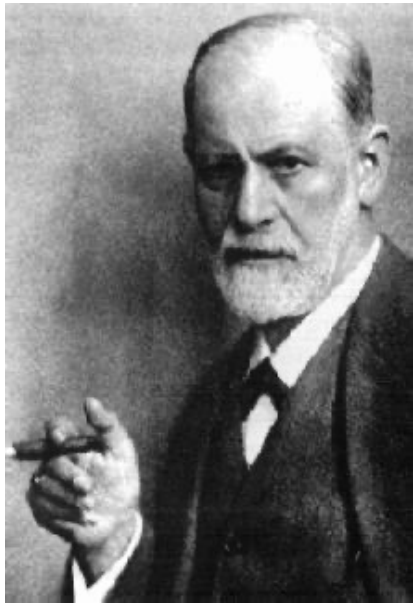
- Making business decisions
- Predictive medicine
- Fraud detection
- Etc.

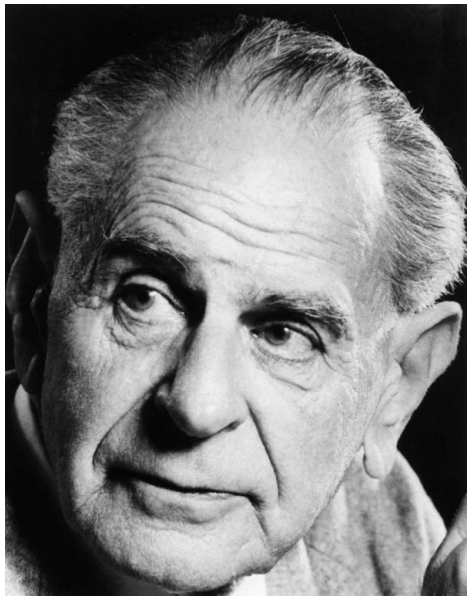
Examples:

- NYC city bike visualization
- Wind map visualization

Ethical concerns around privacy, fairness and other issues

Short paper to read from the book Everybody lies





Much of Freud's theory dealt with the subconscious

- E.g., Freudian slips

Karl Popper claimed that Freud's theories were unscientific because they couldn't be falsified

- i.e., can come up with any 'just so' story to explain a behavior

New data science analyses might make it possible to actually test Freud's theories

Things to do for next class…

1. Complete class survey

2. Do short reading on Canvas from the book "Everybody lies"

3. Try to either install Anaconda on your computer and/or try out Google Colabs.
   - i.e., do homework 0. This does not need to be turned in but you should try to complete it by Sunday 1/22.

Introduction to Python

Programming languages for Data Science

The two most popular languages for Data Science are:

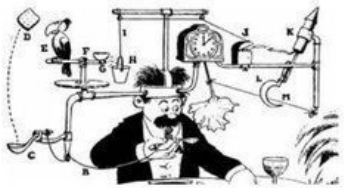
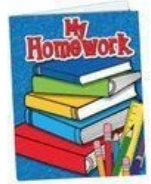
General purpose programming language

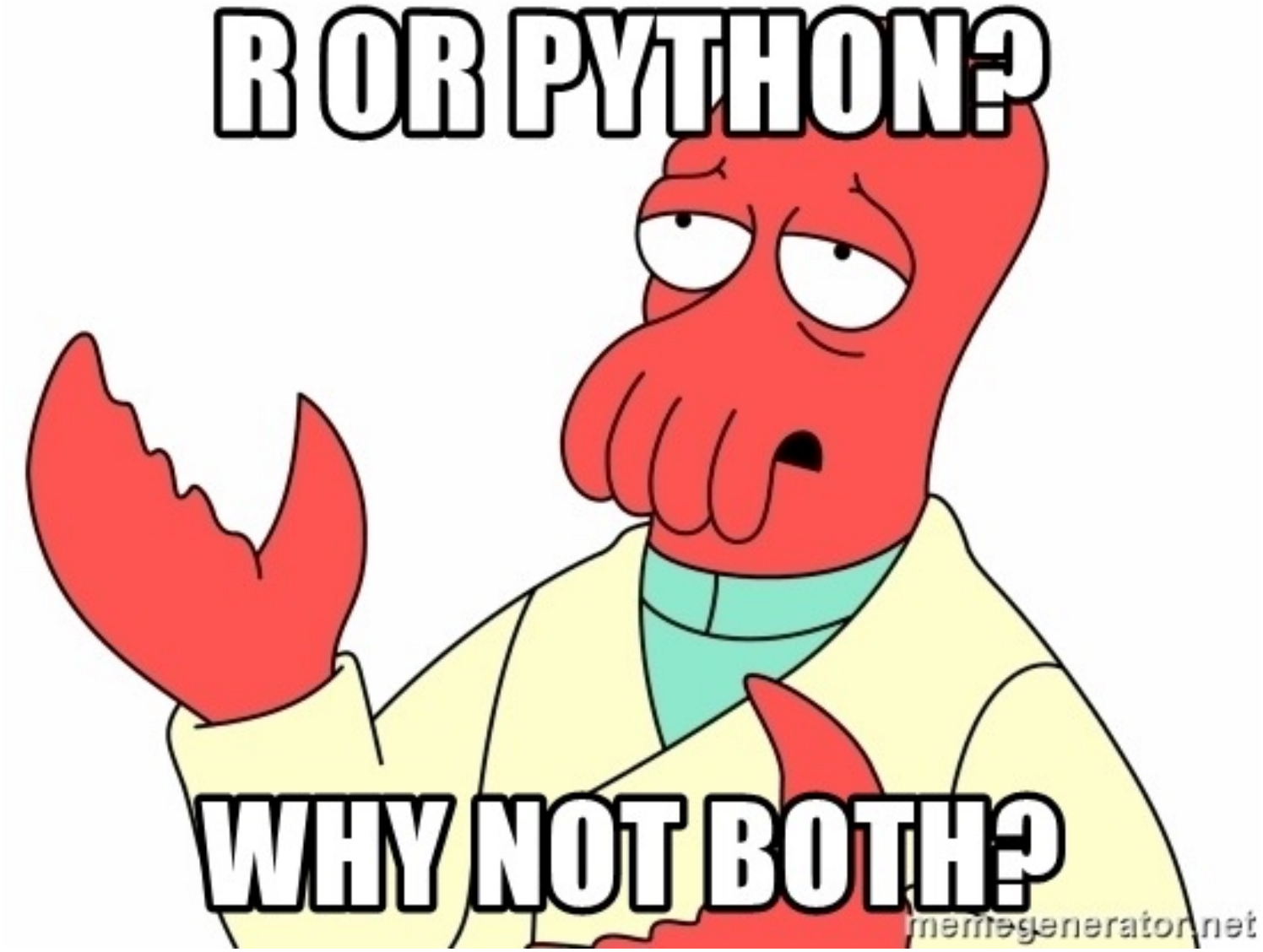
- Can do a lot more than data analysis
- Easy to read
- Easy to write larger software packages
- Good machine learning package (scikitlearn)

Focused on data analysis

- Better for creating pdf reports
- Easy to create interactive apps
- RStudio created a great IDE and support

AS SEEN BY USERS OF …

	STATA	R	SAS	Python	SPSS
STATA					
R					
SAS					
Python					
SPSS					


R OR PYTHON?
WHY NOT BOTH?
memegenerator.net

Terminology: scripts, modules, and packages

A **script** is a piece of code that is run to accomplish a specific task

- E.g., one could write and run a script to download a set of files

A **module** is a piece of code that reused by different scripts and programs

- Modules are ***imported*** by other programs/scripts to add commonly used functionality.
- E.g., import matplotlib.pyplot as plt

A **package** (library) contains several related modules

Packages are an essential building block in programming. Without packages, you would waste a lot of time writing code that's already been written

Environments and conda

Environments allow one to switch between different versions of packages

- This is useful because packages are often updated
  - (and one might want to use an older version)

We will be using **conda** to create an environment that has the data science packages we will use in this class

We will use an environment called *ydata123_2023d*

- It contains packages such as matplotlib, numpy, etc.
- Instructions for creating this environment are on Canvas

Jupyter notebooks

Jupyter notebooks allow one to create an analysis document that contains text, analysis code, and plotted results

Because one can see all the code used to generate results, it allows one to create reproduce analyses

Let’s explore a Jupyter notebook now…

[5]:
```
import matplotlib.pyplot as plt
plt.style.use('classic')
%matplotlib inline
import numpy as np
import pandas as pd
import seaborn as sns
sns.set()
```

[6]:
```
rng = np.random.RandomState(0)
x = np.linspace(0, 10, 500)
y = np.cumsum(rng.randn(500, 6), 0)
```

Next step

Now, create a graph.

[7]:
```
plt.plot(x, y)
plt.legend('ABCDEF', ncol=2, loc='upper left');
```

Things to do for next class…

1. Complete class survey

2. Do short reading on Canvas from the book "Everybody lies"

3. Try to either install Anaconda on your computer and/or try out Google Colabs.
   - i.e., do homework 0. This does not need to be turned in but you should try to complete it by Sunday 1/22.