Data science for everyone*

Prof. Jones-Rooy & Prof. Policastro

Jan. 27, 2020

1.1: Introduction

* All are welcome, no NYU pre-requisites, designed to be accessible to all

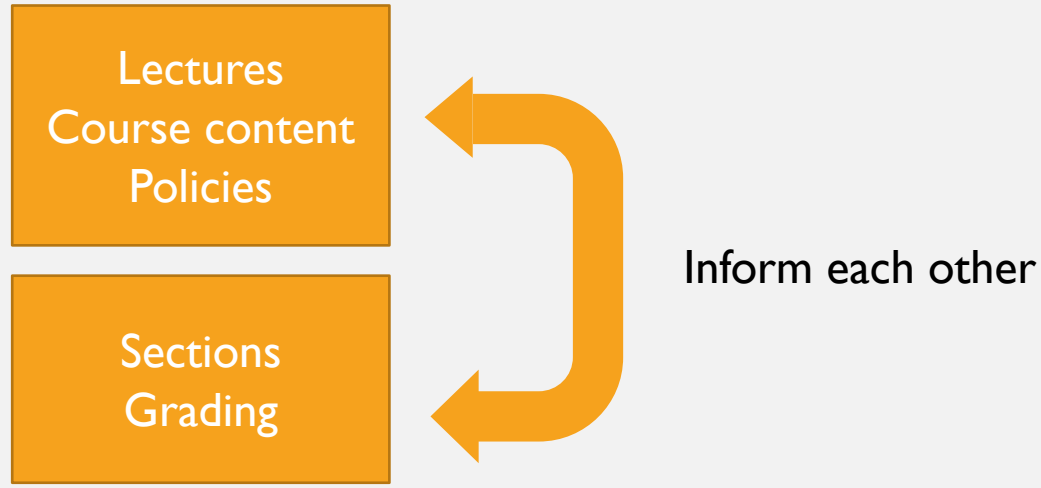
BUT – success in course depends on your investment in it!

Outline

1. Your instructors
2. Course overview
3. What is data science?

Your instructors

- Two professors
  - Prof. Jones-Rooy
  - Prof. Policastro
- Three teaching assistants
  - William Godel (head TA)
  - Angela Lai
  - Jeff Jacobs



Lectures: Introduce course content & concepts
Sections: Review, discuss, practice, ask questions

PROF. JONES-ROOY

Director of Undergraduate Studies

NYU Center for Data Science

andrea.jonesrooy@nyu.edu

DS major/minor inquiries:

cds-undergraduate@nyu.edu

Ph.D.: Political Science, Complex Systems

PROF. POLICASTRO

Clinical Assistant Professor

NYU Center for Data Science

policast@cims.nyu.edu

Also teaches:

DS-UA 112: Intro. to Data Science

DS-GA 1003: Machine Learning

Ph.D.: Mathematics

Berkeley
UNIVERSITY OF CALIFORNIA

TEACHING ASSISTANTS

William Godel

Head TA

Ph.D. student, Politics, NYU

BA: Philosophy, NYU

wpg205@nyu.edu

Angela Lai

Ph.D. student, Data Science, NYU

BS: Data Science, Political Science, Univ. of Rochester

ayl316@nyu.edu

Jeff Jacobs

Ph.D. student, Pol. Sci., Columbia

MS: Computer Science, Stanford

BS: Economics, Mathematics, Computer Science, U. of MD

Jpj251@nyu.edu

Outline

1. Your instructors
2. Course overview
3. What is data science?

LECTURES AND SECTIONS

- Lectures
  - MW, 3:30-4:45p
  - GCASL C95
- Sections
  - Th or F, 50 mins
  - Times & rooms vary (see Albert or syllabus)

YOUR NEW BEST FRIEND

When in doubt, check the syllabus!

Data Science for Everyone
Syllabus

DS-UA 111, Spring 2020
MW, 3.30-4.45p, GCASL C95

Prof. Andrea Jones-Rooy
60 Fifth Ave, 640
ajr348@nyu.edu
Office Hours: W, 1-2p

Prof. Christopher Policastro
60 Fifth Avenue, 704
policast@cims.nyu.edu
Office Hours: Th, 3-4p (room 650)

Find it on Classes > DS-UA 111 > Syllabus

THE SYLLABUS CONTAINS ALL THIS IMPORTANT INFORMATION!

1. Course overview
2. Skills you'll learn
3. Expectations & advice
4. Grading
5. Academic accommodations
6. Readings
7. Course outline: Summary
8. Course outline: Full

Before you email with a question about the course, especially course policy, *check the syllabus*!

We summarize here, but read the whole thing by Lecture 1.2!

1. COURSE OVERVIEW

- Goal: To empower you to understand & use data to explain and (thoughtfully, cautiously, scientifically!) predict the world
- By the end of this course you will be able to:
  - Access & interpret publicly available datasets out in the world
  - Conduct original statistical analyses to test hypotheses and draw meaningful, transparent, and scientific inferences from them
  - Assess the quality of a dataset and its impact on your inferences
  - Evaluate data-driven arguments or conclusions in the news and other outlets

2. SKILLS YOU'LL LEARN

How to:

- Program in Python
- Conduct statistical tests
- Evaluate strengths & weaknesses of a dataset
- Demystify common data science buzzwords & jargon (hint: it's probably not actually AI)
  - And while you're thinking about that, think about this.

3. EXPECTATIONS & ADVICE

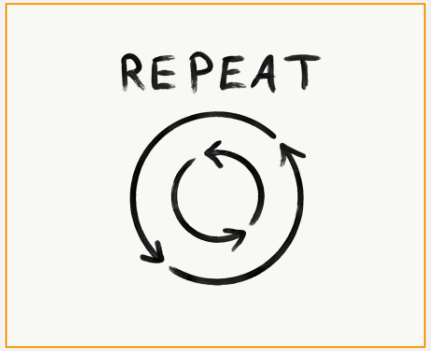
- Expectations:
  - Attend all lectures & sections
  - Prepare readings ahead of lecture
  - Complete all assignments & exams
  - Turn in assignments on time

These are all necessary but not sufficient to earn an A

My observation: students who attend class & do readings earn higher grades

We will unpack this causality later!

- Advice:
  - Learning programming & stats is like learning a new language: regular practice is key
  - As with learning languages, you will experience frustration at times!
  - You may encounter code you don't understand, or write something you think should work but doesn't
  - This is all part of the process!
  - Staying on top of labs & homeworks in particular will help!
  - Repetition is your friend



4. GRADING

9 Labs (18%): You will generally have one week for each 2% lab
4 Homeworks (20%): You will have two weeks for each 5% homework
Project (16%): You will have eight weeks to complete one project
Midterm exam (20%): The midterm will be in class on March 11
Final exam (26%): The final will be in class on May 11

More on this soon!

Late policy

- Labs: Zero late policy. If you submit it even one minute late, you earn a 0 for that lab
- HWs & Project: 20% off your grade for every day late, beginning at one minute late
- Midterm & Final: Zero late policy (turn them in when time is up or earn a 0)
  - Exam times are firm – no makeups, no alternative dates
- **Illness**: If you are ill and cannot complete an assignment, you must provide **documentation** from a doctor**. No documentation, no extension.**

Improperly formatted assignments will not be accepted

ACADEMIC HONESTY

1. No assignment on which you receive a grade is collaborative.
2. You may consult with others but all work handed in must be your own.
3. Do not copy another individual's work, answers or ideas.
4. Do not allow another individual to copy your work, answers or ideas.

Plagiarism will not be tolerated. We will punish all academic dishonesty to the fullest extent possible according to University policy.

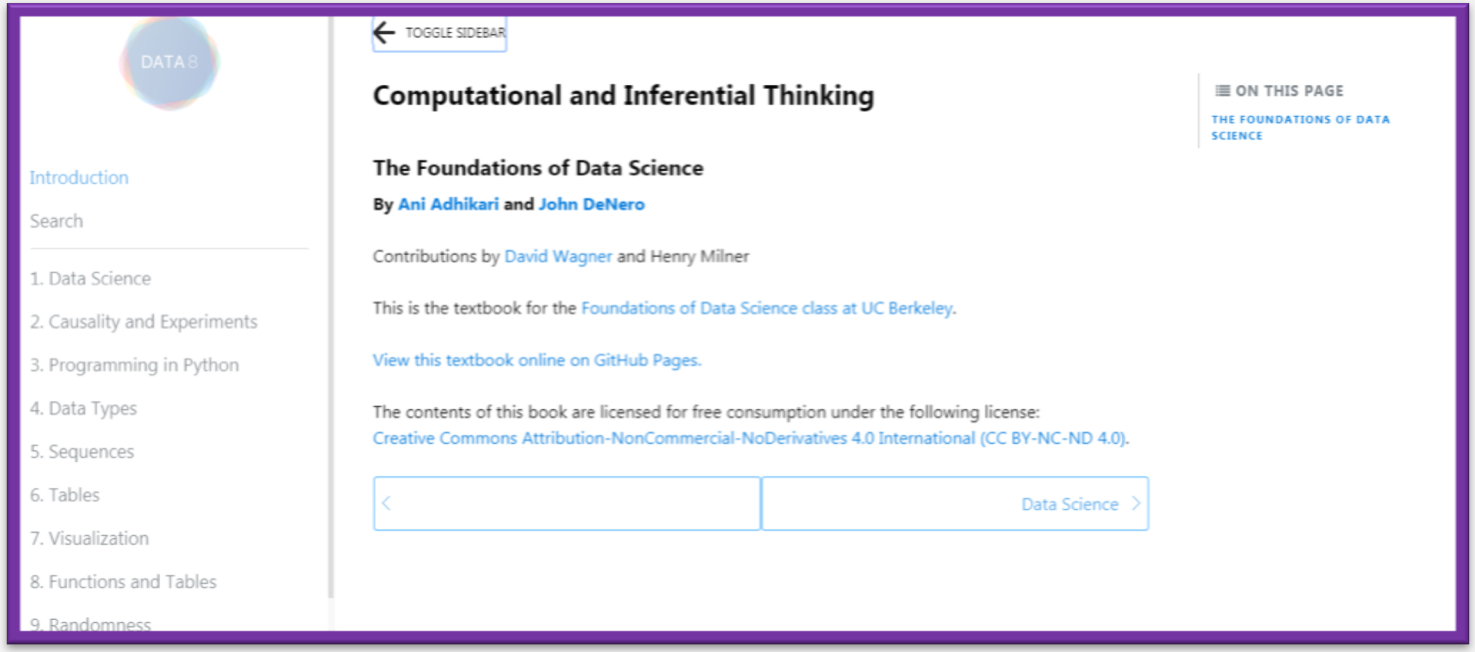
5. ACADEMIC ACCOMMODATIONS

- Accommodations are available for students with disabilities
- Please contact the Moses Center for Students with Disabilities for further information
  - 212-998-4980
  - mosescsd@nyu.edu
- Students requesting accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance

6. READINGS

- Prepare all readings in the syllabus listed in parentheses before lecture meets for that day
- No readings for sections, though you usually will want to have started working on your assignments in order to get the most out of section
- Course text: *Inferential Thinking*
- We will occasionally suggest additional readings and resources. Unless they are in the syllabus, they are optional and for your exploration if you're interested
- We will also reference external data and Jupyter notebooks – and will make it clear when the time comes how to access & use them

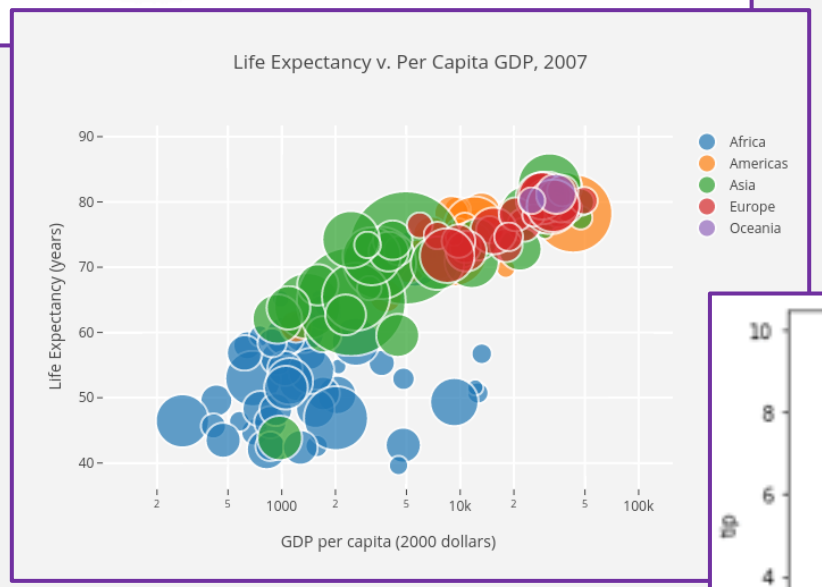
YOUR (FREE) TEXTBOOK



https://www.inferentialthinking.com/

YOUR (FREE) SOFTWARE

python™



jupyter horses_median_mean_CIs Last Checkpoint: 08/12/2019 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

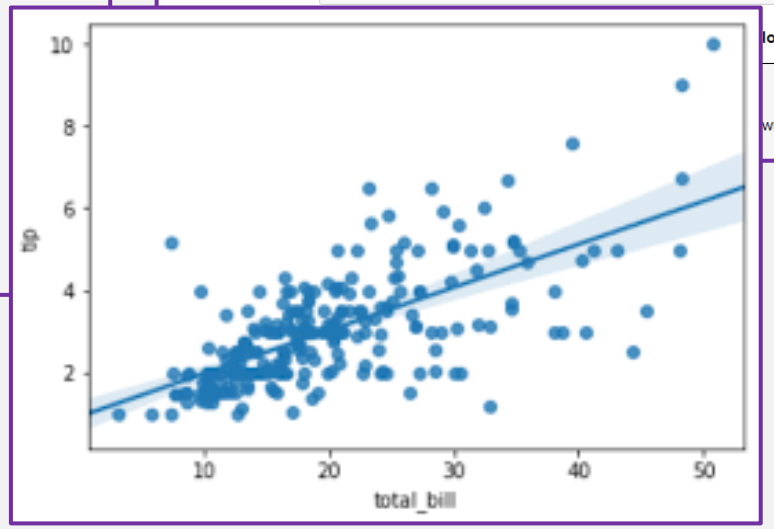
Run Code

```
In [1]: #use CIs for median and mean

from datascience import *
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
import numpy as np
```

```
In [2]: #let's grab the horse data sample
horses = Table.read_table('horse_data.csv')
horses
```

lor	location	markings	weight	foaldate	registrations	disciplines	temperament	
wn	Arthur, Ontario, Canada	nan	1100 pounds	Mar-04	AQHA - American Quarter Horse Association	Hunter Under Saddle (Champion) Equitation (Champion) Hor	2	http://www.equine. for-sa 2



Find it on Classes >
DS-UA 111 >
Jupyter
(but don't worry about it *just* yet)

7. COURSE OUTLINE: SUMMARY

1. **Week 1**: Introduction to data science & thinking like a scientist
2. **Week 2**: Causality & technology for the course
3. **Week 3**: Programming preliminaries
4. **Week 4**: Working with data
5. **Week 5**: Organizing data in Python
6. **Week 6**: Data visualizations & functions
7. **Week 7**: Midterm exam
8. **Week 8**: Randomness, sampling, and distributions
9. **Week 9**: Testing hypotheses
10. **Week 10**: Estimation
11. **Week 11**: Regression
12. **Week 12**: Classification
13. **Weeks 13**: Data & ethics
14. **Week 14**: Next steps in data science & review
15. **Week 15**: Final exam

7. COURSE OUTLINE: ASSIGNMENTS & EXAMS

Assignment	Assigned	Due	Notes
1. Lab 0	Feb. 5	Feb. 12	
2. Lab 1	Feb. 12	Feb. 19	
3. Lab 2	Feb. 19	Feb. 26	
4. Lab 3	Feb. 26	Mar. 4	
5. Lab 4	Mar. 4	Mar. 9	Mon. due to midterm
6. Lab 5	Mar. 25	Apr. 1	
7. Lab 6	Apr. 1	Apr. 8	
8. Lab 7	Apr. 8	Apr. 15	
9. Lab 8	Apr. 15	Apr. 22	
10. HW 1	Feb. 3	Feb. 18	Tues. due to holiday
11. HW 2	Feb. 24	Mar. 9	
12. HW 3	Mar. 23	Apr. 6	
13. HW 4	Apr. 13	Apr. 27	
14. Project	Mar. 2	May 4	Mini-targets throughout
15. Midterm	Mar. 11	Mar. 11	In lecture
16. Final	May 11	May 11	In lecture

COURSE OUTLINE: OVERALL TIMING

- Labs are handed out on Wednesdays after lecture & due one week later
- HWs are handed out on Mondays after lecture and due two weeks later
- The project will be handed out just before the midterm
- The lectures prior to the midterm and final will be review lectures
- Sections meet every week except the week following the midterm
- Lectures labels refer to the week (1, 2, 3, 4, etc.) and whether they are on Monday or Wednesday (1.1, 1.2, 2.1, 2.2)

8. COURSE OUTLINE: FULL

8 Course outline: Full

Each week has three parts: two lectures and one section. Any part below ending in 0.1 refers to a Monday lecture, 0.2 refers to a Wednesday lecture. There are Thursday/Friday sections every week unless otherwise noted.

Week 1: Introduction to data science & thinking like a scientist

Welcome to the first day of the rest of your life! We're going to get acquainted with what we mean by data science, why it's important, and why you're here (in this classroom; broader philosophical questions are unlikely to be resolved, at least not this week).

We particularly focus on the "science" in data science. While "data science" tends to make people think of data, this focus obfuscates the fact that data on its own is just a tool, and it's science where real discovery takes place. It takes principled thinking, including theory development, hypothesis formation, and rigorous testing to actually learn anything from data.

- **1.1** (Jan. 27): Introductions & What is data science?
- **1.2** (Jan. 29): Thinking like a scientist (Ch. 1 Data Science; the course syllabus)

Week 2: Causality

Learning to be a scientist is a lifelong pursuit, but causal inference, the focus of this week, is a wonderful and important place to start. This week will shape your thinking for the rest of the semester – and your life. We will also

Outline

1. Your instructors
2. Course overview
3. What is data science?

DATA SCIENCE IS ABOUT…

Understanding the world using scientific methods + data

Exploration

Identify patterns in information

Distributions, moments, networks, visualizations

Inference

Quantify how reliable a pattern is

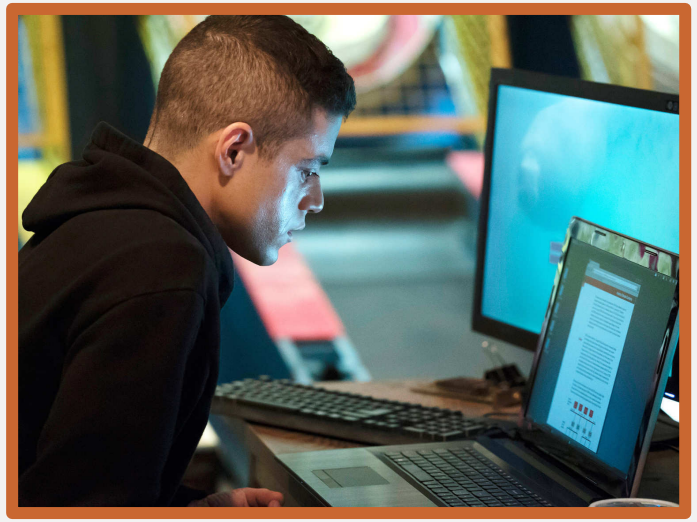
Randomization, experiments

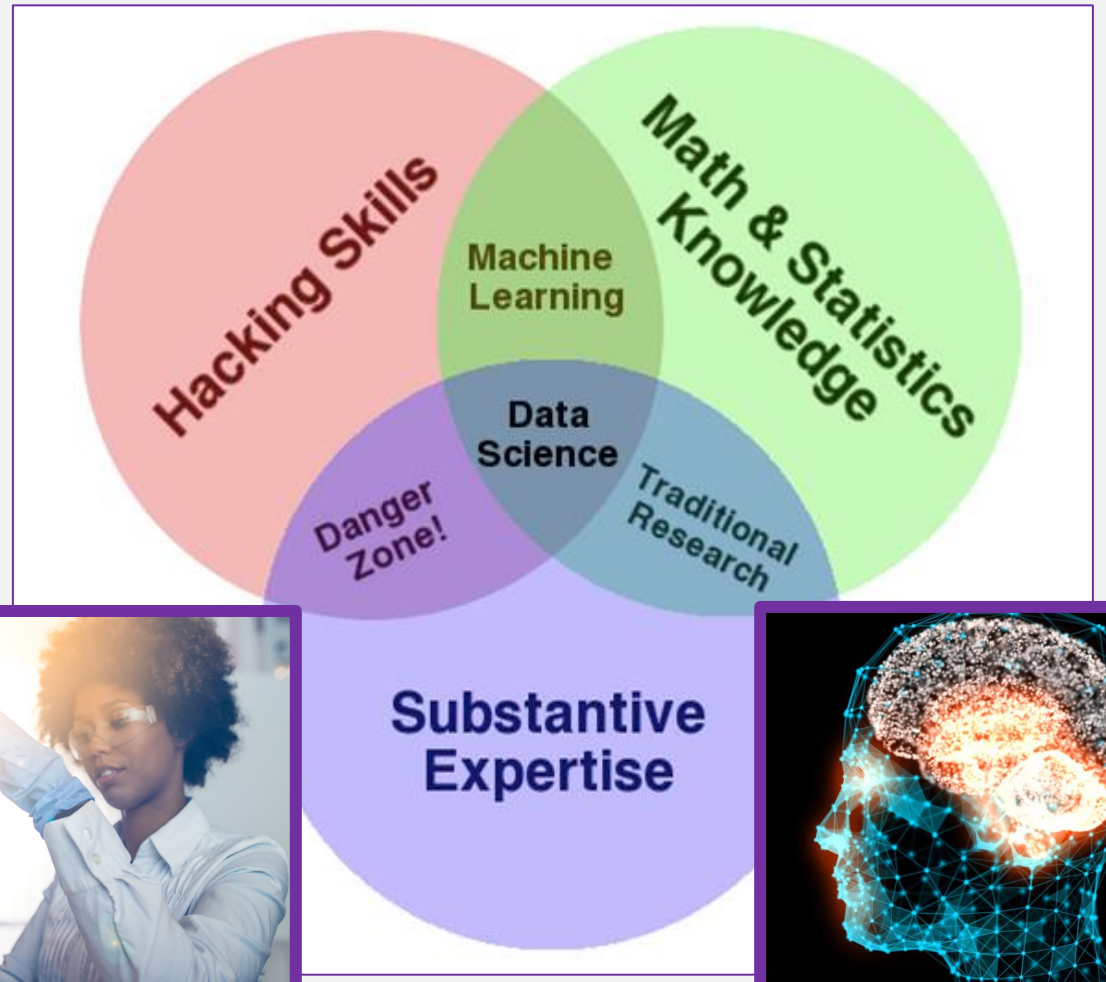
Prediction

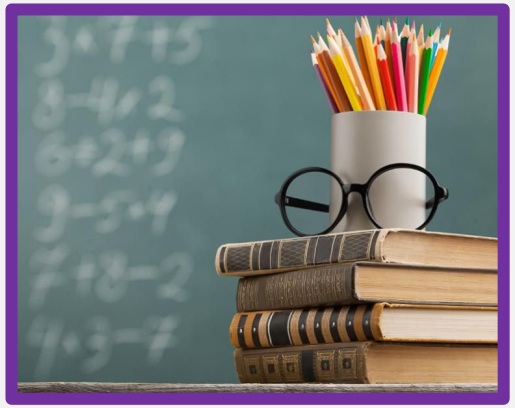
Make informed guesses about the future

Statistical analysis, machine learning

WHERE DOES IT FIT?







WHY HAVE A SPECIAL COURSE IN DATA SCIENCE?

Understanding the world using scientific methods + data

Big data

Examples:

Health

Social media

Climate

Traffic

Computation

Examples:

Screen scraping, web crawlers

Classifiers

Cleaning data

Analyzing

Randomization

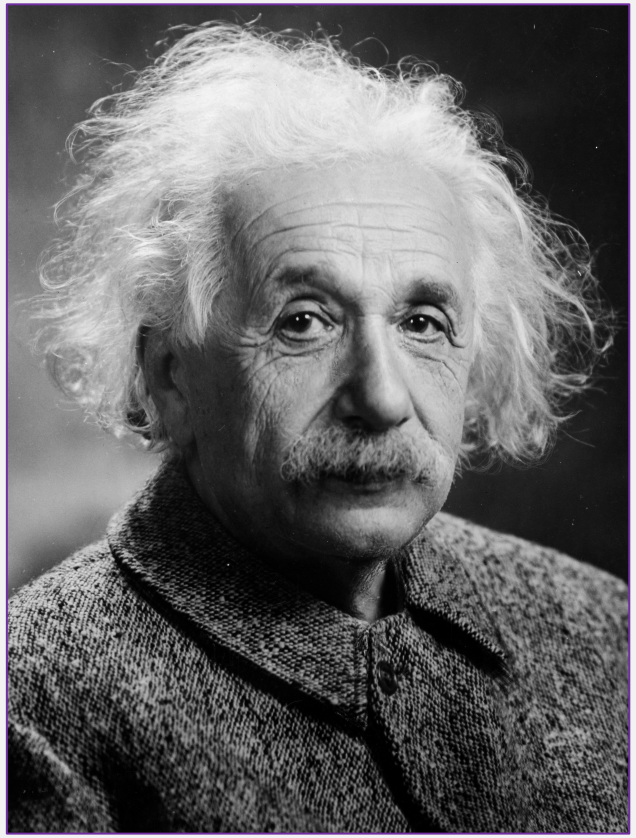
Examples:

Experiments (randomly assign)

A/B Testing

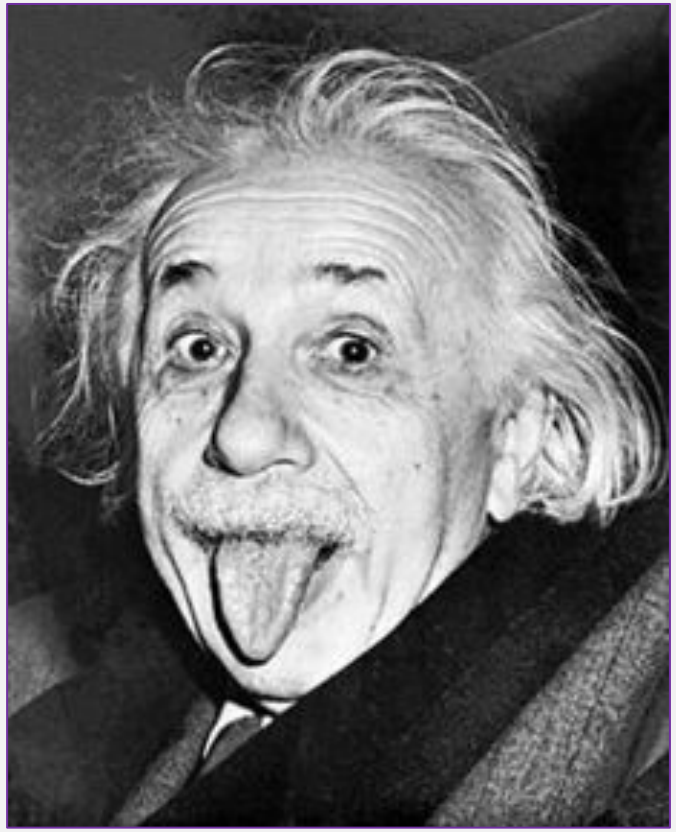
DATA SCIENCE IS A SCIENCE

- Evaluate the *merit* of a *hypothesis* with rigorous empirical evidence and testing.
- We need *statistics*, to help us be *precise* about the route from observations to conclusions.
- Recent *computational* advances allow us to look at huge datasets, with new techniques.



DATA SCIENCE IS AN ART

- We have to *translate* problems into DS goals
- Develop *intuitions* about good and bad data, good and bad models
- Think *creatively* and simply!
- See *similarities* between problems and processes
- Manage DS *process* from end-to-end



WELL, AND…

Harvard Business Review

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

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Summary Save Share 16 Comment Text Size Print $8.95 Buy Copies

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently

Outline

1. Your instructors
2. Course overview
3. What is data science?

Thank you!