**DSFA**
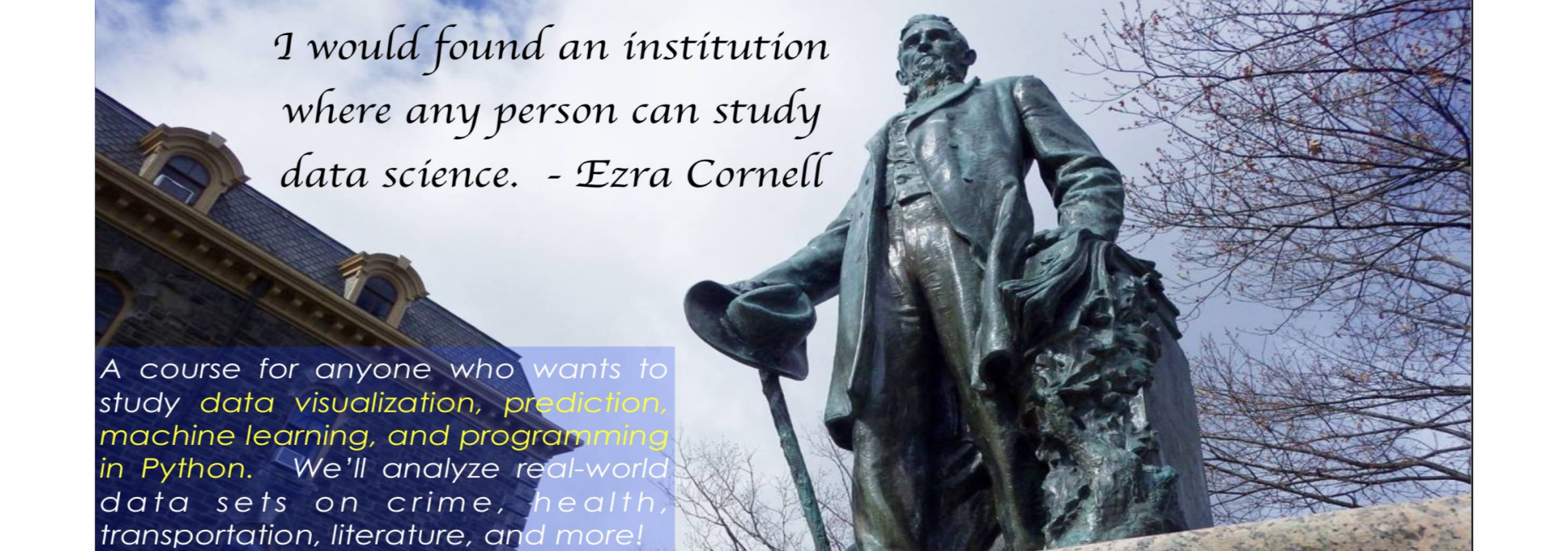Spring 2018

# Lecture 1

CS+ORIE+STSCI 1380
Introduction

*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!

# CS 1380 + ORIE 1380 + STSCI 1380

## Data Science For All

## Spring 2018 MWF 10:10-11:00 am

https://tinyurl.com/datascienceforall

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

# Who are we?



- Professor Udell



- Professor Clarkson

+ TAs: http://www.cs.cornell.edu/courses/cs1380/2018sp/staff.html

# 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)

More info: http://www.cs.cornell.edu/courses/cs1380/2018sp/

# Why Data Science?

# 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 <span style="color:red">average size of a graduate class of data science students is 23 students</span>. With approximately only 110 universities offering data science studies, the growing market will continue to pressure the supply in the US.

January 22, 2016
Data Scientists: The Myth and the Reality
Seamus Breslin

# 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 [some justices] 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.

# WEAPONS OF
# MATH DESTRUCTION

## HOW BIG DATA INCREASES INEQUALITY
## AND THREATENS DEMOCRACY

# CATHY O'NEIL

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

# 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**
  - What is the chance of winning each district?
  - Who might be willing to donate if I asked?
  - How should I ask?
- **Medicine**
  - Which patients are at risk of some disease?
  - Which patients would benefit from surgery?

# Data Science in Action

# Course Structure

# How DSFA works

- Lecture MWF
  - 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 10 total)
  - Projects (3 total)
- Exams:
  - Two prelims + final exam

More info: http://www.cs.cornell.edu/courses/cs1380/2018sp/schedule.html

# Policies, Grading, Etc.

http://www.cs.cornell.edu/courses/cs1380/2018sp/policies.html

# 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)

More info: http://www.cs.cornell.edu/courses/cs1380/2018sp/policies.html

# Now what?

- (Now) If you're not enrolled yet [sign up here](#)
- (Today or tomorrow) Go to section
- (By Friday) Read [Chapter 1](#) 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](#)

# 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.