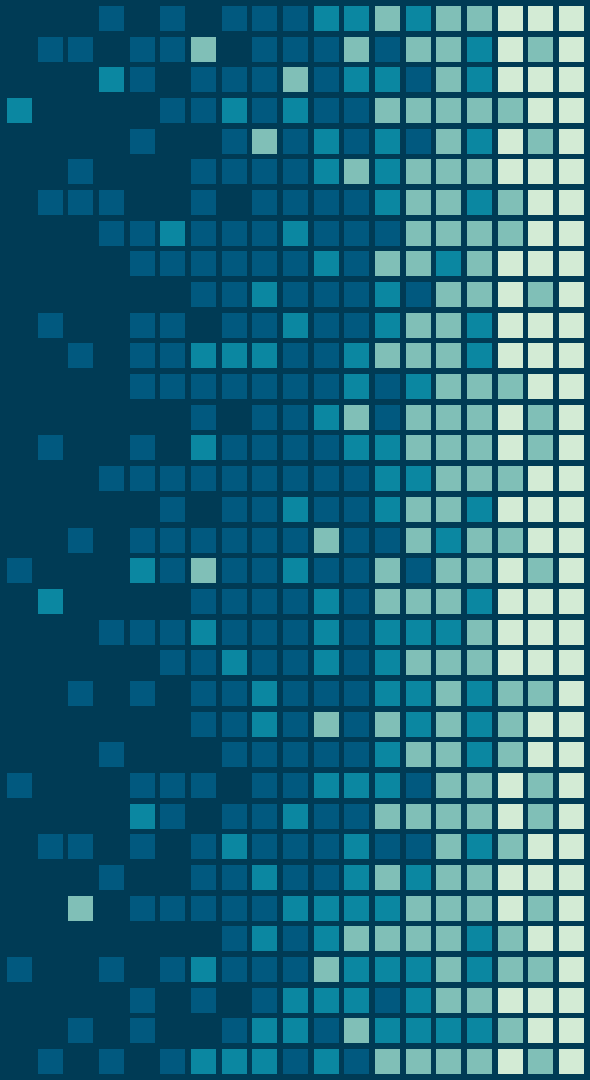
BST 260

Introduction to Data Science

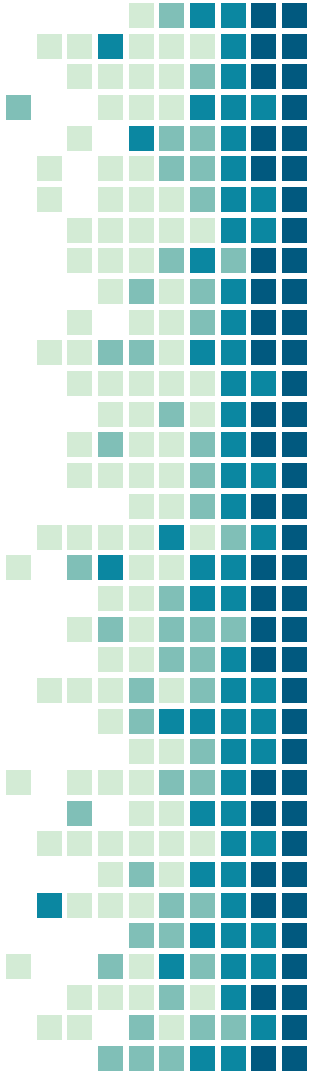
Fall 2019



Today

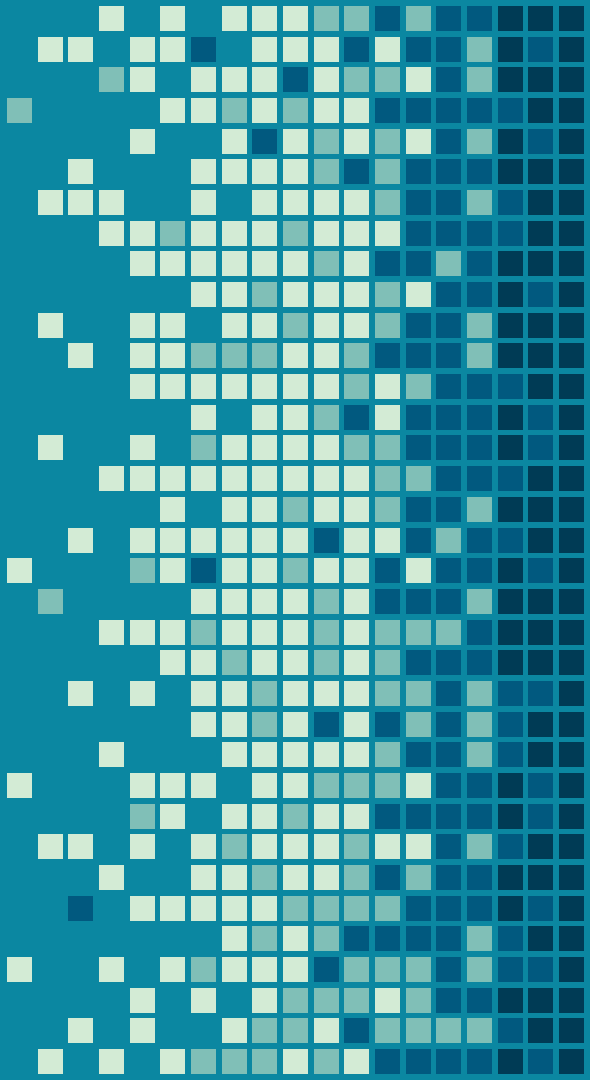
- What *is* data science?
- Why learn data science?
- How do we learn data science?
- Who is helping you learn data science?
- Introduction to R, RStudio and RMarkdown

What *is* data science?



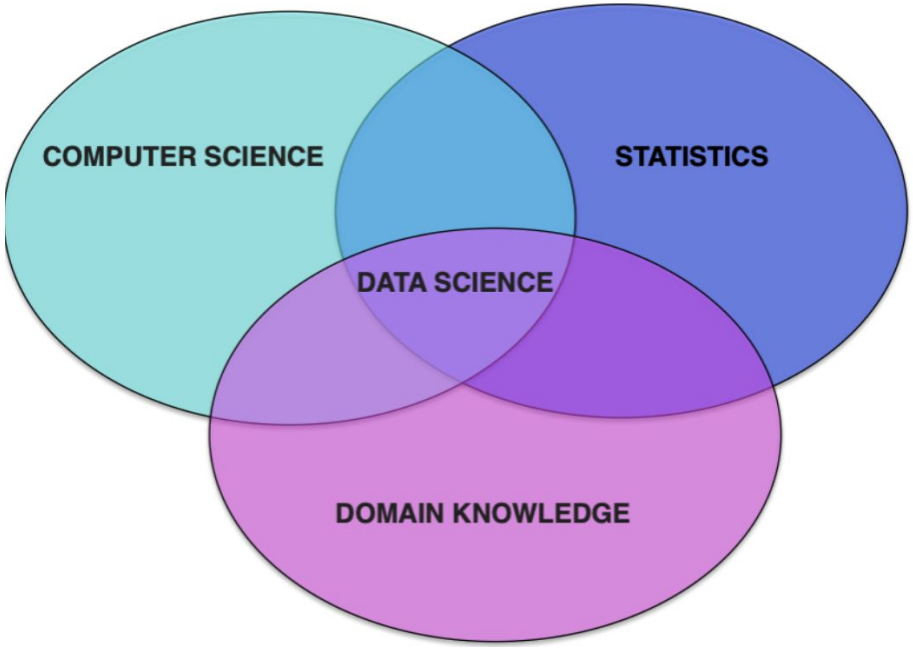
What is data science?

> *Data science is a multidisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and* ***insights*** *from structured and unstructured data*

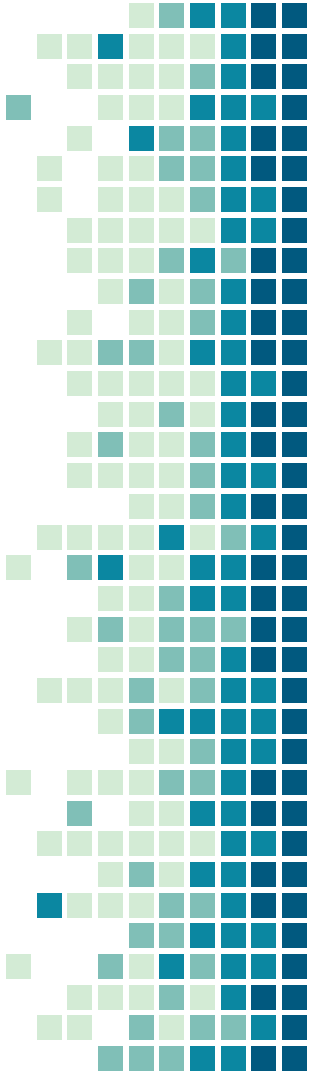


> *Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful **insights** from data*
>
> - *Data Robot*

> *The goal is to turn data into information and information into **insight***
>
> - *Carly Florina, former CEO of Hewlett-Packard*


COMPUTER SCIENCE
STATISTICS
DATA SCIENCE
DOMAIN KNOWLEDGE

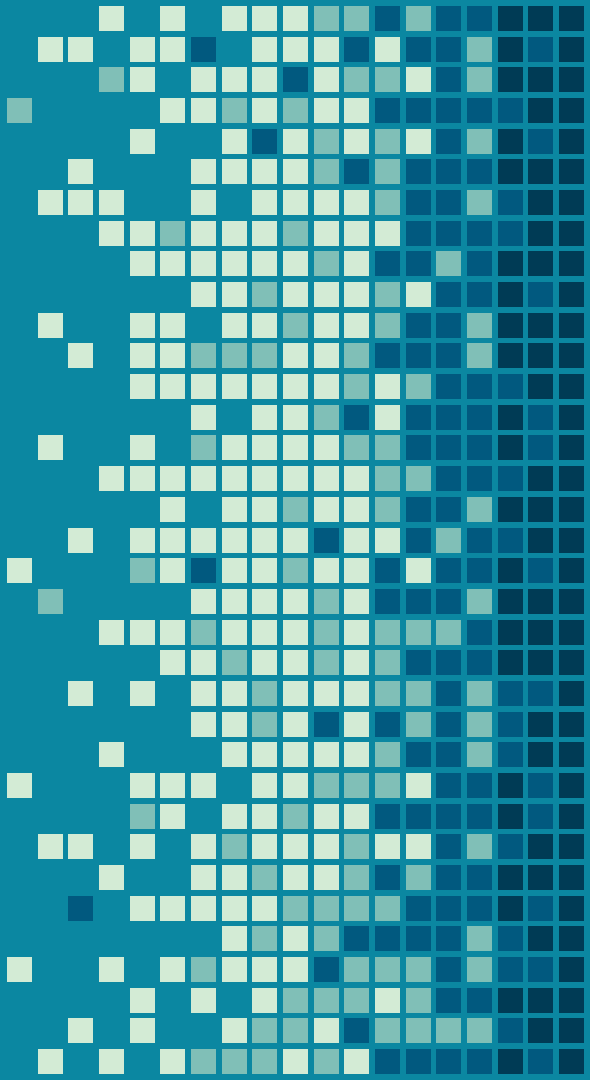
Why learn data science?



“Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world.

-Atul Butte,

Stanford University

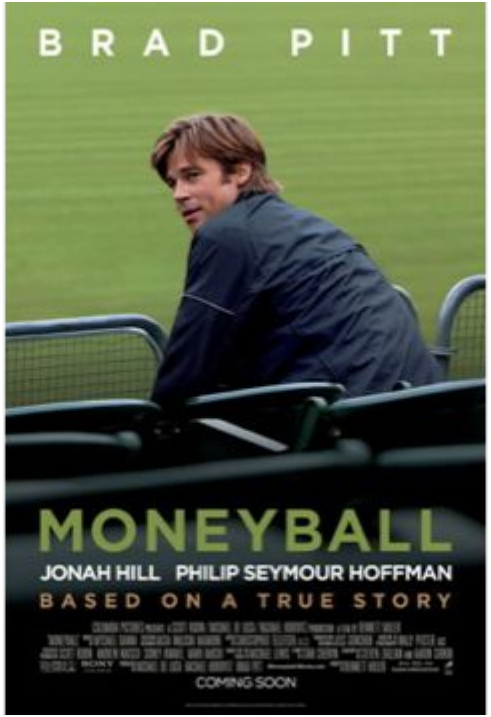


“You can have all the quantitative data you can get, but you still have to distrust it and use your own intelligence and judgement

- *Alvin Toffler, American writer and futurist*

> *The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it is going to be a hugely important skill in the next decades, not only at the professional level, but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data.*
>
> - *Hal Varian, Economist at Google*

Data Science Success Stories

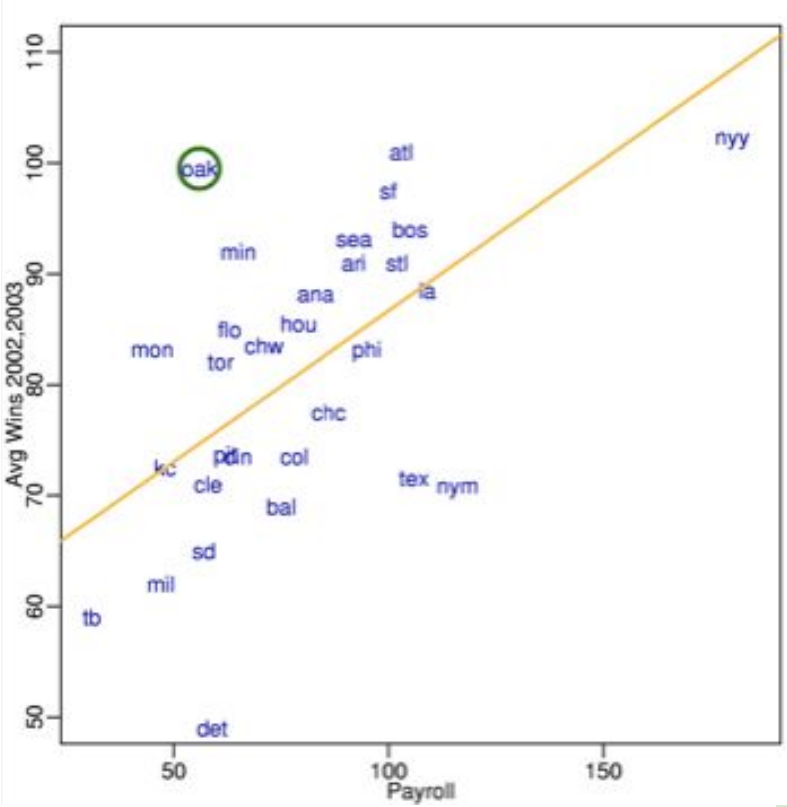


Actual data scientist

Hollywood

Moneyball

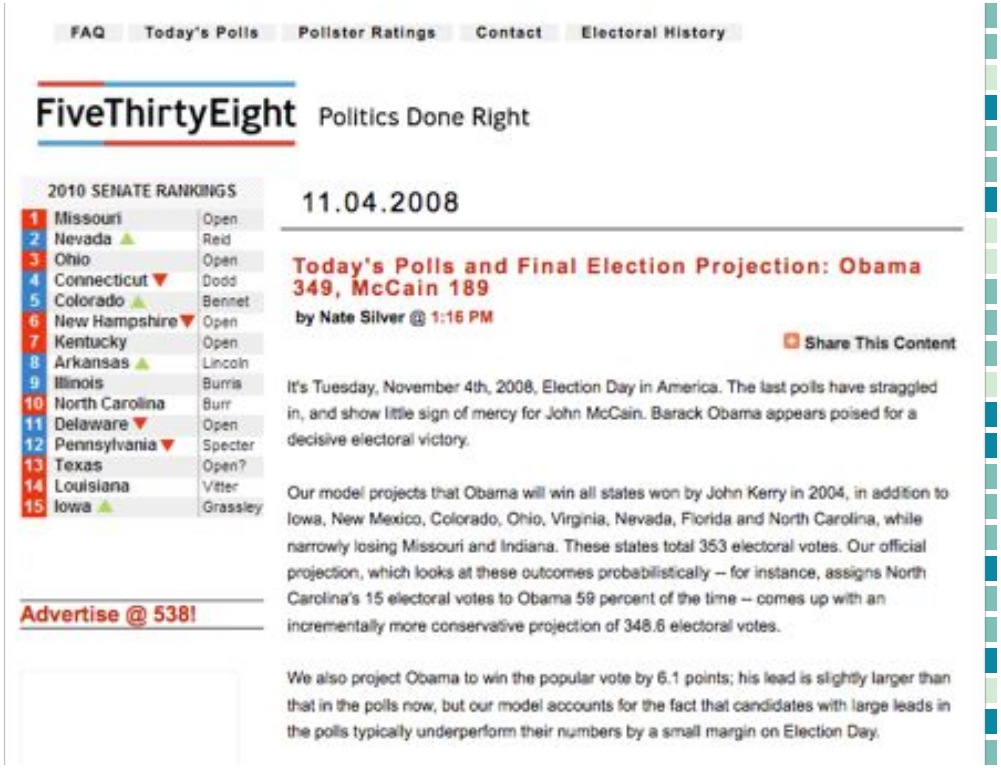
Starting around 2001, the Oakland A's picked players that scouts thought were no good, but data said otherwise



Elections

"Nate Silver won the election"

- Predicted: 349 to 189, 6.1% difference
- Actual: 365 to 173, 7.2% difference



FAQ | Today's Polls | Pollster Ratings | Contact | Electoral History

FiveThirtyEight Politics Done Right

2010 SENATE RANKINGS

1	Missouri	Open
2	Nevada ▲	Reid
3	Ohio	Open
4	Connecticut ▼	Dodd
5	Colorado ▲	Bennet
6	New Hampshire ▼	Open
7	Kentucky	Open
8	Arkansas ▲	Lincoln
9	Illinois	Burris
10	North Carolina	Burr
11	Delaware ▼	Open
12	Pennsylvania ▼	Specter
13	Texas	Open?
14	Louisiana	Vitter
15	Iowa ▲	Grassley

Advertise @ 538!

11.04.2008

Today's Polls and Final Election Projection: Obama 349, McCain 189

by Nate Silver @ 1:16 PM

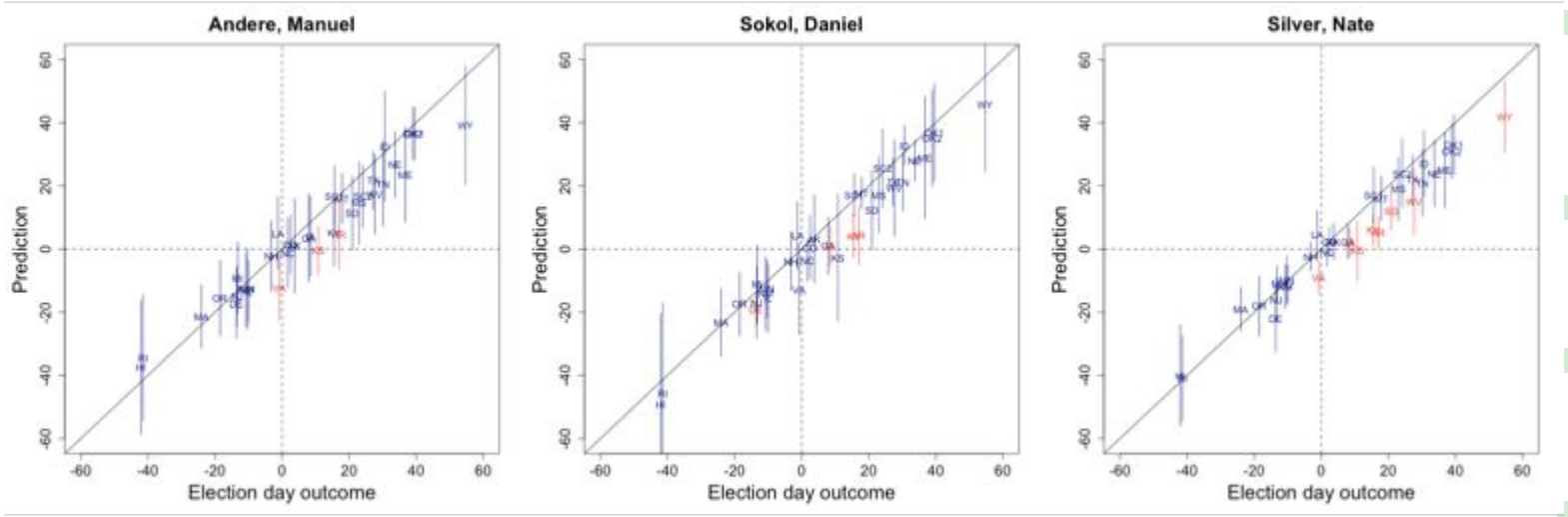
Share This Content

It's Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.

Our model projects that Obama will win all states won by John Kerry in 2004, in addition to Iowa, New Mexico, Colorado, Ohio, Virginia, Nevada, Florida and North Carolina, while narrowly losing Missouri and Indiana. These states total 353 electoral votes. Our official projection, which looks at these outcomes probabilistically -- for instance, assigns North Carolina's 15 electoral votes to Obama 59 percent of the time -- comes up with an incrementally more conservative projection of 348.6 electoral votes.

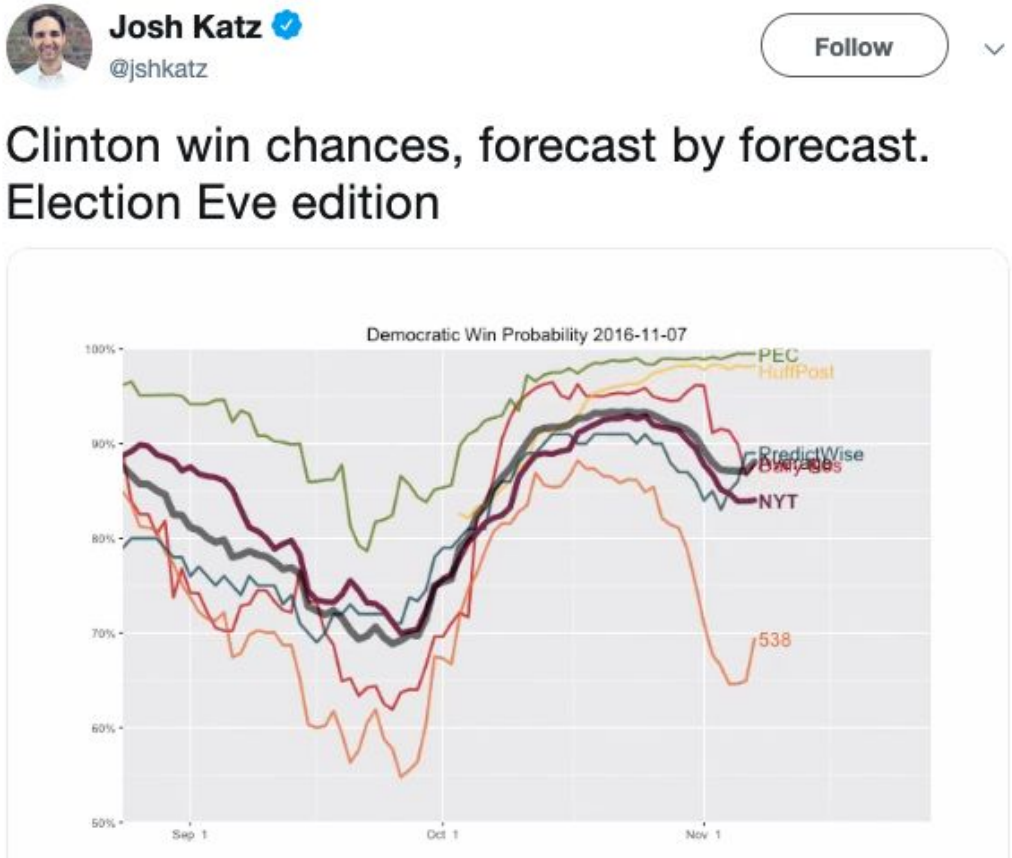
We also project Obama to win the popular vote by 6.1 points; his lead is slightly larger than that in the polls now, but our model accounts for the fact that candidates with large leads in the polls typically underperform their numbers by a small margin on Election Day.

2014 Senate Race



2016 Presidential Election

[Twitter gif]



Many other examples

- Spell checkers
- Speech recognition
- Language translators
- Digitizing books
- Medical diagnostics
- Personalized medicine
- Etc.

Who is helping you learn data science?

Instructor

- Heather Mattie
- Instructor of data science
- PhD Biostatistics, Harvard GSAS
- Advisor: JP Onnela
- Research centers on statistical network science, machine learning, algorithmic bias, health disparities

- Email: hemattie@hsph.harvard.edu
- Office: Building 1, 4th floor, room 421A
- Phone: (617) 432-5308
- Office Hour: Tuesdays 1-2pm

TA Team

- Andy Shi (andyshi@g.harvard.edu)
- Greyson Liu (gang_liu@g.harvard.edu)
- Eric Dunipace (edunipace@g.harvard.edu)
- Jane Liang (jwliang@g.harvard.edu)
- Rolando Acosta (roa310@g.harvard.edu)

TA Office Hours

Monday

- 2-3pm, Greyson, Heather's office

Wednesday

- 1-2pm, Jane, Building 2, 4th floor room 428
- 3:30-4:30pm, Rolando, Building 2, 4th floor room 428

Thursday

- 1-2pm, Eric, Kresge LL6

Friday

- 1-2pm, Andy, Building 2, 4th floor room 428

Labs

We will have labs centered around examples related to data and code presented in class. Labs will start next week and will be held almost every week - check canvas and the course website for the schedule

- Wednesday
  - 2-3:30pm, Rolando, Kresge LL6
- Thursday
  - 3:45-5:15pm, Greyson, FXB G11
- Friday
  - 11:30-1pm, Andy, Kresge G13

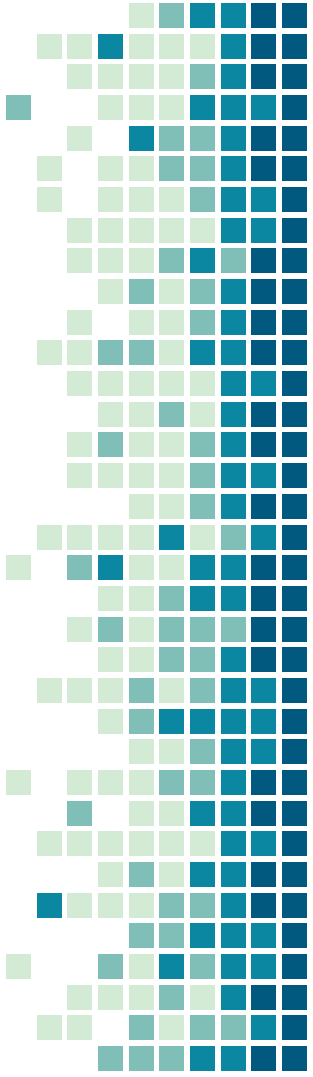
Course Details

Many modes of communication

1. Canvas site
2. GitHub site
3. Slack workspace

Grading

- Homework
  - 35% of final grade
  - 5 homework assignments
- Two midterms
  - 30% of final grade; 15% each
  - Multiple choice questions on canvas
  - Some questions will require writing code
- Final Project
  - 35% of final grade
  - Can work alone or on a team of up to 4 people

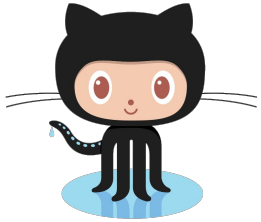


Grading

- You are welcome to discuss the course material and homework questions with others, but the work you turn in must be your own

- You are not allowed to discuss during the midterms

Homework



GitHub

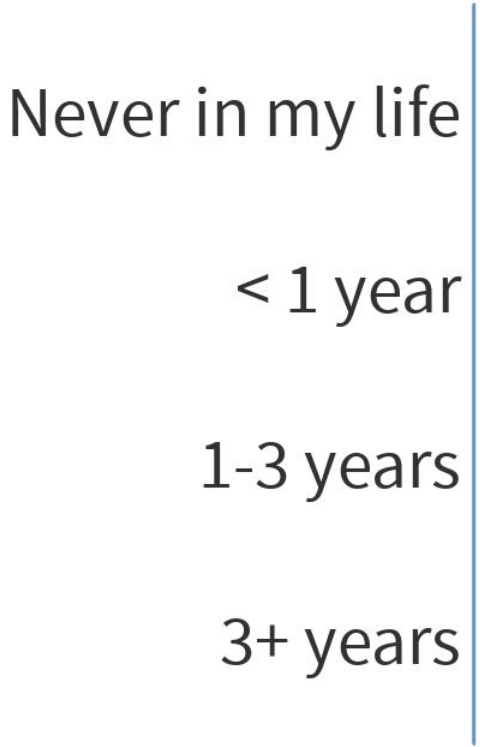
- Real-world focus
- Scrape and wrangle (clean) messy data
- Explore data
- Visualize data
- Apply statistical analyses
- Communicate results
- Will be written in R and submitted via GitHub

Final Project

- Individual or teams of up to 4 students
- Choose your own data and project
- Part 1: describe your project question and plan for answering it
- Part 2: present results and conclusions
  - RMarkdown file
  - Screencast
  - Website

How do we learn data science?

How long have you been coding?



How comfortable are you with coding?

I would feel more comfortable in a crowded elevator

Fairly comfortable, like a day in May in Boston

So comfortable it's like I'm wearing an oversized sweater and sipping a latte while looking outside at the rain in a coffee shop during Fall

How will you learn? By doing it!

- You will be coding - a lot
- Skills you'll learn
  - Computer programming
  - Data wrangling
  - Data visualization
  - Statistics
  - Machine learning
  - Communication of results and insights

Class will be centered around **data**

- Homework and lectures will be based on diverse data sets
- Bring your laptop to class!

What you need to do now

- Download and install R and RStudio
- Create a GitHub account
- Complete this survey (also available on the course website and canvas)
- Join the Slack workspace
- Explore the course website
- Make friends with your classmates - you'll need them for the final project 😄