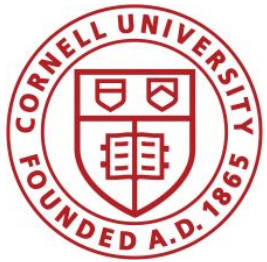


DSFA

Spring 2020

Lecture 1

CS/ORIE/STSCI 1380
Introduction

CS + ORIE + STSCI 1380

Data Science For All

Spring 2020 TR 10:10-11:25am

No experience required – Open to all – Fulfills MQR-AS

Who are we?

Professor Entner

Statistics

Professor Williamson

Operations Research

Who are we?

Artem Bolshakov (atb86)
Physics

Ben Baer (brb225)
Statistics

Who are we?

Julie Barron (jcb468)
ORIE

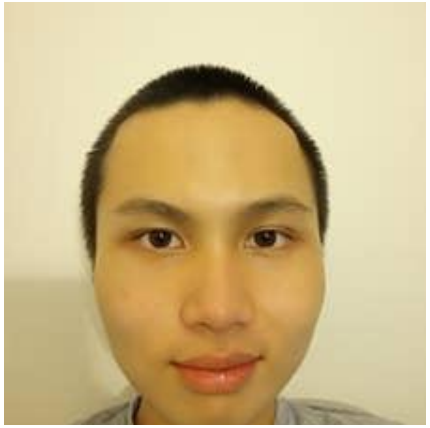
Taeho Kim (tk538)
Economics

Daniel Sanky (ds869)
Information Science

Who are we?

Kate Schrage (krs273)
English

Anders Wikum (aew236)
ORIE



Yao Yu Yeo (yy826)
Biological Sciences

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:

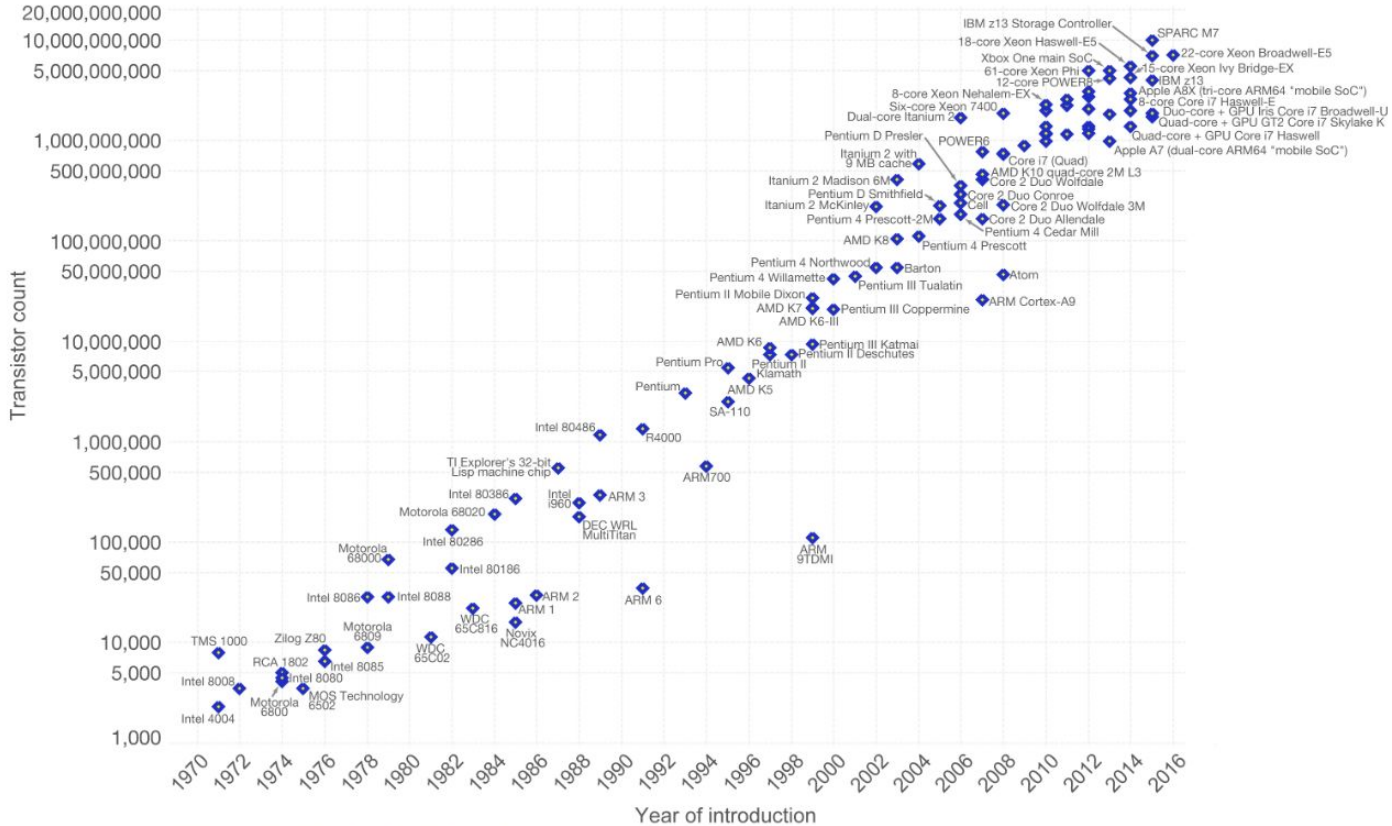
- have already taken both CS & Stats intro classes (it will be too slow for you)

Why Data Science?

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Our World in Data

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.

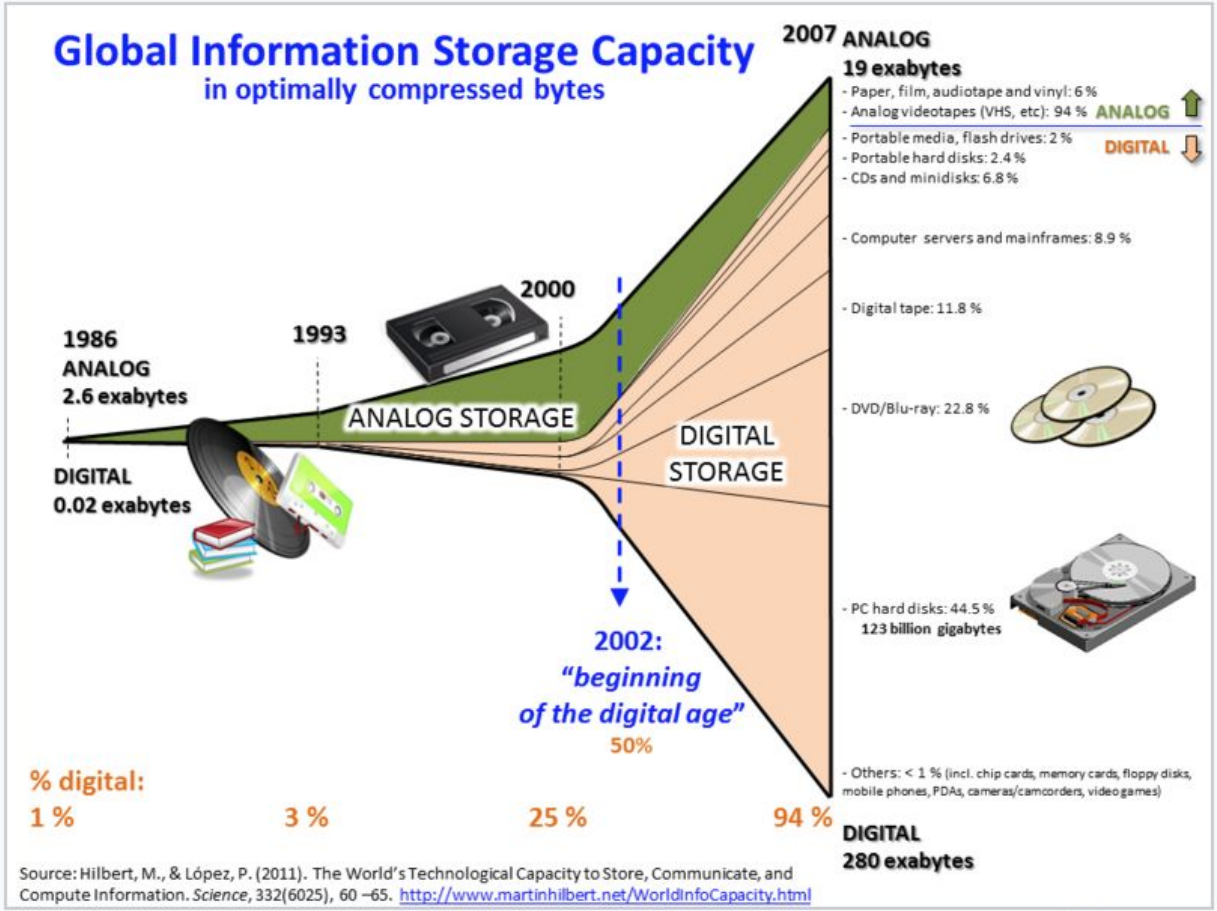


Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)

The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Kaypro II (1982)
64K RAM
2.5 Mhz processor
191K floppy drives

iPhone 7 (2016)
256GB RAM
(4,000,000 times as much)
2.34 Ghz processor
(1000 times as fast)

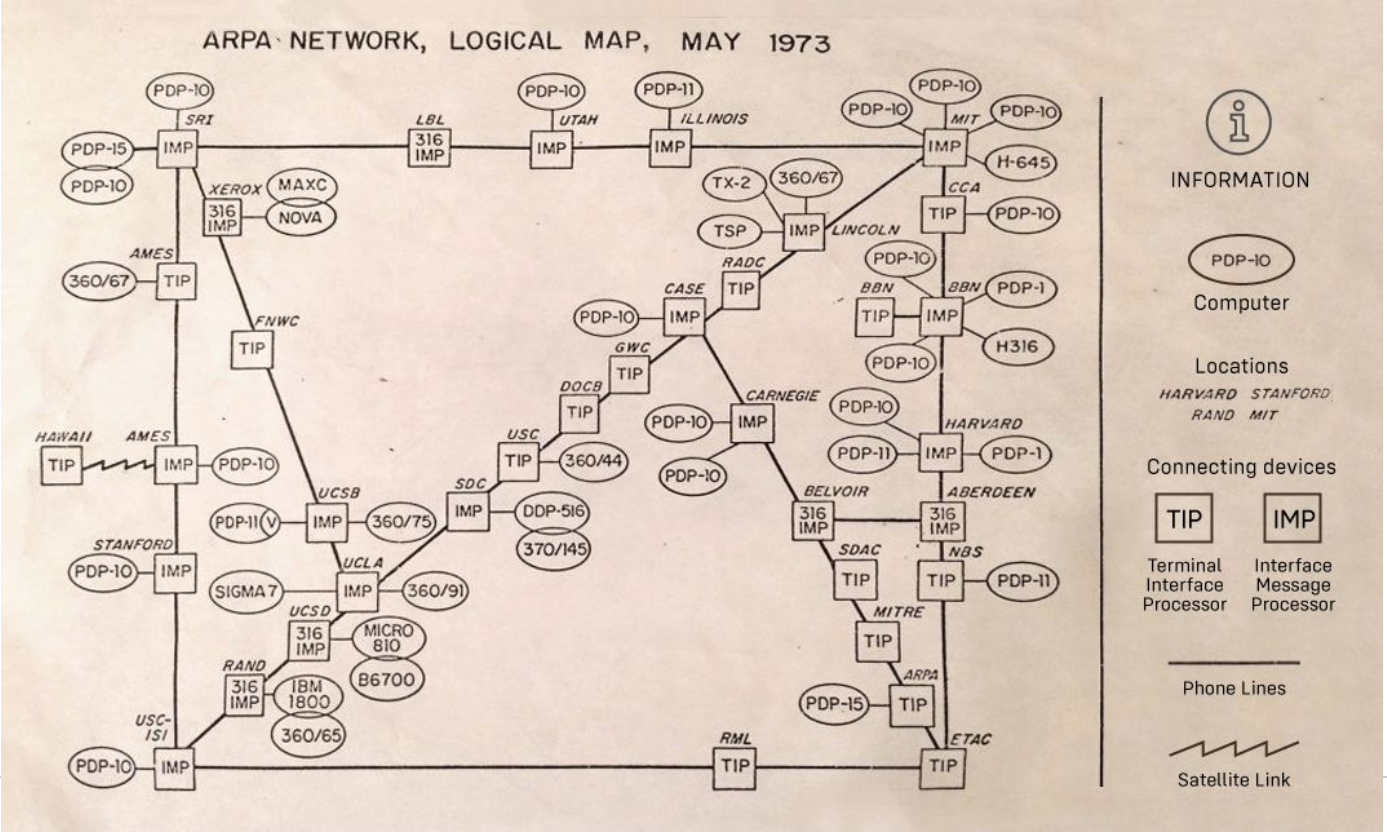


Growth of and digitization of global information-storage capacity[1]

Digital Data Terminology

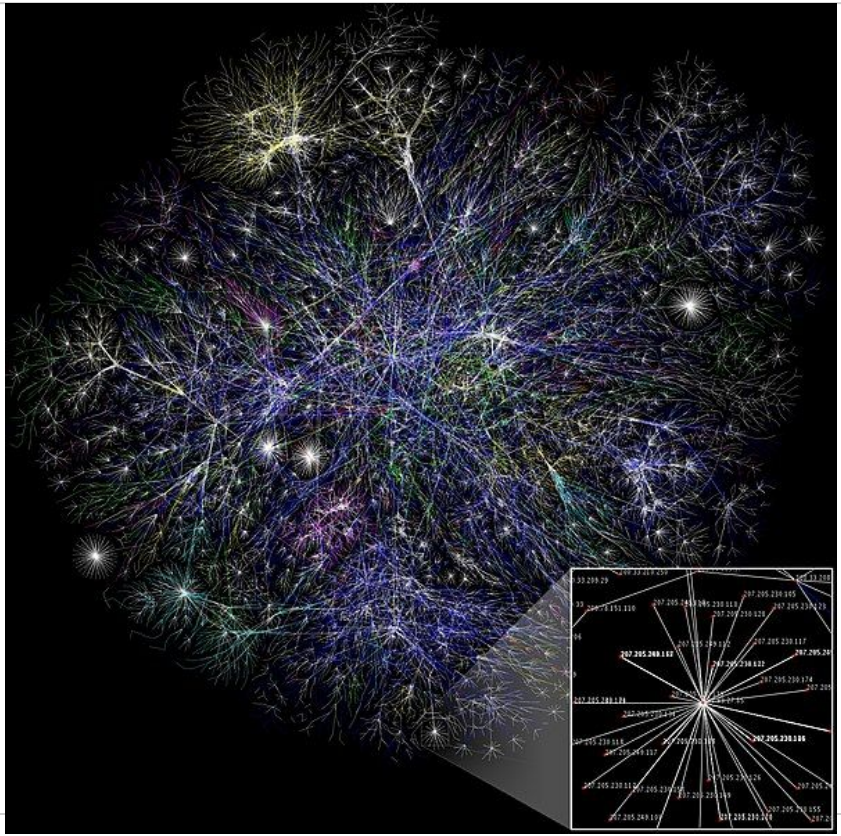
- **Bit** - binary unit: 0/1
- **Byte** - eight bits
- **Kilobyte** - 2^10 or 1024 bytes
- **Megabyte** - 2^20 bytes or 1024 kilobytes
- **Gigabyte** - 2^30 bytes or 1024 megabytes
- **Terabyte** - 2^40 bytes or 1024 gigabytes
- **Petabyte** -2^50 bytes or 1024 terabytes
- **Exabyte** - 2^60 bytes or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

Arpanet, 1973



Wikipedia

Internet now



Wikipedia

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.

OCT. 17, 2017 AT 6:00 AM

The Supreme Court Is Allergic To Math

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 your mind as well as your body

It seems to promote cognitive performance

ECONOMIST.COM

Science News *from research organizations*

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

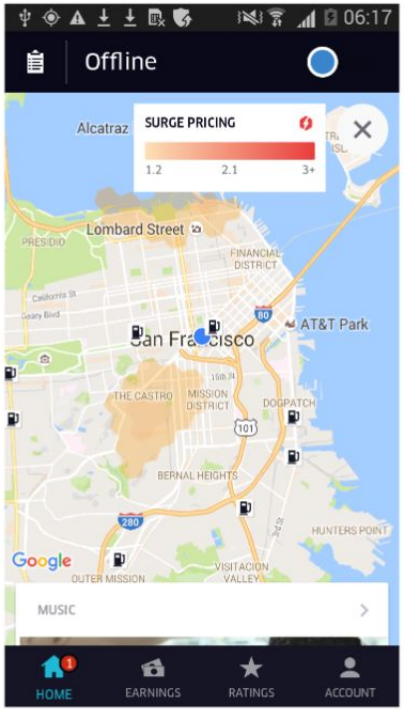
- **Exploration**
  - Identifying patterns in information
  - Uses visualizations
- **Inference**
  - Quantifying whether those patterns are reliable
  - Uses randomization
- **Prediction**
  - Making informed guesses
  - Uses machine learning

Data Science Stories

- Citibike sharing program in Manhattan
  - Where/when is there demand for bikes?

Data Science Stories

Do Uber drivers respond to surge pricing?



Lu, Frazier, Kislev 2018

What we'll do: Citibike visualization

Learn enough computing to do our own visualizations and observations to identify patterns in big data sets.

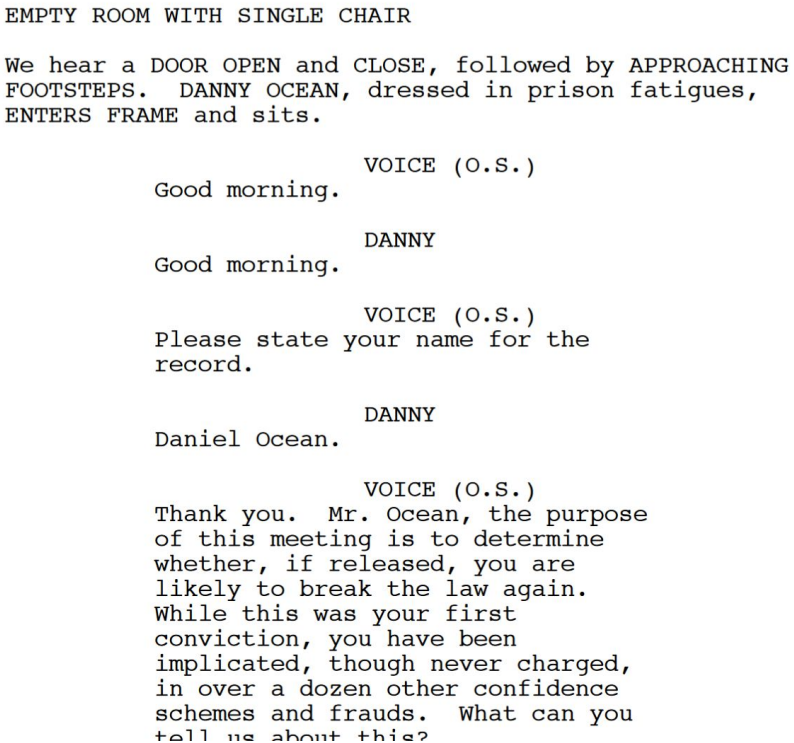
What we'll do: Deflategate

Deflategate refers to the alleged deflation of footballs below the required amount by the New England Patriots during a January 18, 2015 playoff game against the Indianapolis Colts.

What does statistics say about how likely this deflation was?

What we'll do: Classify movies

Program a computer to predict the genre of a movie just from its script.



EMPTY ROOM WITH SINGLE CHAIR

We hear a DOOR OPEN and CLOSE, followed by APPROACHING FOOTSTEPS. DANNY OCEAN, dressed in prison fatigues, ENTERS FRAME and sits.

VOICE (O.S.)
Good morning.

DANNY
Good morning.

VOICE (O.S.)
Please state your name for the record.

DANNY
Daniel Ocean.

VOICE (O.S.)
Thank you. Mr. Ocean, the purpose of this meeting is to determine whether, if released, you are likely to break the law again. While this was your first conviction, you have been implicated, though never charged, in over a dozen other confidence schemes and frauds. What can you tell us about this?

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?

Data Science in Action

Course Structure

Everything you want is here:

www.cs.cornell.edu/courses/cs1380

How DSFA works

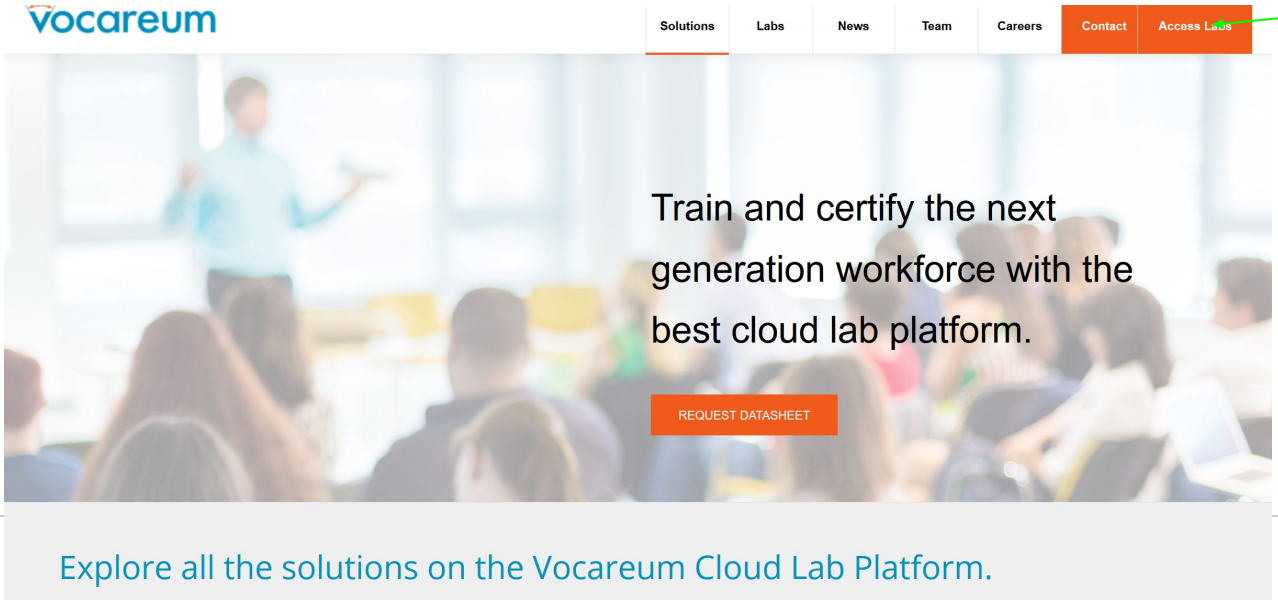
- Lecture Tuesday and Thursday
  - Participation counts for (small part) of grade
- Section every week on W or Th
  - Including this week!
  - Attend the one you signed up for.
  - Bring a laptop! (Talk to me if you don’t have one)

How DSFA works

- Assignments
  - Labs (weekly)
  - Homework (mostly weekly)
    - Generally out Friday, due following Friday
  - Projects (about 3)
    - Longer HW, longer time to work on them.
  - You may work with a partner on projects if they are in the same section as you.
- Exams
  - Two prelims: February 27 and April 7
  - Final exam: May 15, 9AM

How DSFA works

Labs, HW, and projects use Vocareum (www.vocareum.com), a site with Jupyter notebooks built in. It is free until February 5, will cost $30 to buy access then.



Click here to log in

How DSFA works

If you were registered for the course as of noon yesterday, you should have gotten a Vocareum email to register for the website.

If not, go to https://tinyurl.com/dsfa2020, leave your netid/email.

How DSFA works

We'll be using iClickers and REEF polling to ask multiple-choice questions during lectures and get answers.

You can get an iClicker from the Cornell Store. REEF polling allows you to answer these questions from an app on your phone; sign up at https://app.reef-education.com/#/login using your Cornell netid. There is a fee.

If you get an iClicker, it must be registered on Canvas.

Other resources

- Piazza, a website for class announcements and asking questions piazza.com/cornell/spring2020/csoriestsci1380/home.
- Canvas, where we'll be keeping track of grades.
- The course website also links to an online textbook at www.cs.cornell.edu/courses/cs1380/2018sp/textbook.

Section Schedule

Section	Time	Room	TA
DIS201	W 12:20-2:15pm	Thurston Hall 202	Ben
DIS202	W 2:30-4:25pm	Thurston Hall 202	Ben
DIS203	W 7:30-9:25pm	Thurston Hall 202	Artem
DIS205	R 12:20-2:15	Hollister Hall 362	Artem

Policies, Grading, Etc.

- Details are posted on the course website: www.cs.cornell.edu/courses/cs1380/.

.

Academic Integrity

- Labs:
  - Work together in section 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)
  - Don't look at others' code!
- iClickers:
  - Don't click in on anyone else's clicker.

In particular:

- Don't post code on Piazza
- Cite your sources (including other students)

Getting help

Questions about material:

- Ask a friend
- Ask on Piazza
- Go to section
- Go to office hours

Logistical questions:

- Ask your section TAs

Now what?

- (Now) If you're not enrolled yet sign up
- (Tomorrow or Thursday) Go to section
- (By Thursday) Read Chapters 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, or REEF Polling
- (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 Canvas (and you don’t need a Reef account)

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.