Lecture #1: Introduction to Data Science

Asso. Prof. Huynh Trung Hieu

Sources: Introduction to Data Science

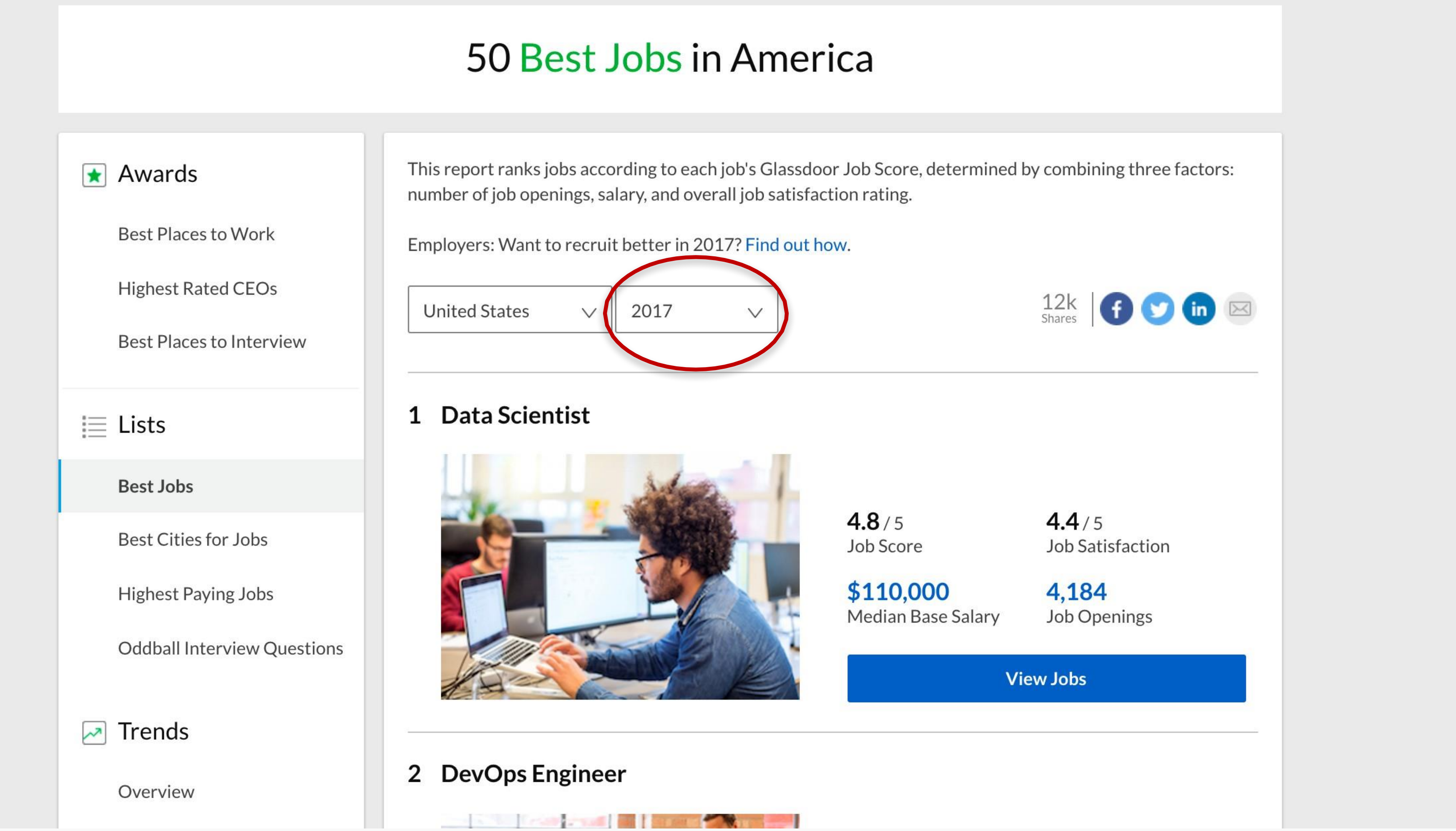
By Pavlos Protopapas, Kevin Rader and Chris Tanner

Lecture Outline

- Why data science? Why taking this course?
- 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

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]

1 Data Scientist

4.8 / 5 Job Score

4.4 / 5 Job Satisfaction

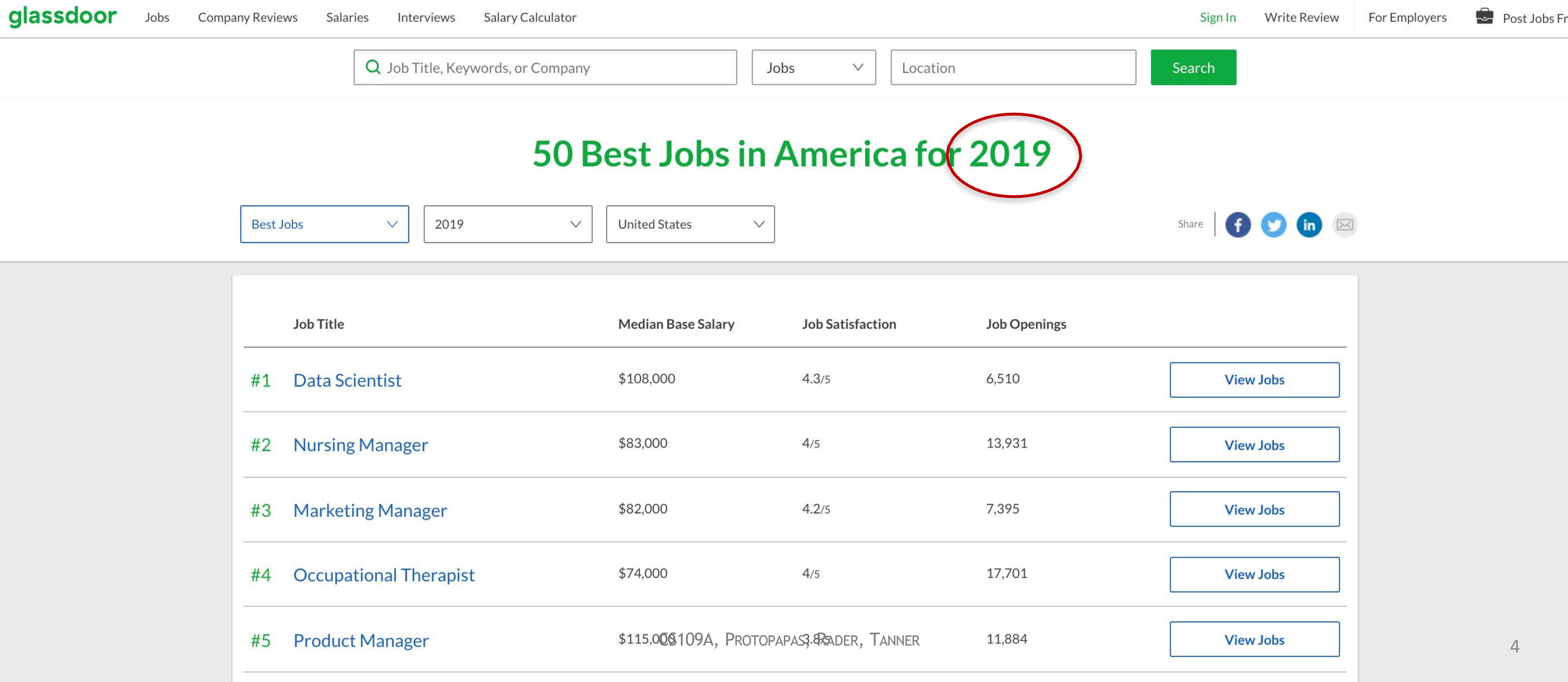
$110,000 Median Base Salary

4,184 Job Openings

2 DevOps Engineer

Why?

Jobs!



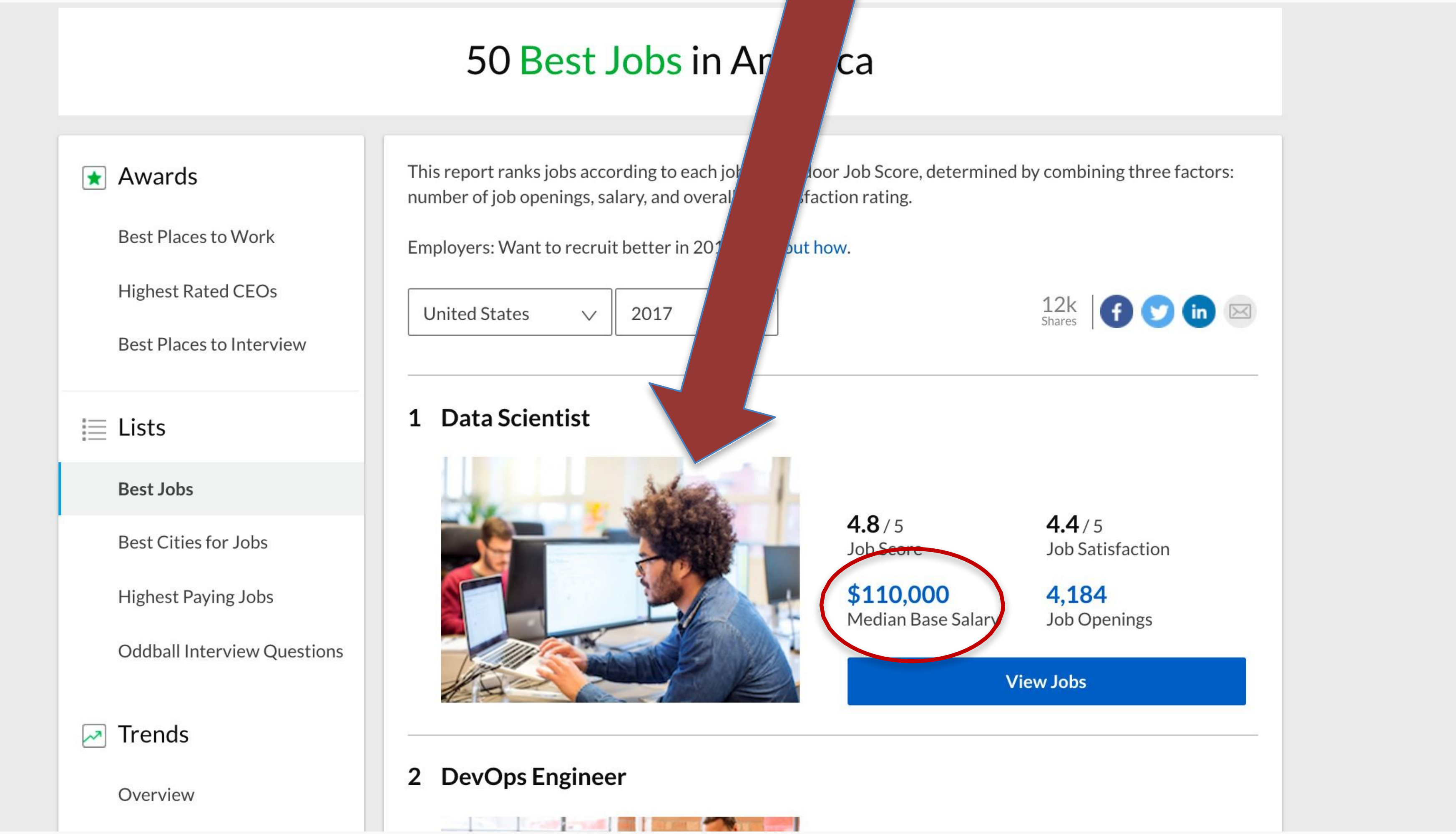
50 Best Jobs in America for 2019

Best Jobs | 2019 | United States

	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

This report ranks jobs according to each job door Job Score, determined by combining three factors: number of job openings, salary, and overal sfaction rating.

Employers: Want to recruit better in 201 out how.

1 Data Scientist

4.8 / 5 Job Score

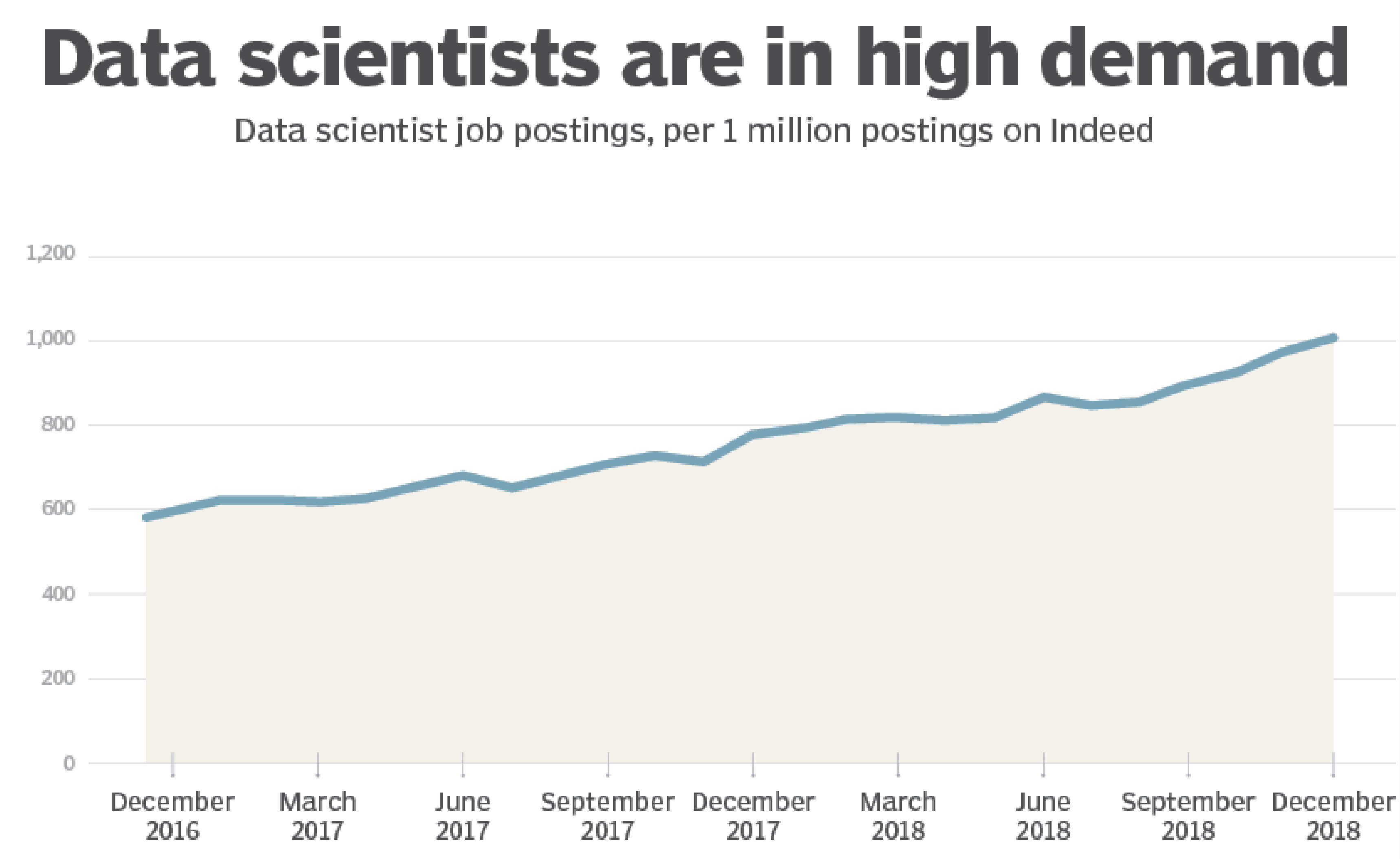
4.4 / 5 Job Satisfaction

$110,000 Median Base Salary

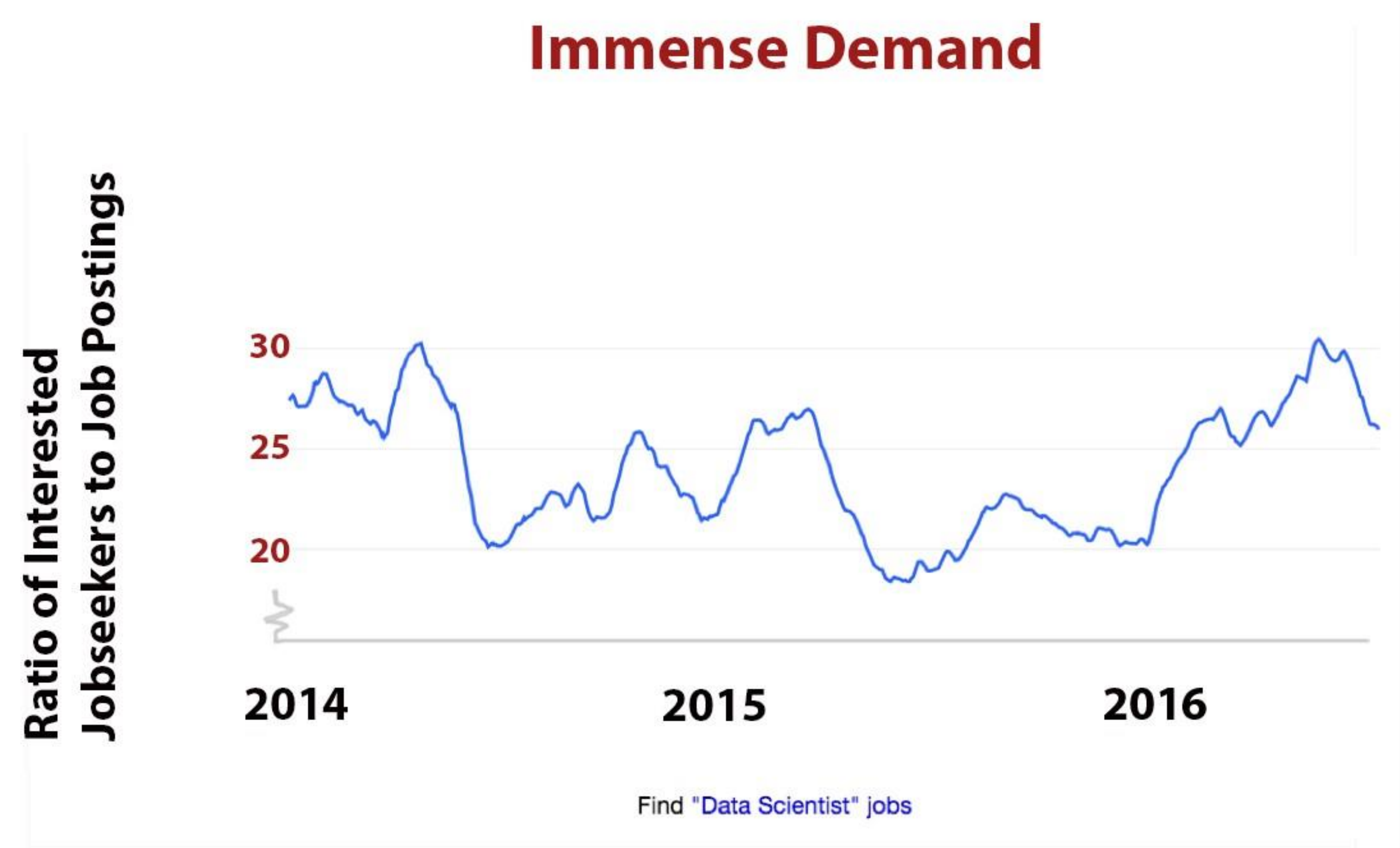
4,184 Job Openings

2 DevOps Engineer

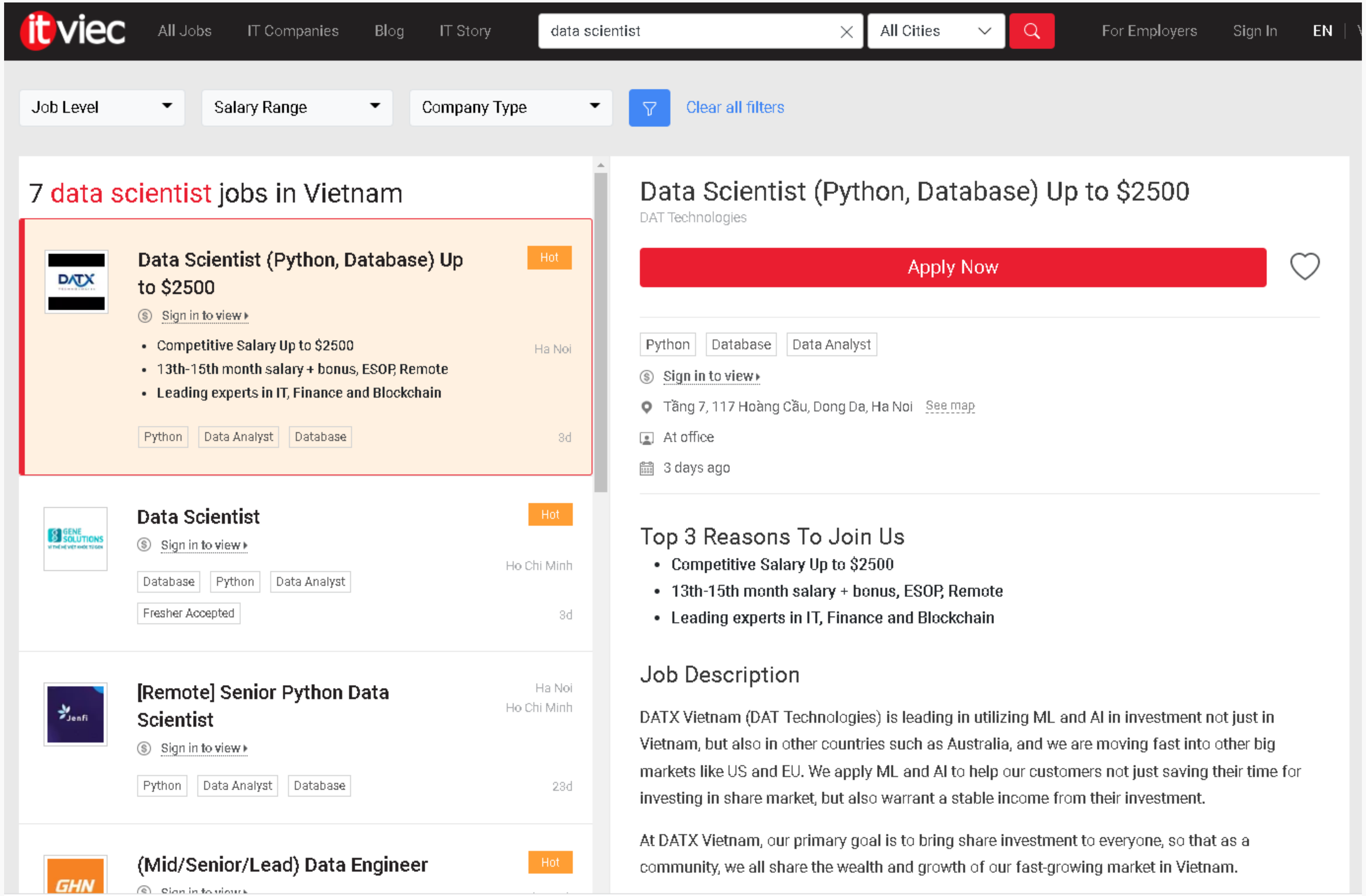
Why?


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?



7 data scientist jobs in Vietnam

Data Scientist (Python, Database) Up to $2500
Hot
Sign in to view
- Competitive Salary Up to $2500
- 13th-15th month salary + bonus, ESOP, Remote
- Leading experts in IT, Finance and Blockchain

Ha Noi
Python | Data Analyst | Database
3d

Data Scientist
Hot
Sign in to view
Database | Python | Data Analyst
Fresher Accepted
Ho Chi Minh
3d

[Remote] Senior Python Data Scientist
Sign in to view
Python | Data Analyst | Database
Ha Noi
Ho Chi Minh
23d

(Mid/Senior/Lead) Data Engineer
Hot

Data Scientist (Python, Database) Up to $2500
DAT Technologies

Python | Database | Data Analyst
Sign in to view
Tầng 7, 117 Hoàng Cầu, Dong Da, Ha Noi See map
At office
3 days ago

Top 3 Reasons To Join Us
- Competitive Salary Up to $2500
- 13th-15th month salary + bonus, ESOP, Remote
- Leading experts in IT, Finance and Blockchain

Job Description

DATX Vietnam (DAT Technologies) is leading in utilizing ML and AI in investment not just in Vietnam, but also in other countries such as Australia, and we are moving fast into other big markets like US and EU. We apply ML and AI to help our customers not just saving their time for investing in share market, but also warrant a stable income from their investment.

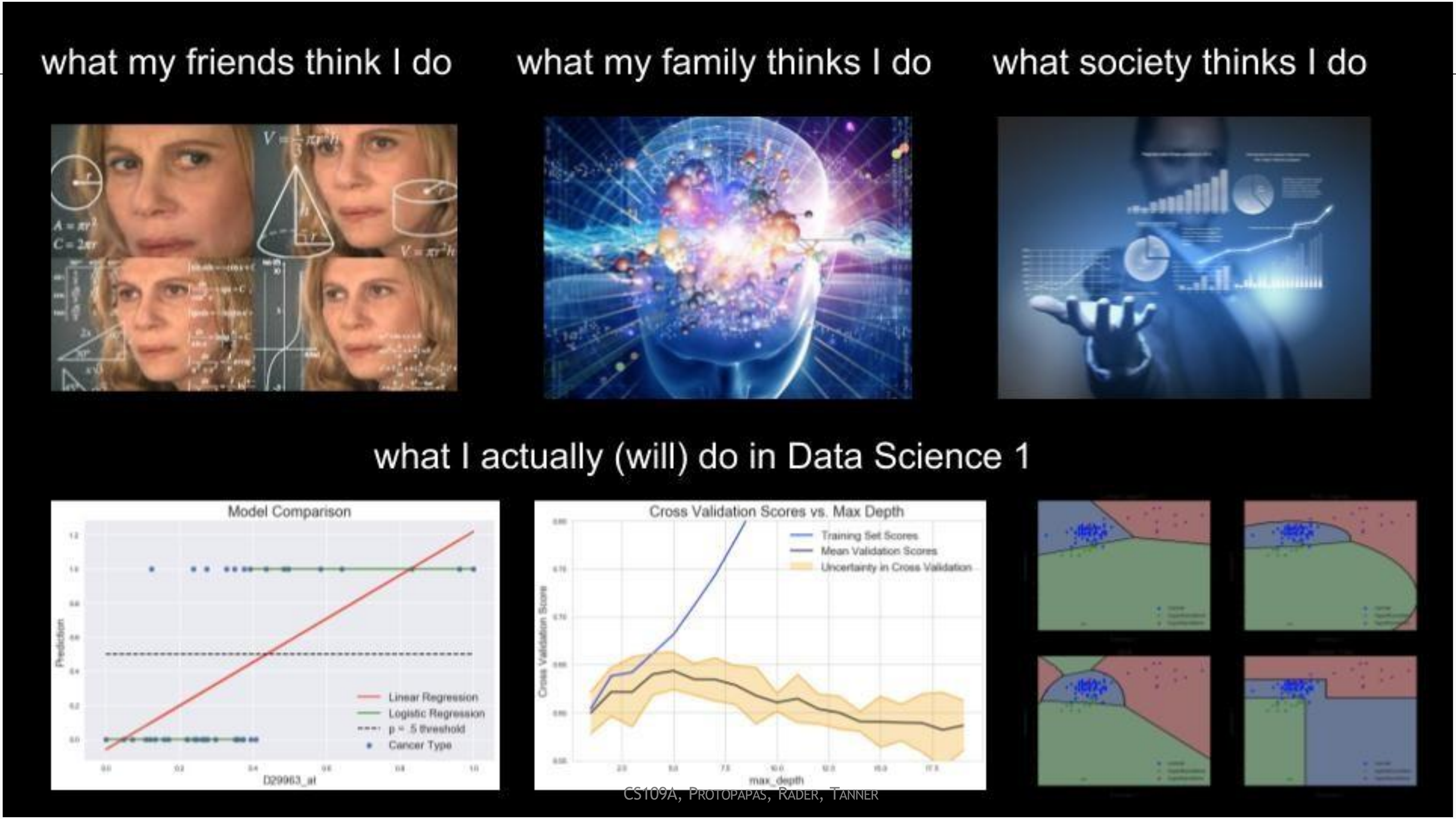
At DATX Vietnam, our primary goal is to bring share investment to everyone, so that as a community, we all share the wealth and growth of our fast-growing market in Vietnam.

https://itviec.com/it-jobs/data-scientist?job_selected=data-scientist-python-database-up-to-2500-dat-technologies-0241

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
Model Comparison
Prediction
D29963_at
Linear Regression
Logistic Regression
p = .5 threshold
Cancer Type
Cross Validation Scores vs. Max Depth
Cross Validation Score
max_depth
Training Set Scores
Mean Validation Scores
Uncertainty in Cross Validation

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)

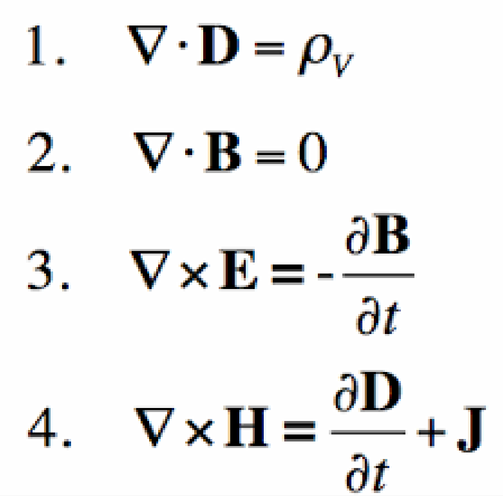
Long time ago (thousands of years) science was only empirical and people counted stars or crops

History (cont)

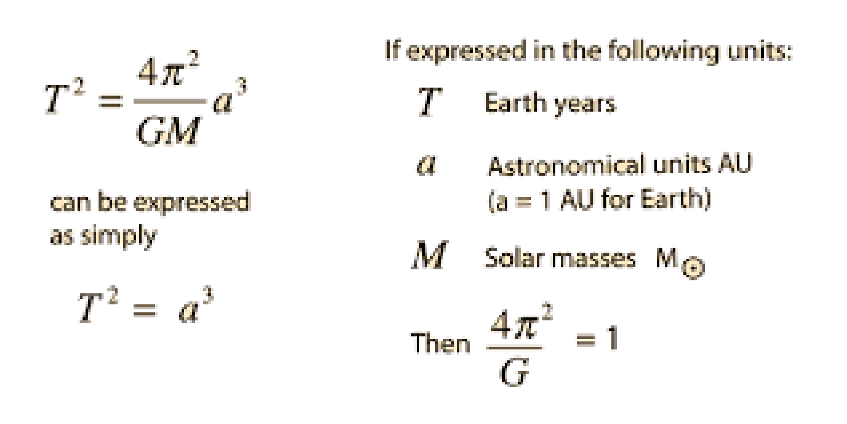
Long time ago (thousands of years) science was only empirical and people counted stars or crops and used the data to create machines to describe the phenomena

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



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

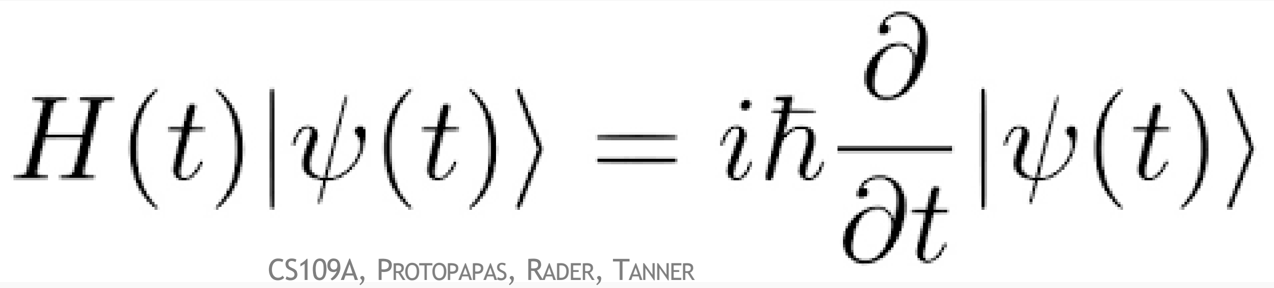
can be expressed as simply

$$T^2 = a^3$$

If expressed in the following units:

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

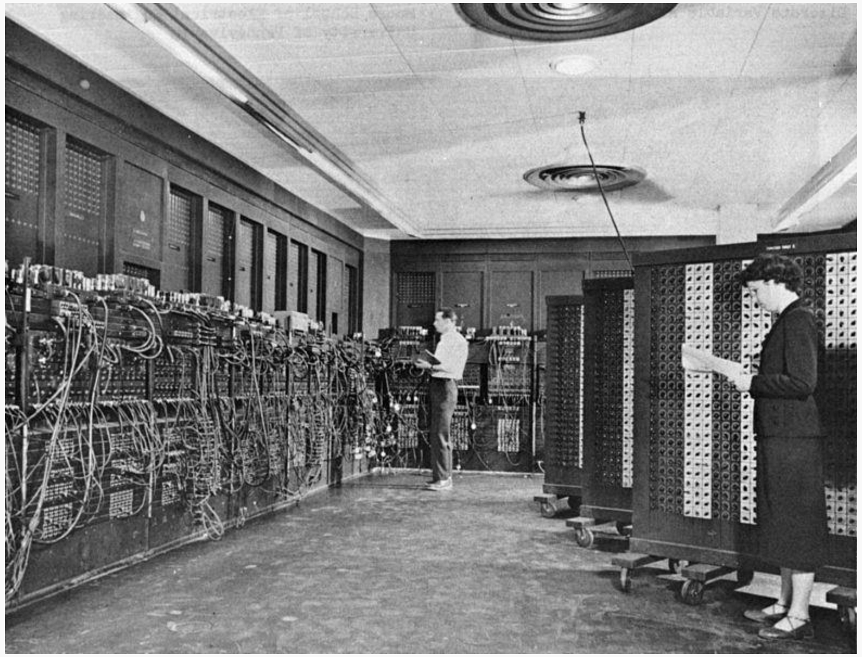
Then $\frac{4\pi^2}{G} = 1$



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

History (cont)

About a hundred years ago: computational approaches



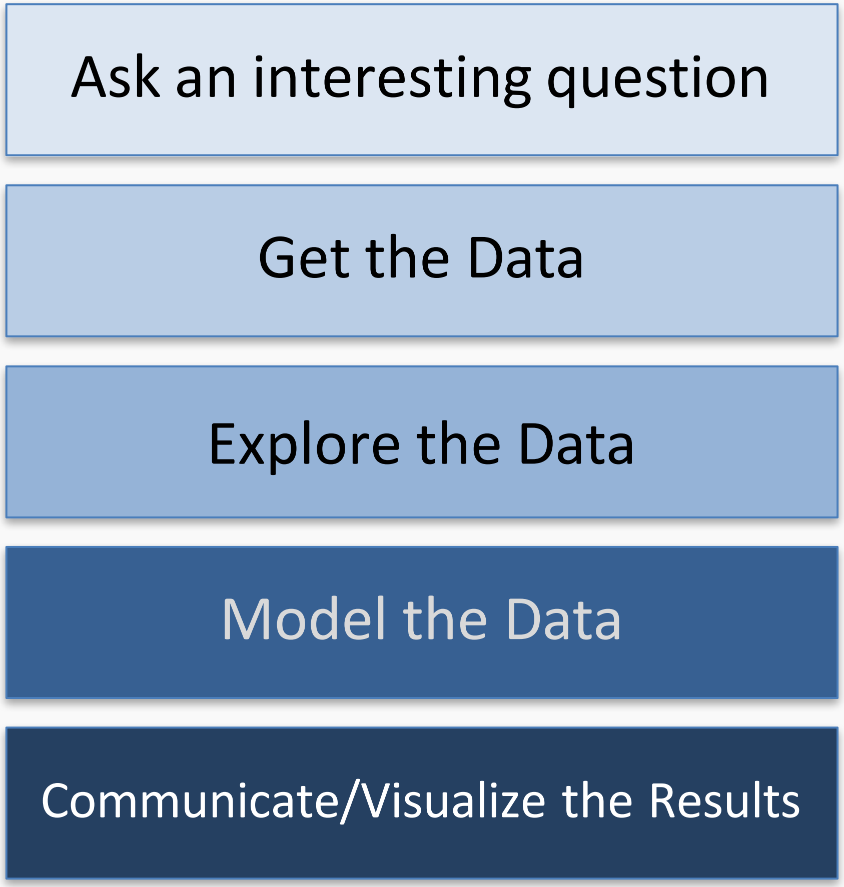
History (cont)

And then …. data science

What is data science?

What?

The Data Science Process



Ask an interesting question

Get the Data

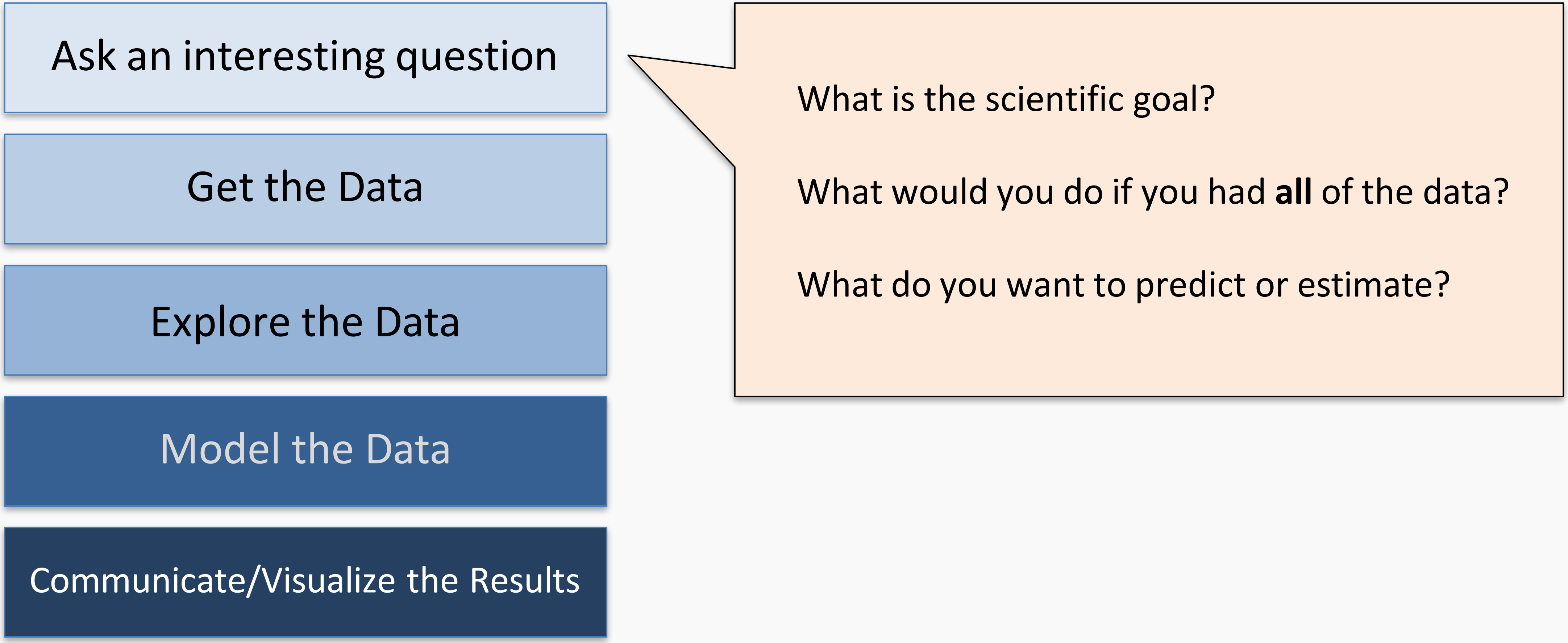
Explore the Data

Model the Data

Communicate/Visualize the Results

What?

The Data Science Process



Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

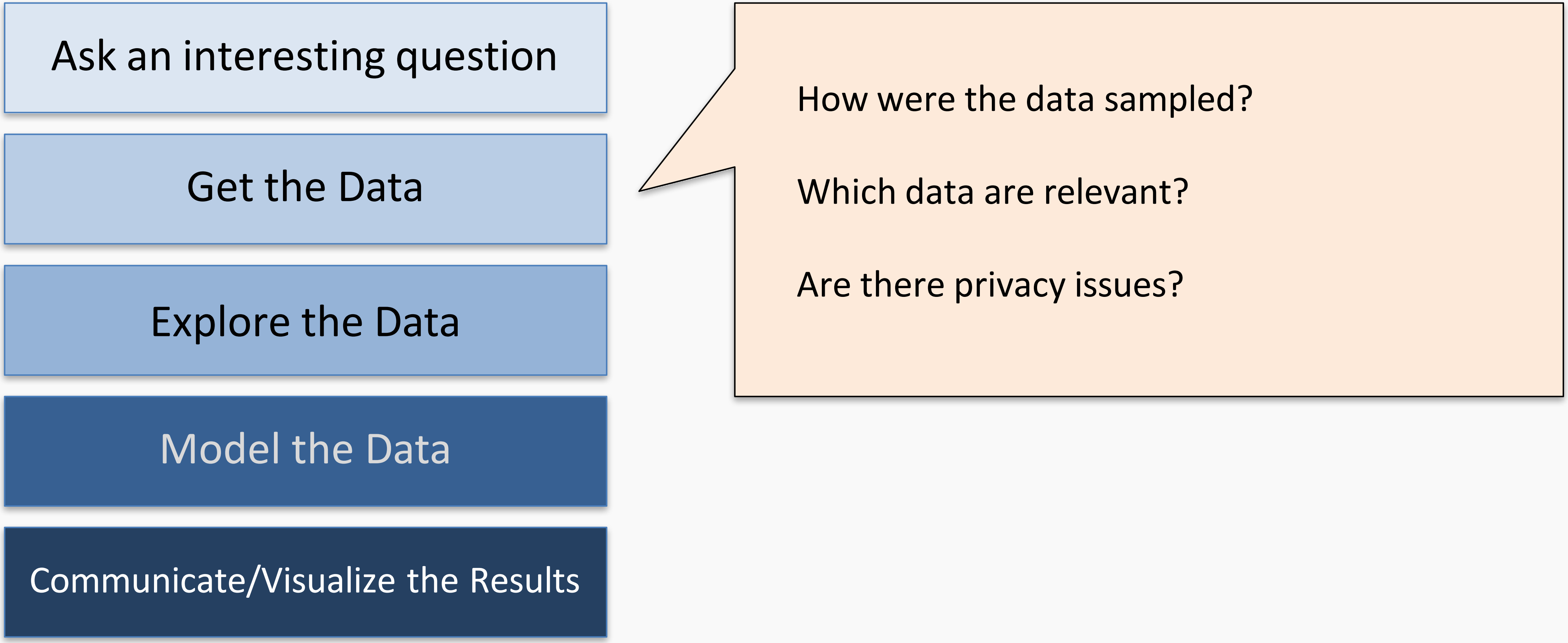
What is the scientific goal?

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

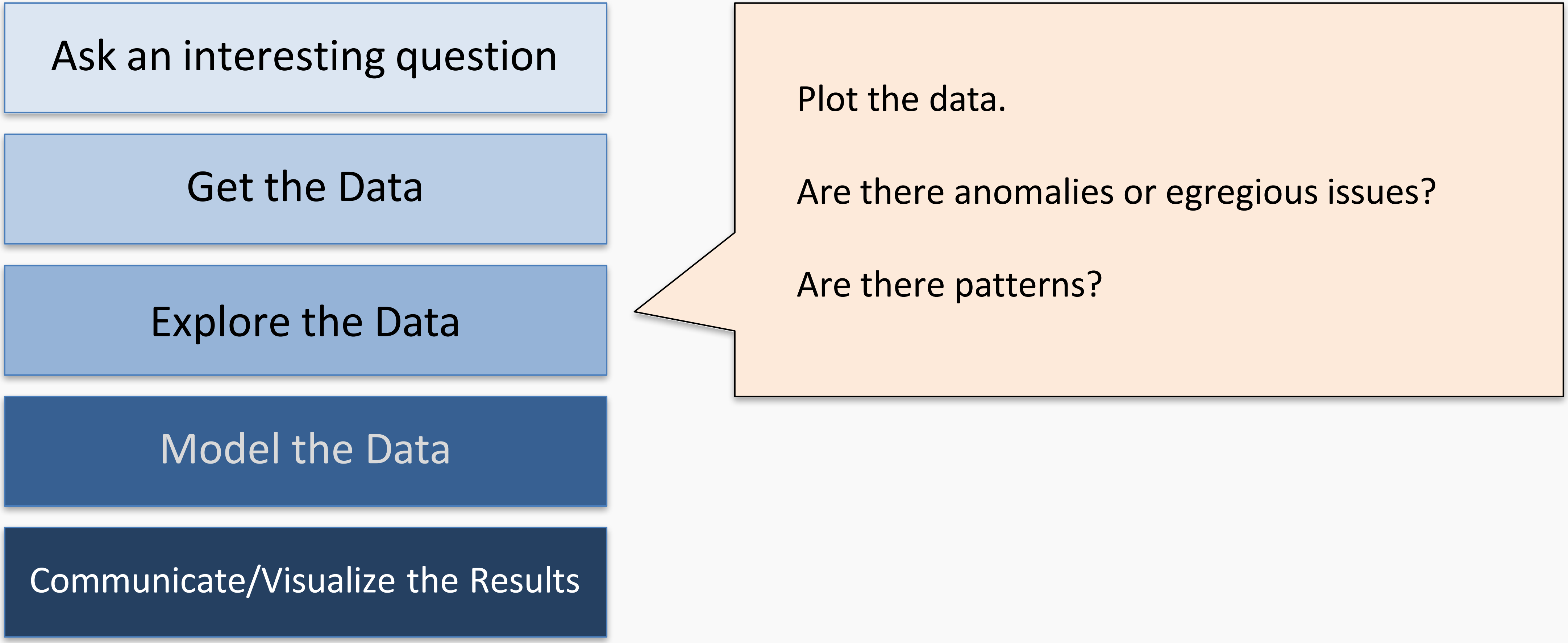
How were the data sampled?

Which data are relevant?

Are there privacy issues?

What?

The Data Science Process



Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

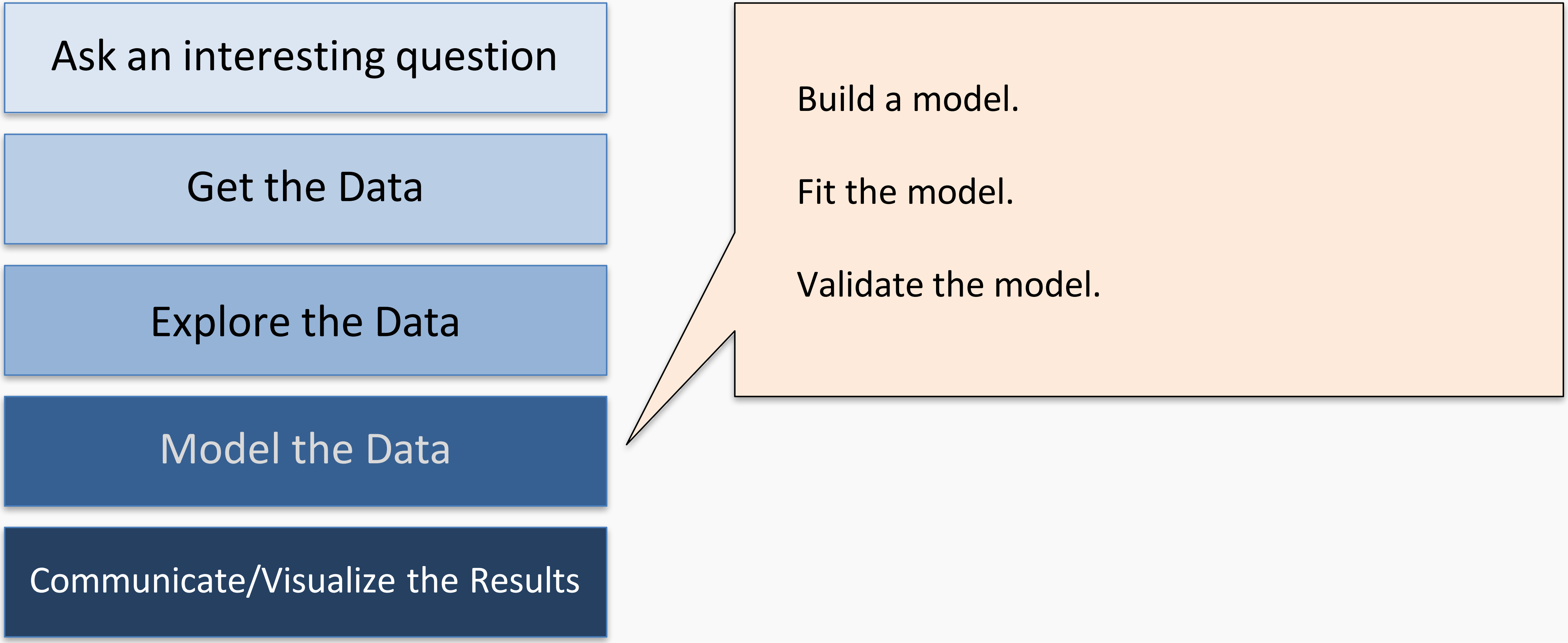
Plot the data.

Are there anomalies or egregious issues?

Are there patterns?

What?

The Data Science Process



Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

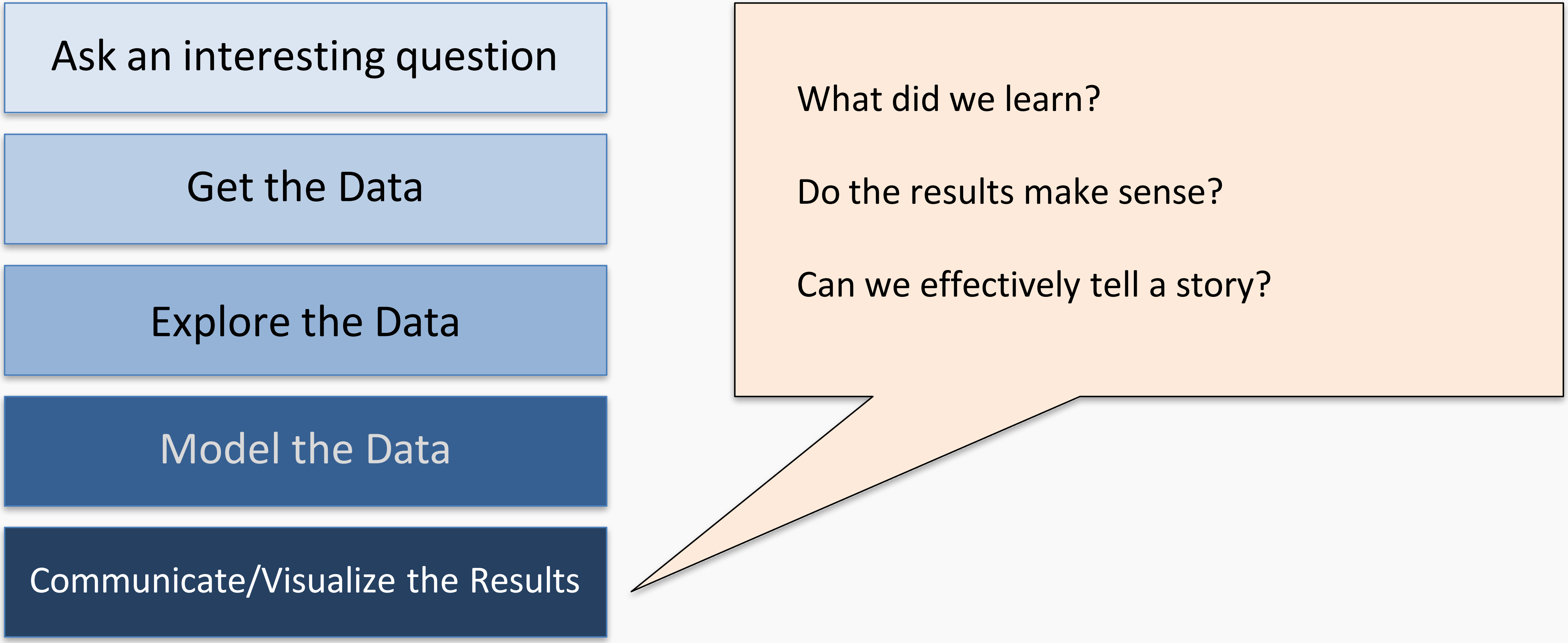
Build a model.

Fit the model.

Validate the model.

What?

The Data Science Process



Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

What did we learn?

Do the results make sense?

Can we effectively tell a story?

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 8-10:

EthiCS

PCA and high dimensionality

Decisions trees and ensemble methods :

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

Other topics relating to this course

A. Neural Networks:

- CNNs
- RNNs
- Generative models

B. Unsupervised Clustering

C. Piecewise Linear Regression

D. Bayesian Modeling

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?

Lecturers:

- Asso. Prof. Huynh Trung Hieu
- Dr. Nguyen Chi Kien

Related Sections

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

Homework(s)

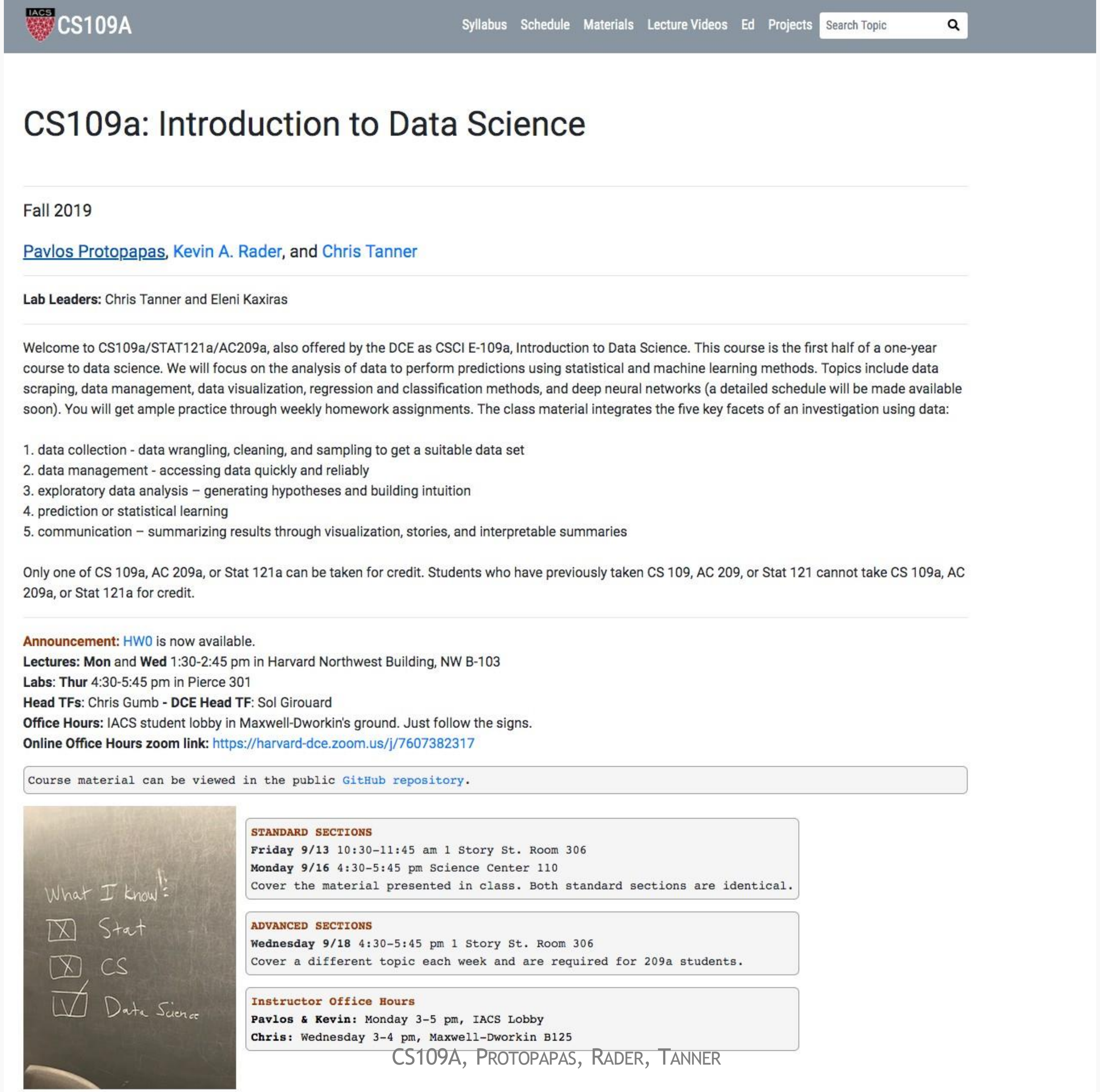
There will be 8 homework (not including Homework 0):

- Homework 0
- 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

Final Project

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

- We will provide some 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)



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.

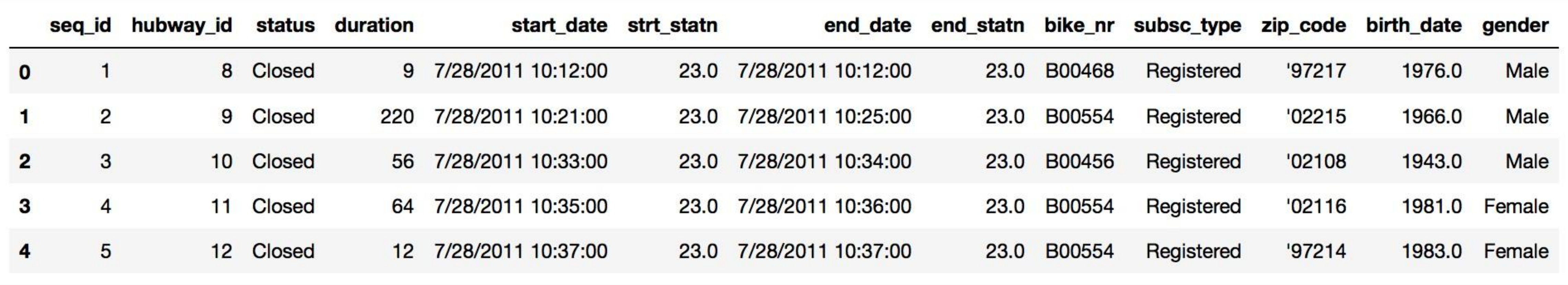
The Data: In April 2017, Hubway held a Data Visualization Challenge at the Microsoft NERD Center in Cambridge, releasing 5 years of trip data.

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?

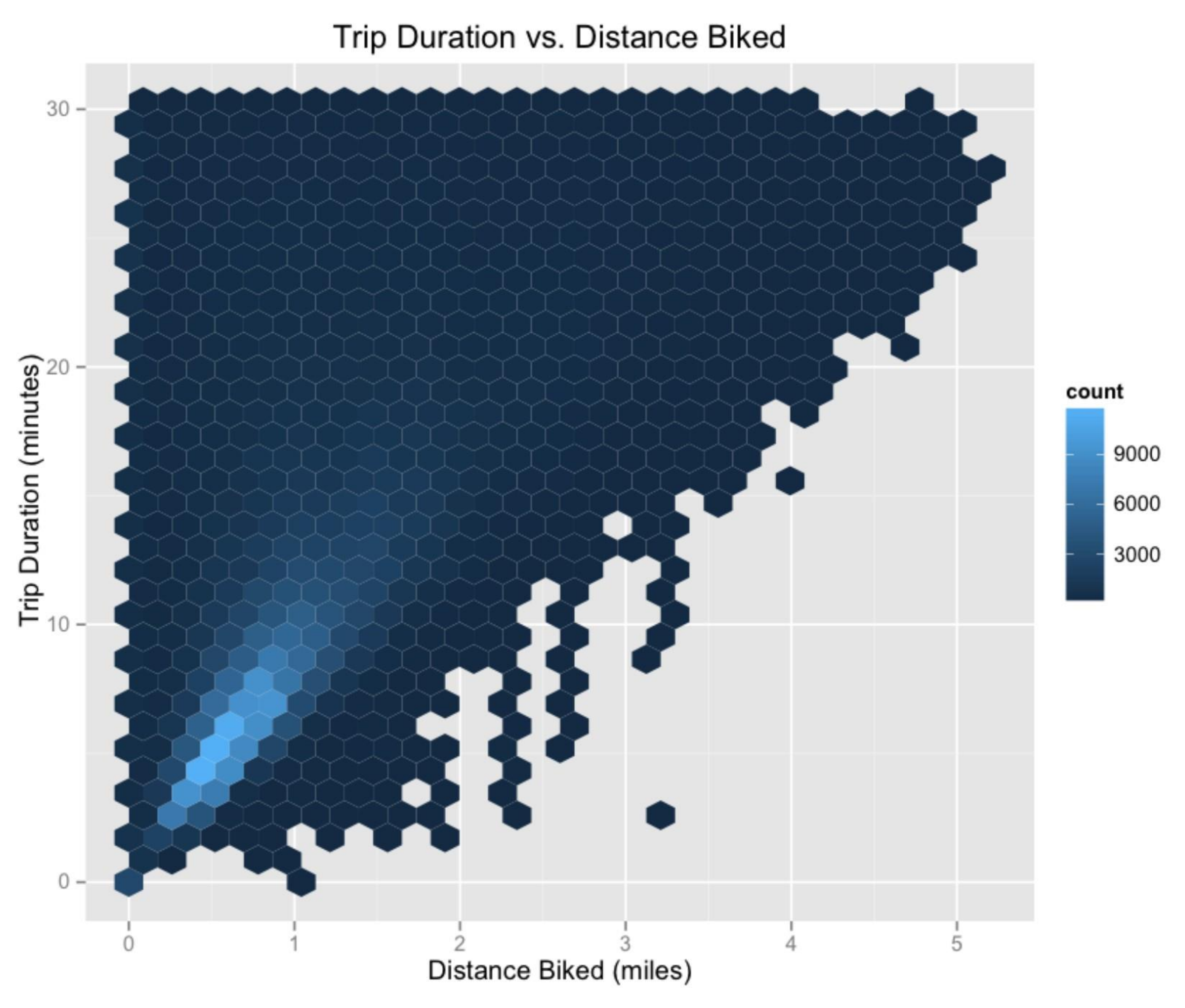
How questions are about modeling relationships between different variables.

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


Trip Duration vs. Distance Biked
Trip Duration (minutes)
Distance Biked (miles)
count