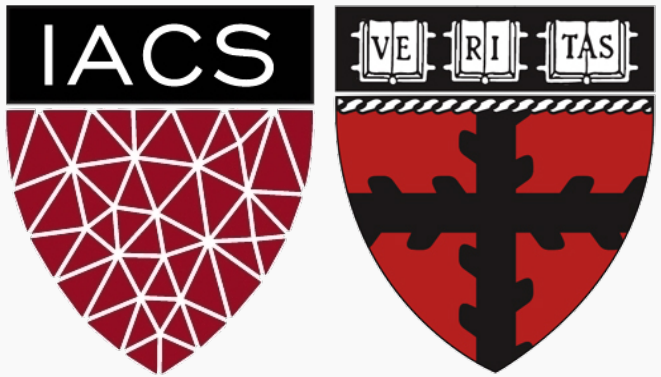
Lecture #1: Introduction to CS109A

aka STAT121A, AC209A, CSCIE-109A

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner



Lecture Outline

- Why data science? Why taking CS109A?
- What is data science?
- What is this class and what it is not?
- The data science process
- Example

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

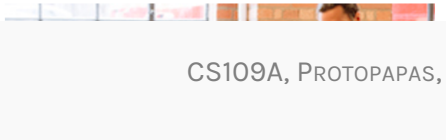
4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

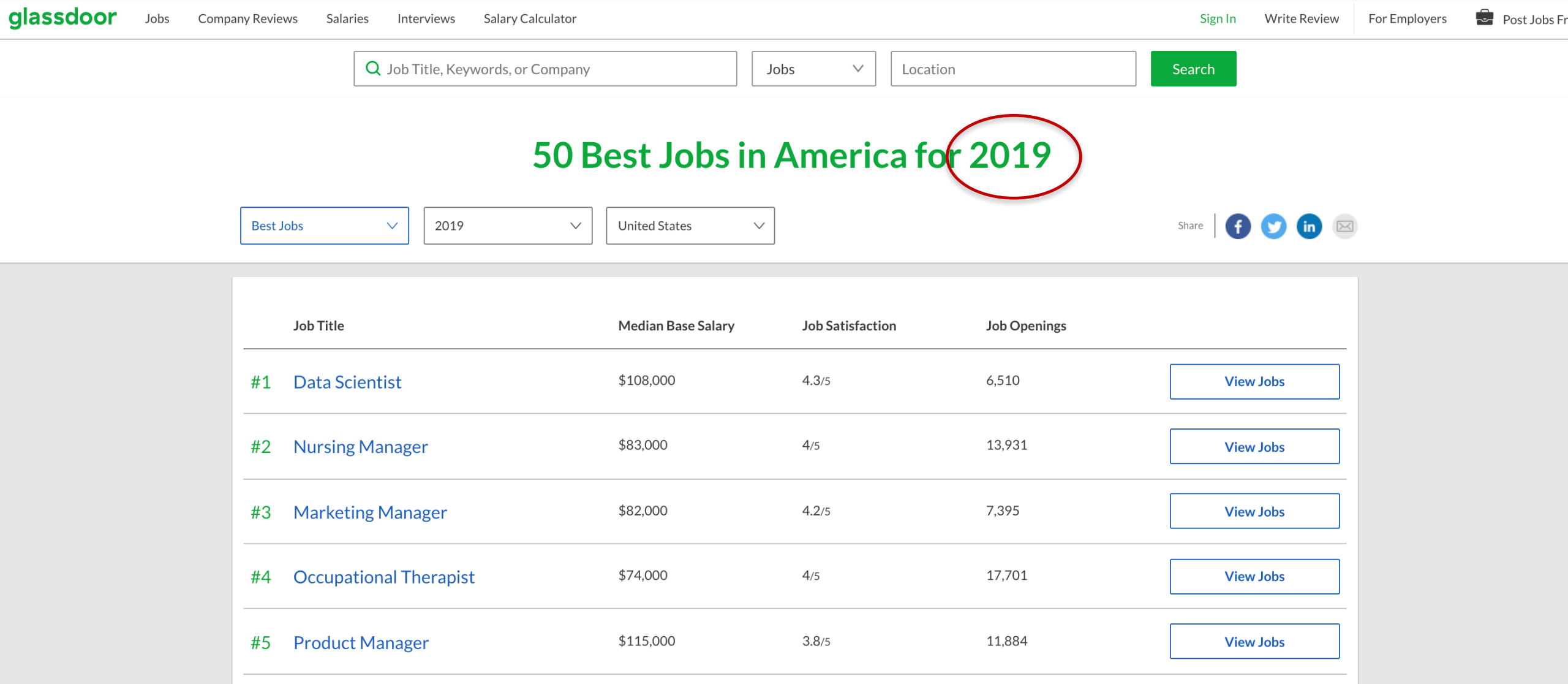
View Jobs

2 DevOps Engineer



Why?

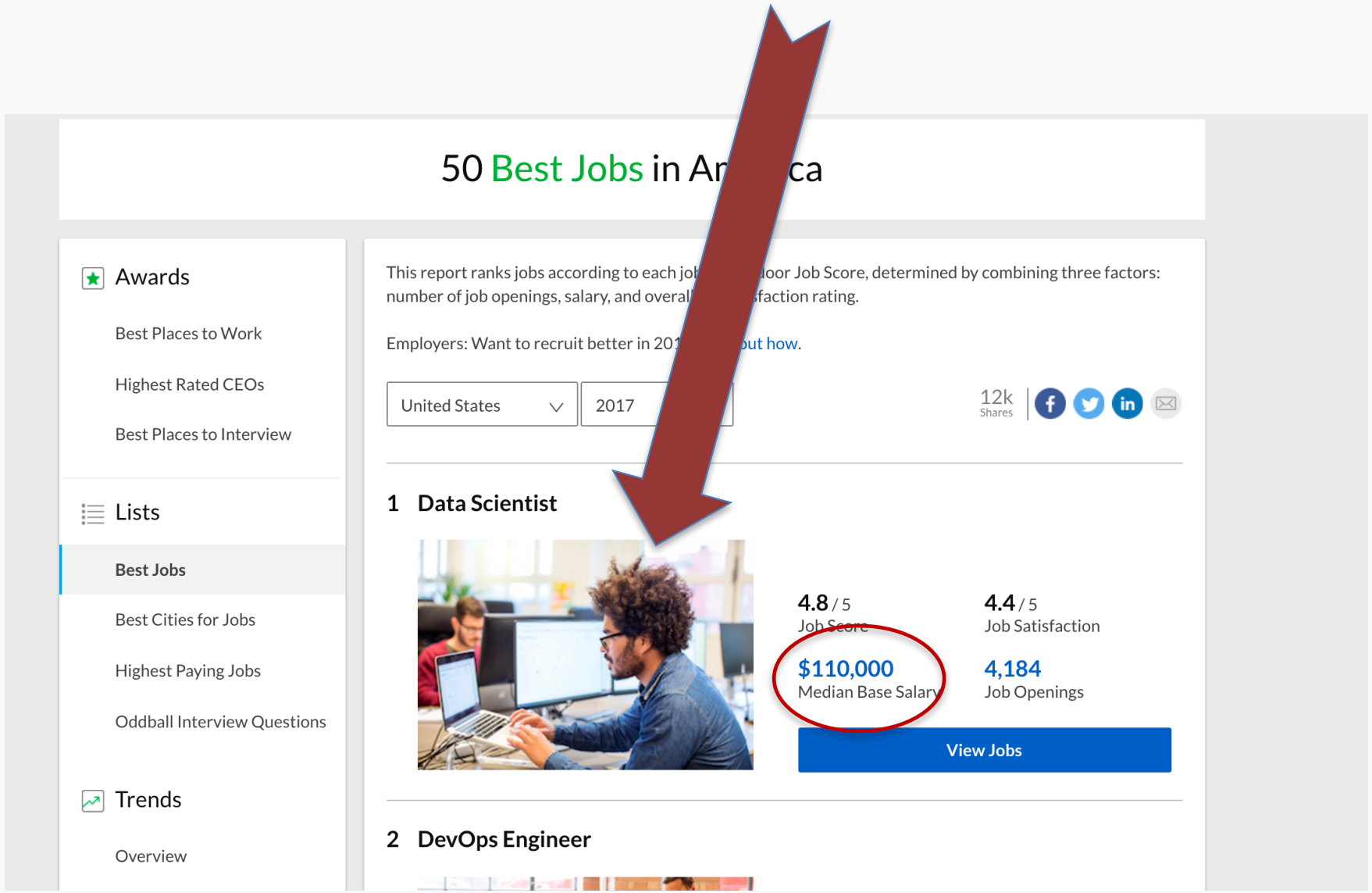
Jobs!



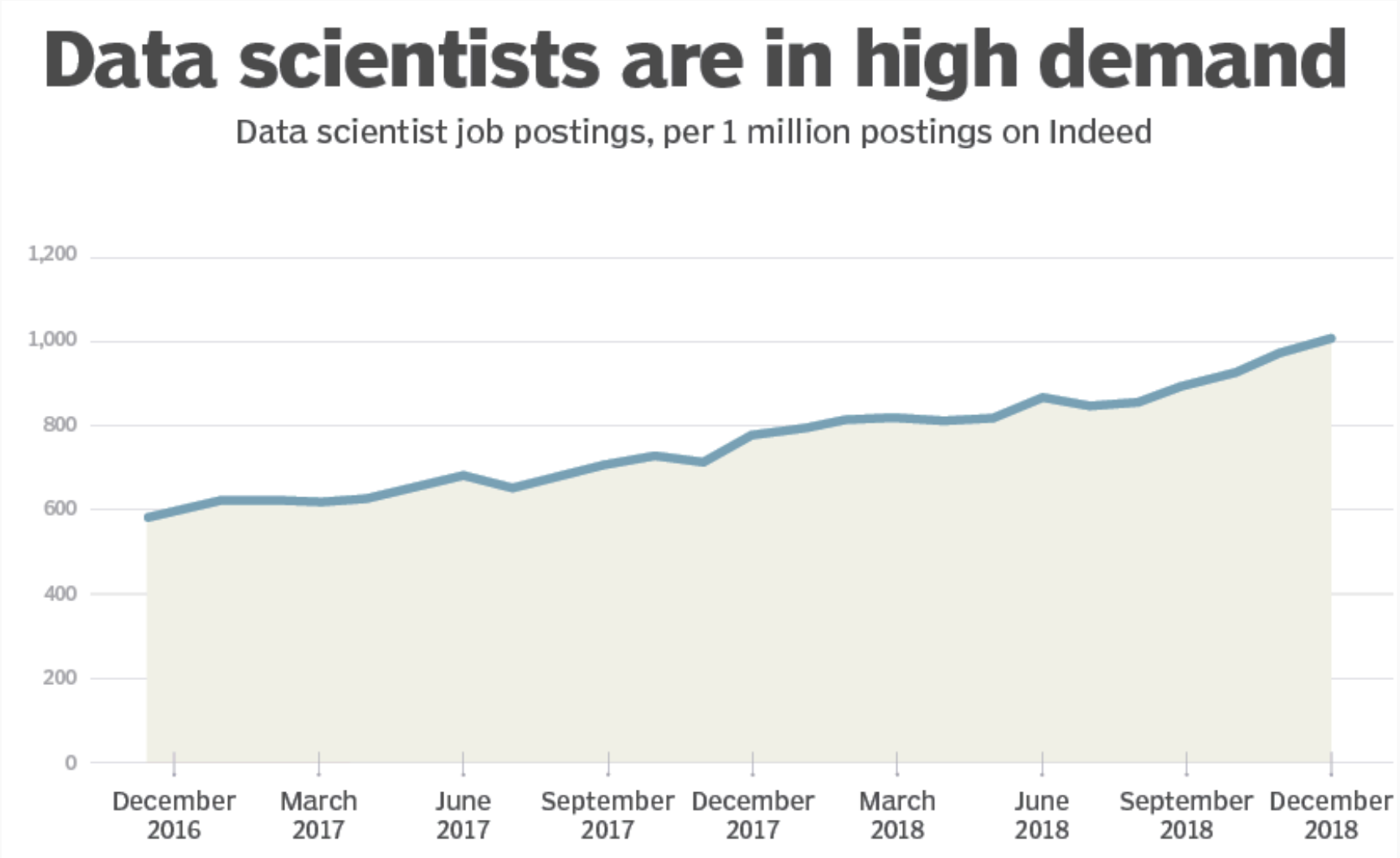
	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Data Scientist	$108,000	4.3/5	6,510	View Jobs
#2	Nursing Manager	$83,000	4/5	13,931	View Jobs
#3	Marketing Manager	$82,000	4.2/5	7,395	View Jobs
#4	Occupational Therapist	$74,000	4/5	17,701	View Jobs
#5	Product Manager	$115,000	3.8/5	11,884	View Jobs

Why?

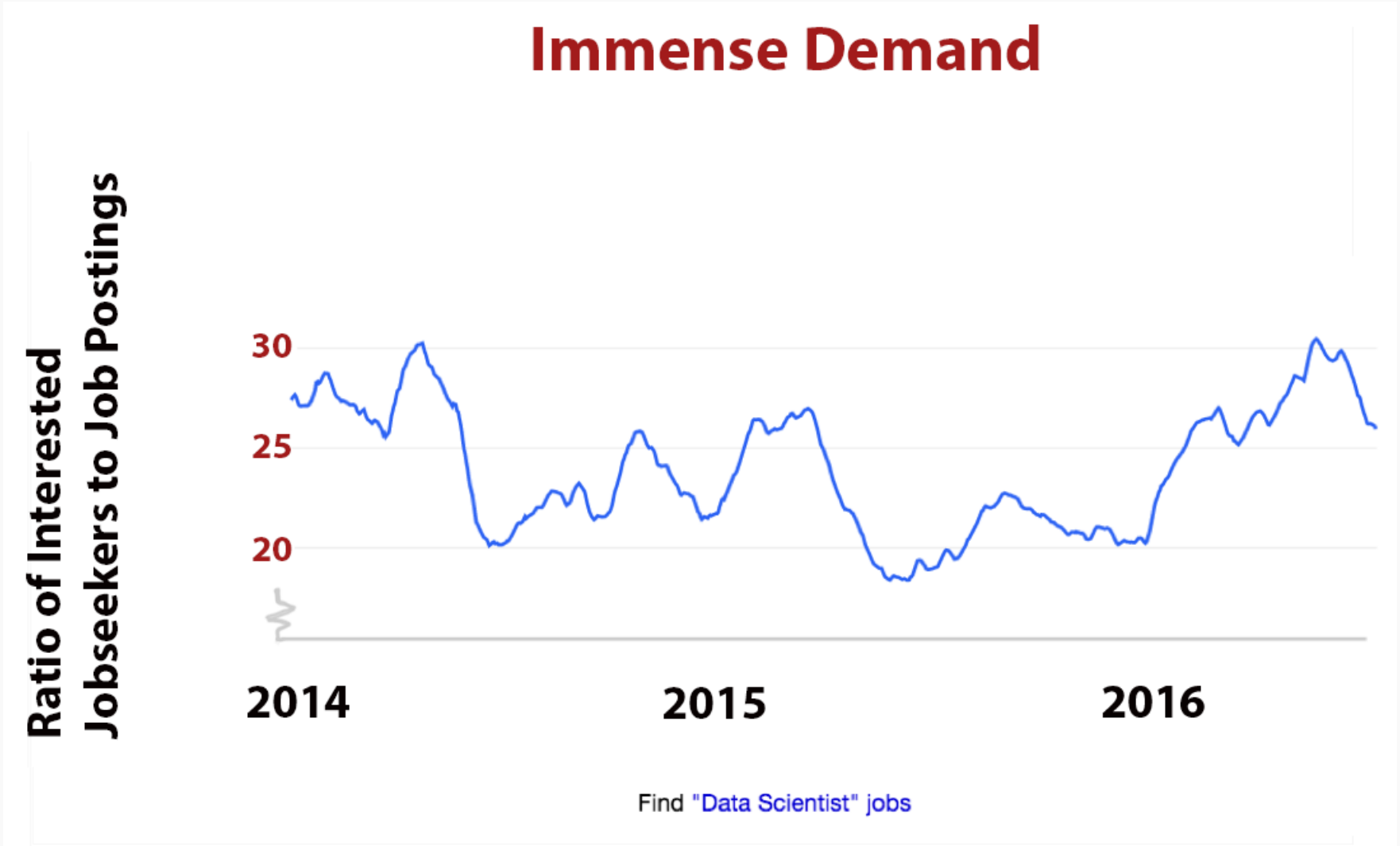
Jobs!


50 Best Jobs in Ar ca
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 doorJob Score, determined by combining three factors: number of job openings, salary, and overall sfaction rating.
Employers: Want to recruit better in 201 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?


Data scientists are in high demand
Data scientist job postings, per 1 million postings on Indeed
1,200
1,000
800
600
400
200
0
December 2016
March 2017
June 2017
September 2017
December 2017
March 2018
June 2018
September 2018
December 2018

Why?


Immense Demand
Ratio of Interested Jobseekers to Job Postings
30
25
20
2014
2015
2016
Find "Data Scientist" jobs

Why?

Why do I love data science?

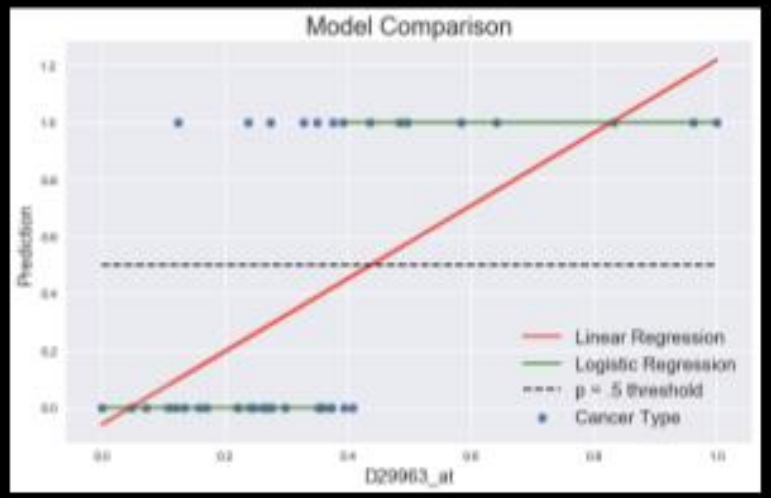
Why are you here?

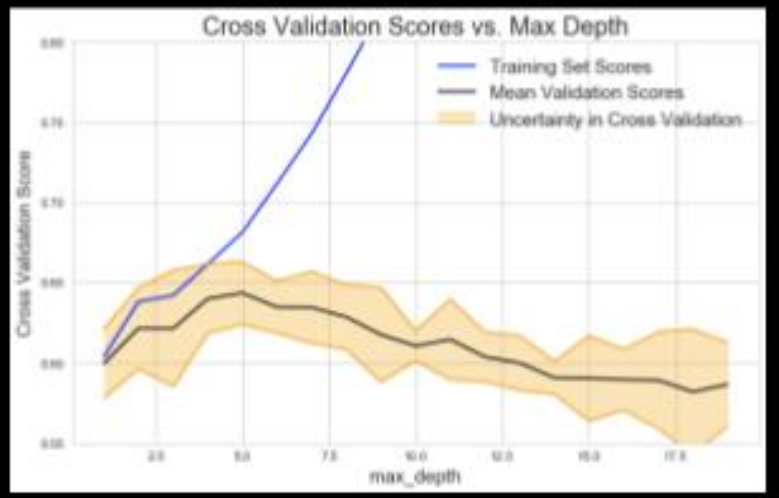
what my friends think I do

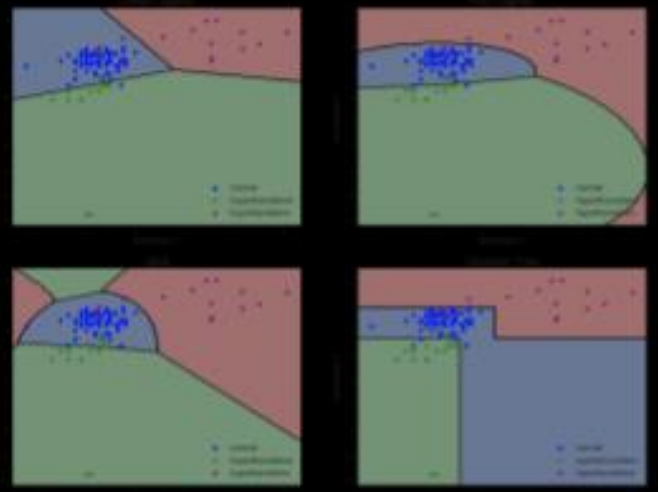
what my family thinks I do

what society thinks I do

what I actually (will) do in Data Science 1







Why?

Why are you here?

A little bit of history

History

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

History (cont)

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

History (cont)

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

History (cont)

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. $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
4. $\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$

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

can be expressed as simply

$$T^2 = a^3$$

If expressed in the following units:

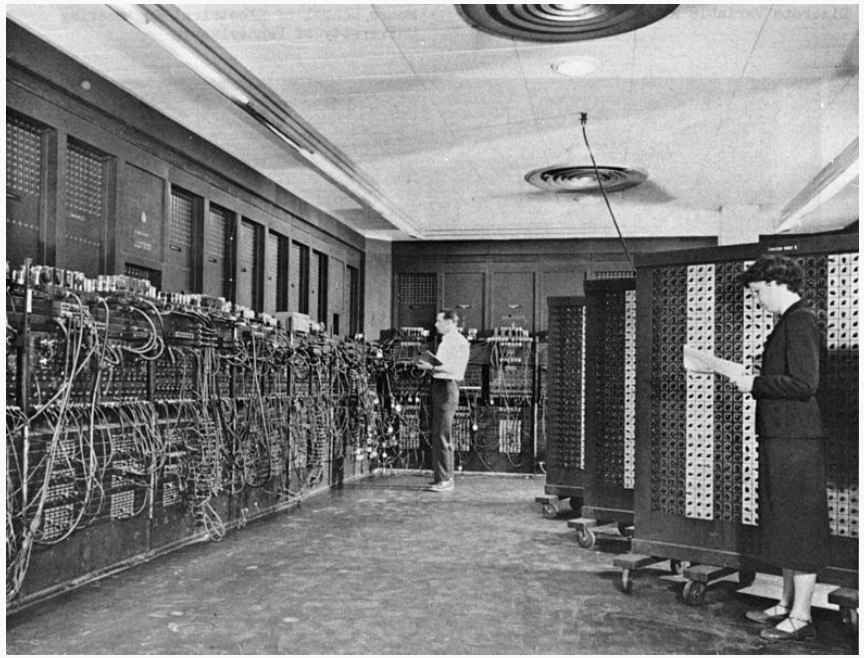
- T Earth years
- a Astronomical units AU (a = 1 AU for Earth)
- M Solar masses $M_\odot$

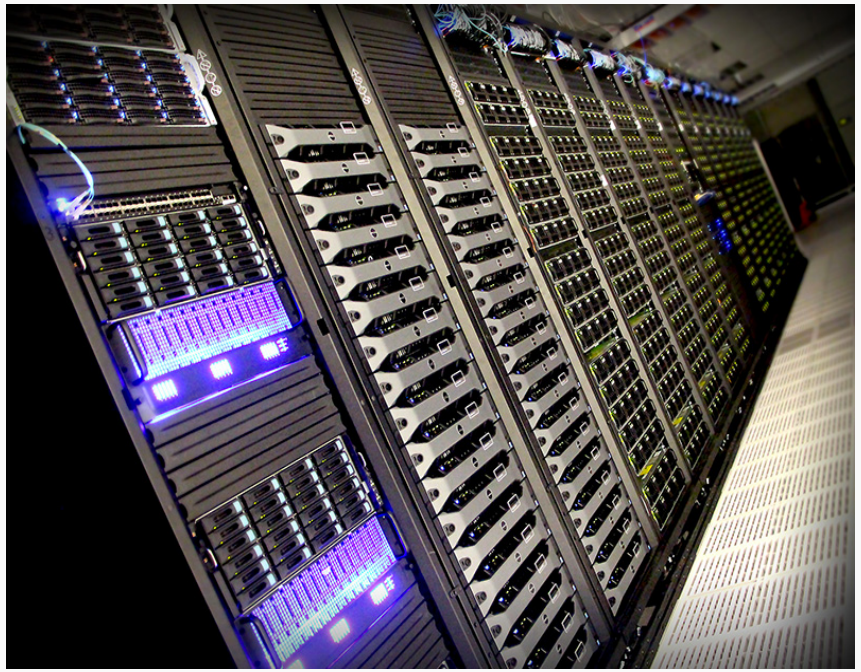
Then $\frac{4\pi^2}{G} = 1$

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

History (cont)

About a hundred years ago: computational approaches





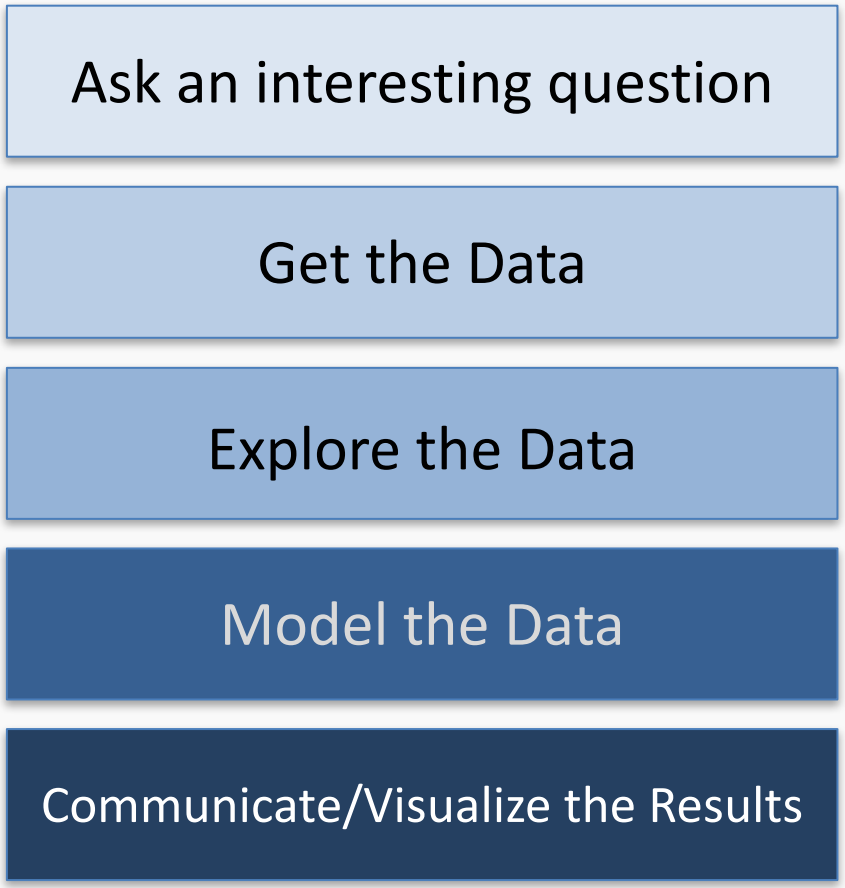
History (cont)

And then …. data science

What is data science?

What?

The Data Science Process



What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

What is the scientific goal?

What would you do if you had **all** of the data?

What do you want to predict or estimate?

What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

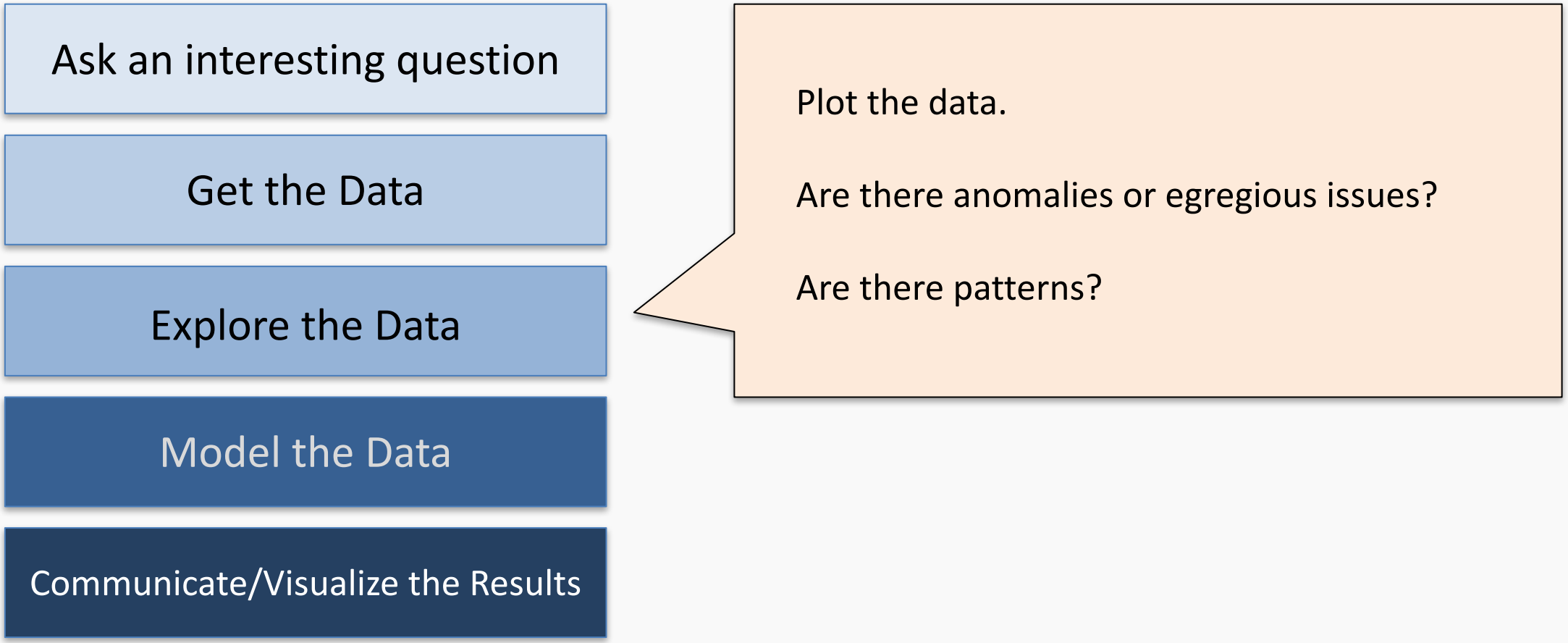
How were the data sampled?

Which data are relevant?

Are there privacy issues?

What?

The Data Science Process



What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

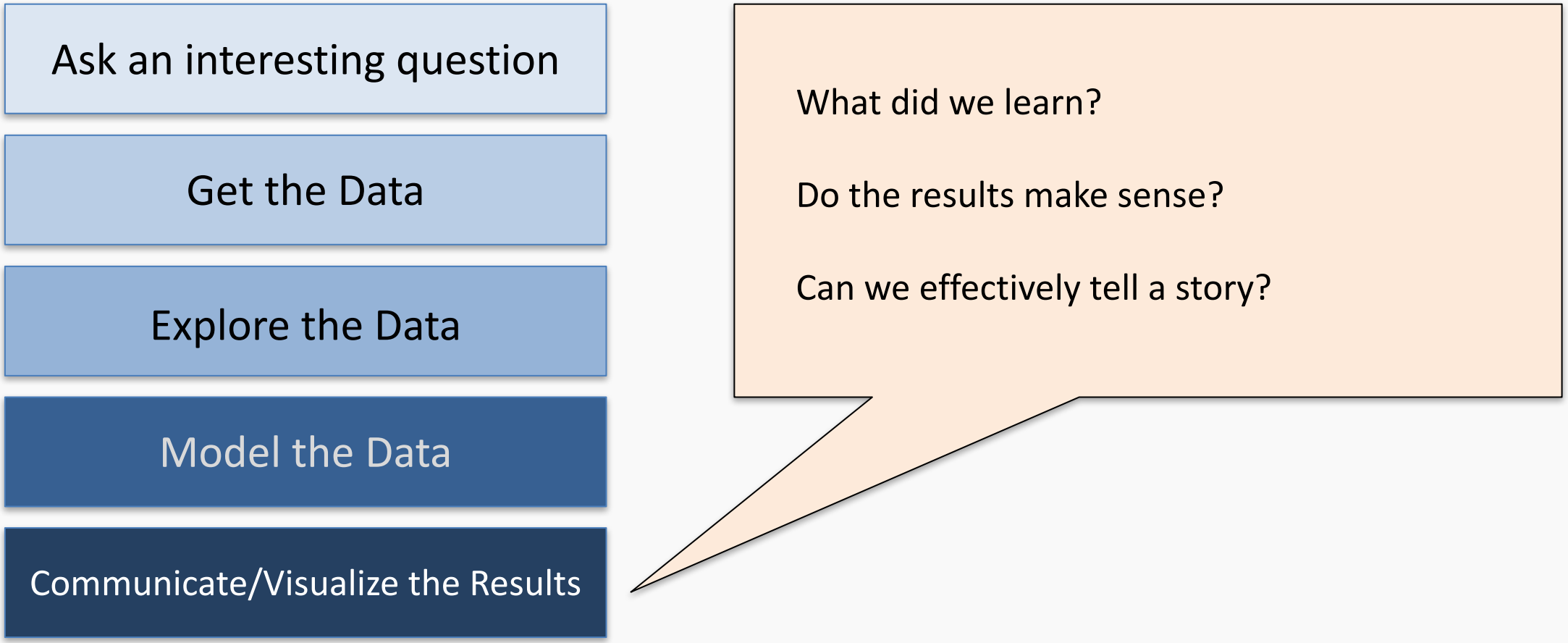
Build a model.

Fit the model.

Validate the model.

What?

The Data Science Process



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?

Week 1:

Getting ready with python, jupyter notebooks, environments and numpy.

What?

Week 2:

Basic statistics, visualization, pandas and data scraping

What?

Week 3 and 4:

Regression, and sklearn using transportation data:

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

What?

Week 5:

Exploratory Data Analysis, matplotlib and seaborn:

- Basic concepts of EDA
- Basic concepts of Visualization and Communications

What?

Week 6-7:

Classification, data imputations on Health Data:

- Logistic Regression (linear and polynomial)
- Multiple Logistic Regression
- Missing data and knn classification

What?

Week 8:

EthiCS

PCA and high dimensionality

What?

Week 9 and 10:

Decisions trees and ensemble methods :

- Simple Decision Trees for classification and Regression
- Bagging
- Random Forest
- Boosting
- Stacking

What?

Week 10-12:

Neural Networks:

- Perceptron, Back Propagation and SGD
- MLP and design choices
- Advanced MLP, regularization, dropout, batch normalization
- Neural Network solvers

What?

Week 12:

More visualization and model interpretation

What?

Week 13:

Experimental Design:

- AB testing
- Causal inference
- Randomization testing
- Adaptive and multi-arm bandit designs

CS109B

A. Neural Networks:

- CNNs
- RNNs
- Generative models

B. Unsupervised Clustering

C. Piecewise Linear Regression

D. Bayesian Modeling

CS109C – Advanced Practical Data Science

A. Productions Data Science, from notebooks to the cloud

B. Big models, transfer learning and architecture learning

C. Visualization tools for interpreting models

D. Sequential data, seq2seq with attention, transformers, NLP and time series modeling

Who?

Pavlos Protopapas

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

Teaches CS109(a/b/c) and the data science capstone course.

Research in astrostatistics: machine learning, statistical learning, big data for astronomical problems. 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? Instructor

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 causal inference.
Hobbies include the outdoors, sports (especially the aquatic variety), and of course, farming.

Who? Instructor

Chris Tanner

Lecturer at IACS, teaching CS109A and AC297R (capstone) now, and CS109B in the Spring. Research interests are within Natural Language Processing and Deep Learning. Hobbies include hiking and camping, designing/sewing hiking bags, and photography.

Who? Lab instructors

Eleni Kaxiras

Eleni is the assist. Director for Data Science and Computation at SEAS. She has been this course's Head TF for the last 3 years and she is now a lab instructor. She is currently a doctoral student. She is interested in the application of deep learning in analyzing biological signals. She owns olive trees in the island of Crete.

Who? Head TFs

Chris Gumb

Chris is currently working towards a graduate degree in Data Science from Harvard Extension School with a particular focus on NLP. His other interests and hobbies include: music theory & jazz improvisation; and film history.

Sol Girouard

She has been a head TF for 109B and she is a Quant, Math-Econ and Data Scientist who channels her applied interdisciplinary background in the intersection of financial markets and technology. Tae kwon full contact second degree black belt.

Who? Teaching Fellows

Advanced Section (the 209 part):

Cedric Flamant

Section leaders:

Marios Mattheakis

Robbert Struyven

Abhimanyu (Abhi) Vasishth

Who? Teaching Fellows

Rashmi Banthia
Evan Mackay
Brandon Walker
Rachel Moon
Nicholas Stern
Pat Sukhum
Zheyu Wu

Yun Bin (Matteo)Zhang
Marcus Heijer
Nathan Hollenberg
Maddy Nakada
Tim Pugh
Alex Yu
JavierMachin

Lectures, Labs, Advanced Sections, Sections and Office Hours

During lecture will cover the material which you will need to complete the homework, and to survive the rest of your life in CS109A. *Attending* lectures is required - quizzes during and at the end of each lecture (drop 50% 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 GitHub.
2. Lectures will be video taped (and live streamed for DCE students) and posted approximately within 24 hours on web page.

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

Lectures, **Labs**, Advanced Sections, Sections and Office Hours

Labs are meant to help you better understand the lecture materials via examples.

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

Thursdays 4:30-5:45 pm @Pierce 301.

Lectures, Labs, **Advanced Sections, Sections** and Office Hours

Lectures and labs are supplemented by 1.5 hour sections led by teaching fellows. There are two types of sections:

- *Standard Sections* will be a mix of review of material and practice problems similar to the homework

  Friday 0:30-11:45 am at 1 Story St. Room 306 and Mon 4:30-5:45 pm in Science Center 110

- *Advanced Sections* ***(A-Sections)*** will cover advanced topics like the mathematical underpinnings of the methods seen in lectures and labs.

  Weds 4-5:15 pm at 1 Story St. Room 306

Lectures, Labs, **Advanced Sections**, Sections and Office Hours

Topics

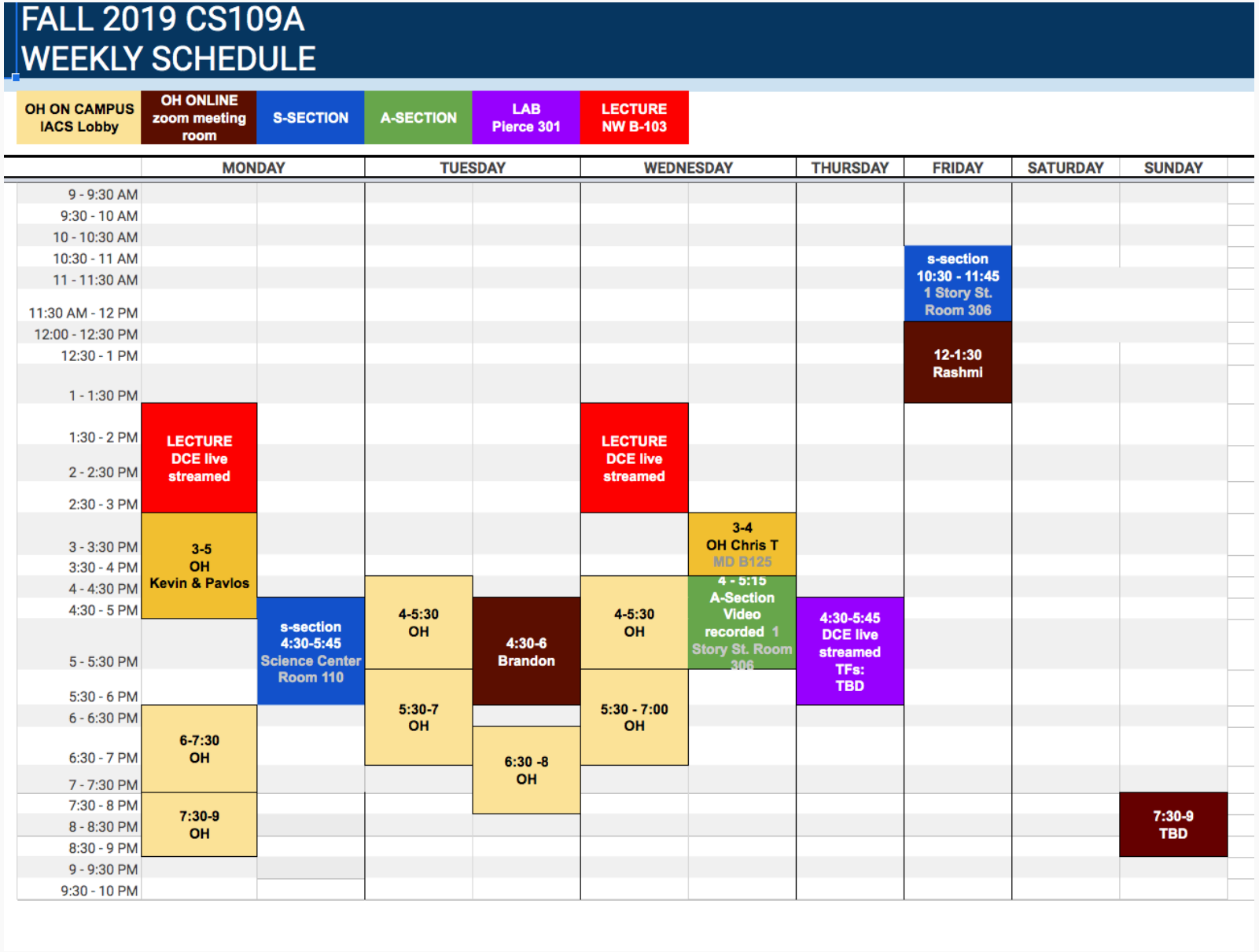
1. Linear Algebra and Hypothesis Testing: The Short Versions
2. Methods of regularization and their justifications
3. Generalized Linear Models
4. Mathematics of PCA
5. Decision trees and Ensemble method;
6. Stochastic Gradient Descent

NOTE 1: The material covered in the Advanced Sections is required for all AC 209A students. There will be one extra question in most homework for AC 209 students which will be based on the A-Section materials.

NOTE 2: No additional quizzes for A-section.

NOTE 3: A-sections and Friday's regular section will be live streamed to everyone.

Lectures, Labs, Advanced Sections, Sections and **Office Hours**



Homework(s)

There will be 8 homework (not including Homework 0):

- Homework 0 (due Sept 11)
- Homework 1: Web scraping, Beautiful Soup
- Homework 2: Regression kNN and LinReg
- Homework 3: Multi-regression, polynomial reg and model selection
- Homework 4*: Log Reg and more
- Homework 5: PCA and ethics
- Homework 6: Random Forest, Boosting and Neural Networks
- Homework 7*: Neural Networks
- Homework 8: Experimental Design

Homework(s)

You are encouraged but not required to submit in pairs, except homework 4 and homework 7, which must work 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 Wednesday 3:00pm.

Final Project

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

- We will provide 7 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)

Help

Help

The process to get help is:

1. Post the question in Ed and hopefully your peers will answer. We monitor the posts and we will respond within 8 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: **cs109a2019@gmail.com**. The Helpline is monitored by all the instructors and TFs.
4. For personal matters send an email to Pavlos, Kevin and Chris.

Sundays will be slow days, so please be patient!

Grades

Grades

- Homework 0: 1%
- Paired Homework (six): 39%
- Individual Homework (two): 17%
- Quizzes: 10%
- Project: 30%
- Participation: 3%
- **Total: 100%**

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

CS109a: Introduction to Data Science

Fall 2019

Pavlos Protopapas, Kevin A. Rader, and Chris Tanner

Lab Leaders: Chris Tanner and Eleni Kaxiras

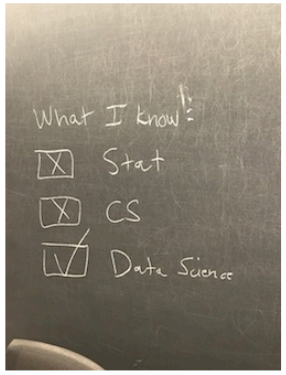
Welcome to CS109a/STAT121a/AC209a, also offered by the DCE as CSCI E-109a, Introduction to Data Science. This course is the first half of a one-year course to data science. We will focus on the analysis of data to perform predictions using statistical and machine learning methods. Topics include data scraping, data management, data visualization, regression and classification methods, and deep neural networks (a detailed schedule will be made available soon). You will get ample practice through weekly homework assignments. The class material integrates 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

Only one of CS 109a, AC 209a, or Stat 121a can be taken for credit. Students who have previously taken CS 109, AC 209, or Stat 121 cannot take CS 109a, AC 209a, or Stat 121a for credit.

Announcement: HW0 is now available.
Lectures: Mon and **Wed** 1:30-2:45 pm in Harvard Northwest Building, NW B-103
Labs: **Thur** 4:30-5:45 pm in Pierce 301
Head TFs: Chris Gumb **- DCE Head TF**: Sol Girouard
Office Hours: IACS student lobby in Maxwell-Dworkin's ground. Just follow the signs.
Online Office Hours zoom link: https://harvard-dce.zoom.us/j/7607382317

Course material can be viewed in the public GitHub repository.



STANDARD SECTIONS
Friday 9/13 10:30–11:45 am 1 Story St. Room 306
Monday 9/16 4:30–5:45 pm Science Center 110
Cover the material presented in class. Both standard sections are identical.

ADVANCED SECTIONS
Wednesday 9/18 4:30–5:45 pm 1 Story St. Room 306
Cover a different topic each week and are required for 209a students.

Instructor Office Hours
Pavlos & Kevin: Monday 3–5 pm, IACS Lobby
Chris: Wednesday 3–4 pm, Maxwell–Dworkin B125

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