Lecture #0: Introduction to CS109A

CS 109A, STAT 121A, AC 209A

Pavlos Protopapas Kevin Rader

Lecture Outline

What is Data Science

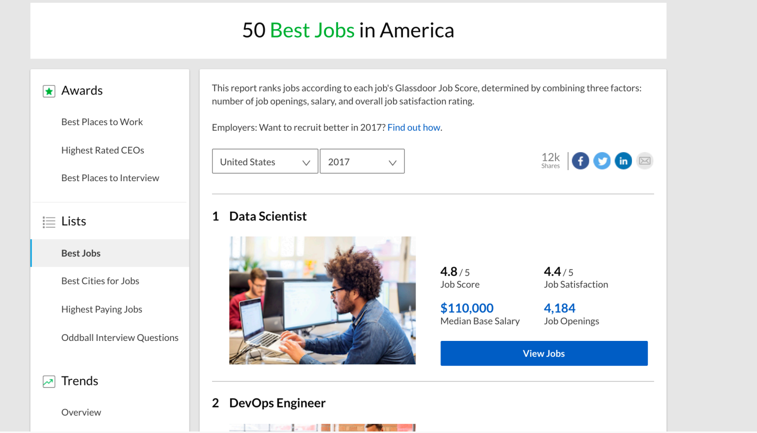
What is This Class?

The Data Science Process

What is Data Science

Why?

Jobs!!!



50 Best Jobs in America

Awards

Best Places to Work

Highest Rated CEOs

Best Places to Interview

Lists

Best Jobs

Best Cities for Jobs

Highest Paying Jobs

Oddball Interview Questions

Trends

Overview

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.

Employers: Want to recruit better in 2017? Find out how.

United States | 2017

12k Shares

1 Data Scientist

4.8 / 5 Job Score

4.4 / 5 Job Satisfaction

$110,000 Median Base Salary

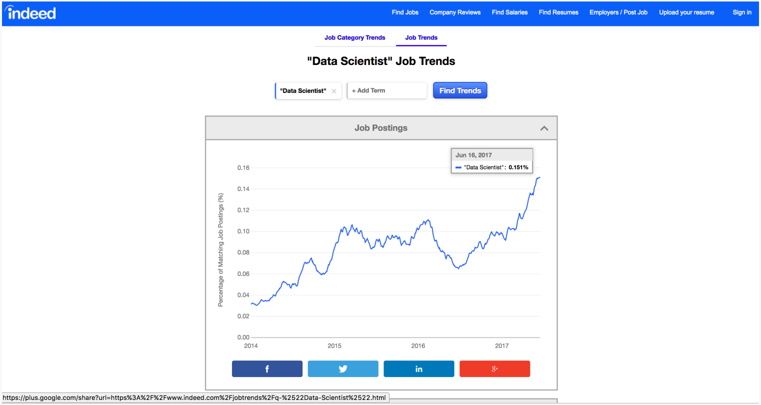
4,184 Job Openings

View Jobs

2 DevOps Engineer

Why?

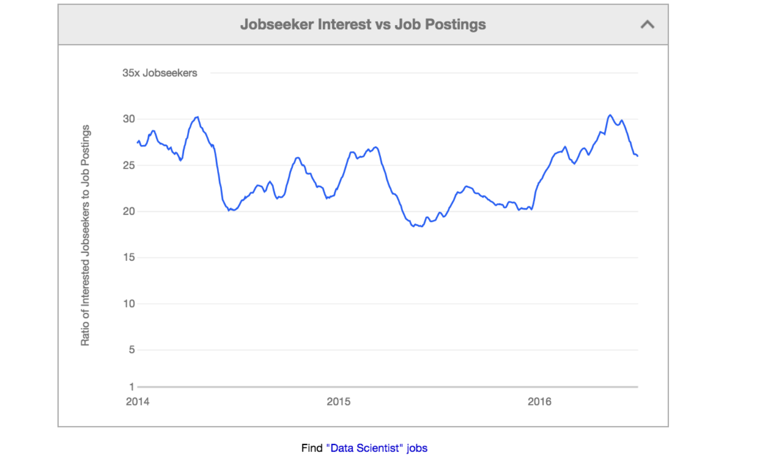
Jobs!!!



"Data Scientist" Job Trends

Why?

Jobs!!!



Jobseeker Interest vs Job Postings
35x Jobseekers
30
25
20
15
10
5
1
Ratio of Interested Jobseekers to Job Postings
2014
2015
2016

Find "Data Scientist" jobs

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?

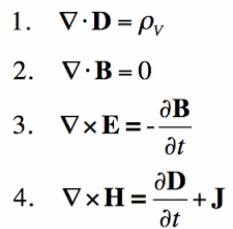
Long time ago (thousands of years) science was only empirical and people counted stars or crops.

How?

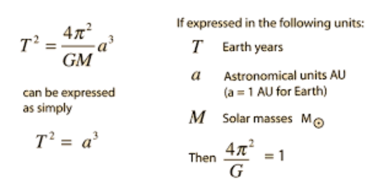
Long time ago (thousands of years) science was only empirical and people counted stars or crops and use the data to create *machines* to describe the phenomena

How?

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



1. $\nabla \cdot \mathbf{D} = \rho_V$
2. $\nabla \cdot \mathbf{B} = 0$
3. $\nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}$
4. $\nabla \times \mathbf{H} = \dfrac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$



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

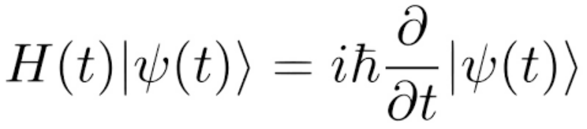
can be expressed as simply

$$T^2 = a^3$$

If expressed in the following units:

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

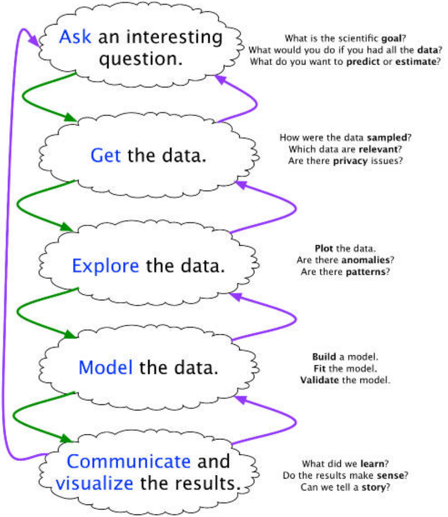
Then $\dfrac{4\pi^2}{G} = 1$



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

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 is This Class?

What

Four modules. The material of the course is divided into 4 modules. Each module (except module 0) 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 module 0 will 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, Presentations Stack and Large Data Management):

- Logistic Regression (linear and polynomial)
- Multiple Log-Regression
- Regularization
- Classification with decision trees
- Missing data and knn classification

What

Module 3 (Ensemble Methods, Natural Science data, Web Site building and report writing, large code skills):)

- Random Forrest
- Bagging
- Boosting
- Stacking
- Support Vector Maching

Who

Kevin

Who

Kevin Rader

Who

Kevin Rader

Senior preceptor in Statistics. Teaches CS 109A & Stat 139 this fall and Stat 102 and Stat 98 in the spring. Research interests include complex survey analysis and casual inference. Hobbies include the outdoors, sports (especially the aquatic variety), and of course, farming.

Who

Rahul Dave

Rahul Dave, lab guru and python guru, is a lecturer at the IACS. He teaches AM207 in the spring, and has taught labs for cs109 in both 2013 and 2015. He loves mountains and Bayesian Stats.

Who

Margo Levine

Margo is the Associate Director of Undergraduate Studies and a lecturer in Applied Mathematics. She has taught AM 21a, 21b, 50, 105, 108, 115, and 201, and she's excited to be working on a CS / Stats course this semester.

Who

Pavlos Protopapas

Who

Pavlos Protopapas

Teaches CS109 and the Capstone course for the Data Science masters program. Research in astrostatistics and excited about the new telescopes coming online in the next few years. He has absolutely no hobbies or interests except teaching CS109.

Teaching Fellows etc

(Head TF) Eleni Kaxiras: eleni@seas.harvard.edu

Teaching Fellows etc

Section Leaders
Nick Hoernle *
Patrick Ohiomoba *
Ryan Lee *
Matt Holman
Nathaniel Burbank
Zona Kostic
Albert Wu

Teaching Fellows etc

Lab Assistants
Ted Zhu
Chin Hui Chew
Xindi Zhao
Rohan Thavarajah
Chris Siviy
Russell Kunes

Lectures, Labs, Sections, Office hours

Lectures: Mondays and Wednesdays 1:00-2:30pm @ 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.
We will use a mix of notes and examples via notebooks

1. Lecture notes and associated notebooks will be posted before lecture on Canvas
2. Lectures will be video taped and posted approximately in 24 hours on Canvas

Lectures, Labs, Sections, Office hours

Labs: Thursdays 4:00-5:30pm and Fridays 10:00-11:30am at the red couch area outside the lecture hall.
Labs are meant to help you understand the lecture materials better via examples.

1. These two labs will be the same and therefore you need to only attend one of the two
2. Thursday lab will be video taped and posted approximately in 24 hours on Canvas

Lectures, Labs, Sections, Office hours

Sections: Lectures and labs are supplemented by 1 hour sections led by teaching fellows. There are two types of sections:

1. *Standard Sections* will be a mix of review of material and practice problems similar to the homework
2. *Advanced Sections (A-Sections)* will cover advanced topics like the mathematical underpinnings of the methods seen in lectures and labs.

NOTE: The material covered in the Advanced Sections is **required** for all AC 209A students. There will be one extra question in each homework for AC 209 students which will be based on the A-Section materials.

Lectures, Labs, Sections, Office hours

Office Hours:

Lectures, Labs, Sections, Office hours

Instructors Office Hours:

- **Margo:** Monday 2:30-4:00pm, IACS student lobby MD ground floor
- **Kevin:** Tuesday 1:00-3:00 pm, IACS student lobby MD ground floor
- **Pavlos:** Tuesday 3:00-5:00 pm, IACS student lobby MD ground floor
- **Rahul:** Wednesday 2:30-4:00pm, IACS student lobby MD ground floor

Lectures, Labs, Sections, Office hours

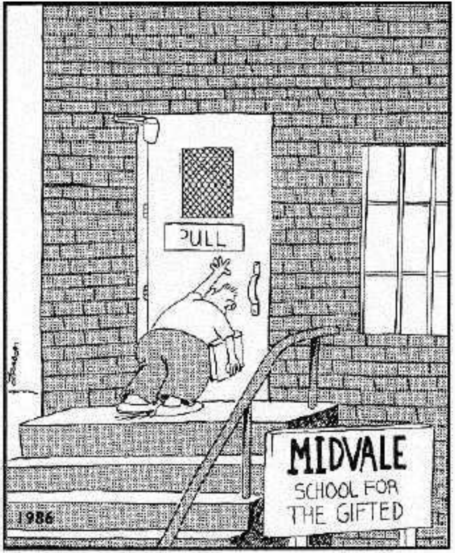
TF Office Hours:

Open Office Hours where TFs are present to help you. You do not need to sign up. Just show up.

Mondays and Thursdays 7:00pm-8:30pm room in the Red couch area in NW basement

Tuesdays 4:00-5:30pm in the Red couch area in NW basement.

AC 209 Students



AC 209 Students

Students enrolled for the AC 209A course have the following extra requirements:

1. Attend A-Sections
2. Complete an extra question in homework 2-8
3. Complete extra questions in midterm
4. Expand the scope of the final project beyond the methods studied in class

Homework(s)

There will be 8 homework (not including Homework 0)

1. Homework 0
2. Homework 1 (module 0)
3. Homework 2, 3, 4 (module 1)
4. Homework 5, 6, 7 (module 2)
5. Homework 8 (module 3)

Homework(s)

Homework(s)

You are encouraged but not required to submit in pairs.

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

All assignments will be posted on Wed. at 6pm and will be due on next week’s Wed. at 11.59pm.

Midterm

There will be one midterm (take-home) to be done individually which it counts for 30% of the final grade.

- Published Nov. 1 due on 9:00am Nov. 6
- 36 hours to complete it
- Extra questions for the AC209 students

Final Project

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

- We will provide 5-10 datasets which you could use for your final project
- We will also provide a project definition for each of the data set
- You can create your own project definition but must use one of the data sets provided (to be approved by the instructors)
- 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 different expectations for the AC209 students

More details to come early November

Help

Help

The process to get help is:

1. Post the question in Piazza and hopefully your peers will answer. We monitor the posts but we will respond no earlier than 24 hours from the posting time
2. Go to Office Hours, **this is the best way to get help**
3. For private matters send an email to the Helpline: cs109a2017@gmail.com. The Helpline is monitored by all the instructors and TFs
4. For personal matters send an email to Pavlos and/or Kevin

Grade

Grade

- Homework 40%
- Quizzes 10%
- Midterm 30%
- Final 20%

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?

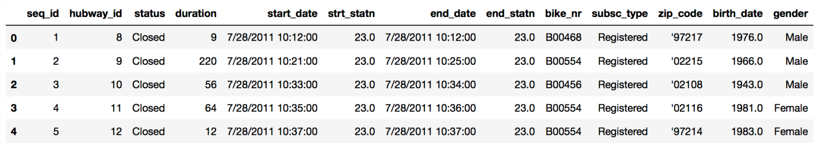
The Data Exploration/Question Refinement Cycle

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:

The Data Exploration/Question Refinement Cycle

- **Who?** Who’s using the bikes?

  Refine into specific hypotheses:

  - More men or more women?

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?

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:

The Data Exploration/Question Refinement Cycle

- **Where?** Where are bikes being checked out?

  Refine into specific hypotheses:

  - More in Boston than Cambridge?

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?

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:

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?

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?

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?

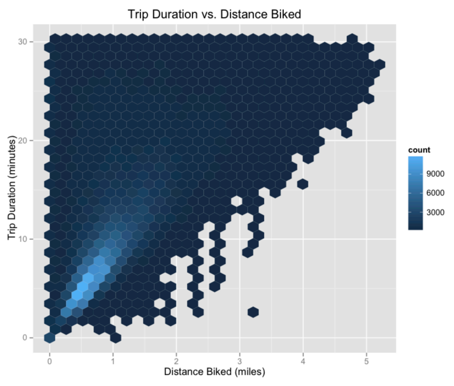
How questions are about modeling relationships between different variables.

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



Jupyter Notebooks