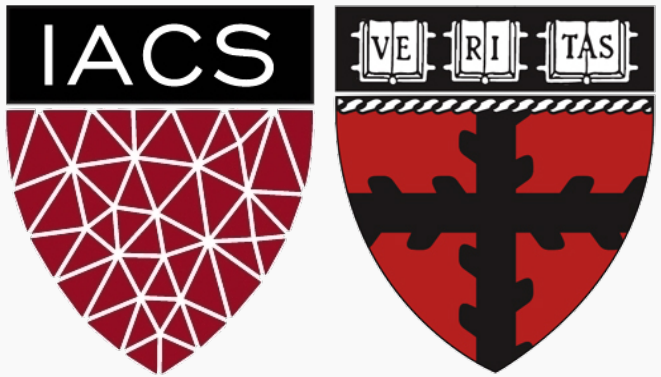
# Lecture #1: Introduction to CS109A

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

# 50 Best Jobs in America for 2020

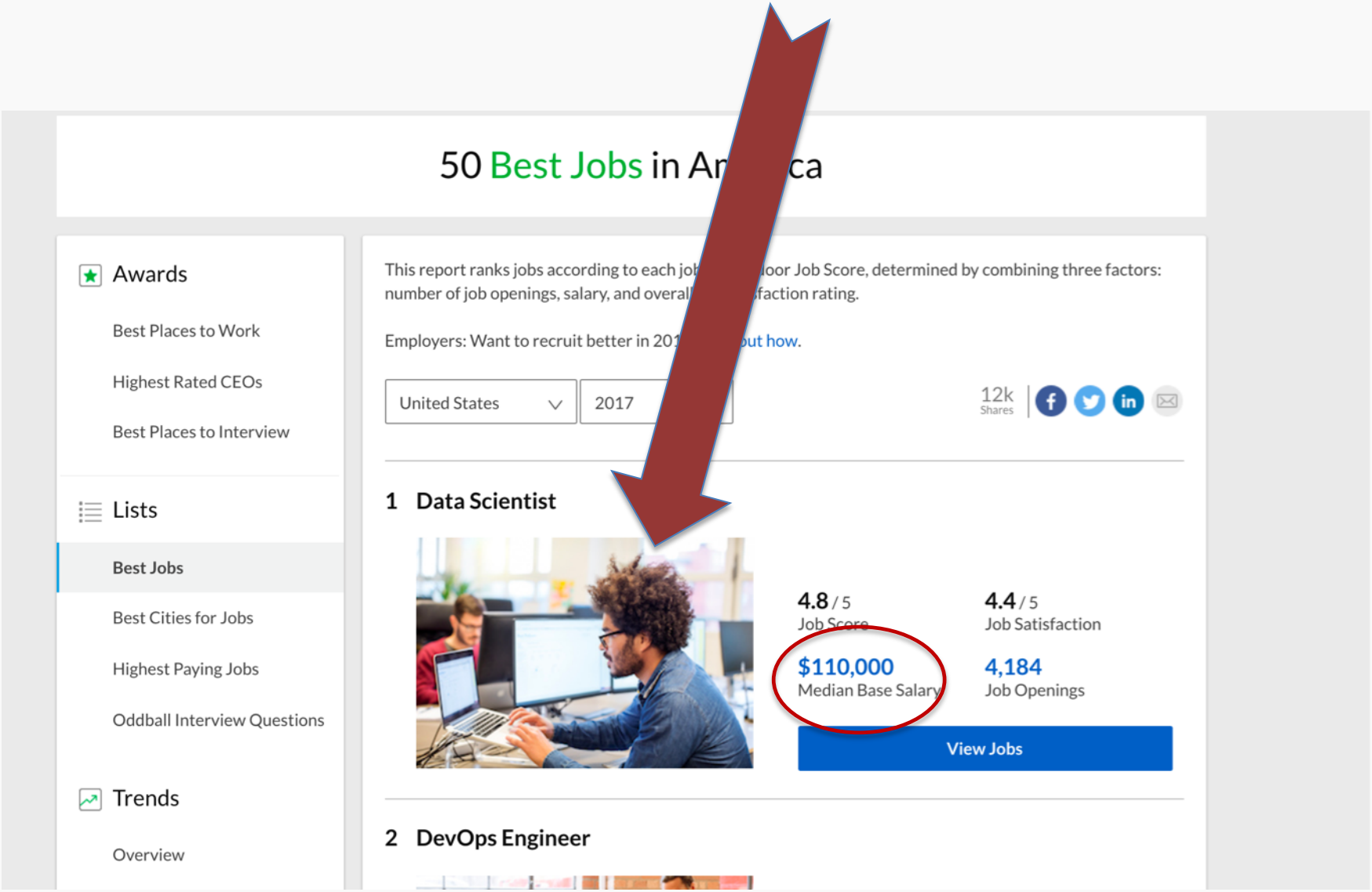
Best Jobs | 2020 | United States

Share

| | Job Title | Median Base Salary | Job Satisfaction | Job Openings | |
|---|---|---|---|---|---|
| #1 | Front End Engineer | $105,240 | 3.9/5 | 13,122 | View Jobs |
| #2 | Java Developer | $83,589 | 3.9/5 | 16,136 | View Jobs |
| #3 | Data Scientist | $107,801 | 4.0/5 | 6,542 | View Jobs |
| #4 | Product Manager | $117,713 | 3.8/5 | 12,173 | View Jobs |
| #5 | DevOps Engineer | $107,310 | 3.9/5 | 6,603 | View Jobs |
| #6 | Data Engineer | $102,472 | 3.9/5 | 6,941 | View Jobs |
| #7 | Software Engineer | $105,563 | 3.6/5 | 50,438 | 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 oor Job Score, determined by combining three factors:
number of job openings, salary, and overall faction rating.
Employers: Want to recruit better in 201 ut 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?

Why do I love data science?

Why are you here?

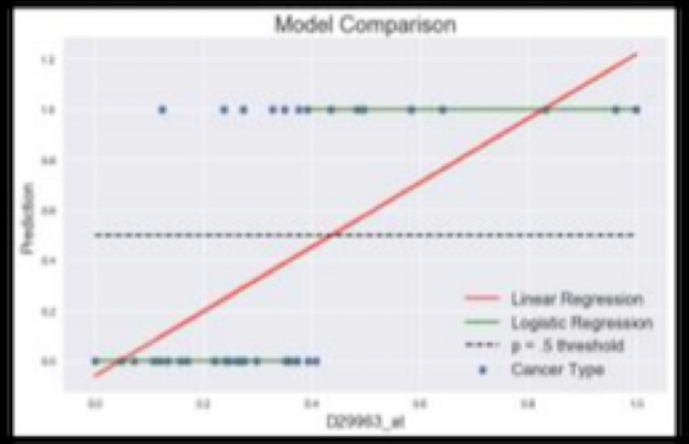
# Why?

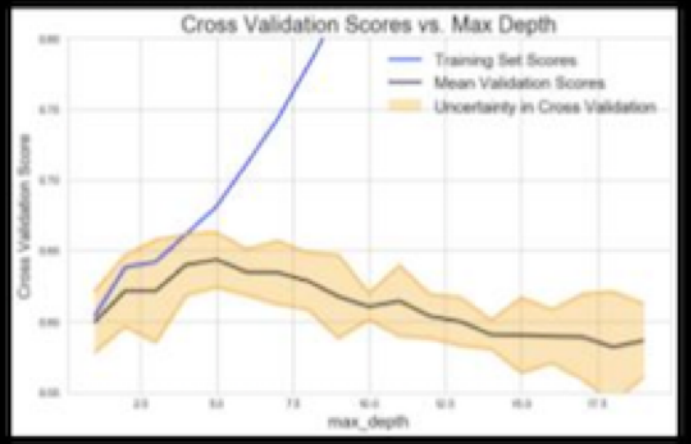
what my friends think I do

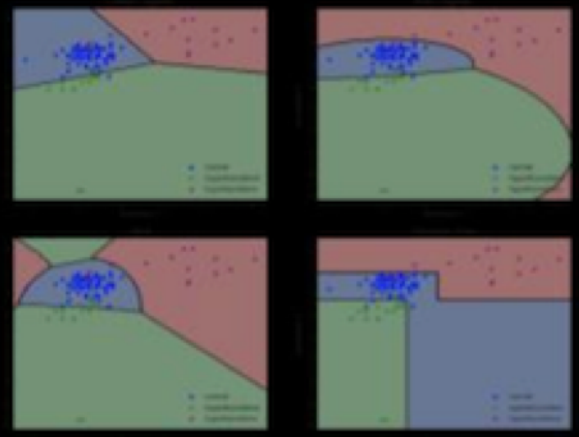
what my family thinks I do

what society thinks I do

what I actually (will) do in Data Science 1



# Why?

Why are you here?

# What is data science?

# 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:

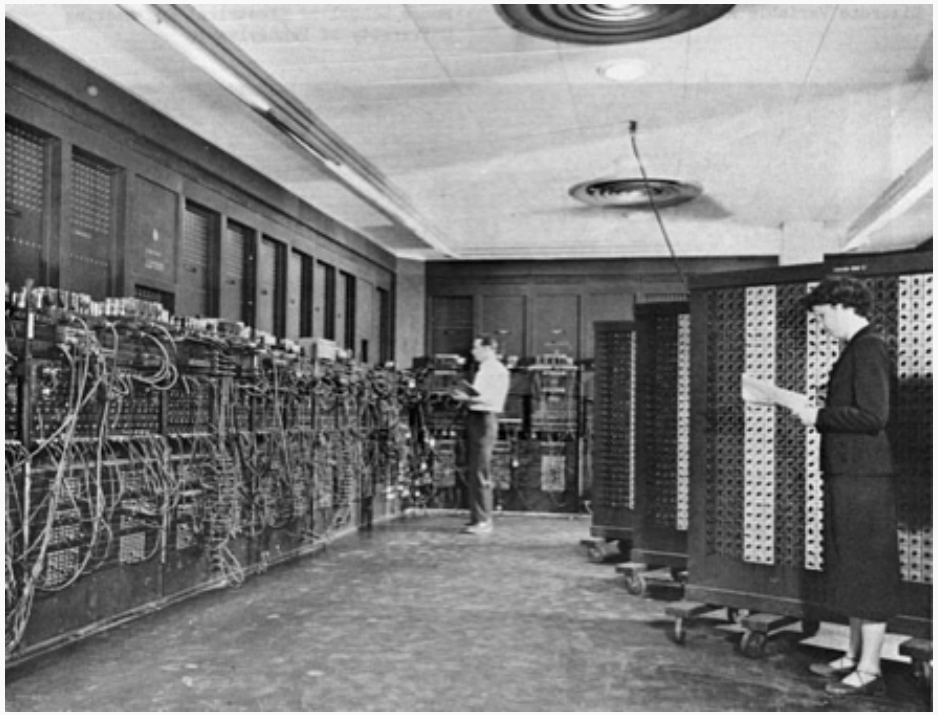
- $T$ Earth years
- $a$ Astronomical units AU (a = 1 AU for Earth)
- $M$ Solar masses $M_\odot$

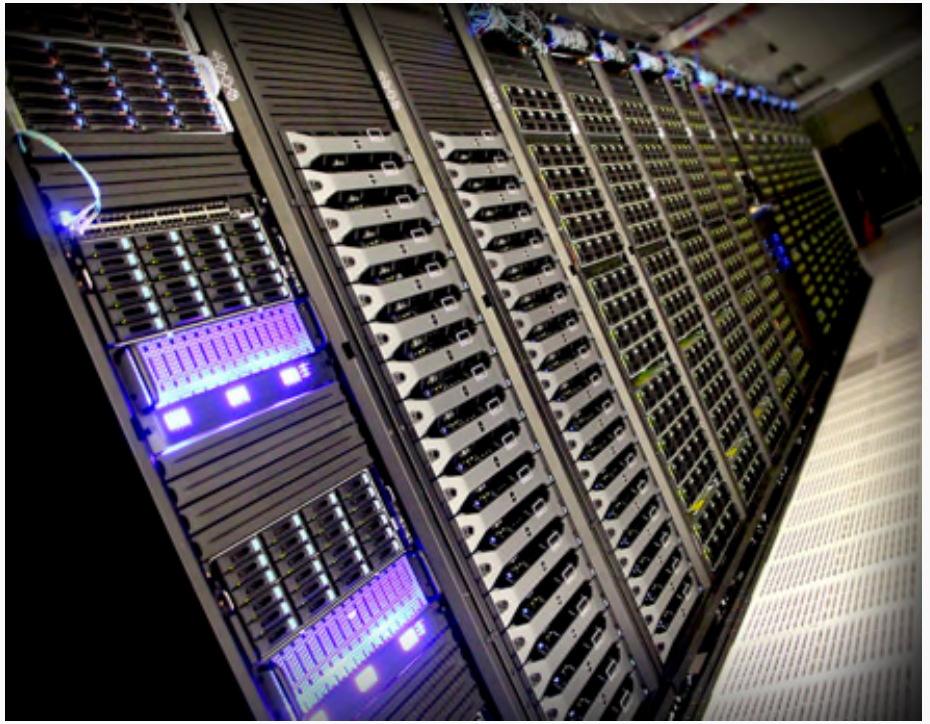
Then $\frac{4\pi^2}{G} = 1$

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

# History (cont)

About a hundred years ago: computational approaches

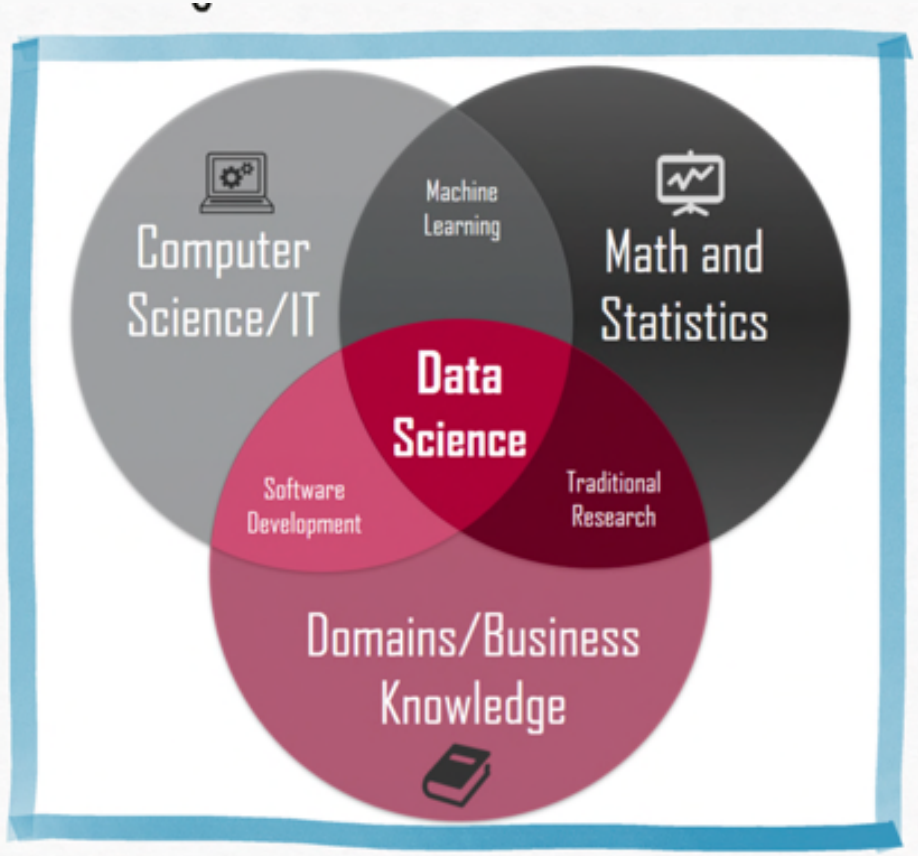




# History (cont)

And then …. data science

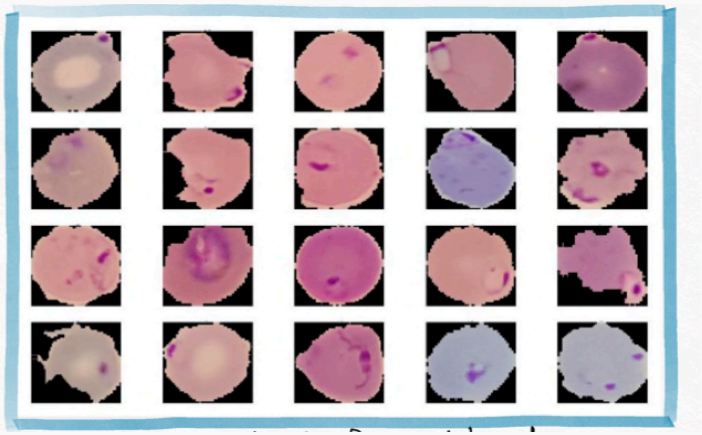
In both data science and machine learning we extract pattern and insights from data.




- Inter-disciplinary
- Data and task focused
- Resource aware
- Adaptable to changes in the environment and needs

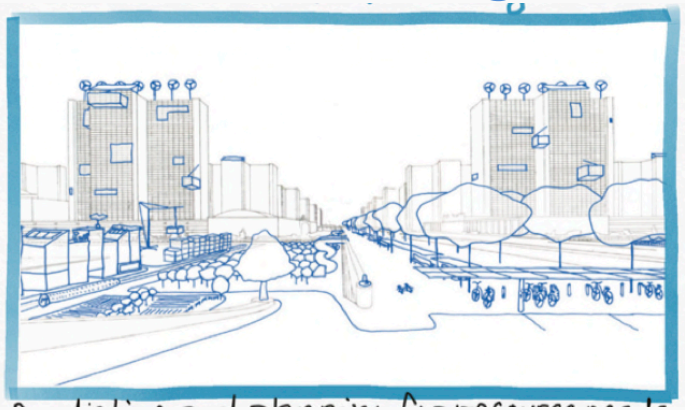
# The Potential of Data Science

## Disease Diagnosis



Detecting malaria from blood smears

## Urban Planning



Predicting and planning for resource needs

## Drug Discovery

Quickly discovering new drugs for COVID

## Agriculture

Precision agriculture

# The Potential of Data Science

## Gender Bias

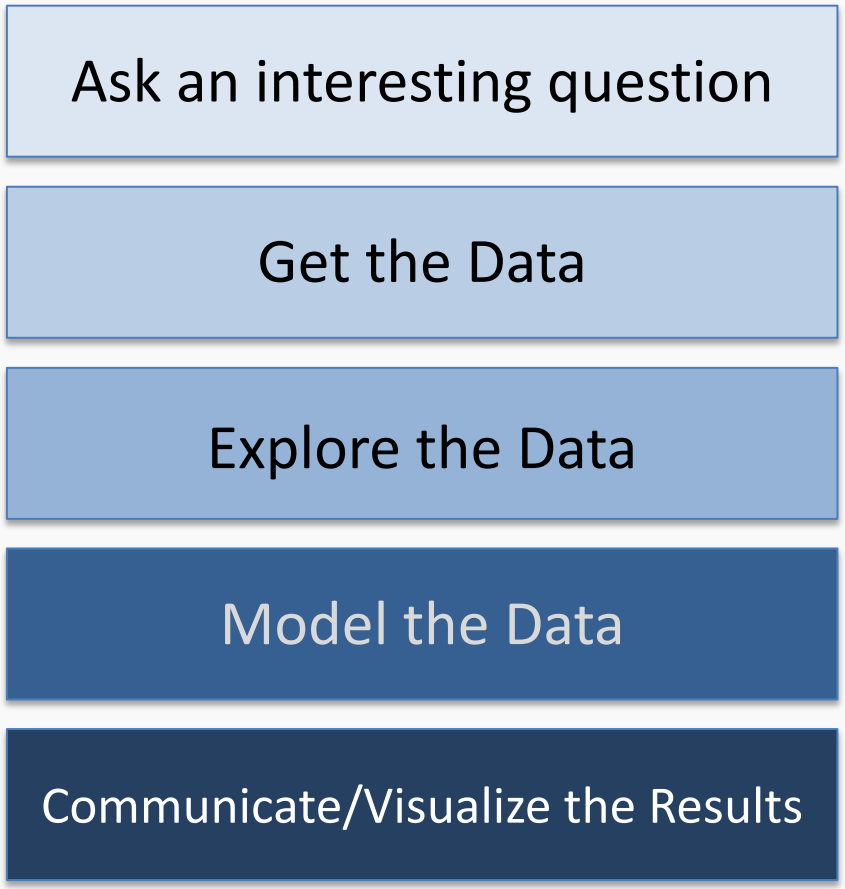
Some DS models for evaluate job applications show bias in favor of male candidate

## Racial Bias

Risk models used in US courts have shown to be biased against non-white defendants

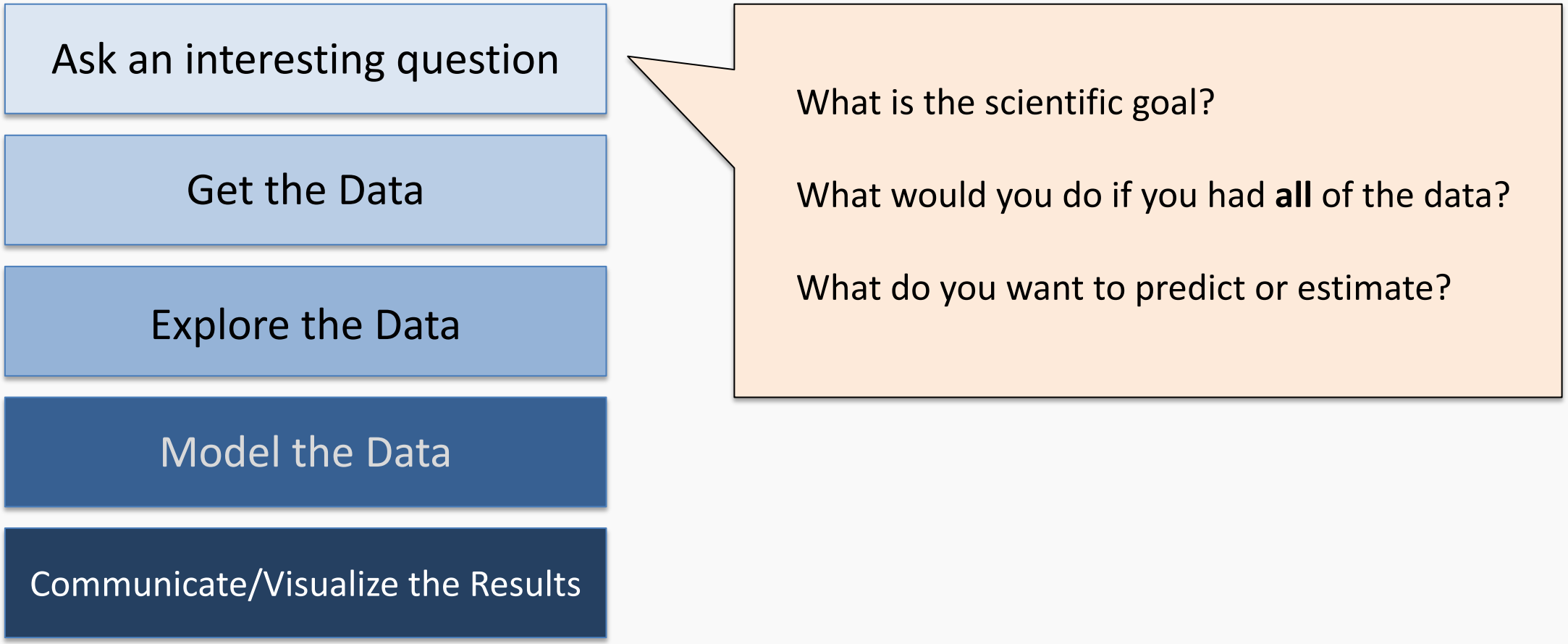
# What?

## The Data Science Process



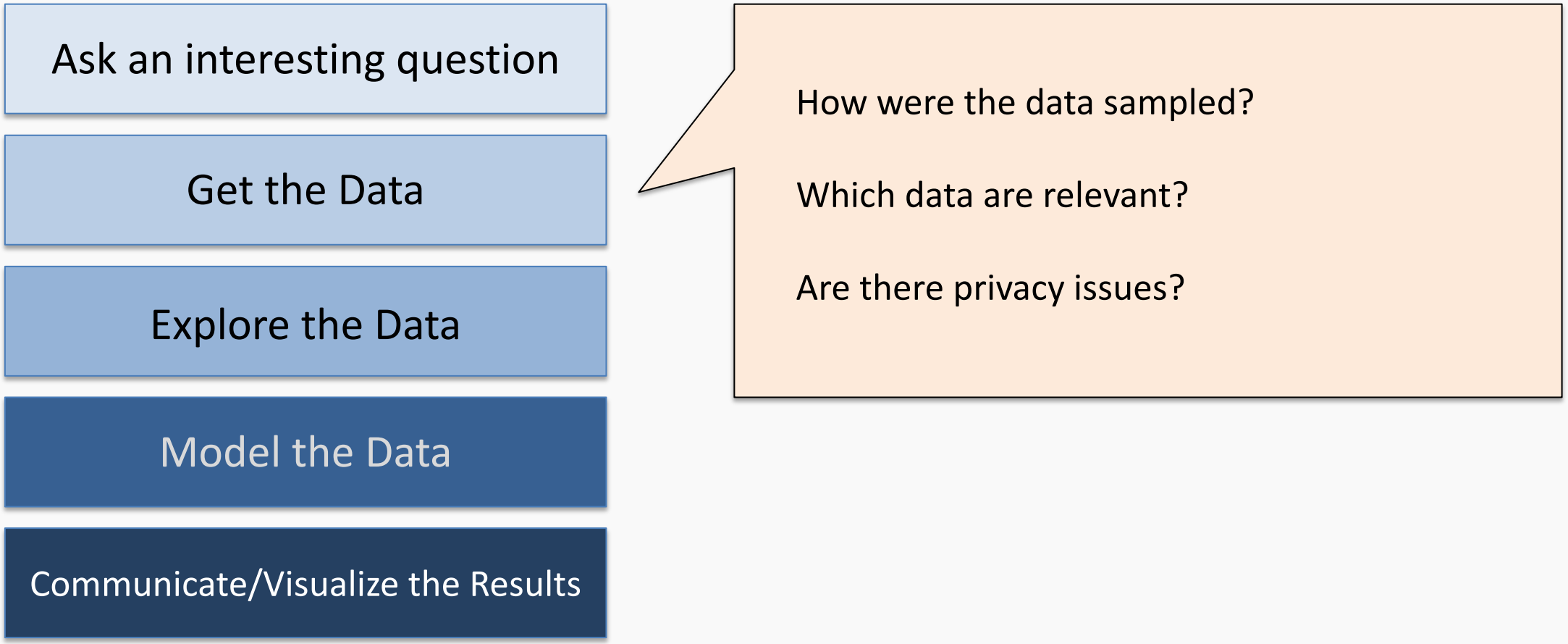
# What?

## The Data Science Process



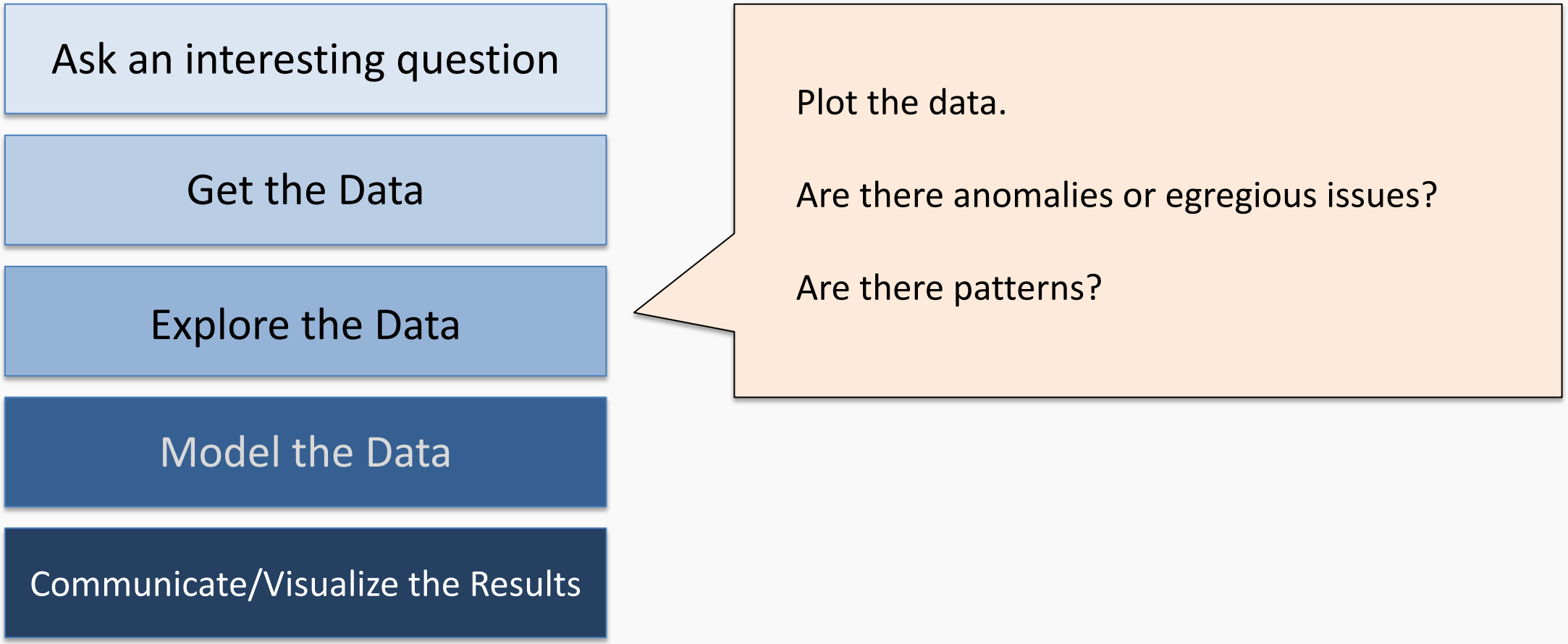
# What?

## The Data Science Process



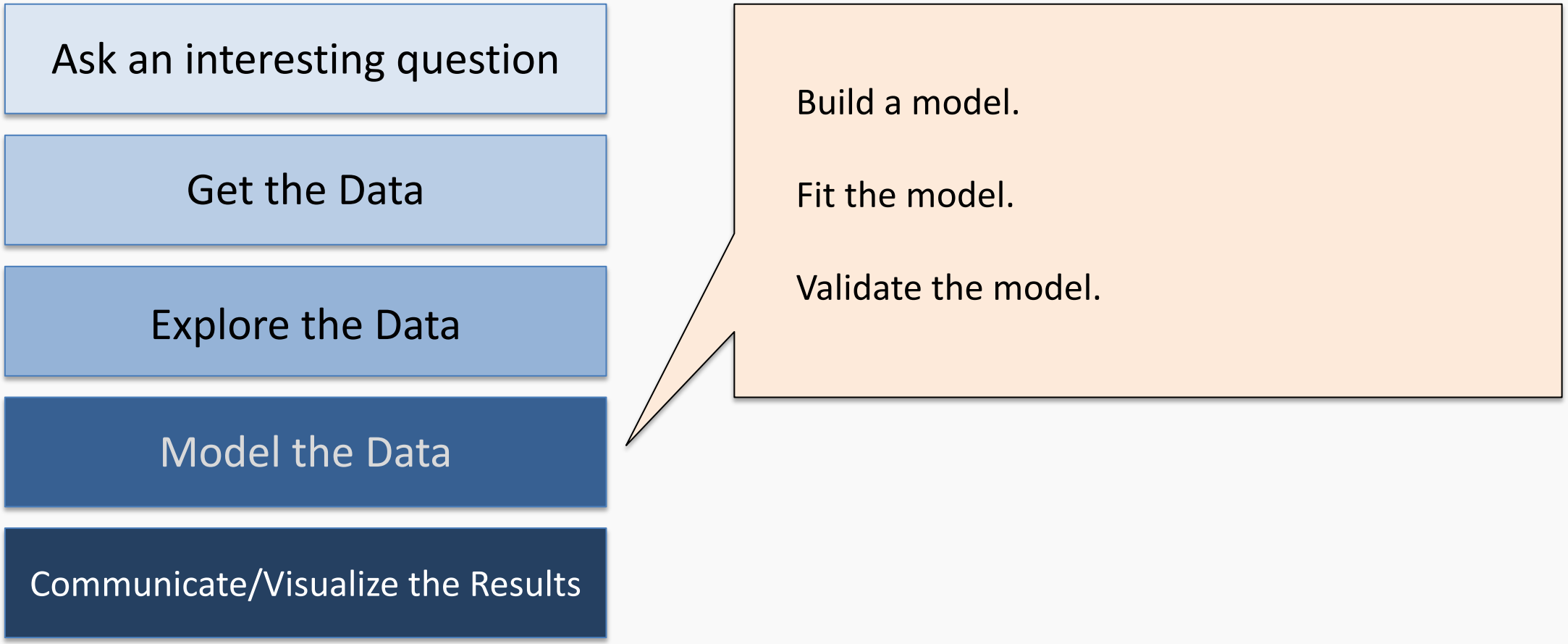
# What?

## The Data Science Process



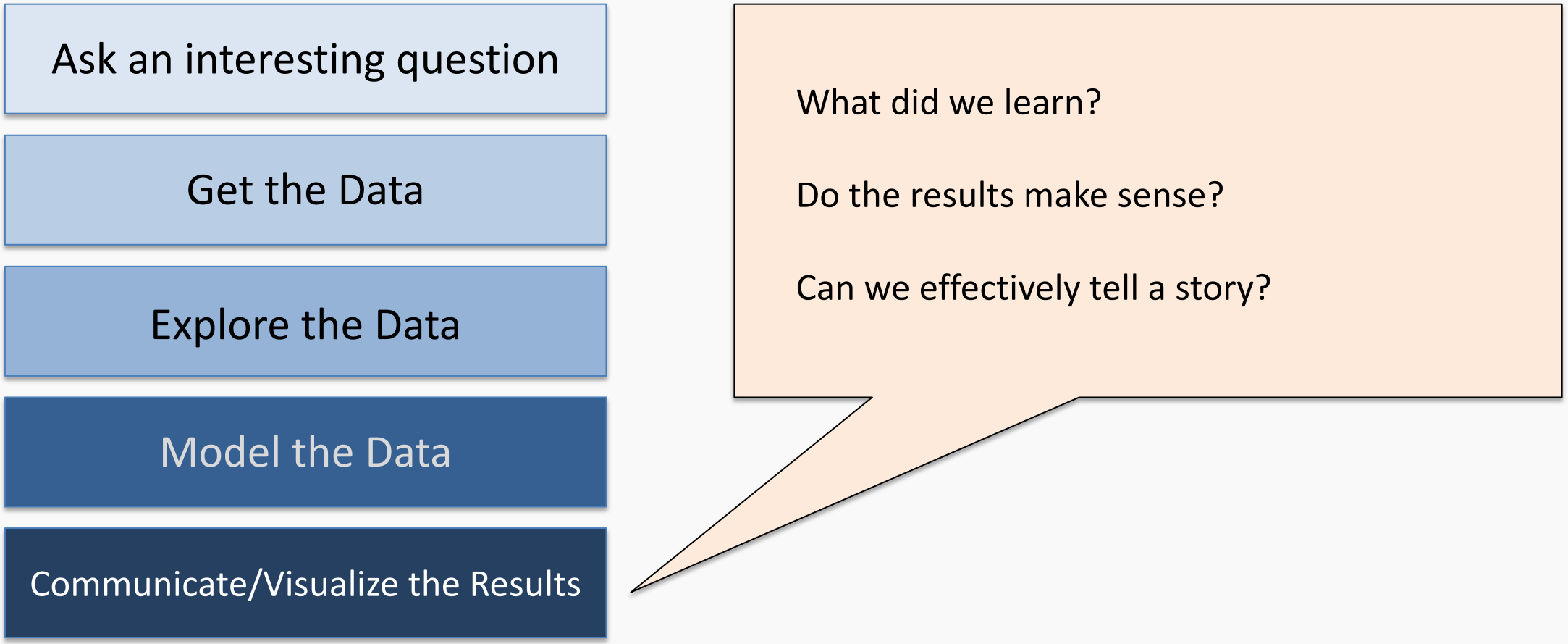
# What?

## The Data Science Process



# 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.

# Goal of the course

## Theory

1. Key Machine Learning concept
2. Important metrics for evaluation
3. Handling different kinds of data
4. Extracting insights from analysis of the models

## Practice

1. Implement ML and deep learning models using python libraries
2. Using free online tools and resources for data science

## Impact

1. Solving real-life problems using DS
2. Evaluating the social impact of DS

## Weeks 1-2: Data

Data Formats + Web Scraping
Pandas

## Weeks 3-5: Regression

kNN Regression
Linear Regression
Multi and Poly Regression
Model Selection and Cross Validations
Inference
Bootstrap
Ridge and Lasso Regularization

## Weeks 6-7: Classification

kNN Classification
Logistic Regression
Multi-class Classification

## Weeks 8: Data

Data Imputation
PCA

## Weeks 9-10: Trees

Decision Trees
Bagging
Random Forest
Boosting Methods

## Weeks 11-12: Neural Networks

Decision Trees
Bagging
Random Forest
Boosting Methods

## Weeks 13-14

Ethics
Model Interpretation

# After CS109A

## CS109B

A. Neural Networks:

- CNNs
- RNNs
- Generative models

B. Unsupervised Clustering

C. Piecewise Linear Regression

D. Bayesian Modeling

## AC295

A. Productions Data Science, from notebooks to the cloud

B. Big models, transfer learning and architecture learning

C. Visualization tools for interpreting models

# Who? Instructors

**Pavlos Protopapas**

Scientific director
Institute of Applied Computational Science

**Kevin Rader**

Senior preceptor in Statistics

**Chris Tanner**

Lecturer at Institute of Applied Computational Science

# Who?

**Eleni Kaxiras**

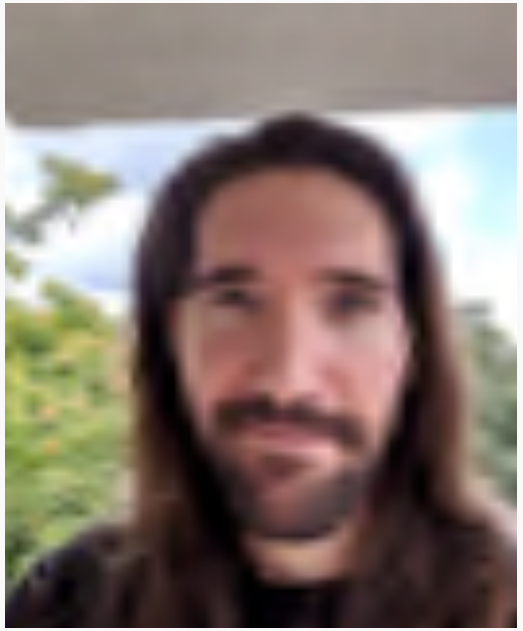
Supportive Instructor

Assistant Director for Data Science and Computation at SEAS

**Marios Mattheakis**

Section Leader

Post-doctoral Fellow IACS



**Chris Gumb**

Head TF

Graduate student of Data Science at Harvard Extension School

# Who? Teaching Fellows

# Course Components

# **Lectures**, 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.

We will use a mix of notes and exercises via edstem.

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

We will have two 'shows' , morning and matinee (A, B)

A: Morning Mon/Wed/Fri 9:00-10:15am @Zoom

B: Matinee Mon/Wed/Fri 3:00-4:15pm @Zoom

# Lecture format

ASYNCHRONOUS

- Quiz
- Finish exercises from previous lecture
- Reading or Video Watching for next lecture

SYNCHRONOUS

Questions from asynchronous material, review of quiz and homework

Live Lecture

Q&A

Hands-on exercises in breakout rooms

Discussion about the exercises

Repeat ⋮

Summary and conclusions

The "hands-on exercises" part will be longer during the Friday lecture as opposed to Mon/Wed.

# **Advanced Sections, Sections** and Office Hours

Lectures are supplemented by 1.25 hours 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 1:30-2:45 am, and Mon 8:30-9:45 pm @zoom

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

  Weds 12:01 pm - 1:15 pm @zoom. A-sections are required for AC209 students.

Note: Sections are not held every week. Consult the course calendar for exact dates.

# **Advanced Sections** topics

**Topics**

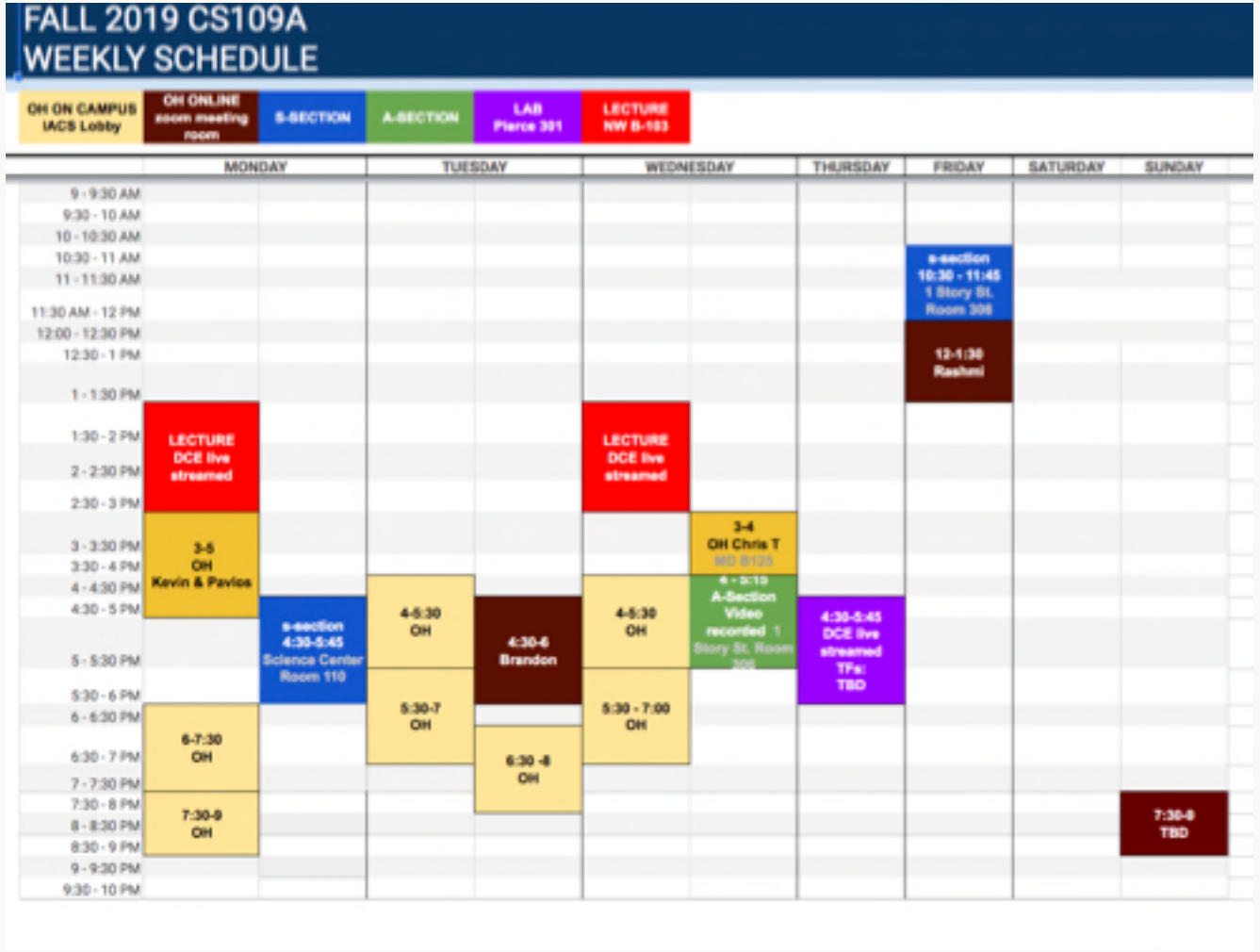
1. Linear Algebra and Hypothesis Testing: The Short Versions
2. Methods of regularization and their justifications
3. Generalized Linear Models
4. Mathematics of PCA
5. Ensemble methods
6. Stochastic Gradient Descent and solvers

**NOTE 1**: The materials in the *Advanced Sections* are 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.

# Office Hours



# Assignments

# Four Graded Components

## Homework: 63%

Homework zero: 1%
Individual Homework (2): 12%
Paired Homework (6): 40%

HW4 and HW6 are the indiv. HW

## Exercises: 6%

During lecture.

All questions are weighted equally.

Due at the beginning of the next morning lecture.

## Quizzes: 6%

End of each lecture.

25% of the quizzes will be dropped from your grade.
All questions are weighted equally.

Due at the beginning of the next morning lecture.

## Projects: 25%

Three milestones plus final presentation and a report in the form of a blog.
More details soon.

# 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 you must work individually.

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

All homework are **due 11:59 pm Wednesdays,** and homework will be released on Wednesdays.

# 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)
- Project topics will be announced September 10$^{th}$.

# 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. Attend the Office Hours; this is the best way to get help.
3. For private matters, send an email to the Helpline: **cs109a2020@gmail.com**. All the instructors and TFs monitor the Helpline.
4. For personal matters, send an email to Pavlos, Kevin, and Chris.

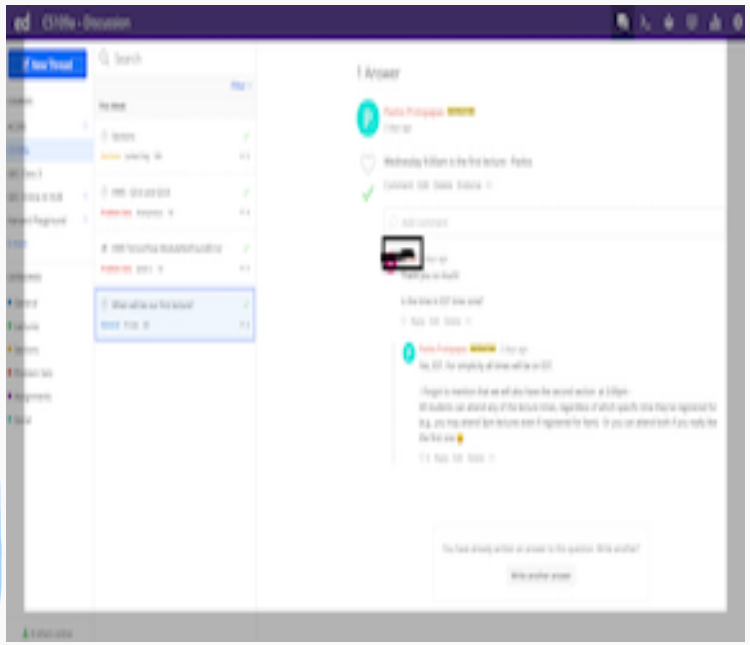
**Sundays will be slow days, so please be patient!**

# Tools for the course

## Web page


- Syllabus
- Calendar
- Link to materials

## edstem



- Forum
  Quizzes
- Reading assignments
- Hands on exercise
- Link to lectures

## Canvas


- Homework
- Grades

# Misc

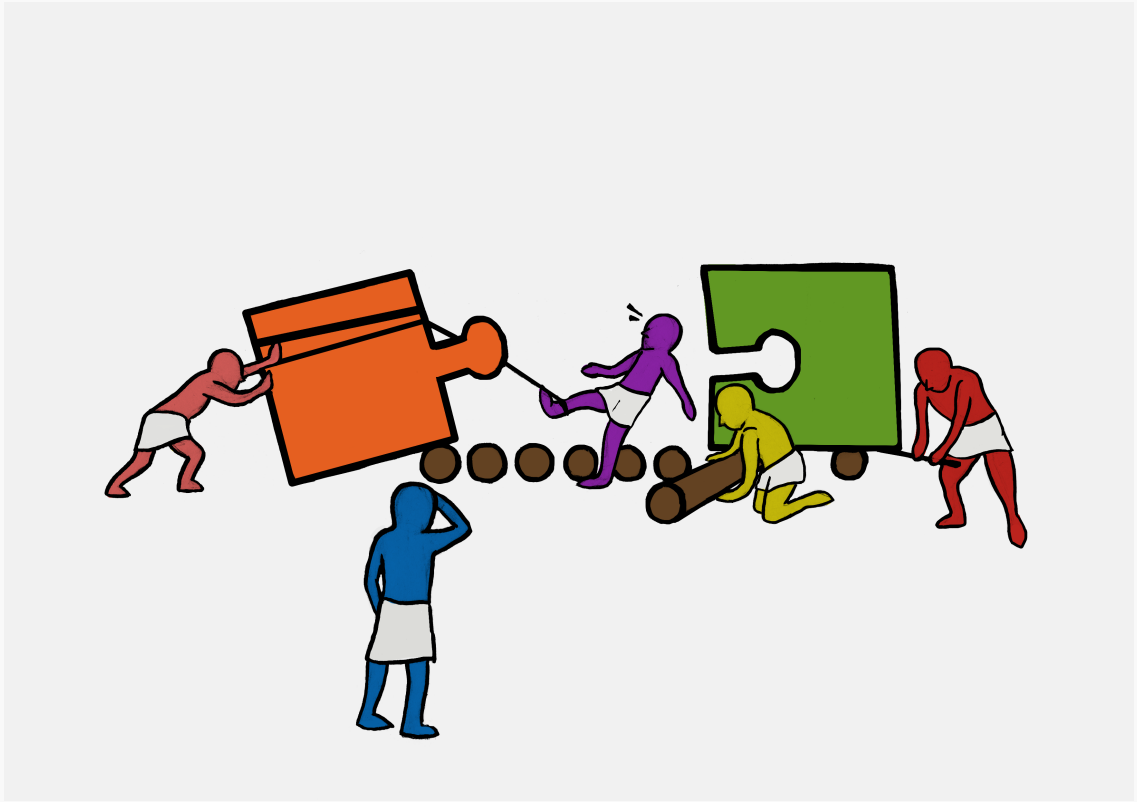
Video release form:
https://canvas.harvard.edu/courses/74056/quizzes/182045

Backup plan if zoom dies:

We will do lecture the next lecture (A or B). If it's still not up by then, we will record and upload to canvas.

Forming groups

https://docs.google.com/forms/d/e/1FAIpQLSfIsQyxdCwUCbJmyWyotp30anrKsuGHfIHt-DQEKMnK8iE4TA/viewform



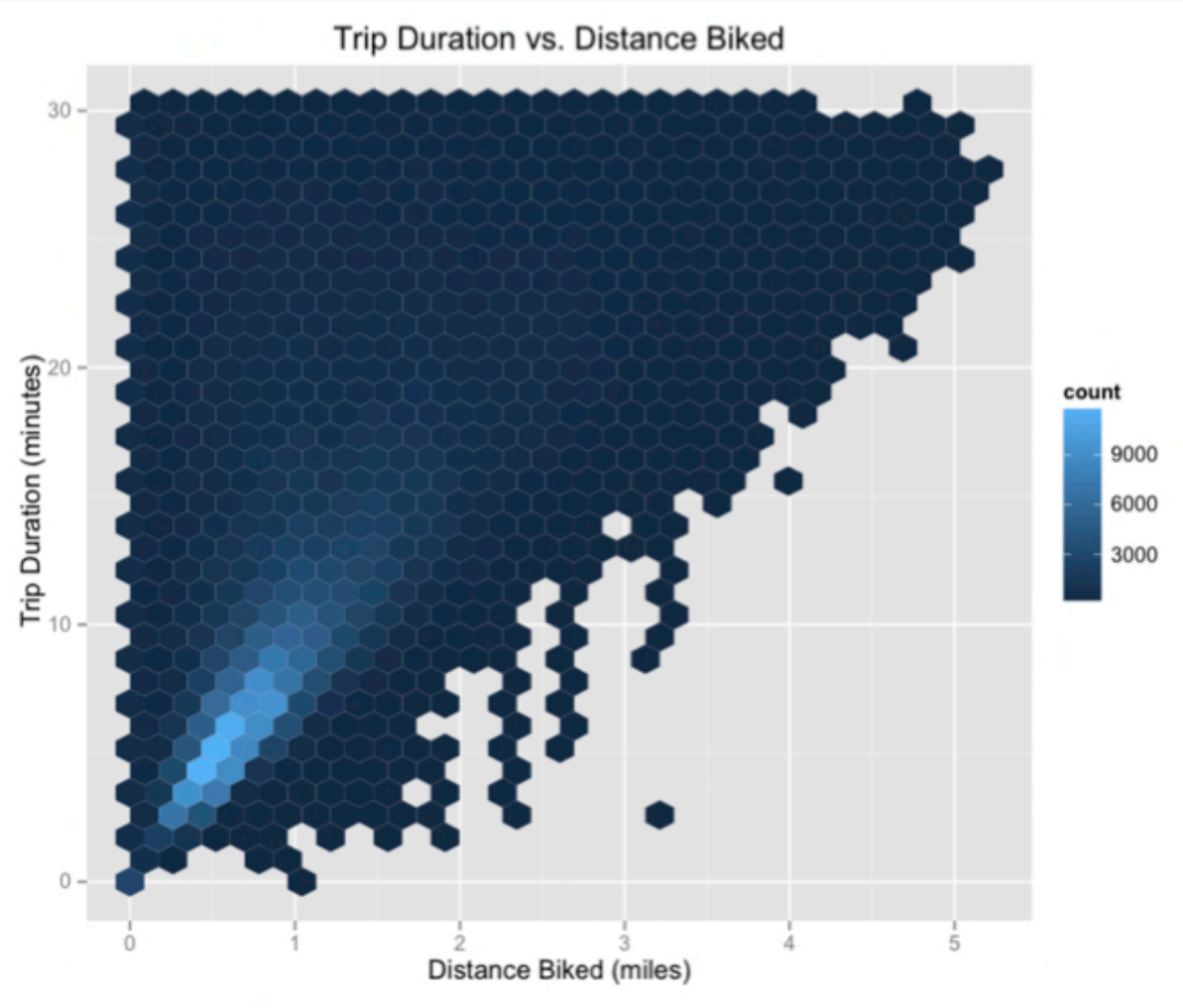
Breakout rooms and in-class exercises

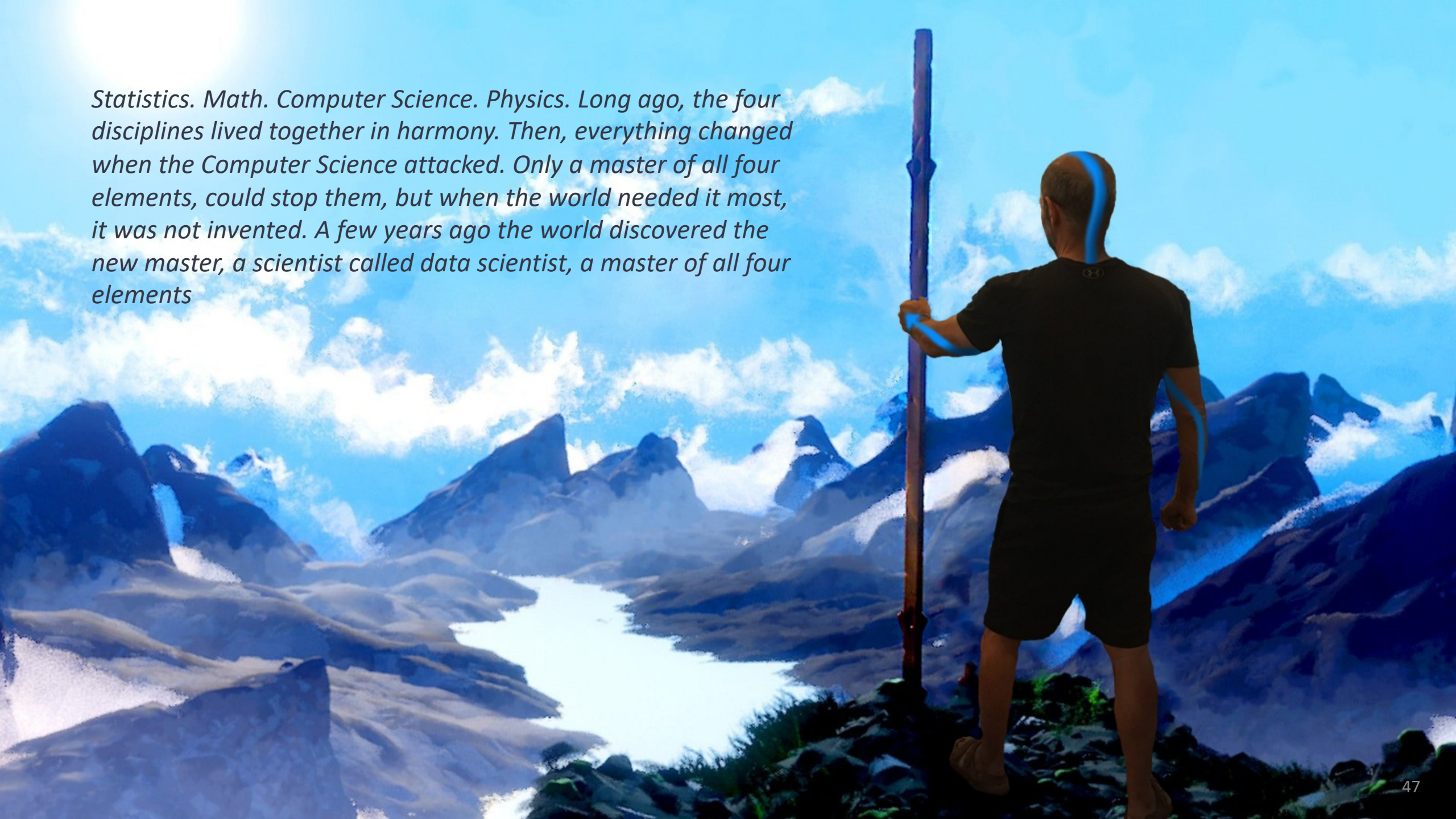
# Inspirations for Data Viz/Exploration

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**




*Statistics. Math. Computer Science. Physics. Long ago, the four disciplines lived together in harmony. Then, everything changed when the Computer Science attacked. Only a master of all four elements, could stop them, but when the world needed it most, it was not invented. A few years ago the world discovered the new master, a scientist called data scientist, a master of all four elements*