I would found an institution where any person can study data science. - Ezra Cornell

A course for anyone who wants to study data visualization, prediction, machine learning, and programming in Python. We’ll analyze real-world data sets on crime, health, transportation, literature, and more!

STSCI + ORIE + CS 1380
Data Science For All
Spring 2019  TR 10:10-11:25am

No experience required – Open to all – Fulfills MQR-AS
Who are we?

- Professor Booth
- Professor Wilson
+ Teaching Assistants
Who are we?

- Polina Kirichenko (pk575)  
  *Operations Research*

- Skyler Seto (ss3349)  
  *Statistics*

- Sean Sinclair (srs429)  
  *Operations Research*

- Antonio Sirianni (ads334)  
  *Sociology*
Who are we?

- Victoria Bao (yb244)  
  Statistics

- Daniel Sanky (ds869)  
  Information Science
Who are you?

Take this class if you:
- are curious about data
- don’t know much/any CS
- don’t know much/any Stats
- don’t know much/any OR

Don’t take this class if you:
- have already taken both CS & Stats intro classes
  (it will be too slow for you)
Why Data Science?
Moore’s Law – The number of transistors on integrated circuit chips (1971-2016)
Moore’s law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore’s law.

The data visualization is available at OurWorldInData.org. There you find more visualizations and research on this topic. Licensed under CC-BY-SA by the author Max Roser.
Global Information Storage Capacity
in optimally compressed bytes

2007
ANALOG
19 exabytes
- Paper, film, audiotape and vinyl: 5%
- Analog videotapes (VHS, etc): 94%
- Portable media, flash drives: 2%
- Portable hard disks: 2.4%
- CDs and mini disks: 8.8%

DIGITAL
280 exabytes
- Computer servers and mainframes: 8.9%
- Digital tape: 11.8%
- DVD/Blu-ray: 22.8%
- PC hard disks: 44.5%
- Others: < 1% (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

1986
ANALOG
2.6 exabytes
1993
DIGITAL
0.02 exabytes
2000
ANALOG STORAGE
DIGITAL STORAGE
2002:
"beginning of the digital age"
50%

% digital:
1%
3%
25%
94%


Growth of and digitization of global information-storage capacity [1]

[1]
Digital Data Terminology

- **Bit** - binary unit: 0/1
- **Byte** - eight bits
- **Kilobyte** - $2^{10}$ or 1024 bytes
- **Megabyte** - $2^{20}$ bytes or 1024 kilobytes
- **Gigabyte** - $2^{30}$ bytes or 1024 megabytes
- **Terabyte** - $2^{40}$ bytes or 1024 gigabytes
- **Petabyte** - $2^{50}$ bytes or 1024 terabytes
- **Exabyte** - $2^{60}$ bytes or 1024 petabytes
Who needs data science?

- Data scientists
- OR, CS, Stats majors
- Lawyers
- Doctors
- Citizens
- Readers of the news

...ALL
National Challenge

In the United States, it is reported that in 2018 there will be more than 490,000 data science positions available, but only 200,000 qualified people to fill the roles. The average size of a graduate class of data science students is 23 students. With approximately only 110 universities offering data science studies, the growing market will continue to pressure the supply in the US.
The Supreme Court is allergic to math

quantify partisan gerrymandering: “It may be simply my educational background, but I can only describe it as sociological gobbledygook.” This was

The Supreme Court does not compute. Or at least would rather not. The justices, the most powerful jurists in the land, seem to have a reluctance — even an allergy — to taking math and statistics seriously.

For decades, the court has struggled with quantitative evidence of all kinds in a wide variety of cases. Sometimes justices ignore this evidence. Sometimes they misinterpret it. And sometimes they cast it aside in order to hold on to more traditional legal arguments. (And, yes, sometimes they also listen to the numbers.) Yet the world itself is becoming more computationally driven, and some of those computations will need to be adjudicated before long. Some major artificial intelligence case will likely come across the court’s desk in the next decade, for example. By voicing an unwillingness to engage with data-driven empiricism, justices — and thus the court — are at risk of making decisions without fully grappling with the evidence.
Standing is good for you, but wait, N=50!!! Why would Psychological Science or The Economist publish a study with such sample size?

Standing is good for your mind as well as your body
It seems to promote cognitive performance

ECONOMIST.COM
Higher coffee consumption associated with lower risk of early death

Date: August 27, 2017

Source: European Society of Cardiology

Summary: Higher coffee consumption is associated with a lower risk of early death, according to new research. The observational study in nearly 20,000 participants suggests that coffee can be part of a healthy diet in healthy people.
What is Data Science?

Answering questions from data using computation

- **Exploration**
  - Identifying patterns in information
  - Uses visualizations

- **Inference**
  - Quantifying whether those patterns are reliable
  - Uses randomization

- **Prediction**
  - Making informed guesses
  - Uses machine learning
Data Science Stories

- **Agriculture**
  - When will the harvest be ready?
  - How large will the harvest be?

- **Political Campaigns**
  - How to summarize information from different polls?
  - What is the chance of winning each state or district?
  - Who might be willing to donate if I asked? How to ask?

- **Medicine**
  - Which patients are at risk of some disease?
  - Which patients would benefit from surgery?
Polls underestimated Trump in red states, Clinton in blue states

2016 election results vs. FiveThirtyEight’s adjusted polling average by state

Republican Vote Margin Relative to Polls

-20 -15 -10 -5 0 +5 +10 +15

Source: David Wasserman
“Genes mirror geography within Europe”

Data Science in Action
Course Structure
How DSFA works

● Lecture Tuesday and Thursday
  ○ Participation counts for grade
● Section every week on W or Th
  ○ Including this week!
  ○ Attend the one you signed up for
  ○ Project partner must be enrolled in same section
● Assignments:
  ○ Labs (about 10 total)
  ○ Homework (about 8 total)
  ○ Projects (3 total)
● Exams:
  ○ Two prelims + final exam
## (Tentative) Section Schedule

<table>
<thead>
<tr>
<th>Section</th>
<th>Time</th>
<th>Room</th>
<th>TA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIS201</td>
<td>W 12:20-2:15pm</td>
<td>Thurston Hall 202</td>
<td>Antonio/Polina</td>
</tr>
<tr>
<td>DIS202</td>
<td>W 2:30-4:25pm</td>
<td>Upson Hall 202</td>
<td>Daniel</td>
</tr>
<tr>
<td>DIS203</td>
<td>W 7:30-9:25pm</td>
<td>Hollister Hall 362</td>
<td>Sean/Skyler</td>
</tr>
<tr>
<td>DIS205</td>
<td>R 12:20-2:15</td>
<td>Thurston Hall 202</td>
<td>Victoria</td>
</tr>
</tbody>
</table>
(Tentative) Office Hours Schedule

<table>
<thead>
<tr>
<th>TA</th>
<th>Time</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonio</td>
<td>Wed 4:30-6:30pm</td>
<td>TBD</td>
</tr>
<tr>
<td>Daniel</td>
<td>Sat/Sun 2:30-3:30</td>
<td>TBD</td>
</tr>
<tr>
<td>Polina</td>
<td>Fri 2-4pm</td>
<td>TBD</td>
</tr>
<tr>
<td>Sean</td>
<td>Mon/Wed 1-2pm</td>
<td>TBD</td>
</tr>
<tr>
<td>Skyler</td>
<td>Tues/Thurs 9:30-10:30</td>
<td>TBD</td>
</tr>
<tr>
<td>Victoria</td>
<td>Mon 12:30-2:30</td>
<td>TBD</td>
</tr>
</tbody>
</table>
Policies, Grading, Etc.

- Details will be posted on the course website
- Blackboard
- Online textbook
Getting help

Questions about material:
- Ask a friend
- Ask on piazza
- Go to section
- Go to office hours

Logistical questions:
- Ask your section TAs
Academic Integrity

- Labs:
  - Work together as much as you’d like
- Homework and projects:
  - All work you submit must be your own
  - Share ideas (eg, in English) not solutions (eg, code)

In particular:
- Don’t post code on Piazza
- Cite your sources (including other students)
Now what?

● (Now) If you’re not enrolled yet sign up
● (Tomorrow or Thursday) Go to section
● (By Thursday) Read Chapter 1 (and 2) of the textbook
● (Constantly) Tell your friends about this class
  ○ Everyone should take this class
  ○ There’s still space
  ○ And it’s not too late
● (Next week) Buy an iClicker at the Cornell Bookstore
● (By the add deadline) Purchase access to Vocareum
Reef Polling

- Answer in-class quiz questions using a smartphone
- Create an account from the login page
- Free 14 day trial. Six month subscription for $15
- Use your Cornell email and NetID to sign in

- If you use an iClicker it must be registered on Blackboard (and you don’t need a Reef account)
Install Anaconda?
Acknowledgement

This course is based on Data 8, a course taught by Ani Adhikari and John DeNero at the University of California, Berkeley. They and their teaching assistants have developed many of the materials we are using in our own course. We are using those materials with their permission, which we gratefully acknowledge.