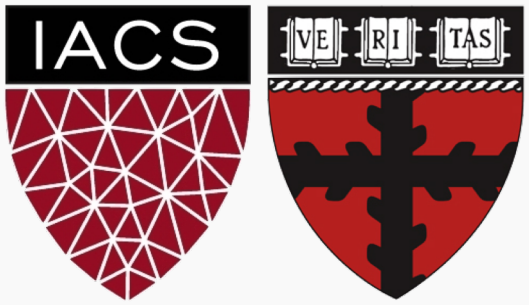
Lecture #0: Introduction to C109a

S-109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader



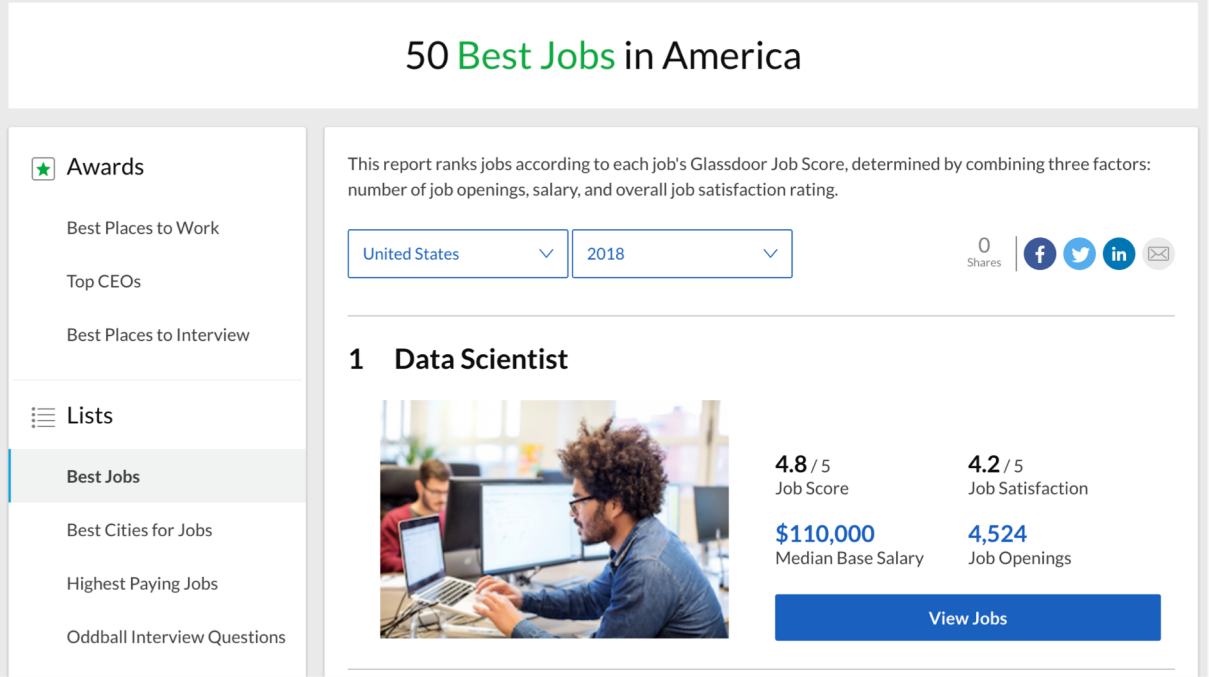
Lecture Outline

- What is Data Science?
- What is This Class?
- The Data Science Process
- Example

What is Data Science?

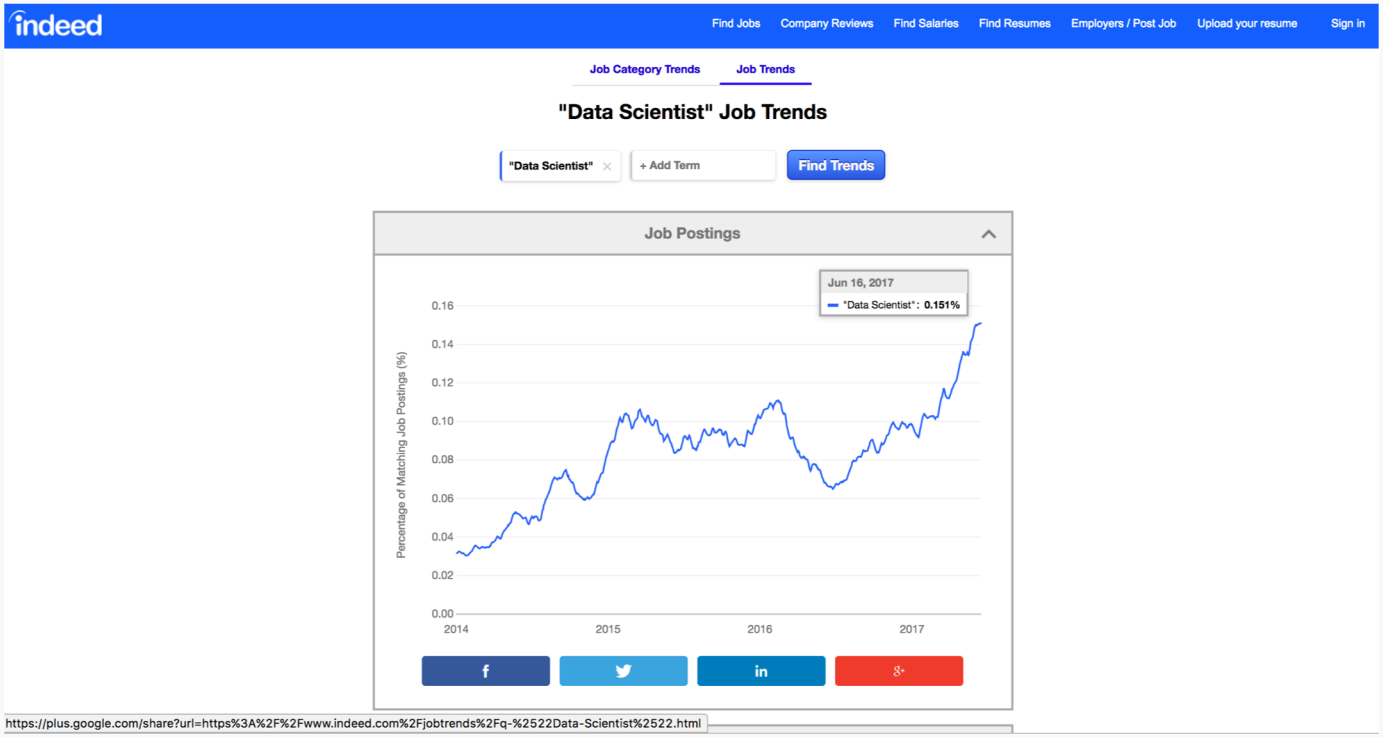
Why?

Jobs!


50 Best Jobs in America
Awards
Best Places to Work
Top CEOs
Best Places to Interview
Lists
Best Jobs
Best Cities for Jobs
Highest Paying Jobs
Oddball Interview Questions
This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.
United States
2018
0
Shares
1 Data Scientist
4.8 / 5
Job Score
4.2 / 5
Job Satisfaction
$110,000
Median Base Salary
4,524
Job Openings
View Jobs

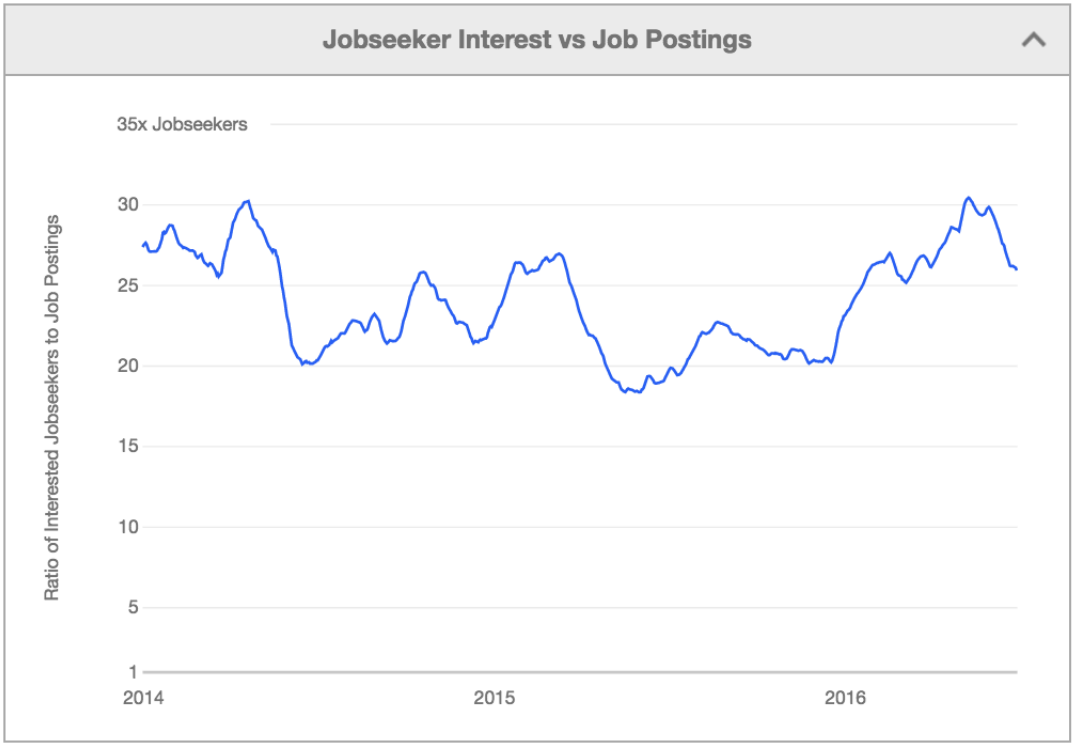
Why?

Jobs demand



Why?

Jobs supply



Why?

Jobs:

By 2018, the US could face a shortage of up to 190,000 workers with analytical skills

McKinsey Global Institute

The sexy job in the next 10 years will be statisticians.

Hal Varian, Prof. Emeritus UC Berkeley Chief Economist, Google

How?

Long time ago (thousands of years) science was only empirical and people counted stars

How?

Long time ago (thousands of years) science was only empirical and people counted stars or crops

How?

Long time ago (thousands of years) science was only

empirical and people counted stars or crops and used the data to create machines to describe the phenomena

How?

Few hundred years: theoretical approaches, try to derive equations to describe general phenomena.

$v = v_0 + at$	$\omega_1 = \omega_0 + \alpha t$
$s = \frac{1}{2}(v_0 + v)t$	$\theta = \frac{1}{2}(\omega_0 + \omega_1)t$
$s = v_0 t + \frac{1}{2}at^2$	$\theta = \omega_0 t + \frac{1}{2}\alpha t^2$
$v^2 = v_0^2 + 2as$	$\omega_1^2 = \omega_0^2 + 2\alpha\theta$
$s = vt - \frac{1}{2}at^2$	$\theta = \omega_1 t - \frac{1}{2}\alpha t^2$

$$T^2 = \frac{4\pi^2}{GM}a^3$$

can be expressed as simply

$$T^2 = a^3$$

If expressed in the following units:

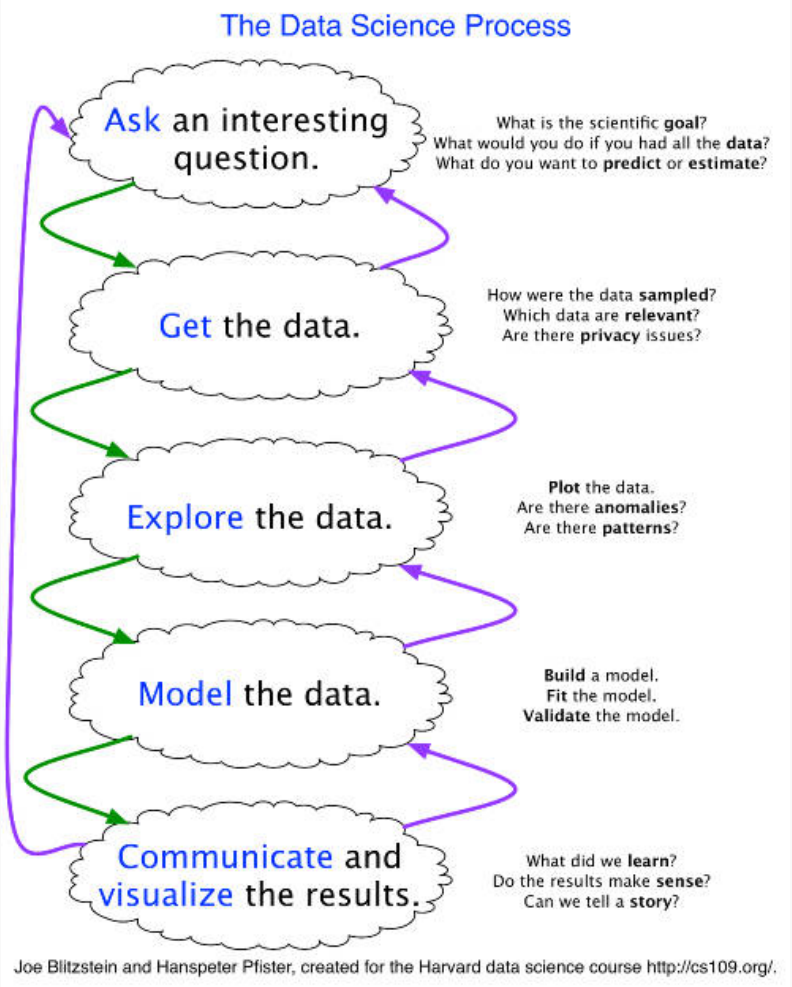
- T Earth years
- a Astronomical units AU (a = 1 AU for Earth)
- M Solar masses $M_\odot$

Then $\frac{4\pi^2}{G} = 1$

$$H(t)|\psi(t)\rangle = i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle$$

What is this class?

What?


The Data Science Process
Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

What?

The material of the course will integrate the five key facets of an investigation using data:

1. data collection; data wrangling, cleaning, and sampling to get a suitable data set
2. data management; accessing data quickly and reliably
3. exploratory data analysis; generating hypotheses and building intuition
4. prediction or statistical learning
5. communication; summarizing results through visualization, stories, and interpretable summaries.

What?

Module 0:

Getting ready with python, jupyter notebooks, some basic statistics, matplotlib (viz) and numpy.

Lectures during this time will often be lab-like.

What?

Module 1:

Regression, Transportation Data, Basic Visualization and sklearn:

- knn regression
- Linear and Polynomial Regression
- Multiple Regression
- Model Selection
- Regularization

What?

Module 2:

Classification, Health Data, Keras and Presentations Stack:

- Logistic Regression (linear and polynomial)
- Multiple Logistic Regression
- Discriminant analysis (LDA and QDA)
- Classification with decision trees
- Missing data and knn classification
- Perceptron, backpropagation and gradient descent

What?

Module 3:

Ensemble Methods, Natural Science data, Web Site Building and Report Writing

- Random Forest
- Bagging
- Boosting
- Stacking
- Support Vector Machine
- Neural Networks
- Reinforcement Learning

Who?

Pavlos Protopapas

Scientific Director of the Institute for Applied Computational Science (IACS)

Teaches CS109(a/b) and the Capstone course for the Data Science masters program.

Research in astrostatistics and he is excited about the new telescopes coming online in the next few years. He has absolutely no hobbies or interests except teaching CS109 and eating.

Who?

Kevin Rader

Senior Preceptor in Statistics. Teaches CS 109A & Stat 139 this fall and Stat 102 & Stat 98 in the spring.

Research interests include complex survey analysis and survival analysis. Hobbies include the outdoors, sports (especially the aquatic variety), and of course, farming.

Who?

Rahul Dave

Rahul, our resident lab guru and python guru, is a lecturer at the IACS. He also teaches AM207, and has taught labs for CS109 in 2013, 2015, and last year. He loves mountains, Bayesian Stats, and determining whether something is a hot dog.

Teaching Fellows

Head TFs:

Eleni Kaxiras
(SEAS, mostly)

Sol Girouard
(DCE, mostly)

Section TFs, Lab Assistants, Graders, oh my!

Nicholas Ruta, Camilo Fosco, Will Claybaugh, Trevor Rhone, Marios Matthaiakis, Mehul Smriti Raje, Ken Arnold, Karan Rajesh Motwani, Cecilia Garraffo, Justin Lee, Filip Michalsky, Yujiao Chen, Rashmi Bathia, Brandon Walker, Yijun Shen, Jiejun Lu, Romil Sirohi, Weihang Zhang, Brian Matejek, Andrea Porelli, Chris Gumb, Jerry Peng, and Boyuan Sun.

Lectures, Labs, Sections, and Office hours

Lectures:

Mondays and Wednesdays 1:30-2:45pm @Northwest Building B103.

During lecture will cover the material which you will need to complete the homework, midterms and to survive the rest of your life. *Attending* lectures is required - quizzes at the end of each lecture (drop 40% of them) .

We will use a mix of notes and examples via notebooks

1. Lecture notes and associated notebooks will be posted before lecture on Canvas and on GitHub.
2. Lectures will be video taped (and live streamed) and posted approximately within 24 hours on Canvas.

Lectures, Labs, Sections, and Office hours

Labs:

Thursdays 4:30-6:00pm and Fridays 10:30-11:45am in Pierce 301. Begin this Friday!

Labs are meant to help you understand the lecture material better via examples…in Python!

Labs will be video taped (and live streamed) and posted approximately within 24 hours on Canvas.

Lectures, Labs, Sections, and Office hours

Sections:

They come in 2 flavors:

- **Advanced Sections** for CS 209a: will be videotaped and posted approximately within 24 hours on Canvas. Required for 209a, anyone is welcome. Covers the mathematics underpinning the models. Begin Sep. 19.
  - **Wednesday 5:00-6:30pm**. Location TBD

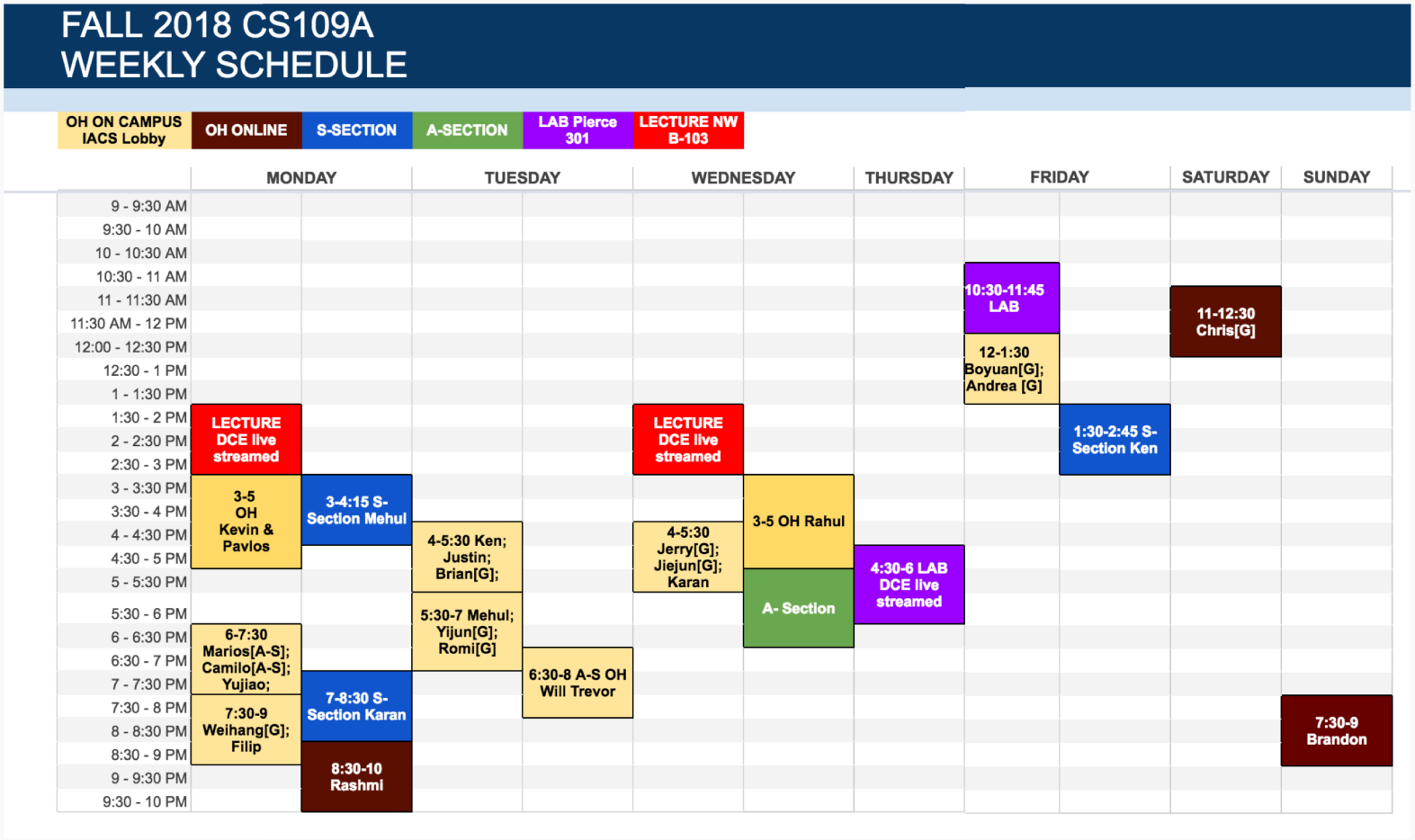
- **Standard Sections** for everyone (not videotaped). Meant to go through practice problems similar to what is on the homework assignments. Begin Sep. 14.
  - **Friday 1:30-2:45pm** Location TBD
  - **Monday 3-4:15pm and 7-8:30pm** Locations TBD

Lectures, Labs, Sections, and Office hours

Office Hours (beginning next week):

- **Instructor:** Mon & Wed 3-5pm in IACS student lobby
- **TF-led:**
  - **On campus** (IACS student lobby):
    Mon 6-9pm
    Tues 4-8pm
    Wed 4-5:30pm
    Fri 12-1:30pm
  - **Online** (via zoom):
    Mon 8:30-10pm
    Sat 11am-12:30pm
    Sun 7:30-9pm.

Weekly Schedule



Homework Assignments

There will be 9 homework (not including Homework 0). **Homework 0 is due Wed, Sep. 12 @ 11:59pm.**

You are encouraged but not required to submit in pairs, except Homeworks 4 & 9, which must submitted individually.

We will be using the 'Groups' function in Canvas to do this, details to be announced later.

All homework are **due 11:59pm Wednesday** and homework will be released on Tuesday by 11:59pm (8 days prior to due date).

Late Days:
You are allowed up to 6 days of late homework submissions, maximum of 1 day on any single assignment, no questions asked. Late homework submissions will not be accepted after 24 hours past the due date. If you exceed your 6 late days, 1 point (20%) will be deducted for late days after that.

Final Project

There will be a final group project (2-4 students) due during exams period.

- We will provide 5+ pre-defined projects which you could use for your final project.
- In some very special cases you can use your own (public) data set and your own project definition (to be approved by the instructors).
- There will be 3 milestones along the way (see course calendar).
- Final submission will be in the format of a website (and python notebook).

Final Projects (and possibly a few more)

Puppies: Using images of puppies, create an image classification system for the classification of pure dog breeds.

Lending Club: Using loan applications and payment histories available online, create a *fair* model to determine a loan recipients ability to repay their loan.

Spotify: Using data available from Spotify, create a Music Recommendation System to suggest songs for users to add to a playlist.

Twitter: Using available tweet data, create a Twitter Bot Detection algorithm to determine whether a tweet was posted by an actual user or an automated bot.

Weather: Using data from published academic project and what is publicly available on weather forecasting websites (like weather underground), build a model for describing patterns over time and predicting meteorological and oceanic temperatures.

Help

Help

The process to get help is:

1. Post the question in Piazza and hopefully your peers will answer. We monitor the posts and we will respond within 12 hours from the posting time.
2. Go to Office Hours, this is the best way to get help.
3. For private matters (regrading, etc.) send an email to the Helpline: **cs109a2018@gmail.com.**
   The Helpline is monitored by all the instructors and TFs.
4. For personal matters (doctor's excuse, etc.) send an email to Pavlos and/or Kevin.

Grades

Grades

- Paired Homework: 49%
- Individual Homework: 20%
- Quizzes: 10%
- Project: 21%
- **Total: 100%**

We do not have predefined cuts for grades. We look for breaks in the cumulative distribution.

The Data Science Process

The Data Science Process

The Data Science Process is similar to the scientific process - one of observation, model building, analysis and conclusion:

- Ask questions
- Data Collection
- Data Exploration
- Data Modeling
- Data Analysis
- Visualization and Presentation of Results

Note: This process is by no means linear!

Analyzing Hubway Data

Introduction: Hubway is metro-Boston's public bike share program, with more than 1600 bikes at 160+ stations across the Greater Boston area. Hubway is owned by four municipalities in the area.

By 2016, Hubway operated 185 stations and 1750 bicycles, with 5 million ride since launching in 2011.

The Data: In April 2017, Hubway held a Data Visualization Challenge at the Microsoft NERD Center in Cambridge, releasing 5 years of trip data.

The Question: What does the data tell us about the ride share program?

The Data Exploration/Question Refinement Cycle

Our original question: **'What does the data tell us about the ride share program?'** is a reasonable slogan to promote a hackathon. It is not good for guiding scientific investigation.

Before we can refine the question, we have to look at the data!

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_date	gender
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0	Male
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0	Male
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0	Male
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0	Female
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0	Female

Based on the data, what kind of questions can we ask?

The Data Exploration/Question Refinement Cycle

Who? Who's using the bikes?

Refine into specific hypotheses:

- More men or more women?
- Older or younger people?
- Subscribers or one time users?

The Data Exploration/Question Refinement Cycle

Where? Where are bikes being checked out?

Refine into specific hypotheses:

- More in Boston than Cambridge?
- More in commercial or residential?
- More around tourist attractions?

Sometimes the data is given to you in pieces and must be merged!

The Data Exploration/Question Refinement Cycle

When? When are the bikes being checked out?

Refine into specific hypotheses:

- More during the weekend than on the weekdays?
- More during rush hour?
- More during the summer than the fall?

Sometimes the feature you want to explore doesn't exist in the data, and must be engineered!

The Data Exploration/Question Refinement Cycle

Why? For what reasons/activities are people checking out bikes?

Refine into specific hypotheses:

- More bikes are used for recreation than commute?
- More bikes are used for touristic purposes?
- Bikes are use to bypass traffic?

Do we have the data to answer these questions with reasonable certainty?
What data do we need to collect in order to answer these questions?

The Data Exploration/Question Refinement Cycle

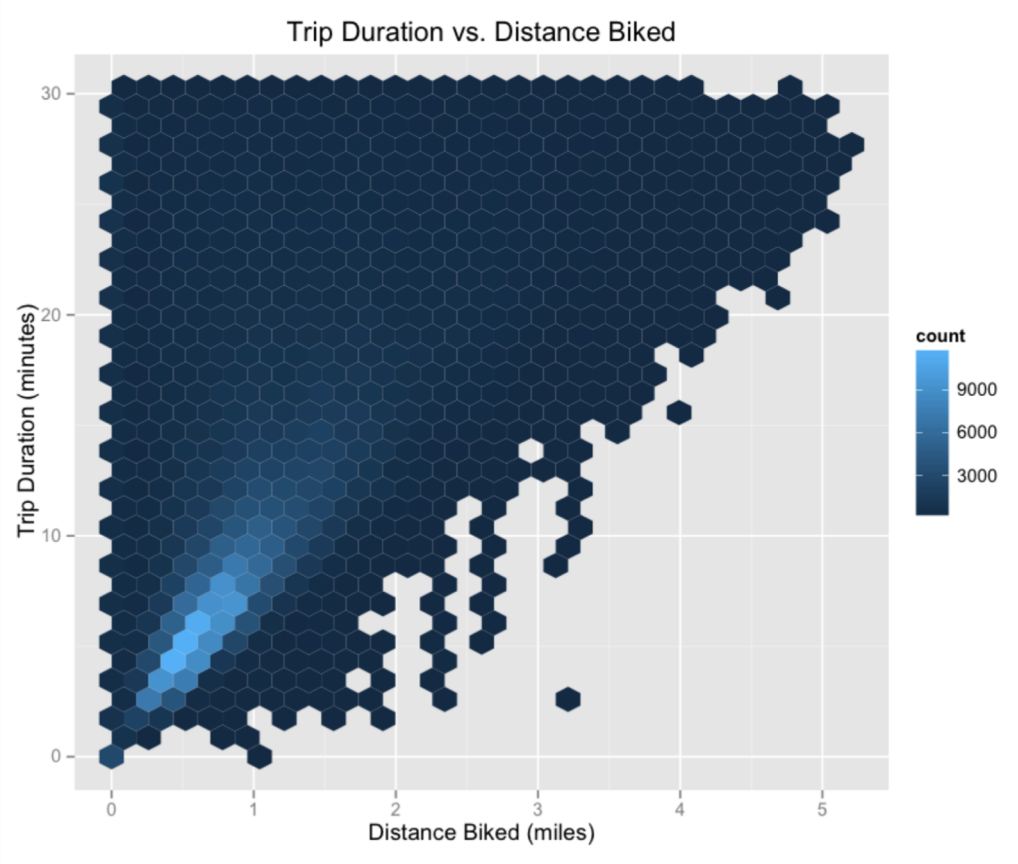
How? Questions that combine variables.

- How does user demographics impact the duration the bikes are being used? Or where they are being checked out?
- How does weather or traffic conditions impact bike usage?
- How do the characteristics of the station location affect the number of bikes being checked out?

How questions are about modeling relationships between different variables.

Inspirations for Data Viz/Exploration

So how well did we do in formulating creative hypotheses and manipulating the data for answers?

Check out the winners of the Hubway Challenge:

http://hubwaydatachallenge.org