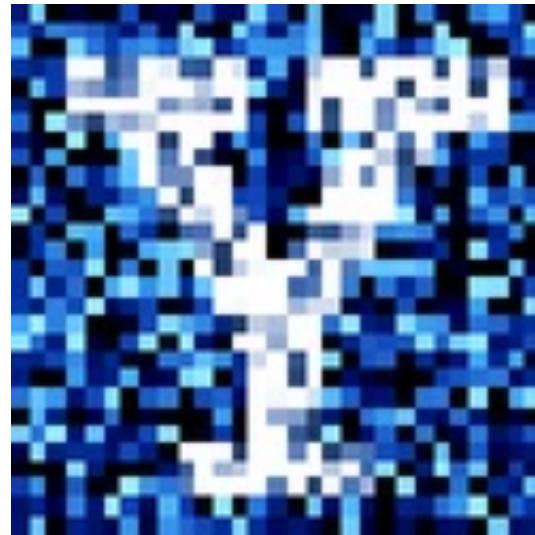


# YData: Introduction to Data Science



Class 01: Introduction

# Overview

## Course overview

- Introductions
- Syllabus and logistics



## What is data science?

- Brief history of data science

If there is time: Intro to Python

# Office hours and contact information

Ethan Meyers (he/him)

Email: [ethan.meyers@yale.edu](mailto:ethan.meyers@yale.edu)

**Note:** I'm going to be a little distracted by a new neural network...

Office hours:

- Mondays and Wednesdays, 2-3pm
  - (subject to change)

Office: Kline Tower, room 1253

- <https://yale.zoom.us/j/96419395464>

# Teaching Assistants

## Preceptor

- Shivam Sharma: [shivam.sharma@yale.edu](mailto:shivam.sharma@yale.edu)



## Teaching Fellows

- Alex Amari: [alex.amari@yale.edu](mailto:alex.amari@yale.edu)

## Undergraduate Learning Assistants

- James Poe: [james.poe@yale.edu](mailto:james.poe@yale.edu)
- Mark Ayiah: [mark/ayiah@yale.edu](mailto:mark/ayiah@yale.edu)
- Sloane Huey: [sloane.huey@yale.edu](mailto:sloane.huey@yale.edu)

TA office hours are on the calendar on Canvas

# Introductions

Let's do some quick introductions

Create groups of ~4 people:

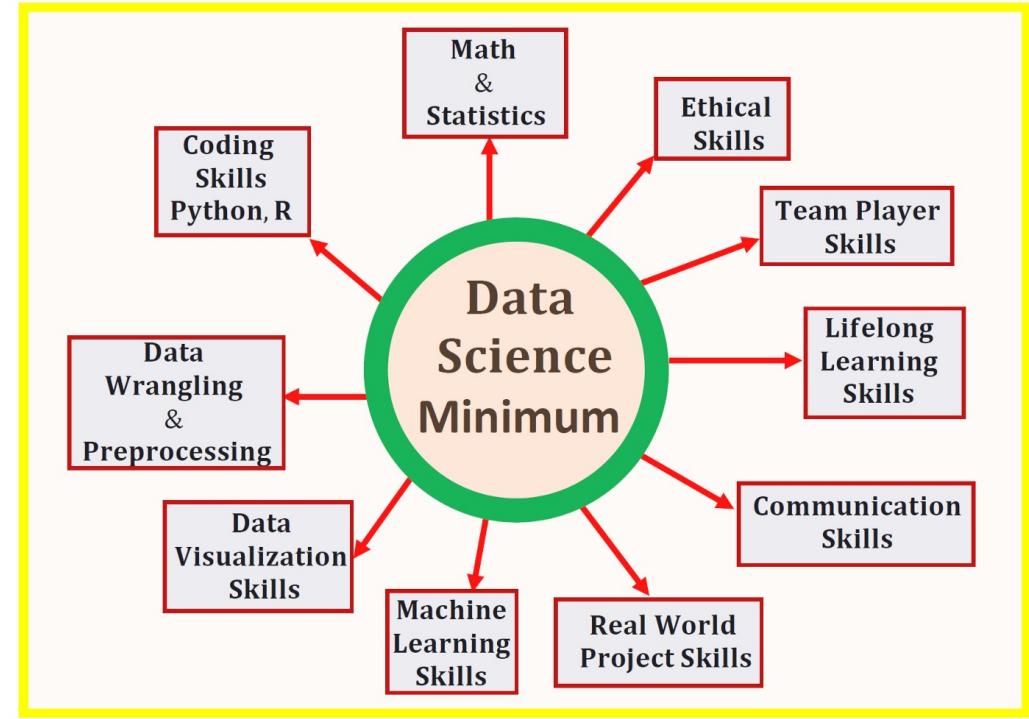
- Your name and preferred gender pronouns
- Your major/grad dept (research area)
- Why you are interested in this class
- Anything else you would like to share with your group



# About this class

# 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



# Tentative plan for the semester: subject to change!

Week	Date	Topic	HW Assigned	HW Due
1	Aug 29	Class overview	0	
2	Sep 3-5	Introduction to Python	1	8-Sep
3	Sep 10-12	Descriptive statistics and plots	2	15-Sep
4	Sep 17-19	Array computations	3	22-Sep
5	Sep 24-26	Tables and data manipulation	4	29-Sep
6	Oct 1-3	Data visualization	5	6-Oct
7	Oct 8-10	Review and midterm exam		
8	Oct 15	Interactive graphics		
	Oct 17	October break		
9	Oct 22-24	Mapping and text manipulation	6	27-Oct
10	Oct 29-31	For loops and writing functions	7	3-Nov
11	Nov 5-7	Statistics perspective: hypothesis tests	Draft of final project	10-Nov
12	Nov 12-14	Statistics perspective: confidence intervals	8	24-Nov
13	Nov 19-21	Machine Learning	9	1-Dec
	Nov 26-28	November recess		
	Dec 3-5	Ethics and review	Final project	8-Dec
Final exam	Dec 16 (Mon) 7pm	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

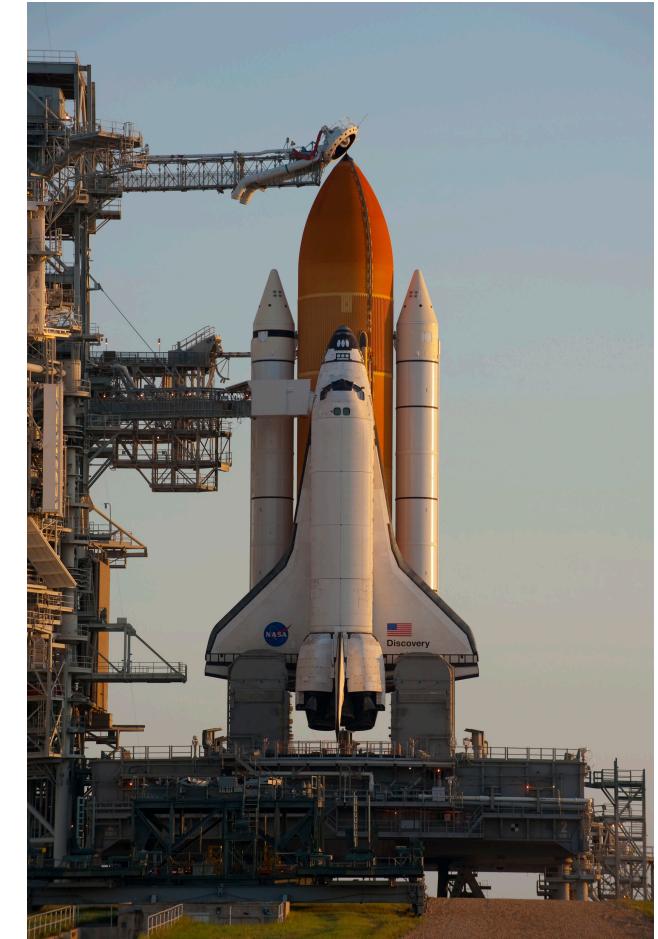
## 2. Gain practical data science skills applicable to any domain

## 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

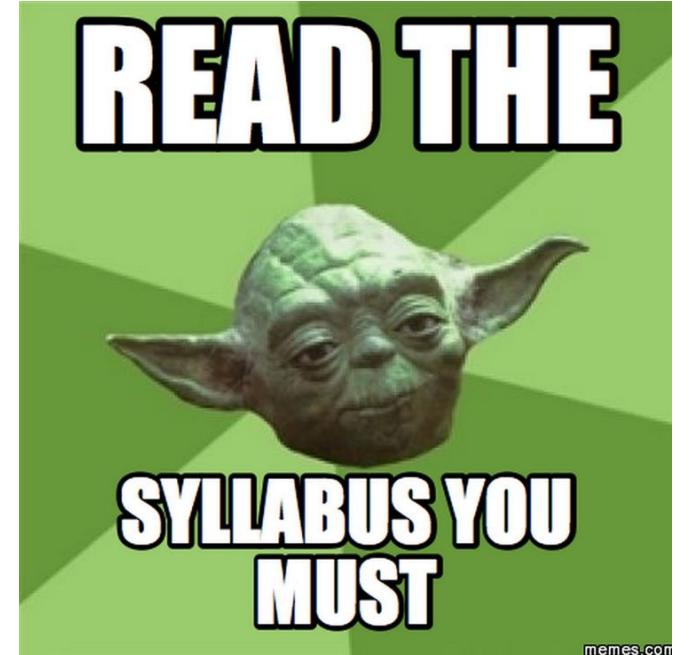
Weekly Python practice sessions (optional)

Weekly homework assignments

A class project

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

Midterm and final exam



# Python practice sessions



Shivam is hosting one-hour practice sessions each week

- Each session will be offered at three different times each week
- Please fill in class survey to let us know what times work for you

Sessions will be a great opportunity to practice Python and get your questions answered!

Highly recommended to attend these sessions

- Attendance is optional
- If you attend all sessions and score below a median cutoff score on the exams, up to 3 points will be added to your score

# Class readings

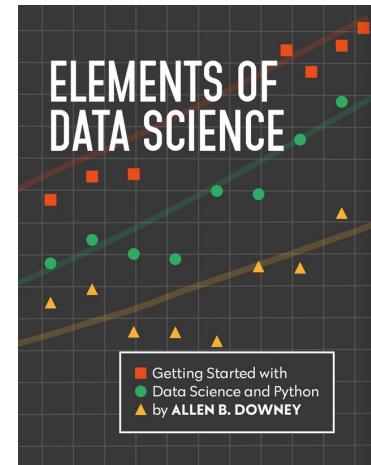
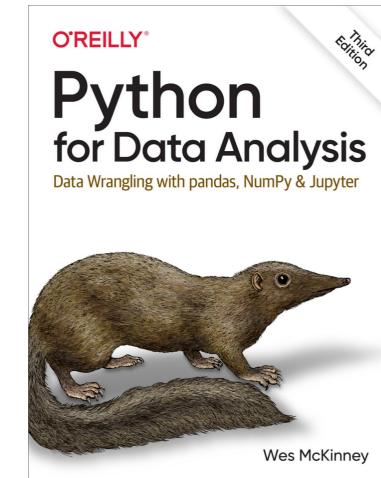
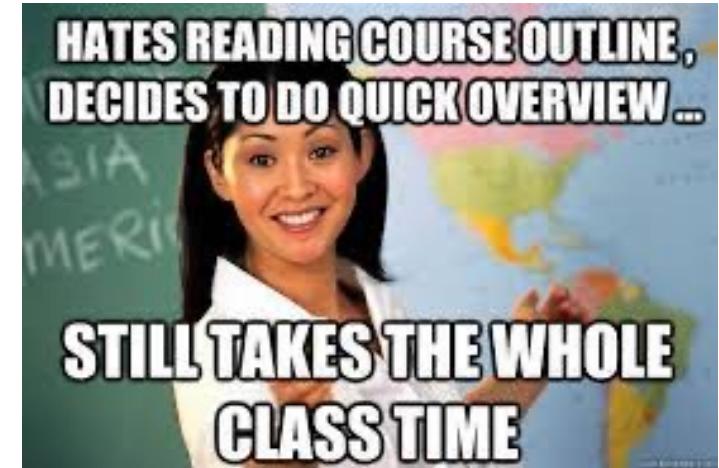
There will be short readings on data science topics approximately every other week

- These will be on Canvas

Readings will also be taken from:

- Brett M (2020). [Data Science for Everyone: course text](#)
- McKinney (2022). [Python for Data Analysis, 3E](#)
- Downey (2024). [Elements of Data Science](#)
- Other sources posted to Canvas/on the Internet

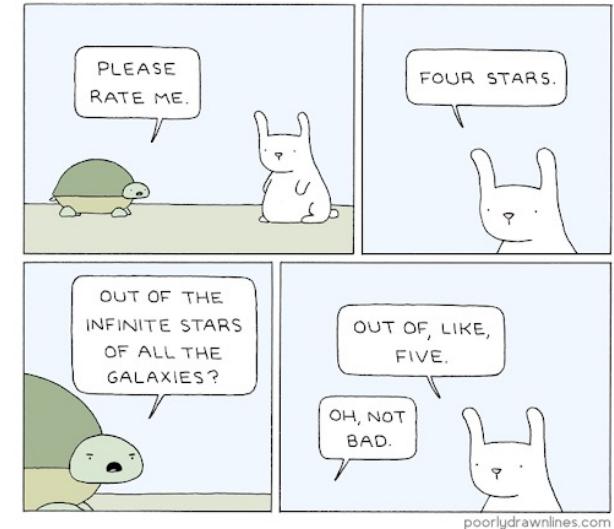
Resources related to programming will also be posted on Canvas under the appropriate class



# Assignments and grades

## 1. Homework problem sets (48%)

- Exploring concepts and analyzing data using Python
- Weekly: 9 in total



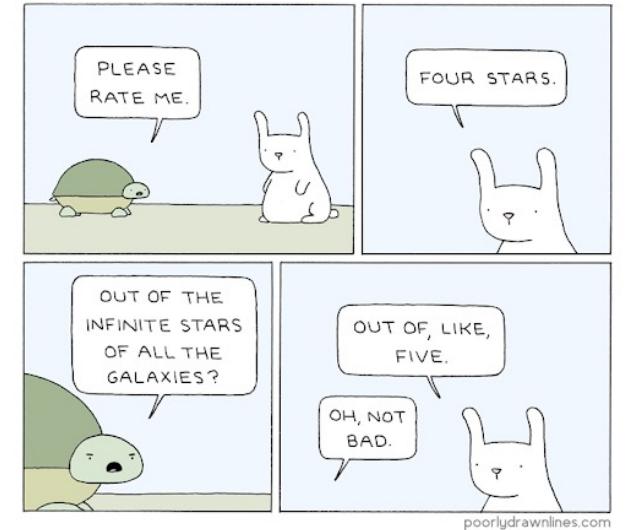
## Homework policies

- 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

## 1. Homework problem sets (48%)

- Exploring concepts and analyzing data using Python
- Weekly: 9 in total



## Homework policies

- You can use chatGPT to answer general questions about Python and/or statistical concepts
  - E.g., It is ok to ask chatGPT “What does the np.sum() function in Python do?”
- You are **not** allowed to use it to directly answer homework questions
  - E.g., It is **not** ok to cut and paste a homework question into chatGPT

# Assignments and grades

## 2. Project (10%)

- A draft of your class project is due 2/3<sup>rds</sup> 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: October 10<sup>th</sup> during the regular class time (15%)
- Final Exam: Monday December 16<sup>th</sup> at 7pm (25%)
- **To take the class you must be able to attend these exams at their scheduled times!**

## 4. Participation (2%)

- Active asking and answering questions on Ed Discussions

# Grade distribution

Grade cut-off are

- A [94-100], A- [90-94), B+ [87-90), B [80-84), etc.
  - I might slightly modify these downward if the class too hard

No strict grade distribution but roughly:

- 25% A, 25% A-, 25% B+, 25% everything else

Students generally score high on the homework (> 90) and exam scores tend to be lower (~80)

If an exam is too hard, I sometimes curve them by adding "free points"

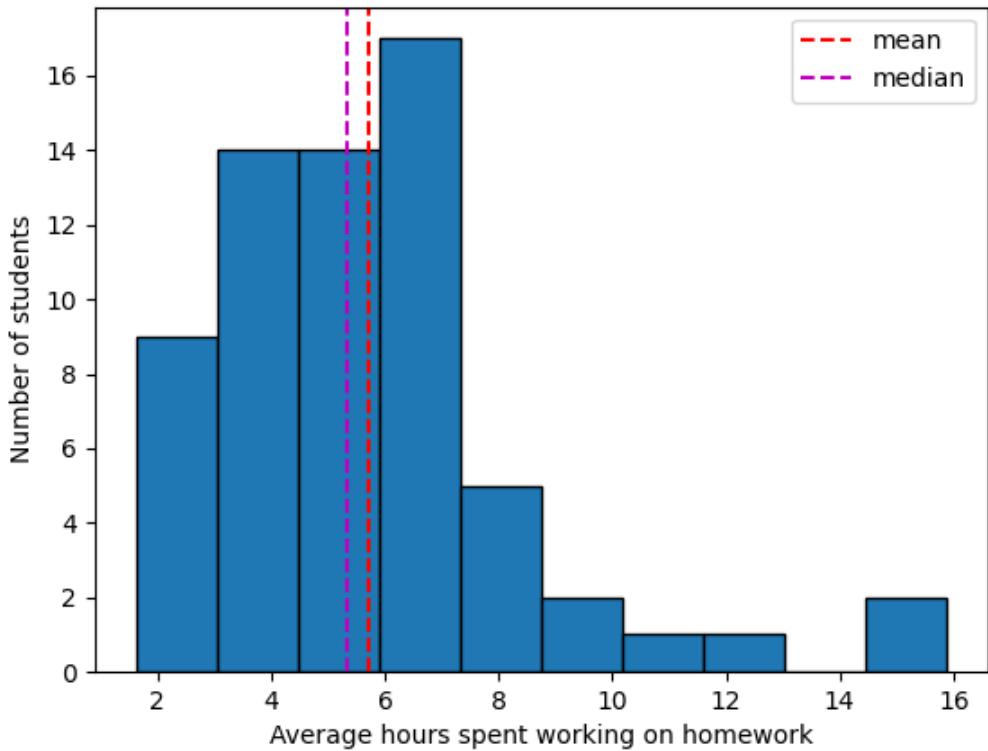
- E.g., if an exam is out of 85 points, I might add a free 15 bonus points so the exam is out of 100



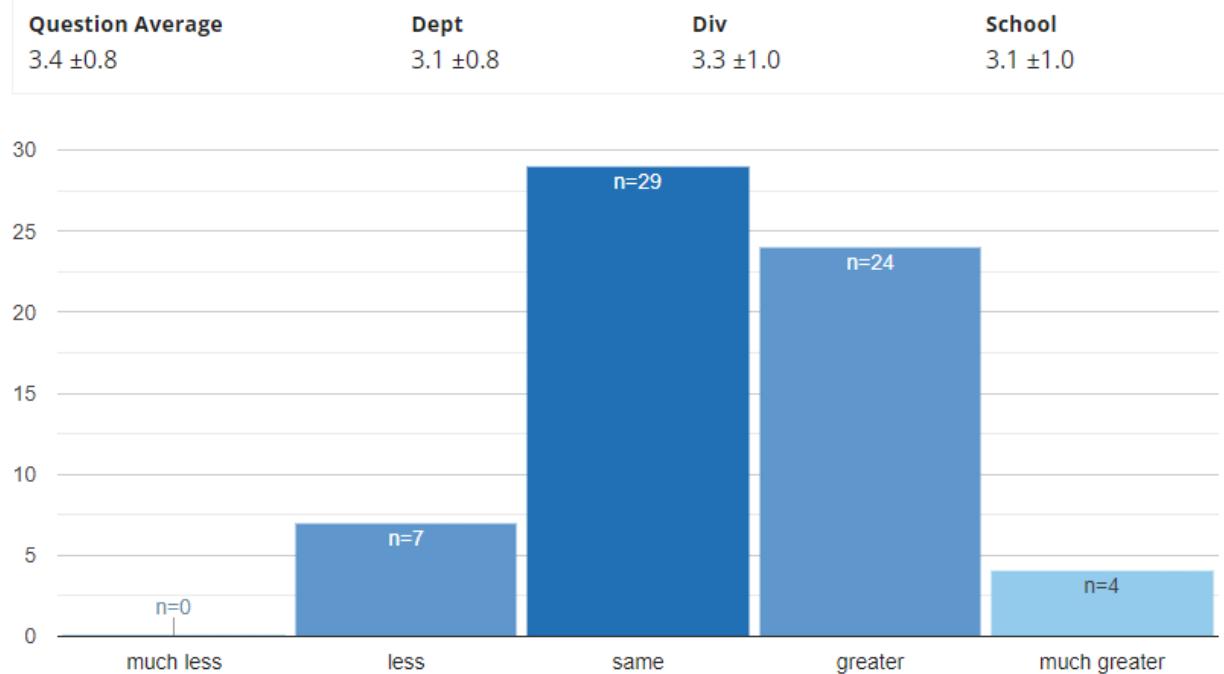
**Please focus on learning rather than on grades!**

# Class workload

Average weekly hours spent working on homework



Relative to other courses you have taken at Yale, the workload of this course was:

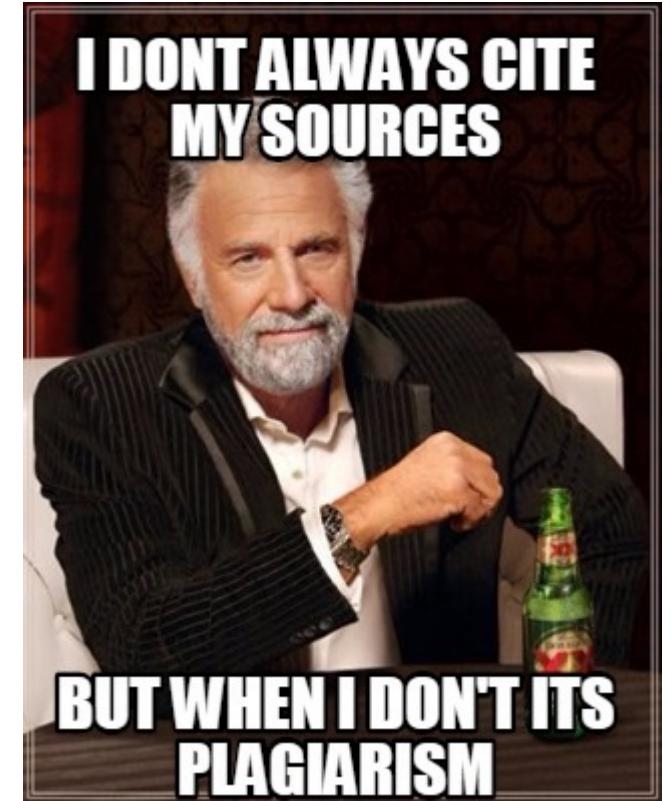


# Policies

**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.
- 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

The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** hw01.ipynb
- Toolbar:** File, Edit, View, Insert, Runtime, Tools, Help, All changes saved
- Code/Text Buttons:** + Code, + Text
- Search/Filter:** 🔎, 🔍, {x}
- Section:** 5. Differences between Universities
- Text:** 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

- Text:** 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.
- Text:** 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.
- Code Cell:** [ ] dissimilarity = ...  
dissimilarity

# Running Jupyter Notebooks

In order to do the homework, you will need to be able to run Jupyter Notebooks

There are a few ways to do this:

- Use the [YCRC Jupyter Server](#)
- Use [Google Colabs](#) with Google drive
- [Install Anaconda on your own computer](#)

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

Grace OnDemand



# Let's test the YCRC Jupyter notebook server...

Before we get started, let's test the [YCRC Jupyter notebook server](#)

A link to the server is at the top of the class Canvas page

**If you can't log in, let us know on the background survey...**

# 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?

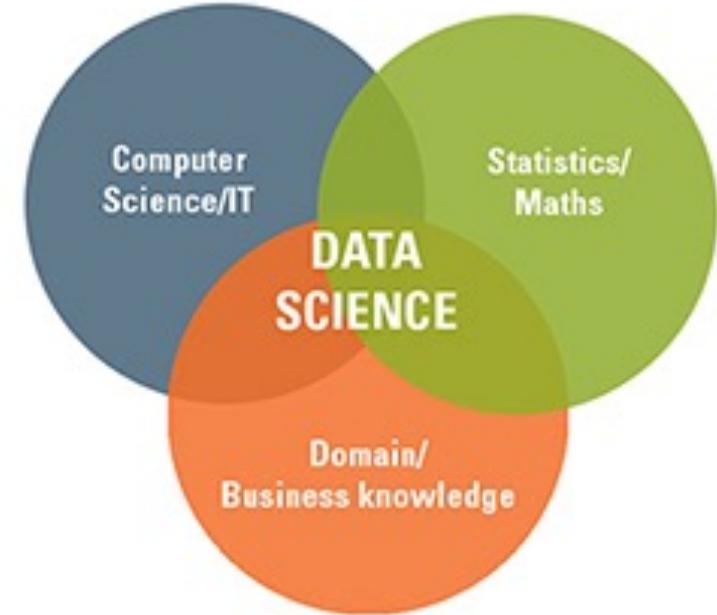
# What is Data Science?

## Thoughts?



Josh Wills  
@josh\_wills

Follow



Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

*"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



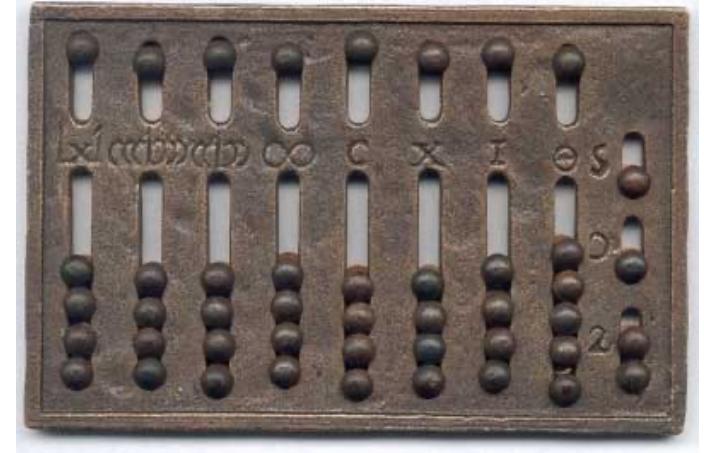
Cuneiform tablets from Uruk, a Mesopotamian settlement 5,000 years old contained transaction data on commodities



# 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 described 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.



CAPTAIN JOHN GRAUNT

The mathematics of probability began to be developed in Europe starting in the 17<sup>th</sup> century

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



# Brief history of Data Science: visualization and math

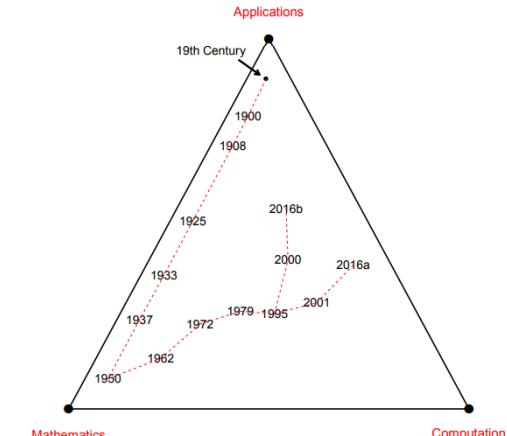
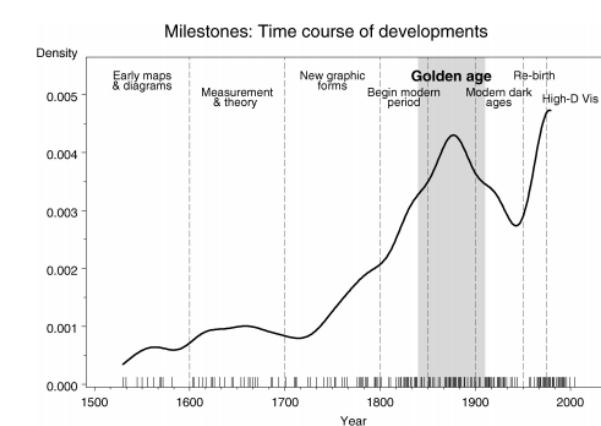
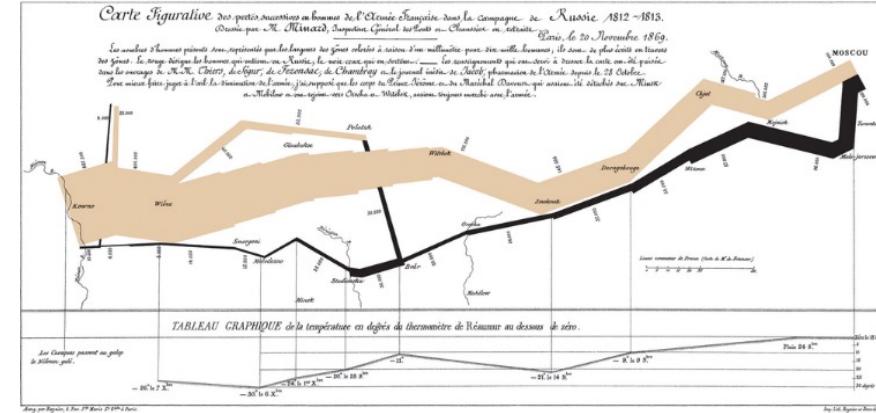
In the second half of the 19<sup>th</sup> century:

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 20<sup>th</sup> century

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

- E.g., Randomized Controlled Trials in Medicine

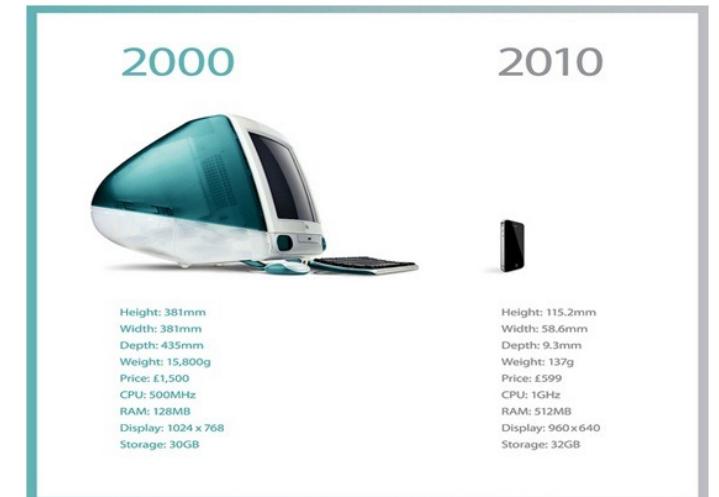


# Brief history of Data Science: the rise of computers

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 20<sup>th</sup> 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
  - E.g., Cleveland 2001

## THE FUTURE OF DATA ANALYSIS<sup>1</sup>

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

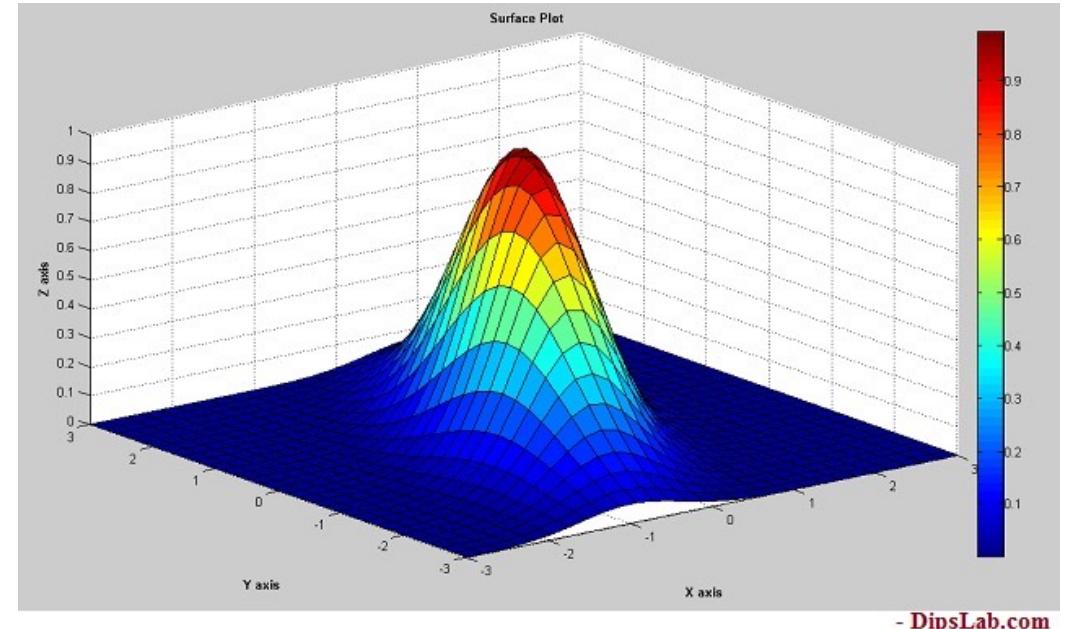
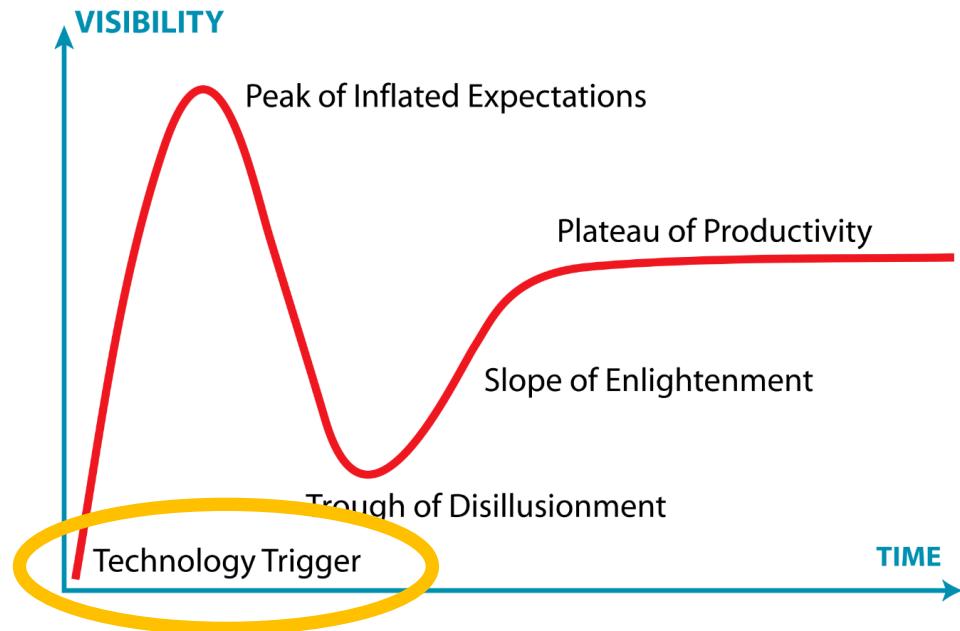
International Statistical Review (2001), 69, 1, 21–26, Printed in Mexico  
© International Statistical Institute

## Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Laboratories, 600 Mountain Avenue, Murray Hill NJ07974, USA  
E-mail: wsc@research.bell-labs.com

# Brief history of Data Science: Technology Trigger

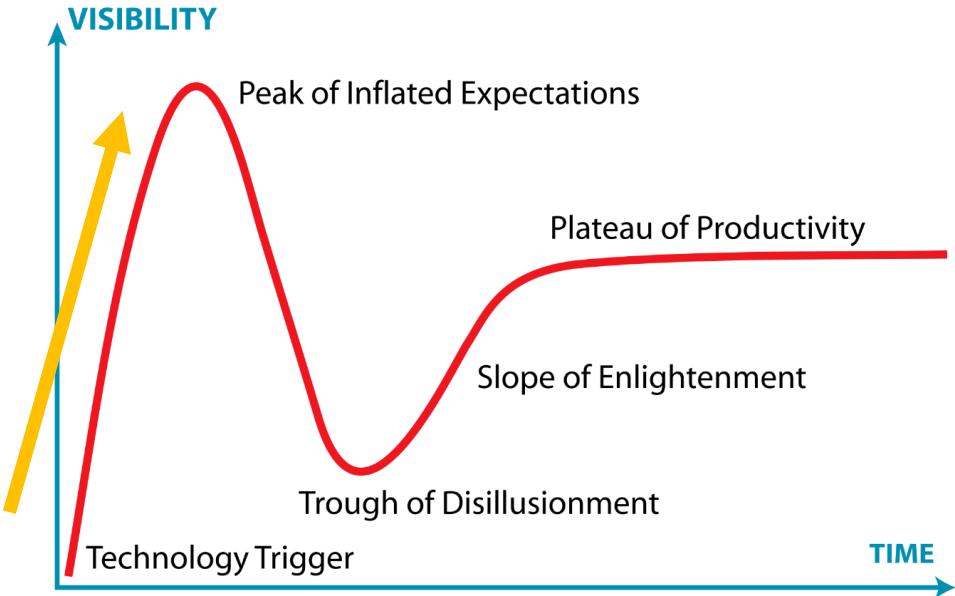


Early 2000's rise of the internet, personal computers and data analysis programming languages changes how data can be analyzed

- MATLAB
- Python: matplotlib 2003, numpy 2005, scikit-learn 2007, pandas 2009

# Brief history of Data Science

2009



Data Science rises with data science competitions, blog posts, and industry jobs

- Data Scientists viewed as “unicorns” because they had to know both statistics and how to program

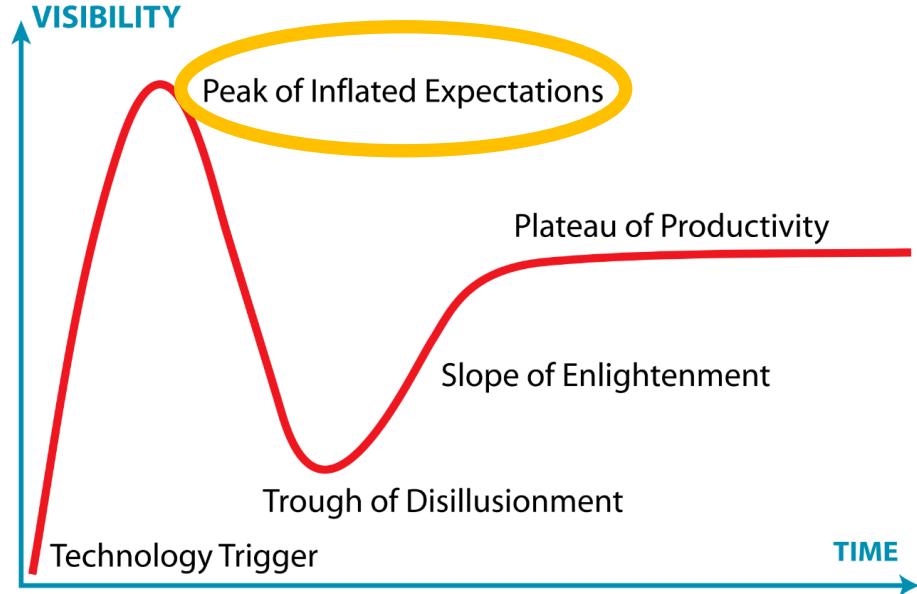
2010

kaggle

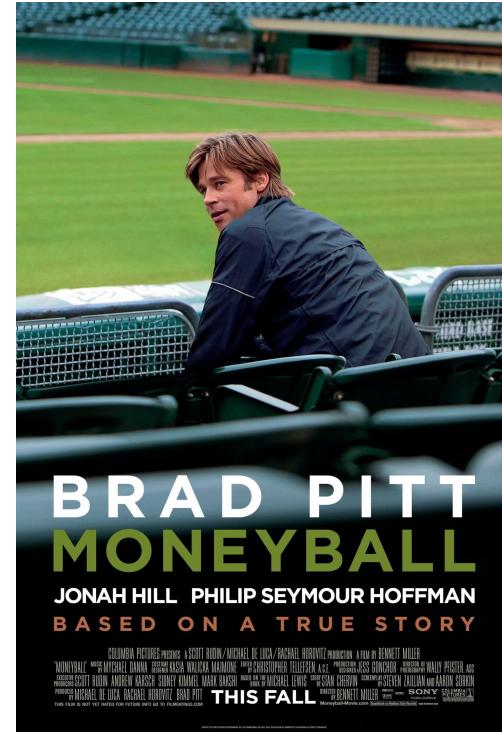
Blog: 2009-2010

okcupid

# Brief history of Data Science



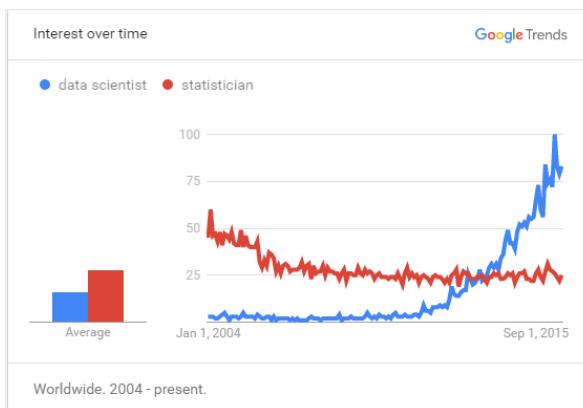
2011



2012



Data Science hits  
“peak of inflated  
expectations”  
around 2012-14



Harvard  
Business  
Review

Latest Magazine Ascend Topics Podcasts Store The Big Idea Data & Visuals Case Selections

2012

## Data Scientist: The Sexiest Job of the 21st Century

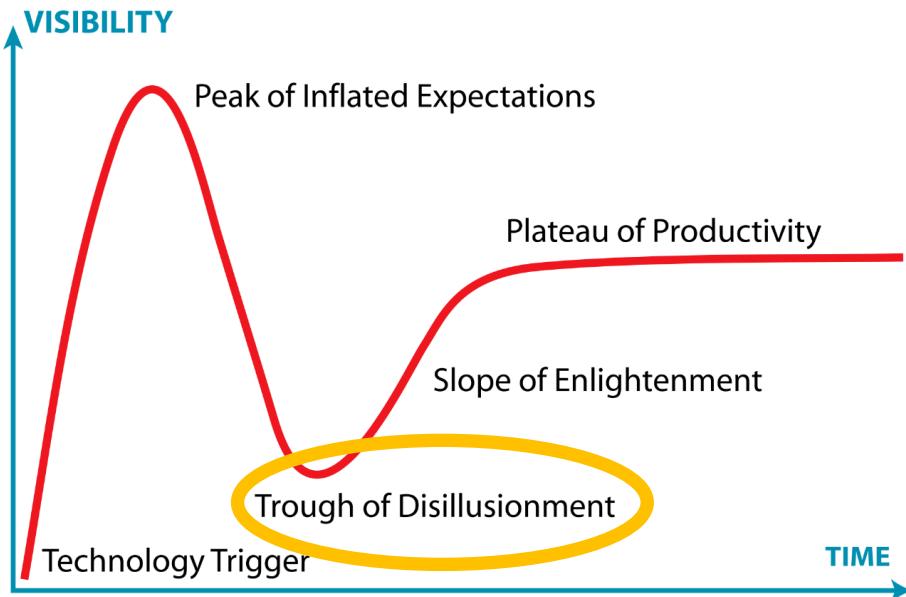
Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

Subscribe

Sign In

# Brief history of Data Science: Technology Trigger



## Negative consequences of predictive models were highlighted circa 2016



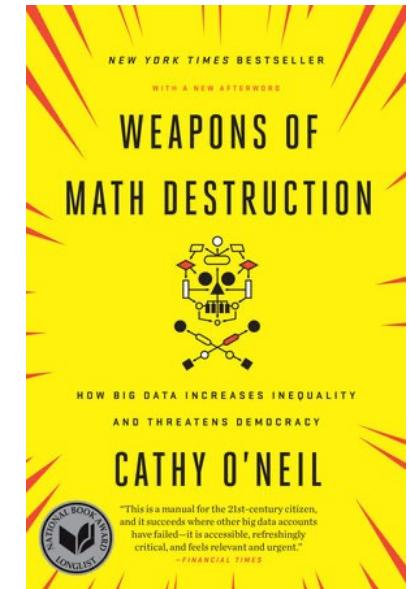
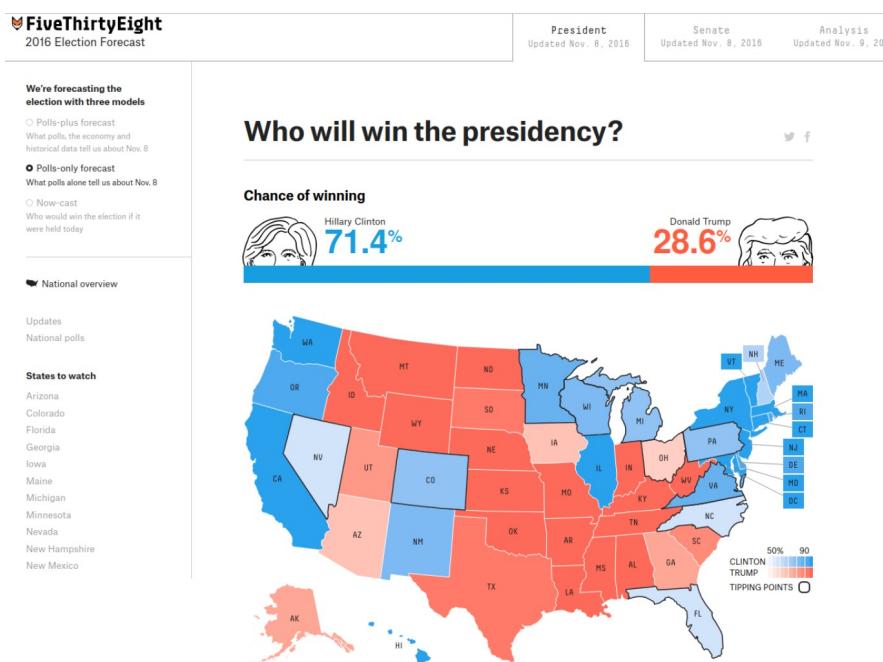
# PROPUBLICA

# Machine Bias

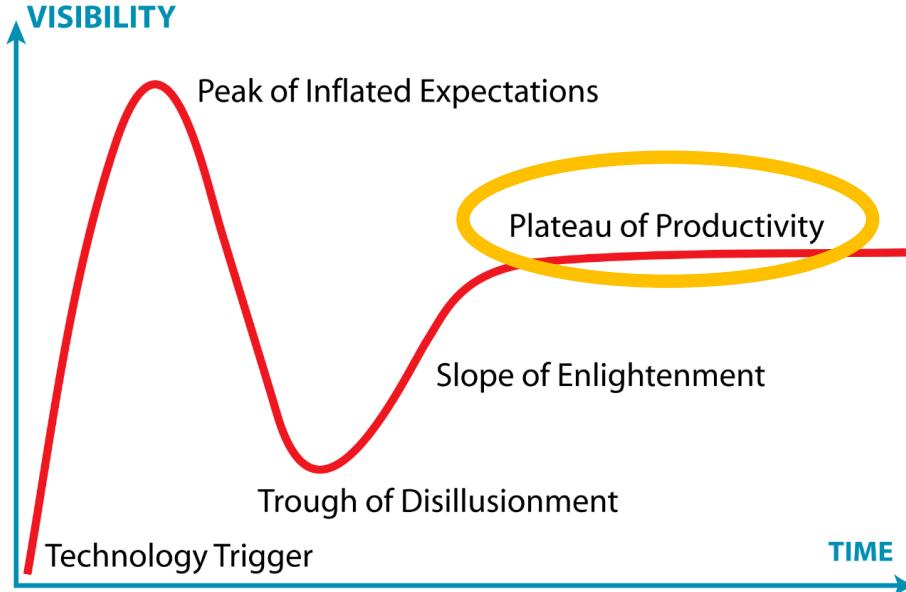
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



# Brief history of Data Science: now



Data Scientists roles in many industries

- E.g., Data journalism ([NYTimes TheUpshot](#))

Many universities have Data Science programs

- In March 2017 Yale renames the Department of Statistics to be the Department of Statistics and Data Science

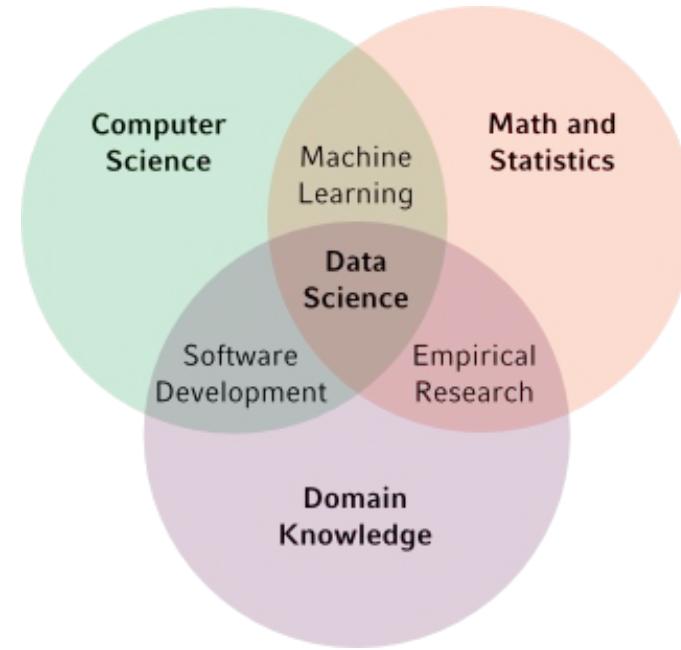
The cover features a dark blue map of the United States with various data visualization icons (charts, graphs, circles) overlaid. The title 'DATA SCIENCE' is at the top, followed by 'IS FOR EVERYONE'. Below the map, the date 'FEBRUARY 2024' and authors 'By Carlo Salerno and Frank Steemers' are listed. Logos for 'THE burningglass INSTITUTE' and 'ExcelinEd' are at the bottom.

Yale  
S&DS

# What is Data Science?

Data Science is a broadening of data analyses:

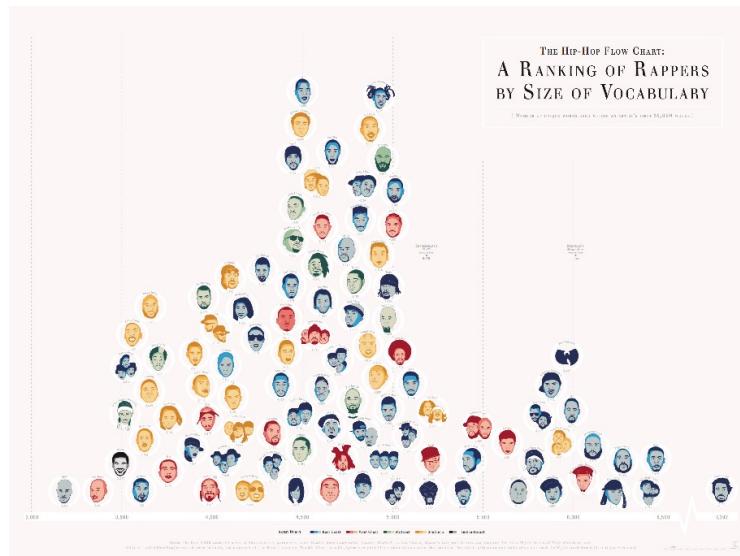
- Beyond what traditional Statistical mathematical/inferential analyses
- To using more computation



## Classical statistical analysis

Descriptives								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Sulfur dioxide	5	18.840	9.6919	4.3344	6.806	30.874	5.1	30.1
Nitrous dioxide	6	6.617	3.9448	1.6105	2.477	10.757	2.2	11.9
Oxygen	4	4.975	3.4092	1.7046	-.450	10.400	2.1	9.3
Total	15	10.253	8.6514	2.2338	5.462	15.044	2.1	30.1

ANOVA					
Bronchial reactivity	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	559.450	2	279.725	6.873	.010
Within Groups	488.408	12	40.701		
Total	1047.857	14			



## Examples:

- [NYC city bikes](#)
- [Wind map visualization](#)

# 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**

Now  $x_1^2 - x_3^2 = \sqrt{p^2 - 4q} \equiv w$  ②  
is a relation in  $R(w)$   
From ①, also  $x_1^2 = x_2^2, x_3^2 = x_4^2$

**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**



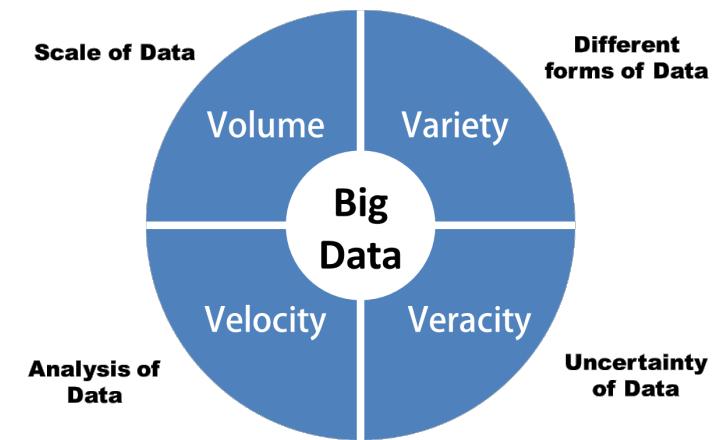
# Big Data

## 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.

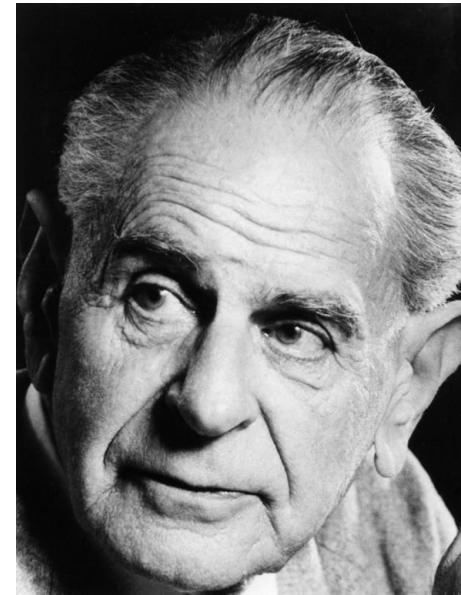
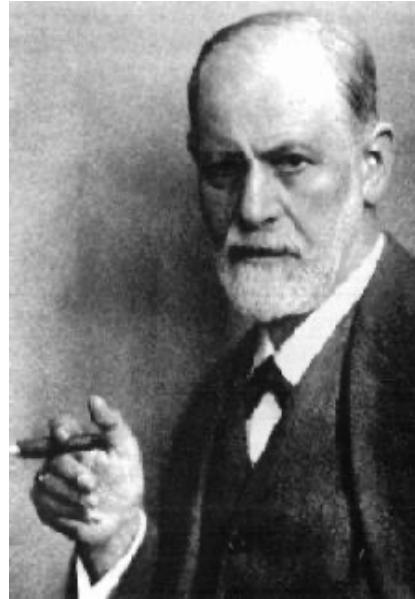
## New analysis and approaches:

- 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



[Data Science vs. Statistician video](#)

# 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. Make sure you can log into the YCRC Jupyter Server