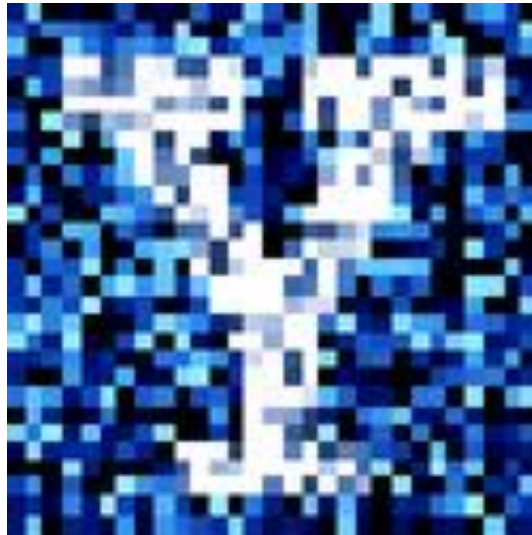


Ydata: Introduction to Data Science



Lecture 01: Introduction

Overview

Course overview

- Introductions
- Syllabus and logistics



What is Data Science?

- Start on the history of Data Science
 - If we run out of time we will finish this next class

Office hours and contact information

Ethan Meyers (he/him)

Email: ethan.meyers@yale.edu

Office hours: 3pm Wednesday

Office: 24 Hillhouse room 206

- <https://yale.zoom.us/j/99455443012>



Teaching Assistants

Course manager

- Hanwen Gu: hanwen.gu@yale.edu

Teaching Fellows

- Shuyu Rao: shuyu.rao@yale.edu
- Daria Bobrova: daria.bobrova@yale.edu

Undergraduate Learning Assistants

- Olivia Probst: olivia.probst@yale.edu
- Clare Chemery: clare.chemery@yale.edu
- Raphael Berz: raphael.berz@yale.edu
- Viraj Shukla: viraj.shukla@yale.edu
- Alden Tan: alden.tan@yale.edu
- Henry Weaver: h.weaver@yale.edu



TA office hours are on the calendar on Canvas

- Zoom link for OH on Canvas

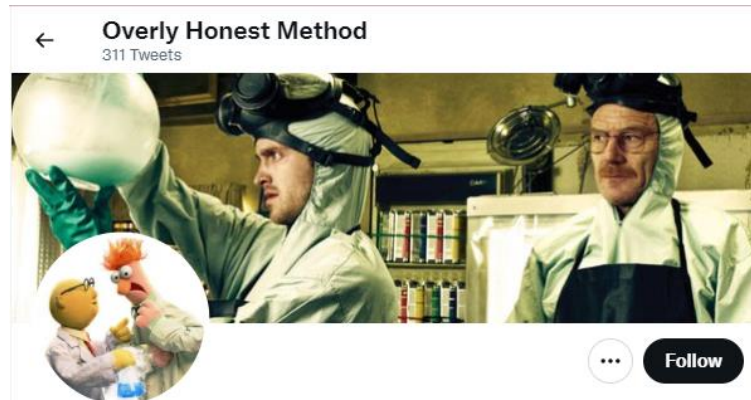
About this class

YData is based on Berkeley's Data 8 course

- It has been adapted at Yale by Jessi Cisewski-Kehe , Elena Khusainova and John Lafferty.

I will be making my own adaptations

- Please bear with me if we hit any rough patches



About this class

Broad survey of topics in Data Science

- smorgasbord



- Cause and Effect
- Data Types
- Building Tables
- Census
- Charts
- Distributions
- Histograms
- Functions
- Groups
- Pivots and Joins
- Iteration
- Chance
- Sampling
- Models
- Comparing Distributions
- Decisions and Uncertainty
- A/B Testing
- Causality
- Confidence Intervals
- Center and Spread
- The Normal Distribution
- Sample means
- Design Experiments
- Correlation
- Linear Regression
- Least Squares
- Residuals
- Regression Inference
- Privacy
- Classification
- Classifiers

Learning goals

1. Understand concepts from Computer Science and Statistics that are useful for analyzing data.

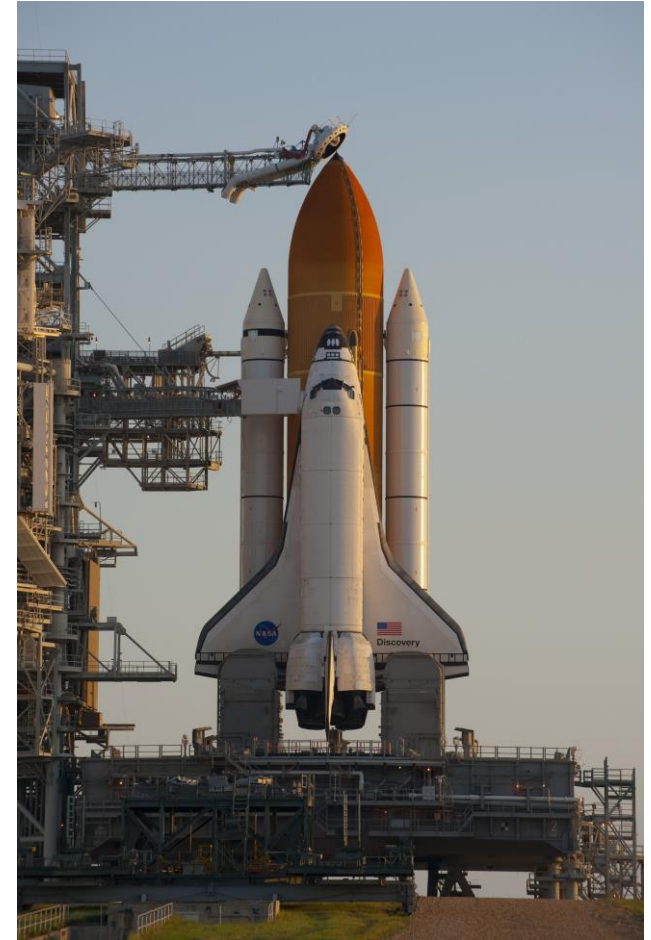
- Understand concepts in statistical inference
 - e.g., hypothesis tests, confidence intervals, etc.
- Learn basics of computation and simulation used for analyzing data.

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 are no prerequisites for this class

- E.g., no prior knowledge of Statistics or Programming is required



Connector seminars

YData seminars are small, independent courses taught by Yale

- Taught by faculty who are excited to share their expertise

YData connectors offered this year:

- YData: Measuring Culture (S&DS 175)
- YData: Humanities and Data Mining (S&DS 176)

Course Materials

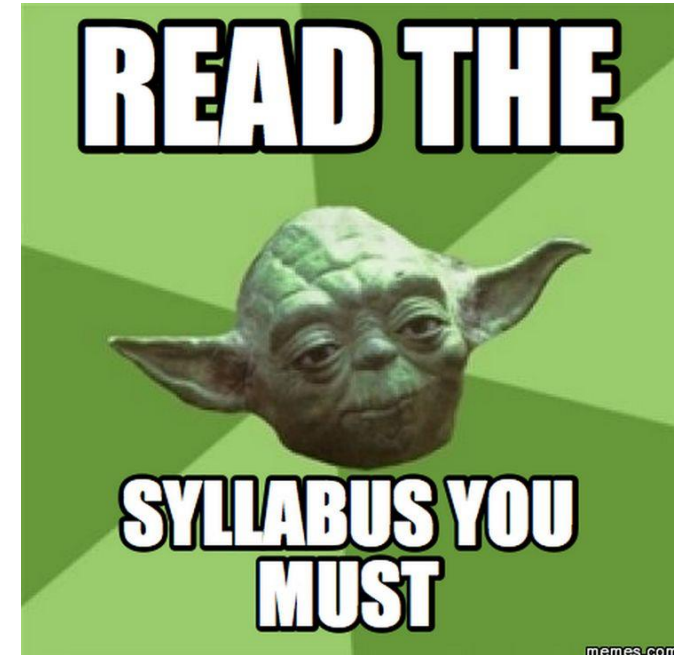
There are two websites for this class that contain relevant information

Canvas

- Syllabus, grades, supplemental reading, etc.

Class calendar

- <https://ydata123.org/sp22/calendar.html>
 - Contains homework, textbook readings, etc.



Class textbook and reading

The class textbook is: [Computational and Inferential Thinking](#)

- Adhikari, A. and DeNero, J. (2018)

This book is free online

Additional reading and other resources will be posted to Canvas



Computational and
Inferential Thinking

Course structure

Three lectures per week

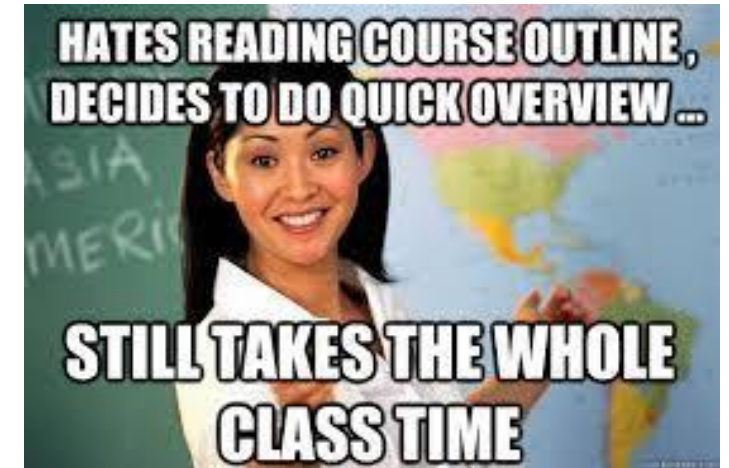
Weekly homework assignments

Three "projects" spread through the semester

- 11 days to complete these

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

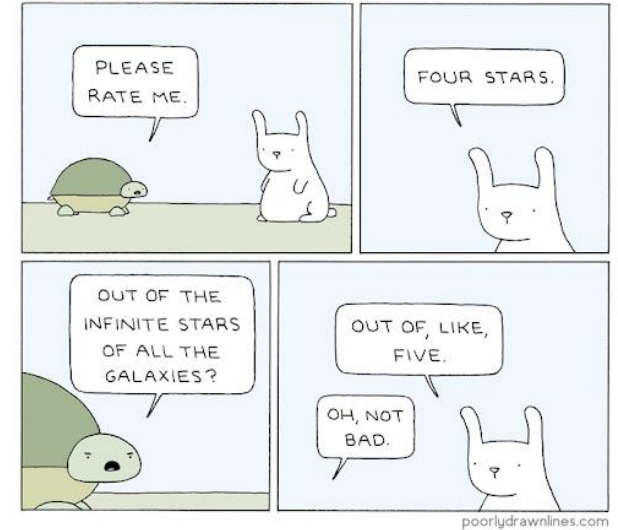
Midterm and final exam



Assignments and grades

1. Homework problem sets (30%)

- Exploring concepts and analyzing data using Python
- Weekly: 11 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 (80%) 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. Projects (24%)

- There are three "projects" that consist of more in depth data analyses
- They will be available on Mondays and the following week on Friday
 - i.e., 11 days later
- There will also be homework on the same week as projects so plan accordingly

3. Exams (44% total)

- Midterm: March 18th during the regular class time (17%)
- Final during finals period (27%)

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!

Things to do for next class...

1. Complete class survey

2. 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 1/30.

Questions?



What is Data Science?

Introduction and Discussion



Introduce yourselves:

- 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

Then briefly discuss:

- What is Data Science?
- How does it differ from Statistics?

What is Data Science?

Thoughts?

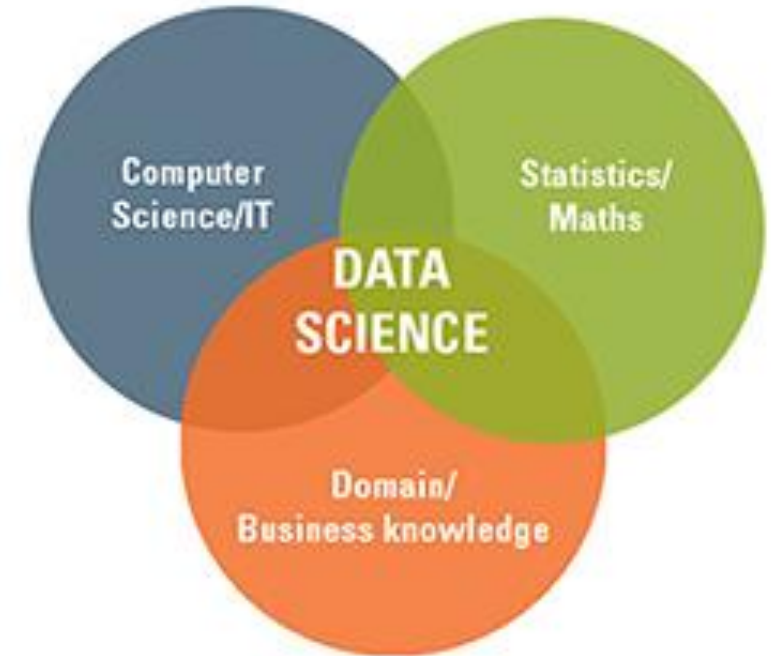


Josh Wills
@josh_wills

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

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



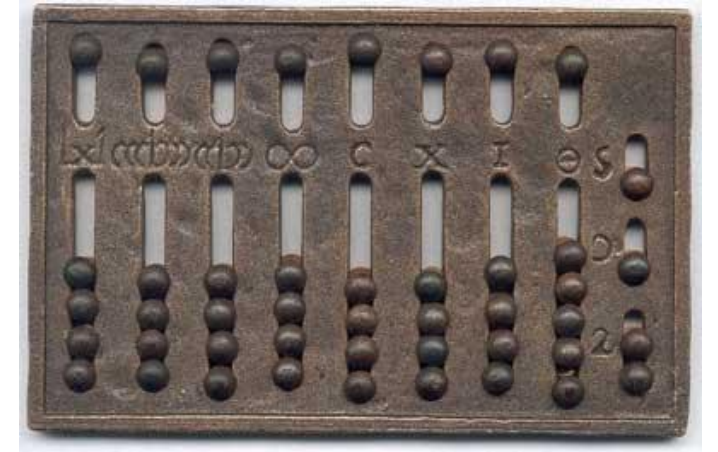
Cuneiform tablets from Uruk, a Mesopotamian settlement 5,000 years old contained transaction data on commodities



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)



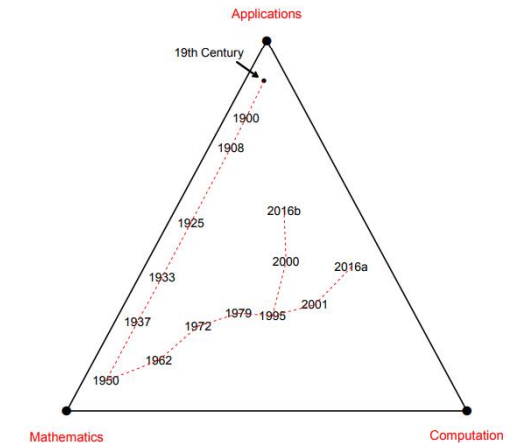
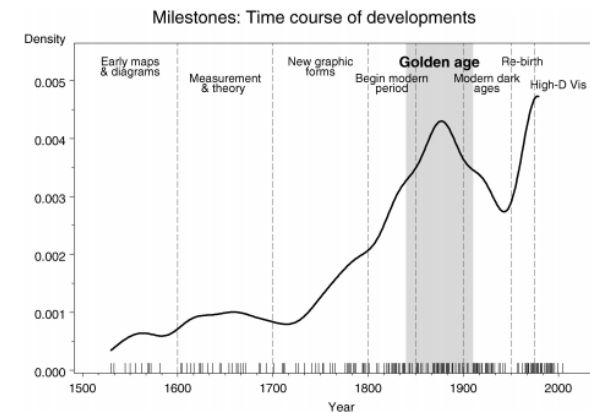
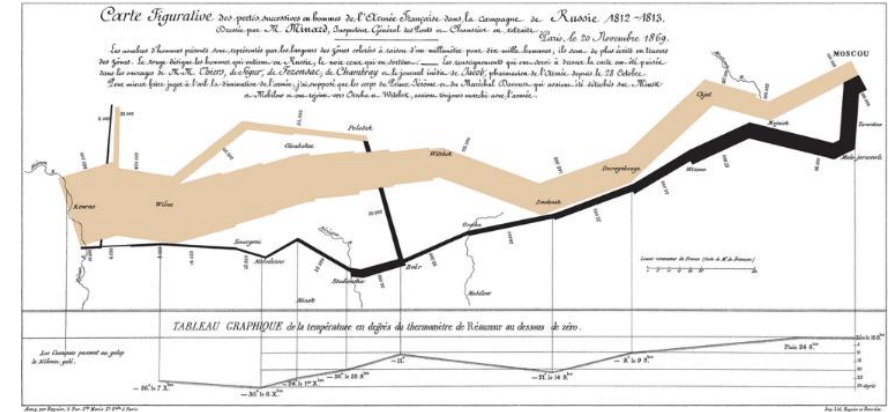
Brief history of Data Science: visualization and math

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 the analysis of data in the first half of the 20th century

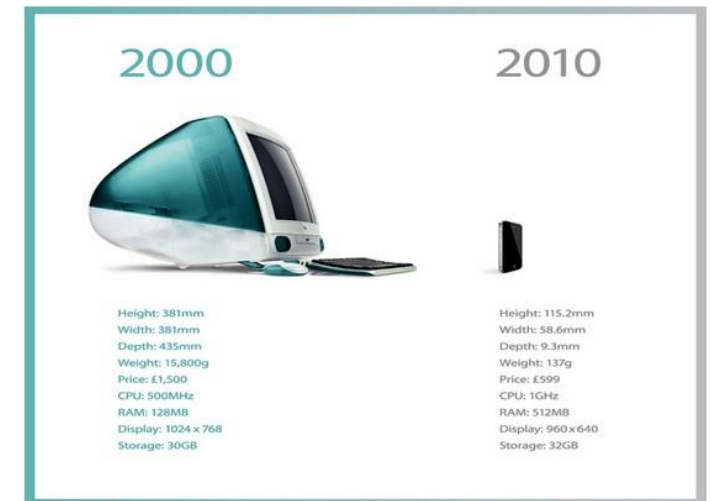


Brief history of Data Science: the rise of computers

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.

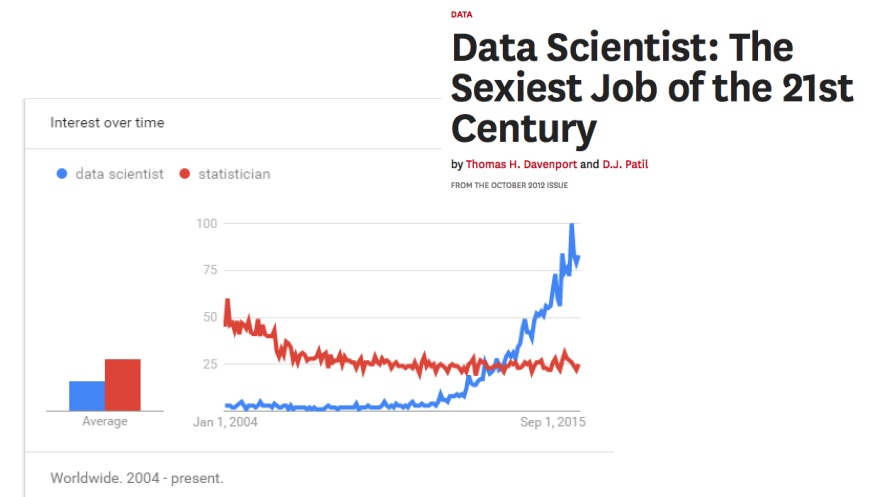
- John Tukey (1962) looks for a broadening of data analysis beyond mathematics
- Breiman (2001) describes a mathematical modeling culture and algorithmic culture
- The term "Data Science" starts being used in the 2000's to describe computational approaches to analyzing data

THE FUTURE OF DATA ANALYSIS¹
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



New ways to choose the best methods

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

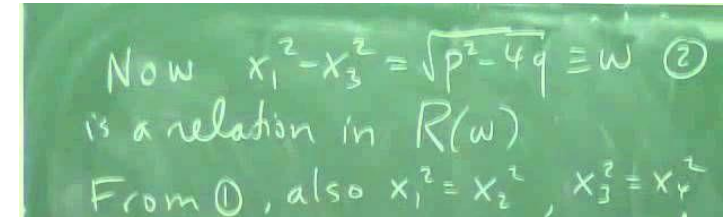
- Best methods are the ones that have mathematical guarantees (proofs)

Data Science empirically evaluates data analysis methods

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

[Data Science vs. Statistician video](#)

The proof is in the math



The proof is in the pudding



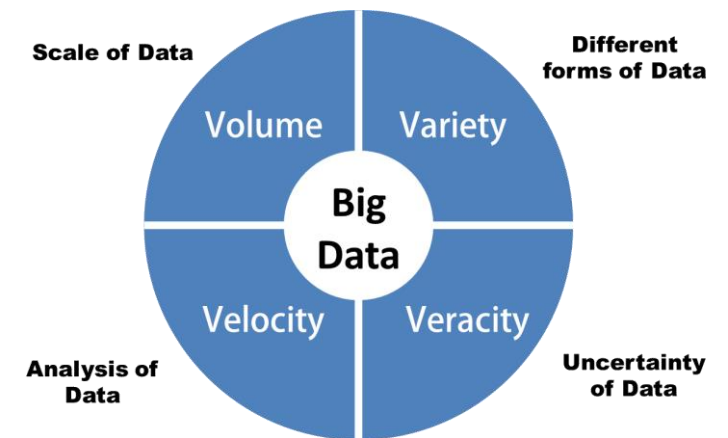
Big Data

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



What is Data Science?

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

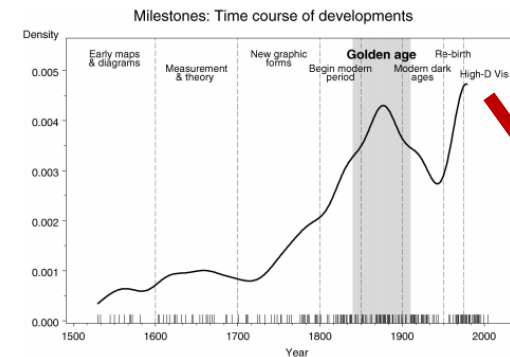
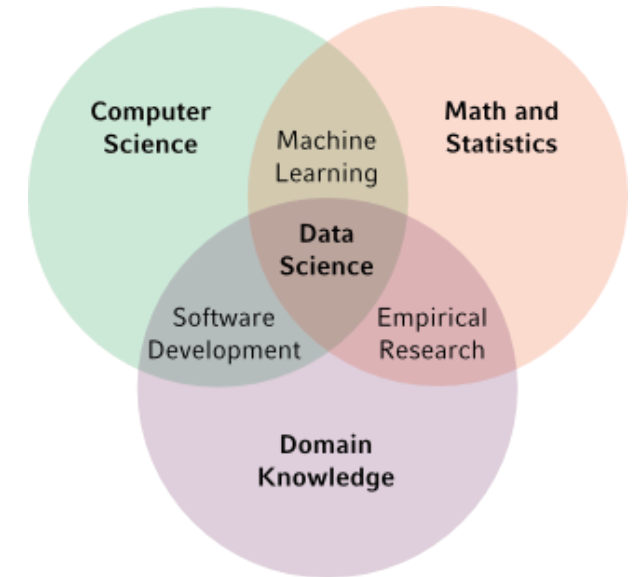
Many other fields impacted by 'Data Science

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

Examples:

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

Ethical concerns around privacy, fairness and other issues



Things to do for next class...

1. Complete class survey

2. 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 1/30.