Lecture Outline

What is Data Science

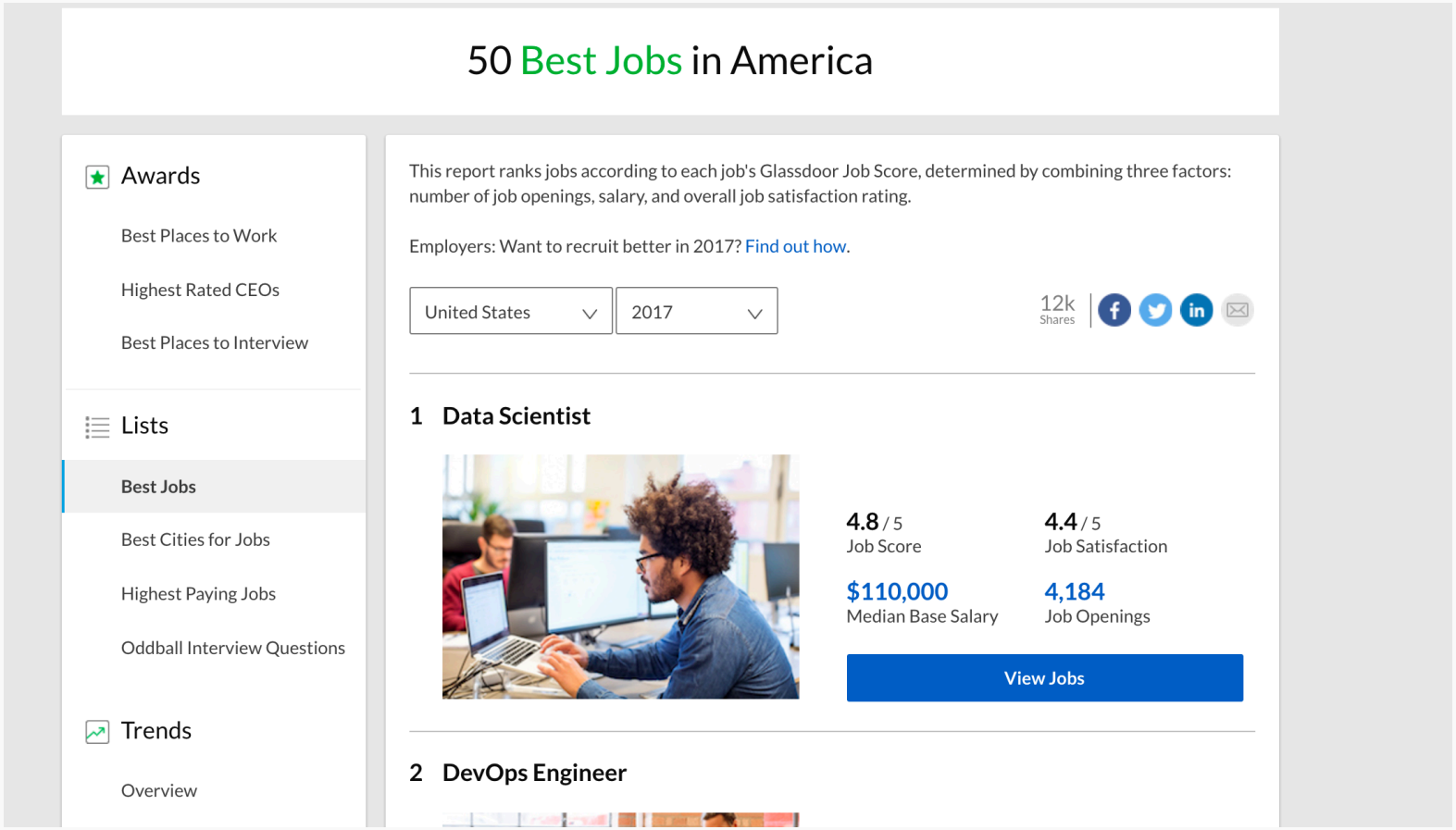
What is This Class?

The Data Science Process

What is Data Science

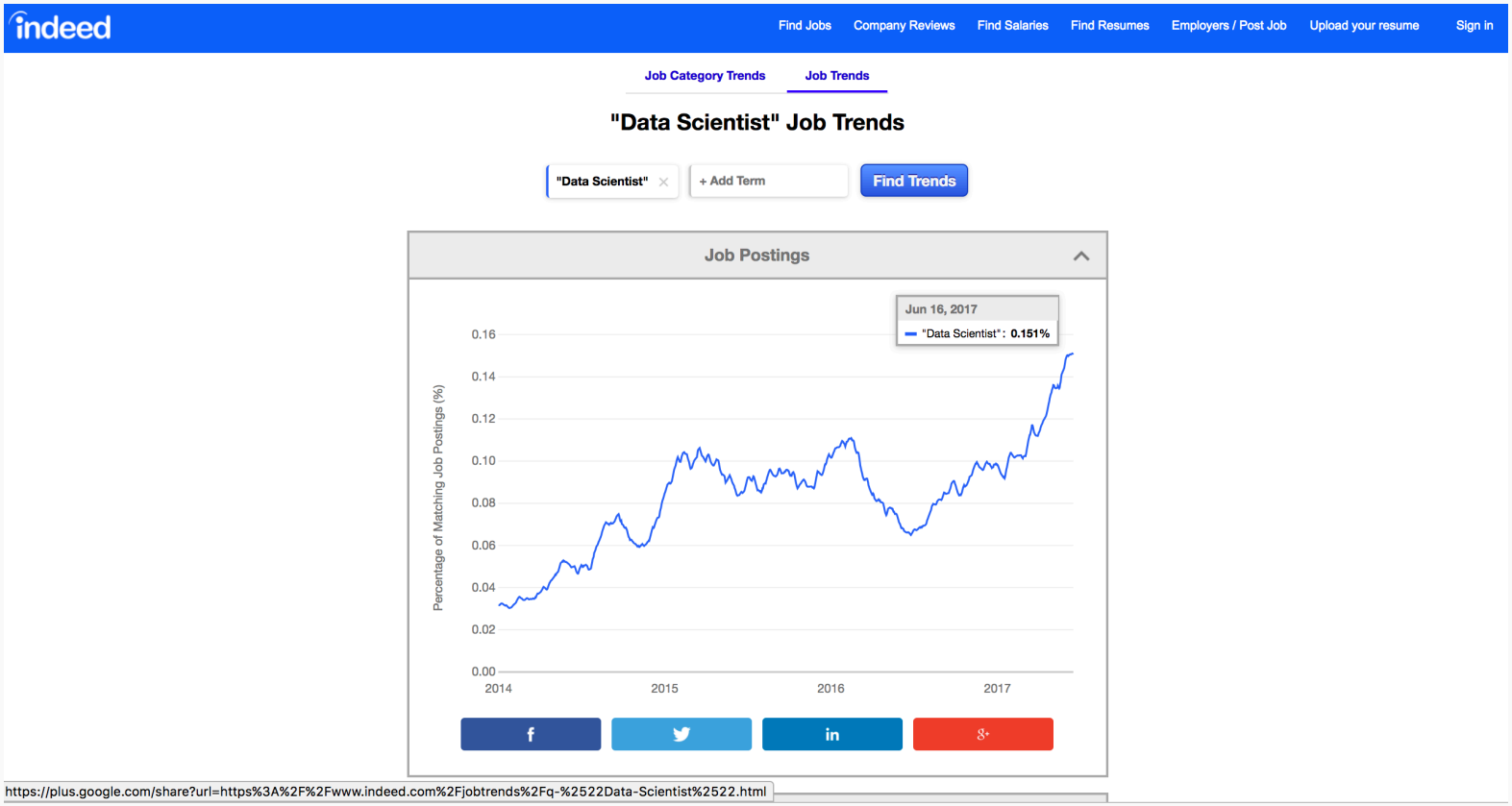
Why?

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

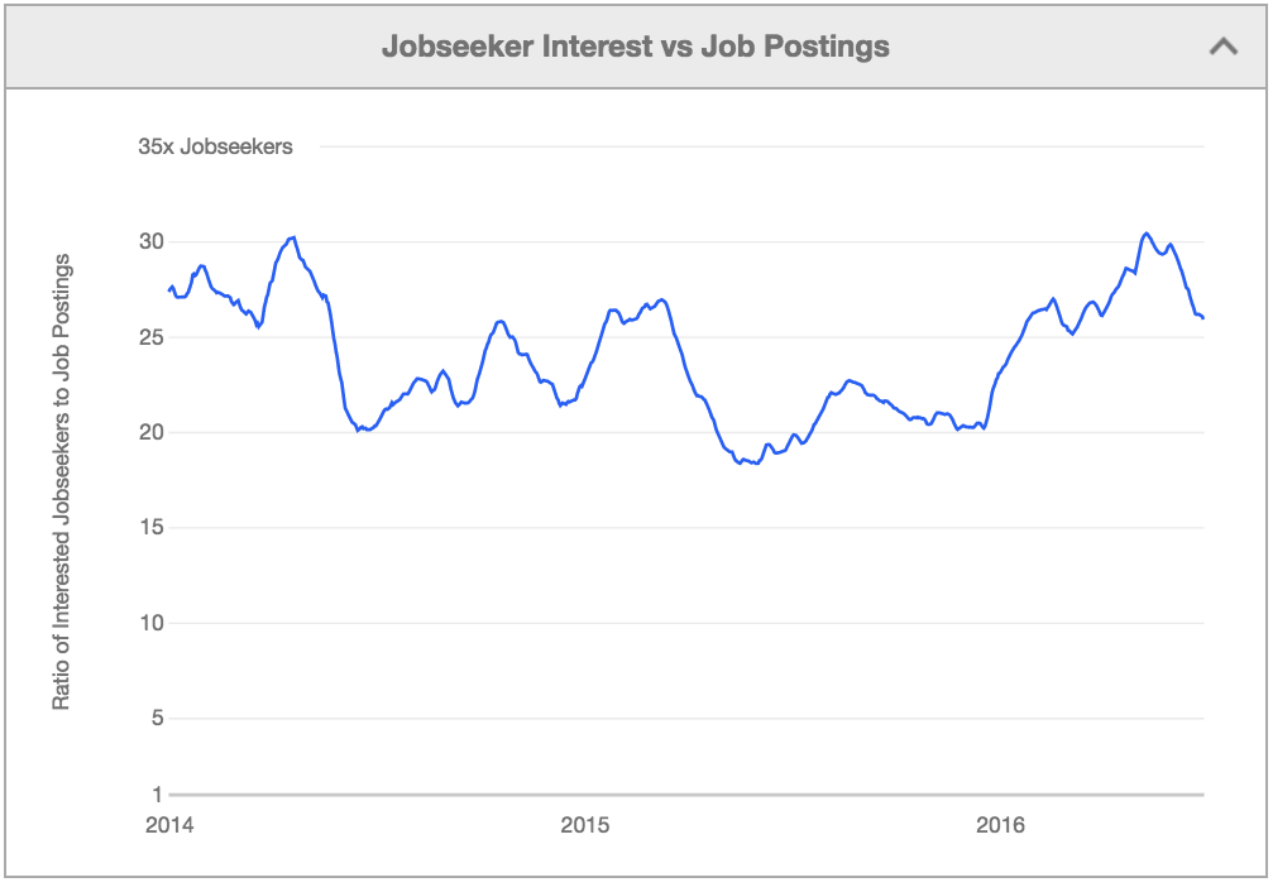
Why?

Jobs!!!


indeed
Find Jobs
Company Reviews
Find Salaries
Find Resumes
Employers / Post Job
Upload your resume
Sign in
Job Category Trends
Job Trends
"Data Scientist" Job Trends
"Data Scientist"
+ Add Term
Find Trends
Job Postings
Jun 16, 2017
"Data Scientist": 0.151%
Percentage of Matching Job Postings (%)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2014
2015
2016
2017
https://plus.google.com/share?url=https%3A%2F%2Fwww.indeed.com%2Fjobtrends%2Fq-%2522Data-Scientist%2522.html

Why?

Jobs!!!


Jobseeker Interest vs Job Postings
35x Jobseekers
Ratio of Interested Jobseekers to Job Postings
30
25
20
15
10
5
1
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. $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
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

$$T^2 = a^3$$

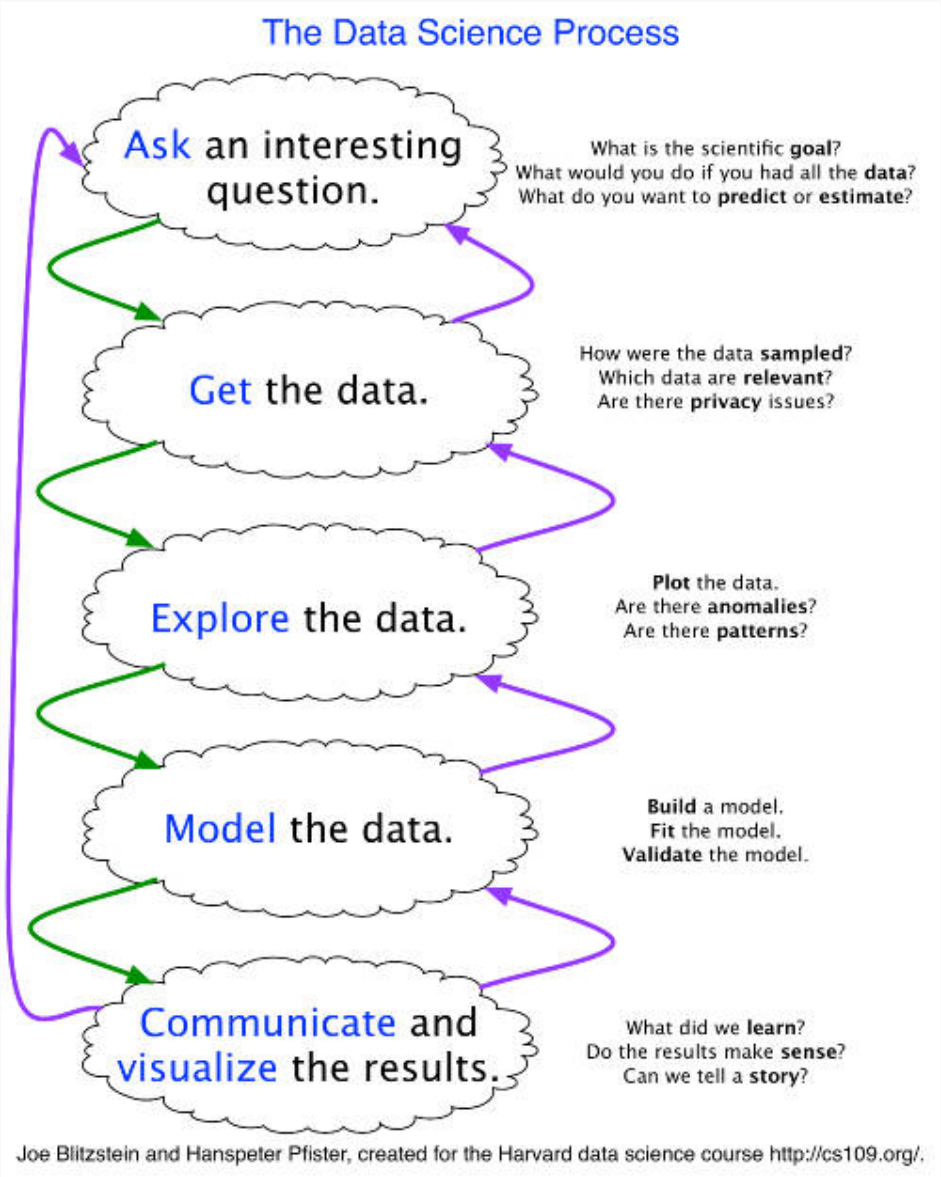
If expressed in the following units:

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

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

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

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

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?

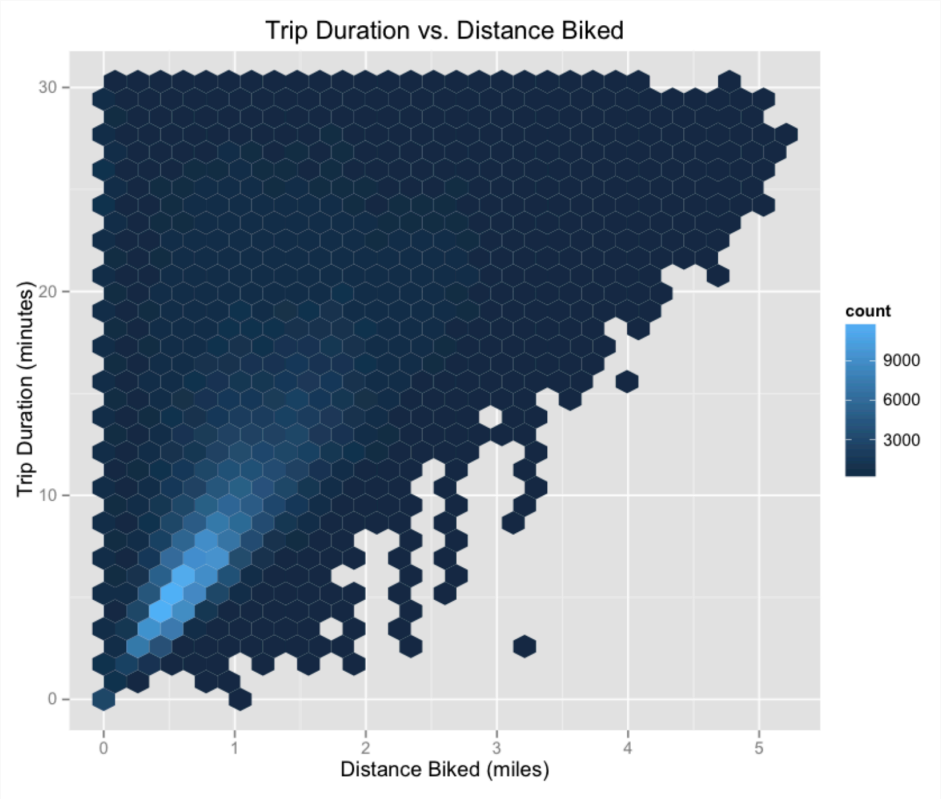
How questions are about modeling relationships between different variables.

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