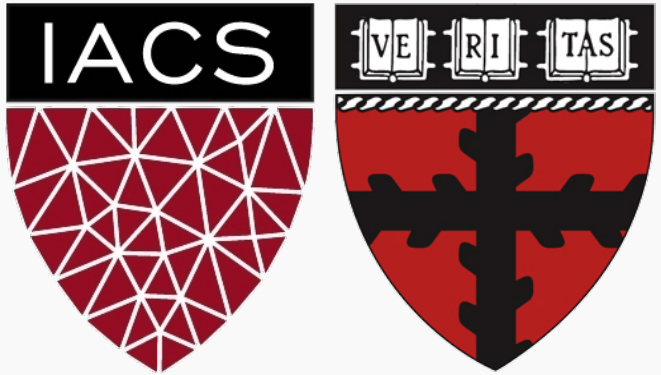
# Lecture #1: Introduction to CS109A

aka STAT121A, AC209A, CSCIE-109A

## CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai



# Lecture Outline

- Why data science?
- Why taking CS109A?
- What is data science?
- What is this class: who, how, what?
- Demo

# Lecture Outline

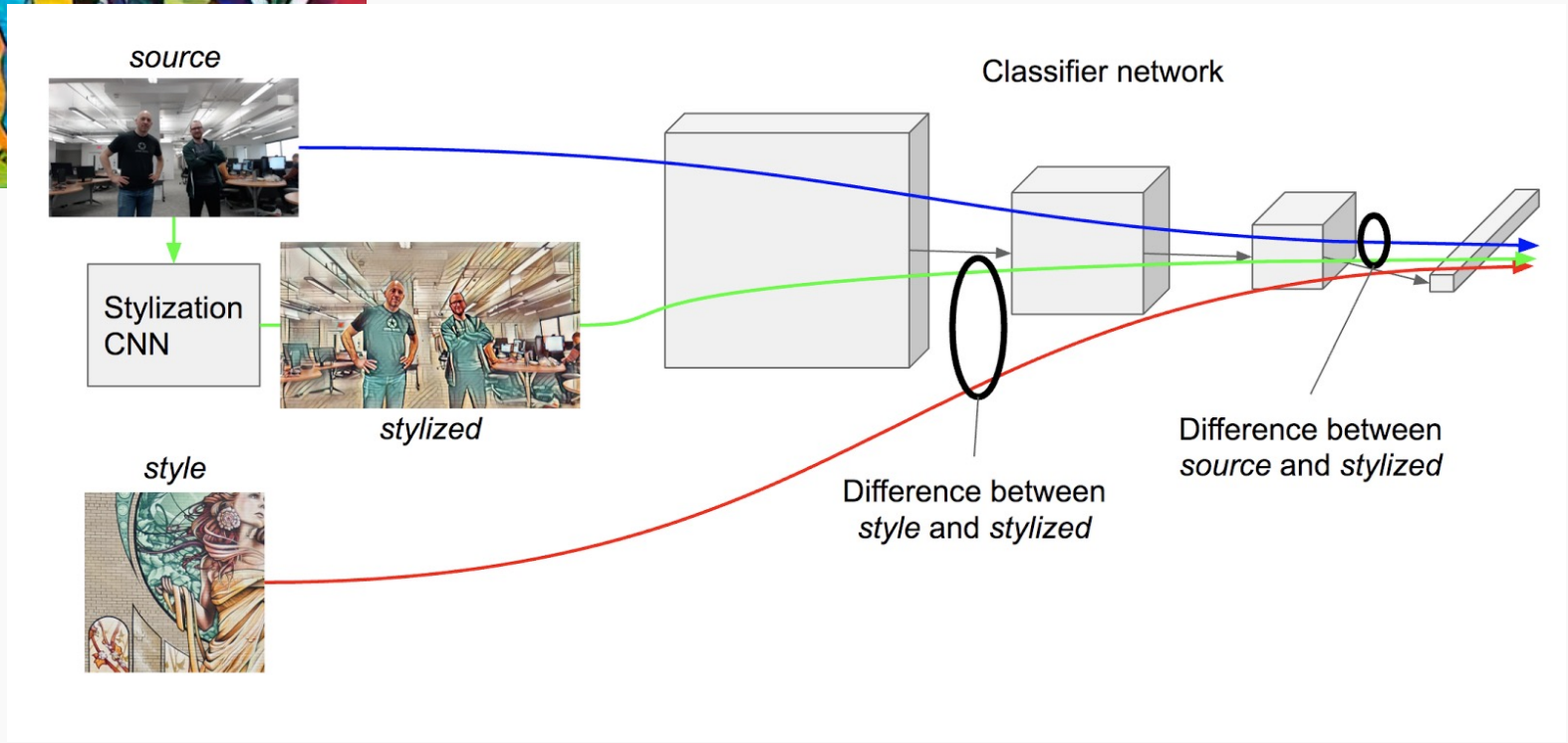
- Why data science?
- Why taking CS109A?
- What is data science?
- **What is this class: who, how, what?**
- Demo

# Why become an AI and Data Science expert?

# But if you decide to do it...

- It's a lot of fun!
- You will be at the cutting edge of research and product
- You will make lots of money doing something you will enjoy.
- It's not that hard to start and do!

# Minimise Loss



# Unsupervised Image-to-Image Translation

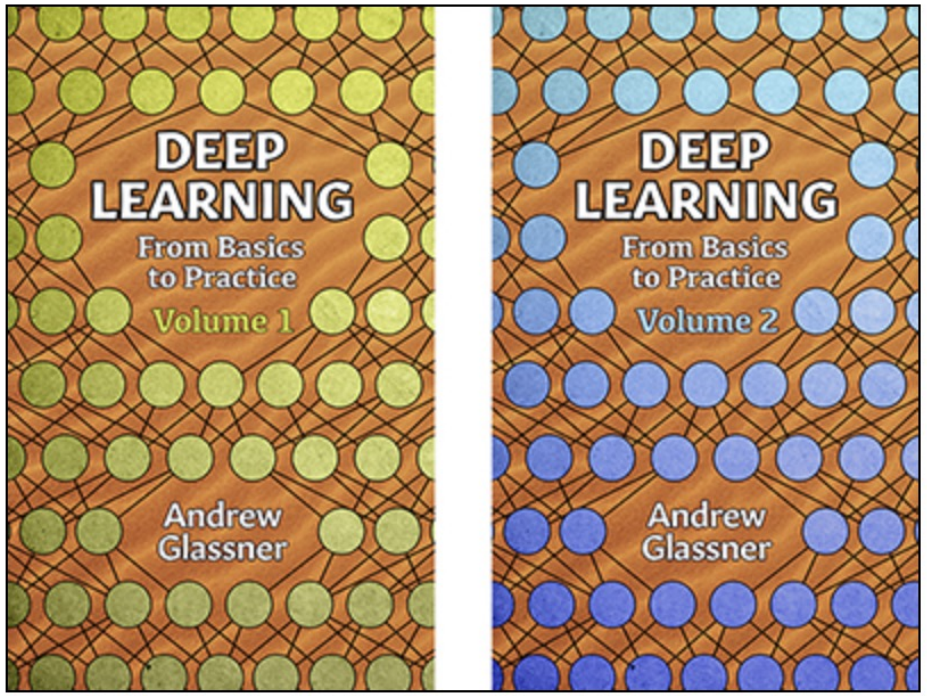
Day to night

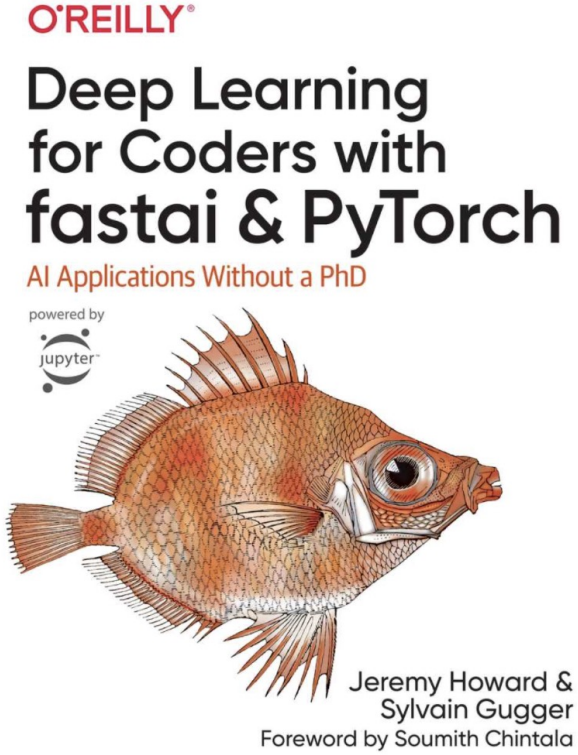
(Liu et al., 2017)

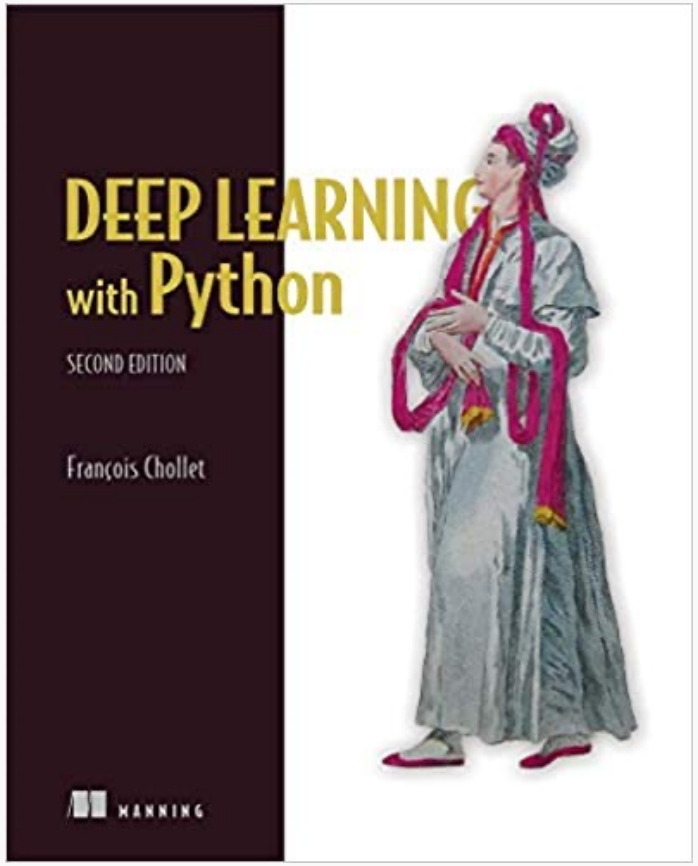




# Resources for learning




# Learn by Reading

Jay Alammar

Visualizing machine learning one concept at a time.

@JayAlammar on Twitter. YouTube Channel

Blog About

# explained.ai

Deep explanations of machine learning and related topics.

Website created by Terence Parr.

Terence is a professor of computer science and was founding director of the MS in data science program at the University of San Francisco. While he is best known for creating the ANTLR parser generator, Terence actually started out studying neural networks in grad school (1987). After 30 years of parsing, he's back to machine learning and really enjoys trying to explain complex topics deeply and in the simplest possible way. Follow `@the_antlr_guy`.

**Lil'Log** ⏳ Archive 🙋 FAQ 🦄 Contact

Jul 11, 2021 generative-model math-heavy

## What are Diffusion Models?

> Diffusion models are a new type of generative models that are flexible enough to learn any arbitrarily complex data distribution while tractable to analytically evaluate the distribution. It has been shown recently that diffusion models can generate high-quality images and the performance is competitive to SOTA GAN.

May 31, 2021 representation-learning long-read language-model

## Contrastive Representation Learning

> The main idea of contrastive learning is to learn representations such that similar samples stay close to each other, while dissimilar ones are far apart. Contrastive learning can be applied to both supervised and unsupervised data and has been shown to achieve good performance on a variety of vision and language tasks.

Mar 21, 2021 nlp language-model safety

## Reducing Toxicity in Language Models

# Learn by Watching

## DEEP LEARNING

**DS-GA 1008 · SPRING 2021 · NYU CENTER FOR DATA SCIENCE**

| | |
|---|---|
| INSTRUCTORS | Yann LeCun & Alfredo Canziani |
| LECTURES | Wednesday 9:30 – 11:30, Zoom |
| PRACTICA | Tuesdays 9:30 – 10:30, Zoom |
| FORUM | r/NYU_DeepLearning |
| DISCORD | NYU DL |
| MATERIAL | 2021 repo |

### 2021 edition disclaimer

Check the repo's `README.md` and learn about:

- Content new organisation
- The semester's second half intellectual dilemma
- This semester repository
- Previous releases

### Lectures

Full Stack Deep Learning | Search | GitHub ☆208 58

Home | Spring 2021 | Fall 2019

**Spring 2021**

Spring 2021 Schedule

✨Course Projects Showcase✨

**Lectures**

Lecture 1: DL Fundamentals

Notebook: Coding a neural net

Lecture 2A: CNNs

Lecture 2B: Computer Vision

Lecture 3: RNNs

Lecture 4: Transformers

Lecture 5: ML Projects

Lecture 6: MLOps Infrastructure & Tooling

Lecture 7: Troubleshooting Deep Neural Networks

Lecture 8: Data Management

Lecture 9: AI Ethics

Lecture 10: Testing & Explainability

## Full Stack Deep Learning - Spring 2021

We've updated and improved our materials for our 2021 course taught at UC Berkeley and online.

**Synchronous Online Course**

We offered a **paid synchronous option** for those who wanted weekly assignments, capstone project, Slack discussion, and certificate of completion.

Enter your email below or follow us on **Twitter** to be the first to hear about future offerings of this option.

And check out the **✨course projects showcase✨**.

email address | Subscribe

### Week 1: Fundamentals

**Table of contents**

Week 1: Fundamentals

Week 2: CNNs

Week 3: RNNs

Week 4: Transformers

Week 5: ML Projects

Week 6: Infra & Tooling

Week 7: Troubleshooting

Week 8: Data

Week 9: Ethics

Week 10: Testing

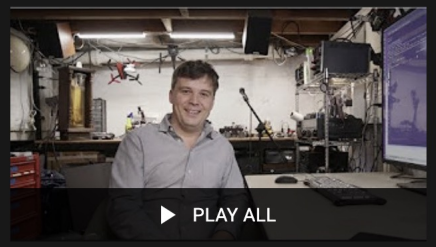
Week 11: Deployment

Week 12: Research

Week 13: Teams

✨Week 14-16: Projects✨

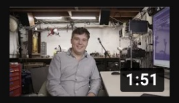
Other Resources




# Introduction to Machine Learning

12 videos • 21,804 views • Last updated on Apr 16, 2019

Weights & Biases SUBSCRIBE

1 


**Intro to ML: Course Overview**

Weights & Biases

2 

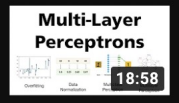

**0. What is machine learning?**

Weights & Biases

3 


**1. Build Your First Machine Learning Model**

Weights & Biases

4 


**2. Mu**

Weights

5 


**3. Cor**

Weights

# Yannic Kilcher

94.3K subscribers

SUBSCRIBE

HOME VIDEOS PLAYLISTS COMMUNITY CHANNELS ABOUT

Uploads PLAY ALL SORT BY


**[ML News] Facebook AI adapting robots | Baidu...**

6.1K views • 1 day ago

CC


**I'm taking a break**

9.2K views • 5 days ago


**[ML News] GitHub Copilot - Copyright, GPL, Patents &...**

14K views • 1 week ago

CC


**Self-driving from VISION ONLY - Tesla's self-driving...**

23K views • 1 week ago

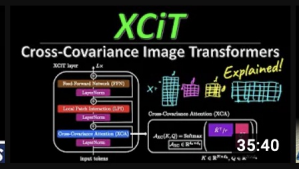

**[ML News] CVPR bans social media paper...**

10K views • 2 weeks ago

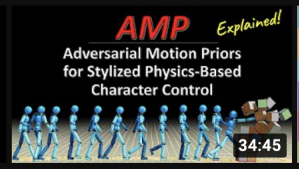
CC


**The Dimpled Manifold Model of Adversarial...**


**[ML News] Hugging Face course | GAN Theft Auto | ...**




**XCiT: Cross-Covariance Image Transformers...**




**AMP: Adversarial Motion Priors for Stylized Physics-...**


**[ML News] De-Biasing GPT-3 | RL cracks chip design |...**

# 50 Best Jobs in America for 2020

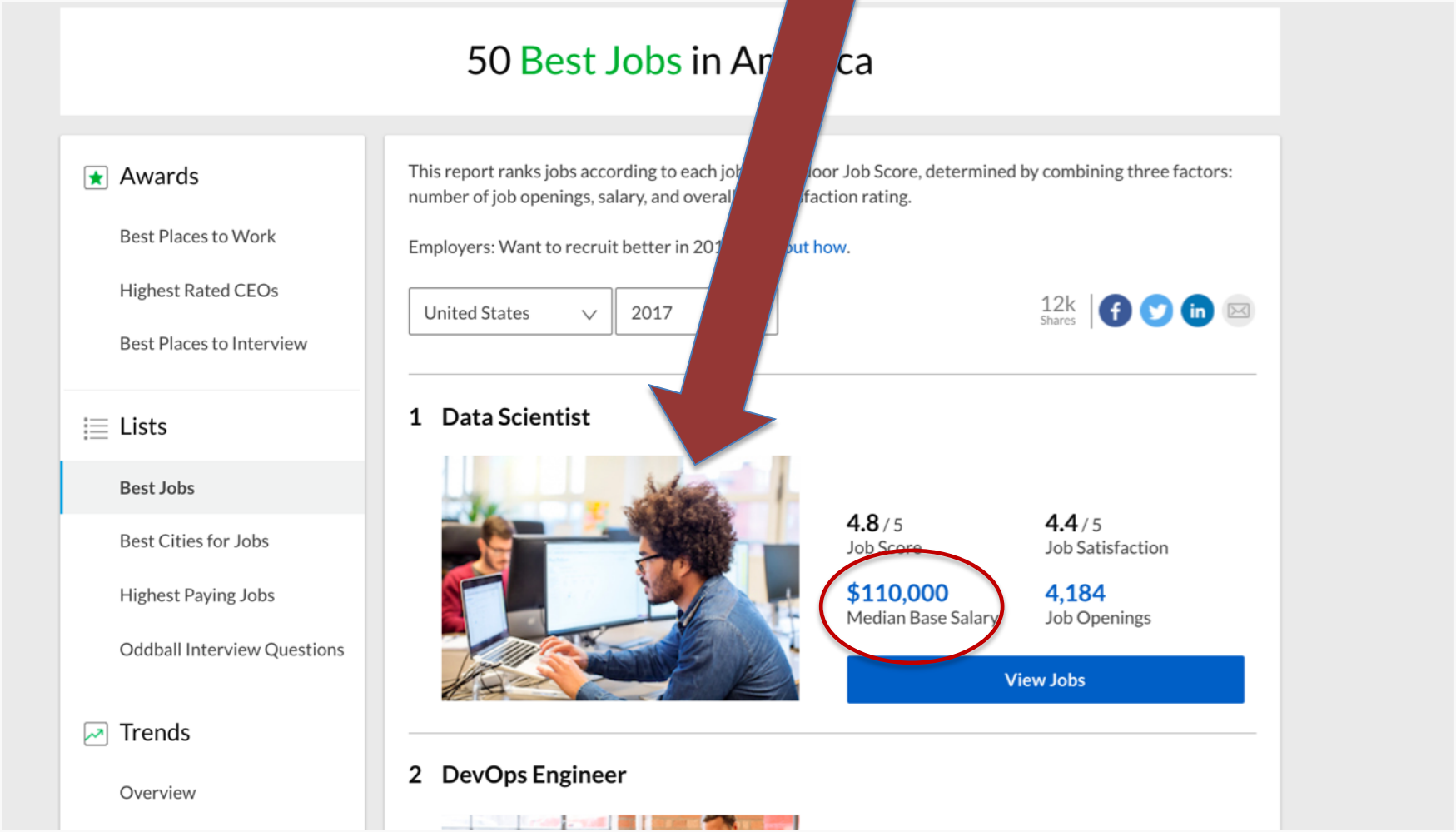
Best Jobs | 2020 | United States

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?

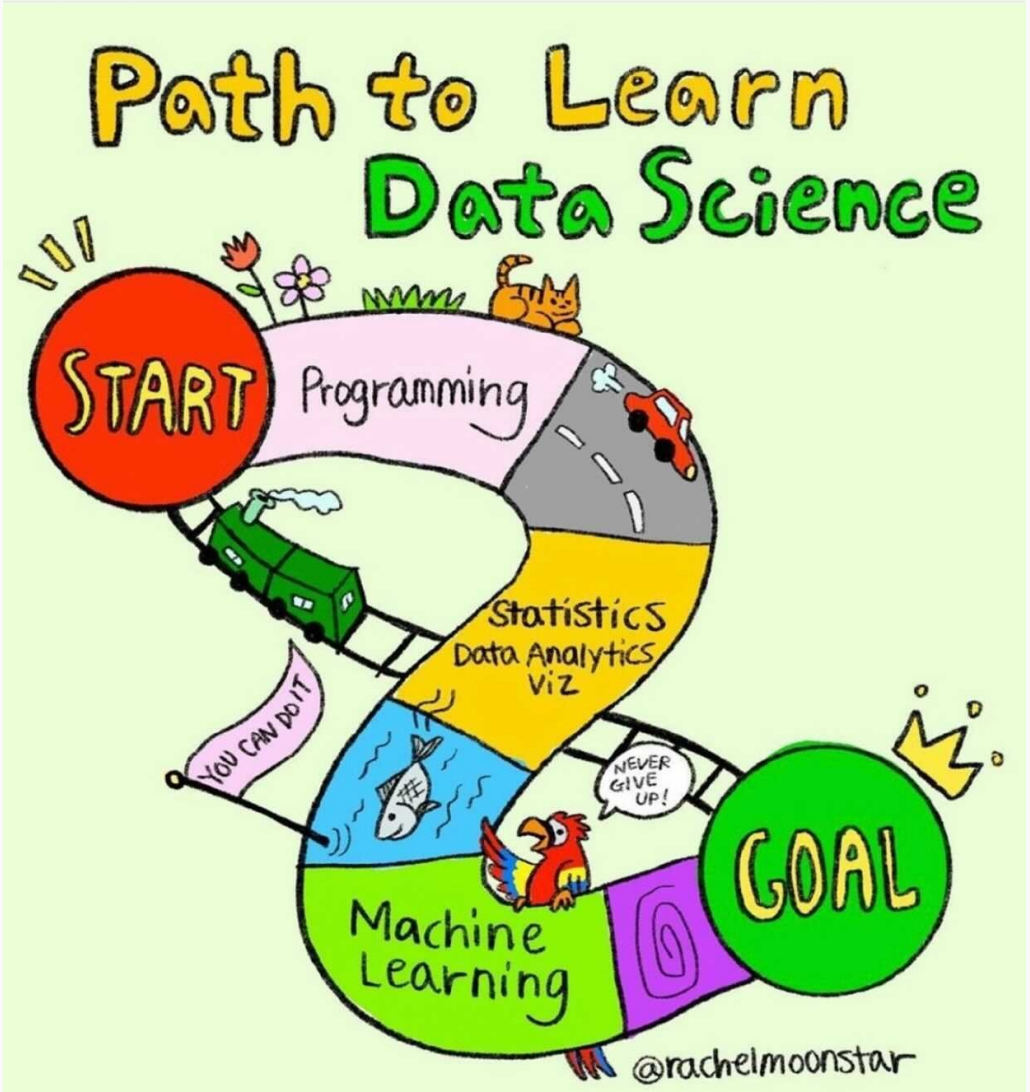
## Jobs!


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

# Lecture Outline

- Why data science?
- **Why taking CS109A?**
- What is data science?
- What is this class: who, how, what?
- Demo

# Memes!


Path to Learn
Data Science
START
Programming
Statistics
Data Analytics
Viz
YOU CAN DO IT
NEVER GIVE UP!
Machine Learning
GOAL
@rachelmoonstar


MOVIECLIPS.com

# Why?

Why are you here?

# Lecture Outline

- Why data science?
- Why taking CS109A?
- **What is data science?**
- What is this class: who, how, what?
- Demo

# 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 ago: theoretical approaches, try to derive equations to describe general phenomena.

$$F = G\frac{m_1 m_2}{d^2}$$

$$i\hbar\frac{\partial}{\partial t} - \Psi = \hat{H}\Psi$$

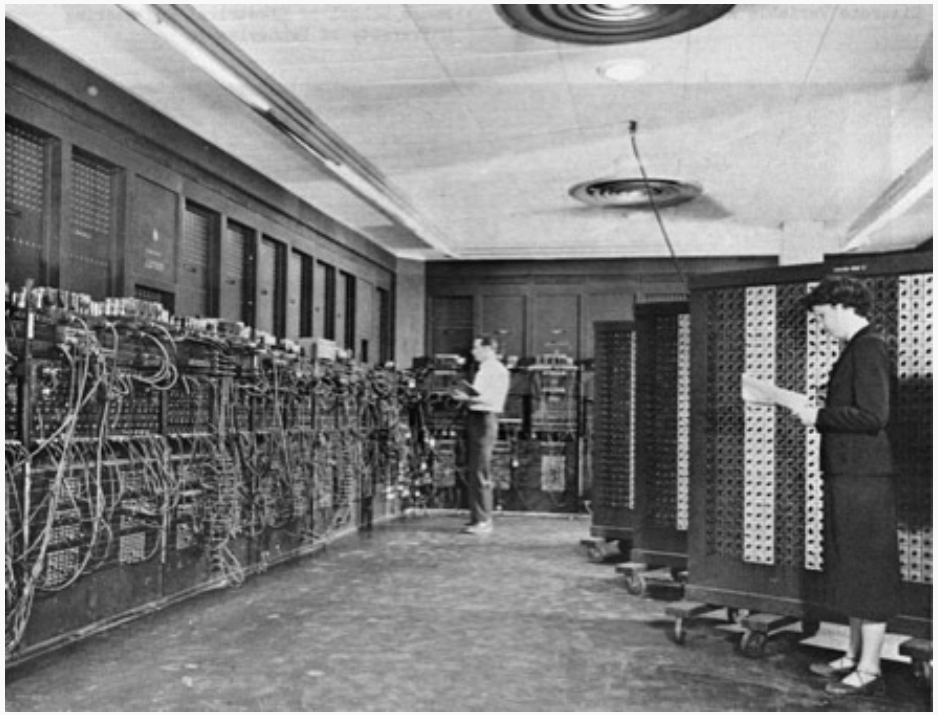
$$\nabla \cdot E = 0 \quad \nabla \times E = -\frac{1}{c}\frac{\partial H}{\partial t}$$
$$\nabla \cdot H = 0 \quad \nabla \times H = \frac{1}{c}\frac{\partial E}{\partial t}$$

$$E = mc^2$$

$$\rho\left(\frac{\partial v}{\partial t} + v \cdot \nabla v\right) = -\nabla p + \nabla \cdot T + f$$

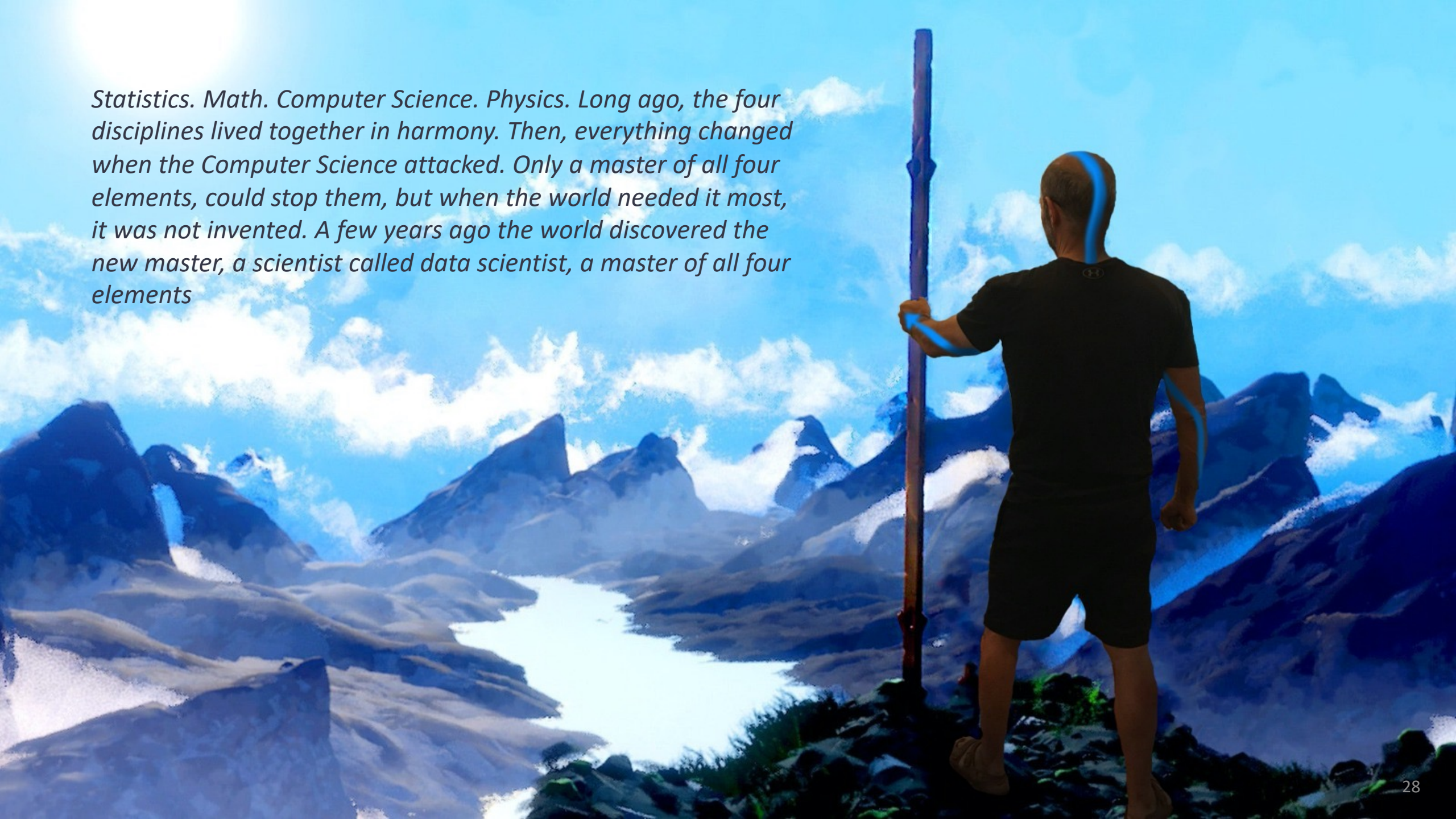
# History (cont.)

About a hundred years ago: computational approaches appeared



# History (cont.)

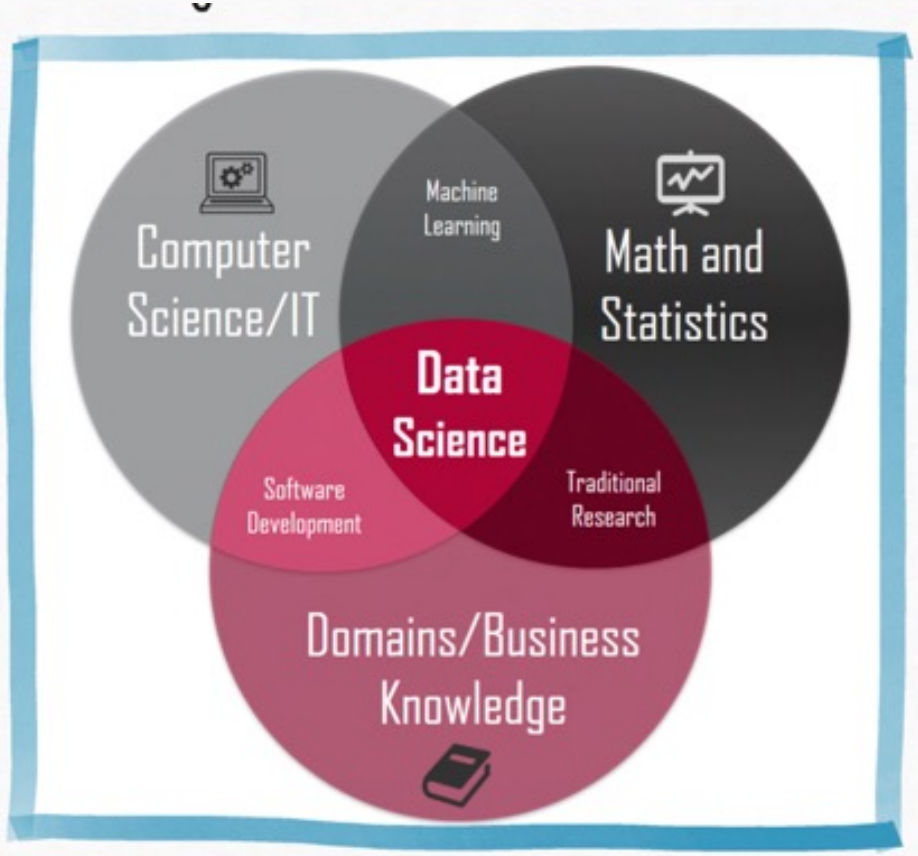
And then it is data science


*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*

# History (cont.)

And then it was data science

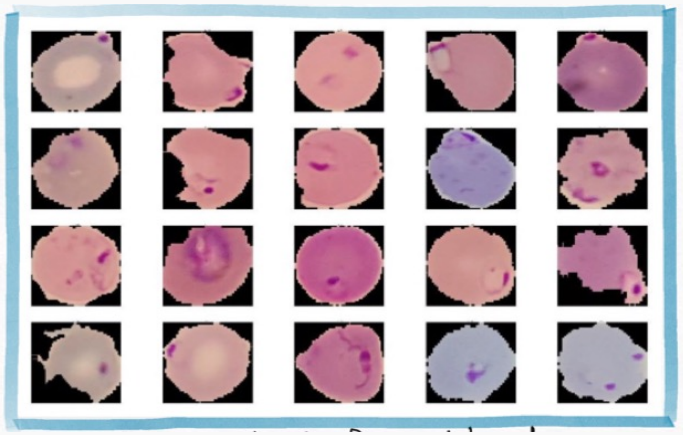
In both data science and machine learning we extract pattern and insights from data.




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

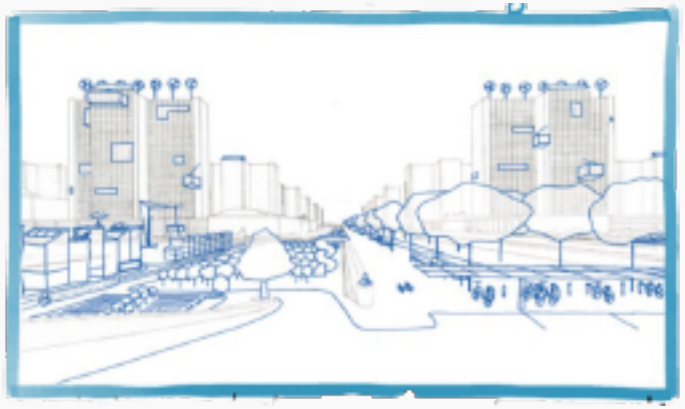
# The Potential of Data Science

## 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

# Lecture Outline

- Why data science?
- Why taking CS109A?
- **What is data science?**
- What is this class: who, how, what?
- Demo

# What?

## The Data Science Process

Ask an interesting question

Get the Data

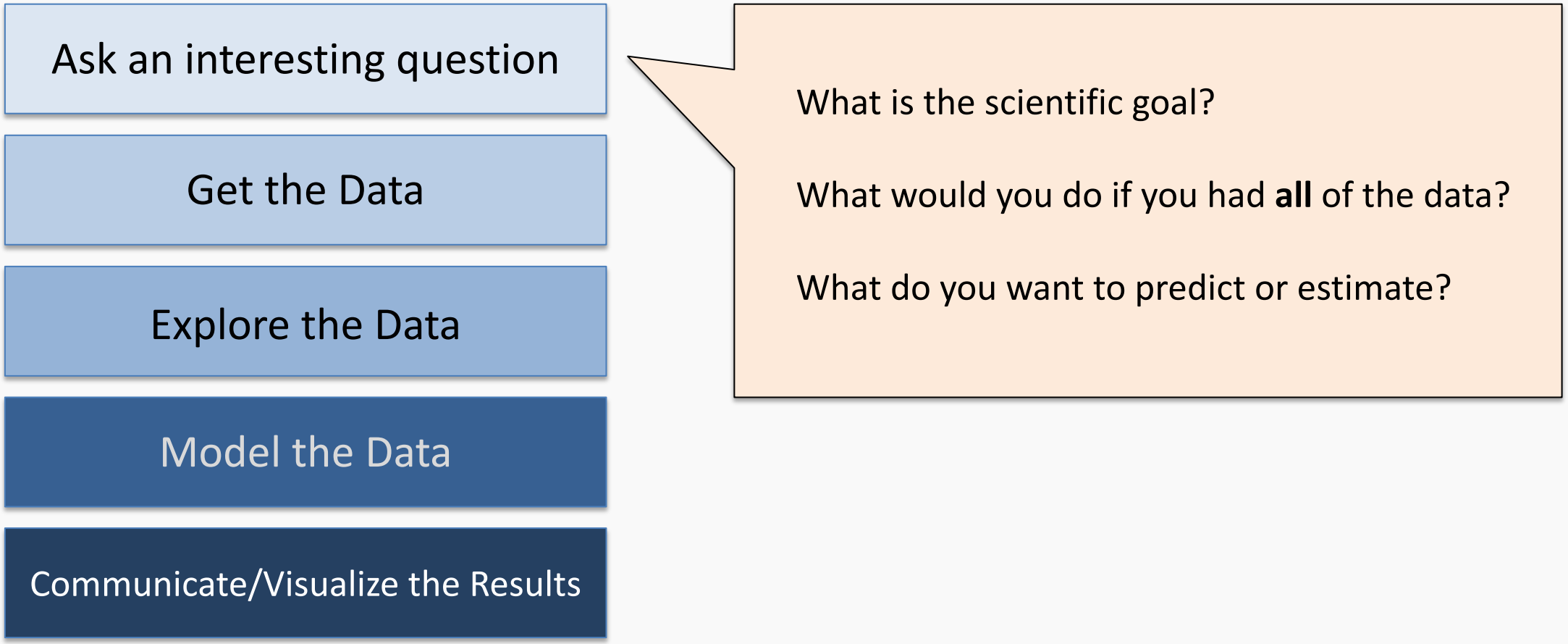
Explore the Data

Model the Data

Communicate/Visualize the Results

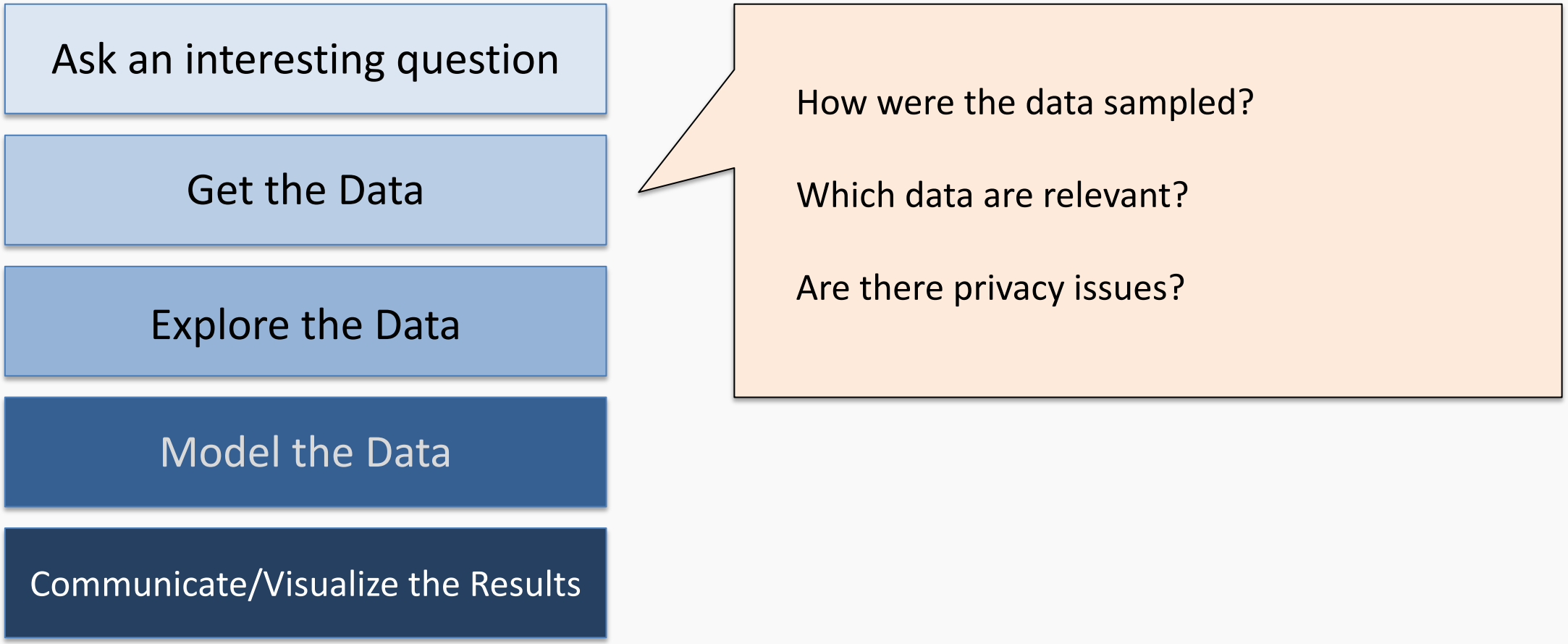
# What?

## The Data Science Process



# 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

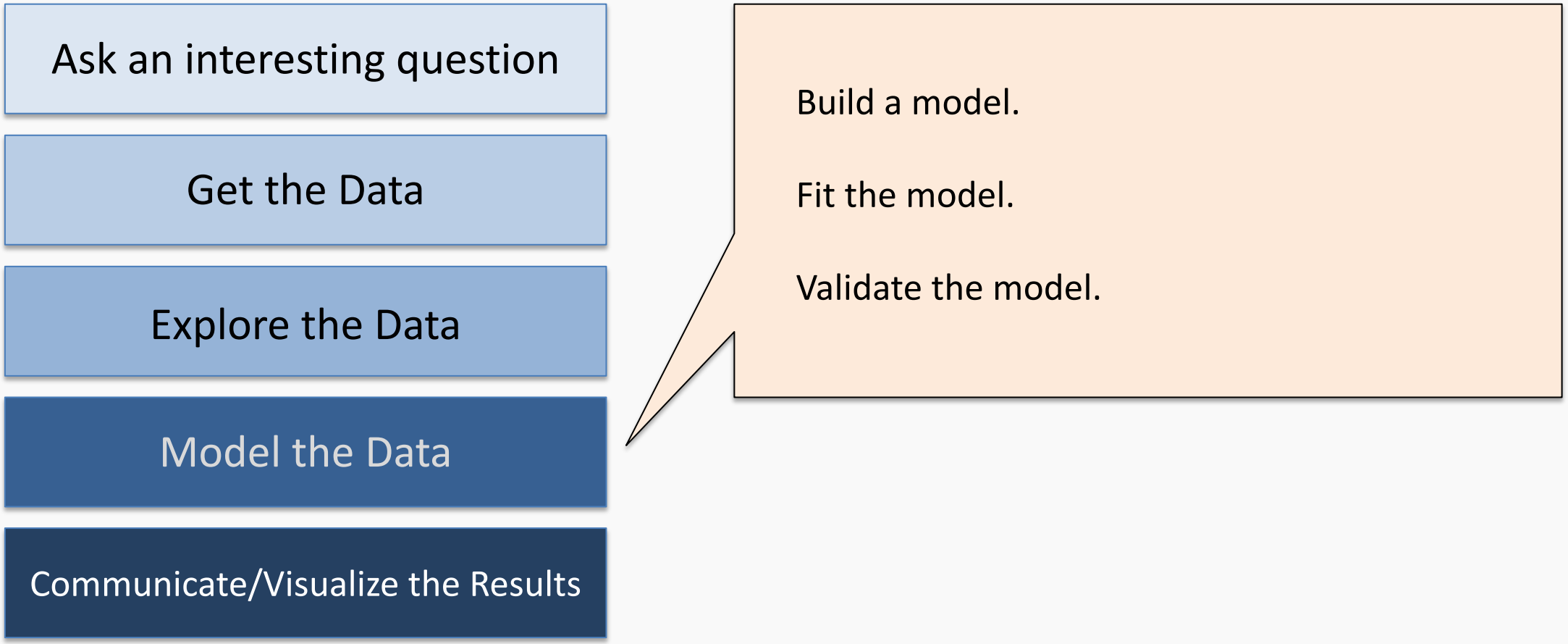
Plot the data.

Are there anomalies or egregious issues?

Are there patterns?

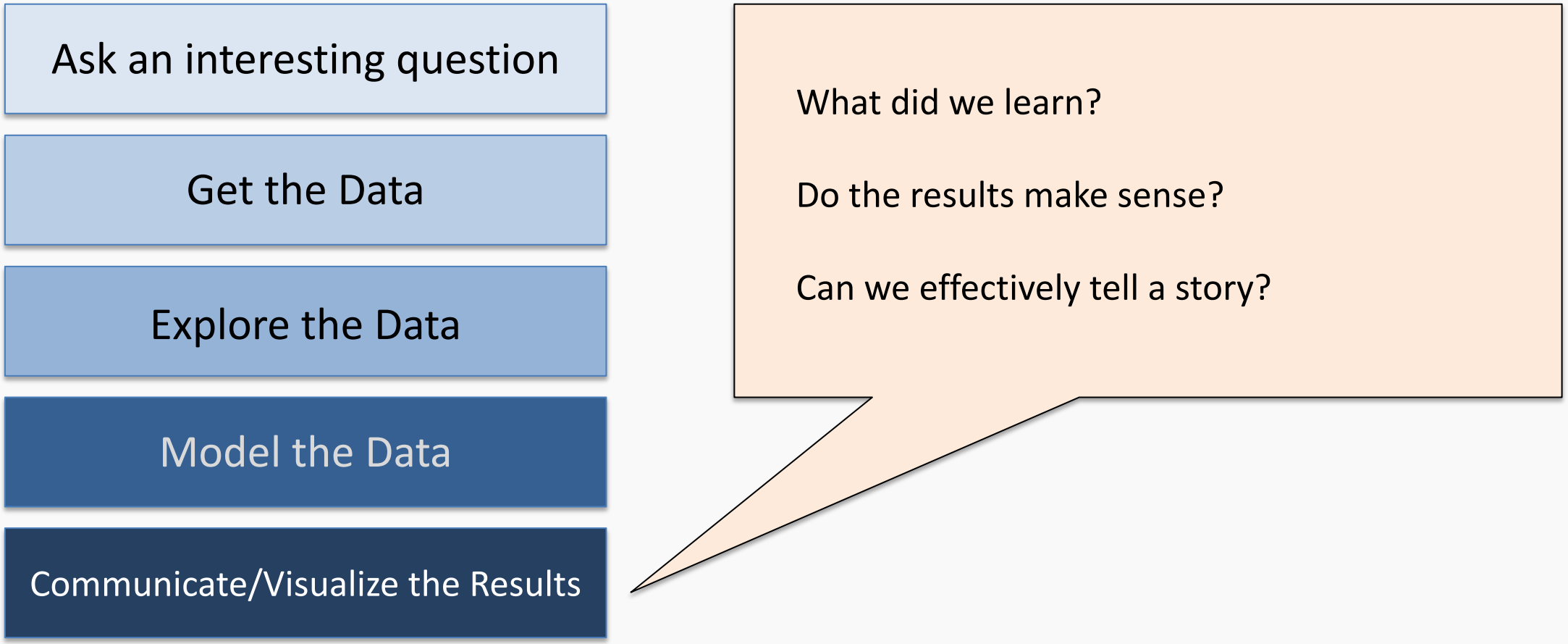
# What?

## 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.

# Goals of the course

## Theory

1. Key Machine Learning concepts
2. Important metrics for evaluation
3. 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
3. Handling different kinds of data

## 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: Data Issues

Data Imputation
PCA

## Weeks 7: Data Issues

Visualization
Ethics

## Weeks 9: Classification

kNN Classification
Logistic Regression
Multi-class Classification

## Weeks 10-11: Trees

Decision Trees
Bagging
Random Forest
Boosting Methods

## Week 13

Ethics
Model Interpretation

## Weeks 14-15: NLP

Language models
Tokenization
N-grams, tf-idf

# After CS109A

## CS109B

A. Neural Networks:
- MLP
- CNNs
- RNNs
- Generative models
- Deep RL

B. Unsupervised Clustering

C. Bayesian Modeling

## AC215

A. Productionize Data Science, from notebooks to the cloud

B. Big models, transfer learning and architecture learning

C. Design and Development

D. Deployment, Scaling, & Automation

# Not an exclusive list

- CS171/CS271 (Visualization)
- CS181 (ML)
- CS18A (AI)
- CS 187 (NLP)
- Stat 110 (Probability)
- Stat 111 (Inference)
- Stat 139 (Linear Models)
- Stat 149 (Generalized Linear Models)
- Stat 131 (Time Series)
- Stat 171 (Stochastic Processes)
- Stat 195 (Statistical Machine Learning).
- CS208 (Privacy)
- CS282R (ML: Generative Models)
- CS282BR (Sequential Learning)
- AC295/CS287 (DL for NLP)

# Who? Instructors

**Pavlos Protopapas**

Scientific director
Institute of Applied Computational Science

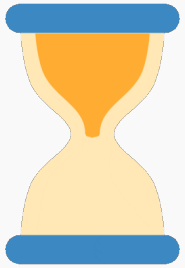
Principle Investigator of StellarDNN, a research lab within IACS/SEAS. Research in the intersection of astronomy, ML and statistics. Recently he is interested in solving differential equations for physical systems using deep NN, inference in DNN, and applying NLP techniques in astronomical time series analysis.

He loves classical music and opera, and he often visits the BSO.

A certified cook from *Le Cordon Bleu*, loves eating as much as cooking.

Funny fact: During a failed military service he was declared the worst soldier in NATO.

tiktok: @pavlosprotopapas



# Digestion Time

# Who? Instructors

**Natesh Pillai**
Professor of Statistics

He graduated from Duke University in 2008 and did his post-doctoral research at Warwick University.

His interests are the interface of applied probability and statistics, with a particular research focus on climate.

Natesh is also part of the Harvard Data Science Initiative. He was awarded the young scientist award by the International Indian Statistical Association in 2018. He is currently an Amazon Scholar. Prior to that, he was a chief scientist at Correlation One, where he developed a data science curriculum for professionals and trained a few cohorts of students across the world.

In his free time, he dabbles in chess.

# Who?

**Marios Mattheakis**

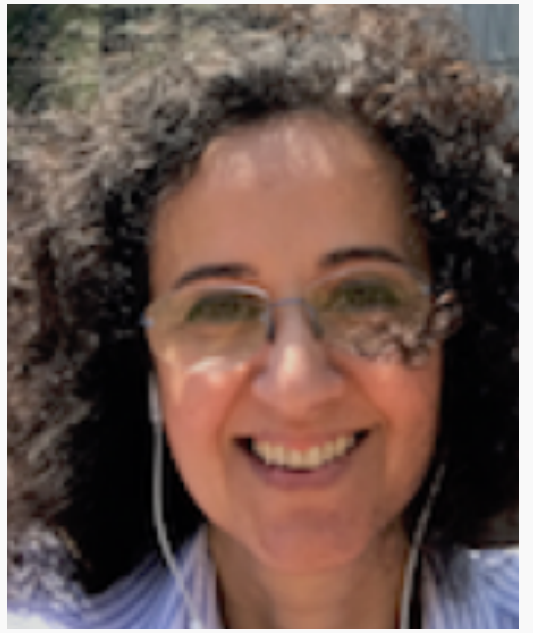
Lab Leader

Post-doctoral Fellow
IACS

**Chris Gumb**

Head TF

Graduate student of Data Science at Harvard Extension School



**Eleni Kaxiras**

Lab Instructor

Assistant Director for Data Science and Computation at SEAS

# Who? Teaching Fellows

| | |
|---|---|
| Kamran Ahmed | Nabib Ahmed |
| Tale Lokvenec | Yuen Ting Chow |
| Diego Zertuche | Javier Machin |
| Mark Penrod | Mike Sedelmeyer |
| Henry Jin | Joel Zhang |
| Varshini Reddy | Vivek Bhatia |
| Shuheng Liu | Kacper Krasowiak |
| Hayden Joy | Moni Radev |
| Tao Tsui | Vlad Ivanchuk |
| Angela Garabet | Abhishek Malani |
| Patrick DeKelly | Aqdas Kamal |

# Course Components

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

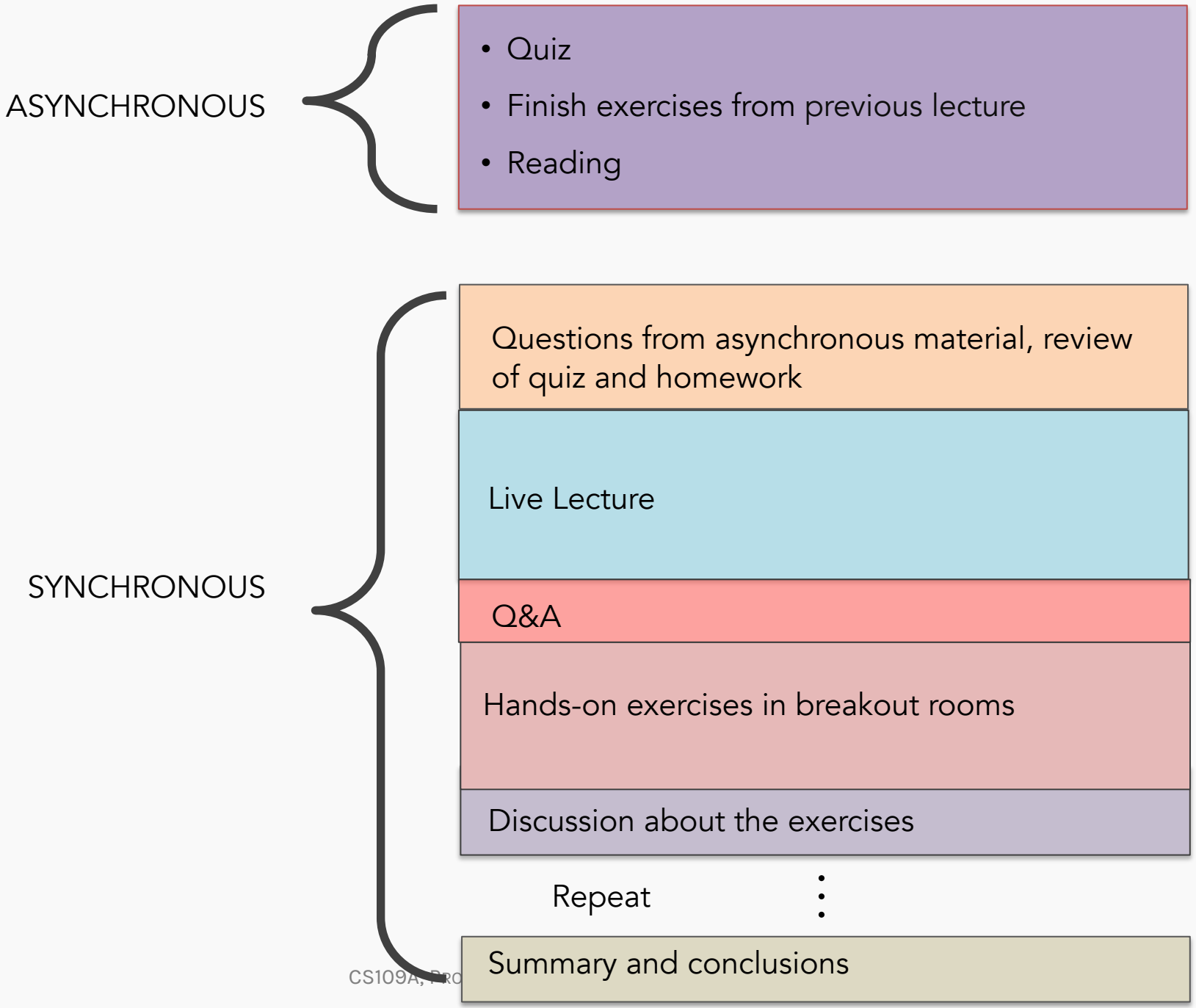
During lecture will cover the material which you will need to complete the homework, and to survive the rest of your life in CS109A.

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.

Mon/Wed 9:45-11:00am in person @ SEC 1.321 and @Zoom for Extension School Students (zoom link is on canvas under zoom).

# Lecture format



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

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

Weds 12:45-2pm  pm @TBD. A-sections are required for AC209 students.

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

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

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

Weds 1:00-1:15 pm @TBD. A-sections are required for AC209 students.

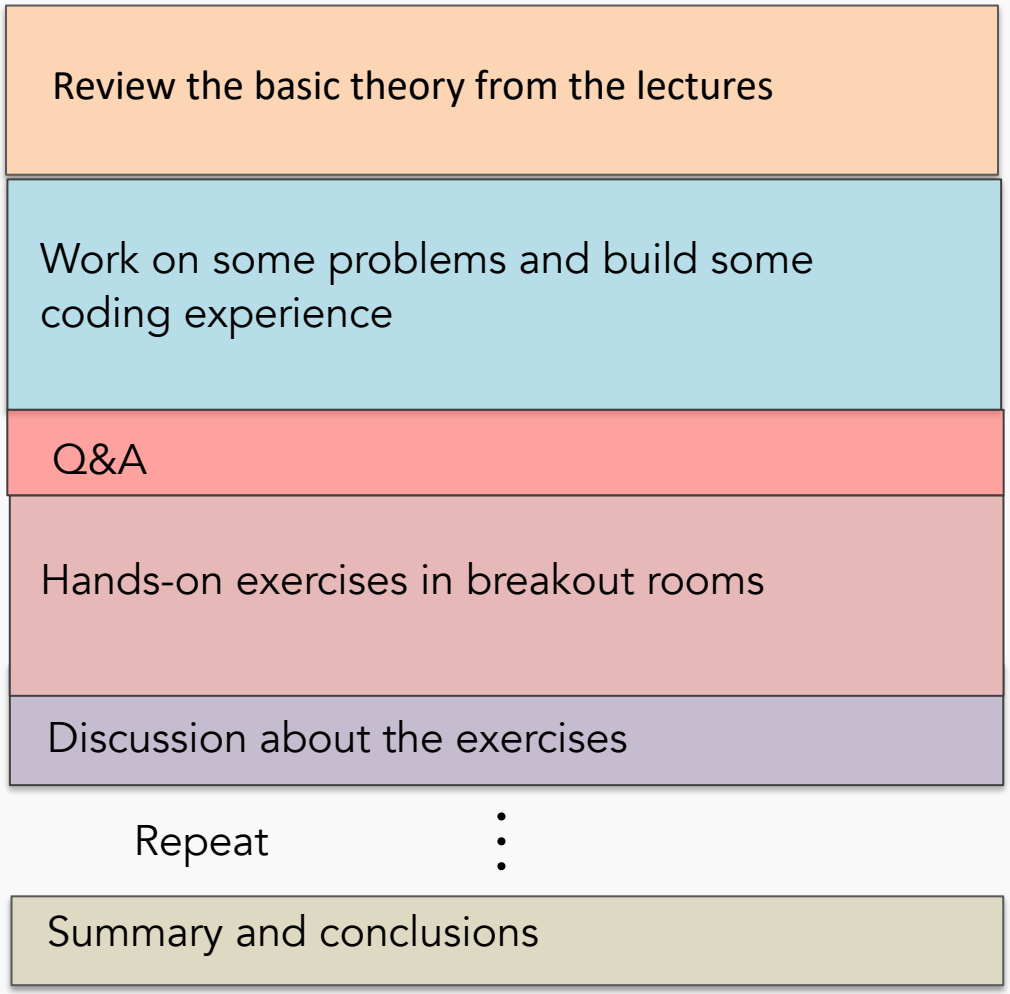
Note: Sections are not held every week. Consult the course calendar for exact dates.

*Labs* will be a mix of review of material and practice problems like the homework.

Friday 9:45-11:00am in person @ SEC 1.321 and @Zoom for Extension School

Students' attendance at labs is required

# Lab format


Review the basic theory from the lectures
Work on some problems and build some coding experience
Q&A
Hands-on exercises in breakout rooms
Discussion about the exercises
Repeat
Summary and conclusions

# **Advanced Sections** topics

**Topics**

1. Linear Algebra and Hypothesis Testing: The Short Versions
2. Methods of regularization and their justifications
3. Mathematics of PCA
4. Generalized Linear Models
5. Ensemble methods
6. Advanced Experimental Design

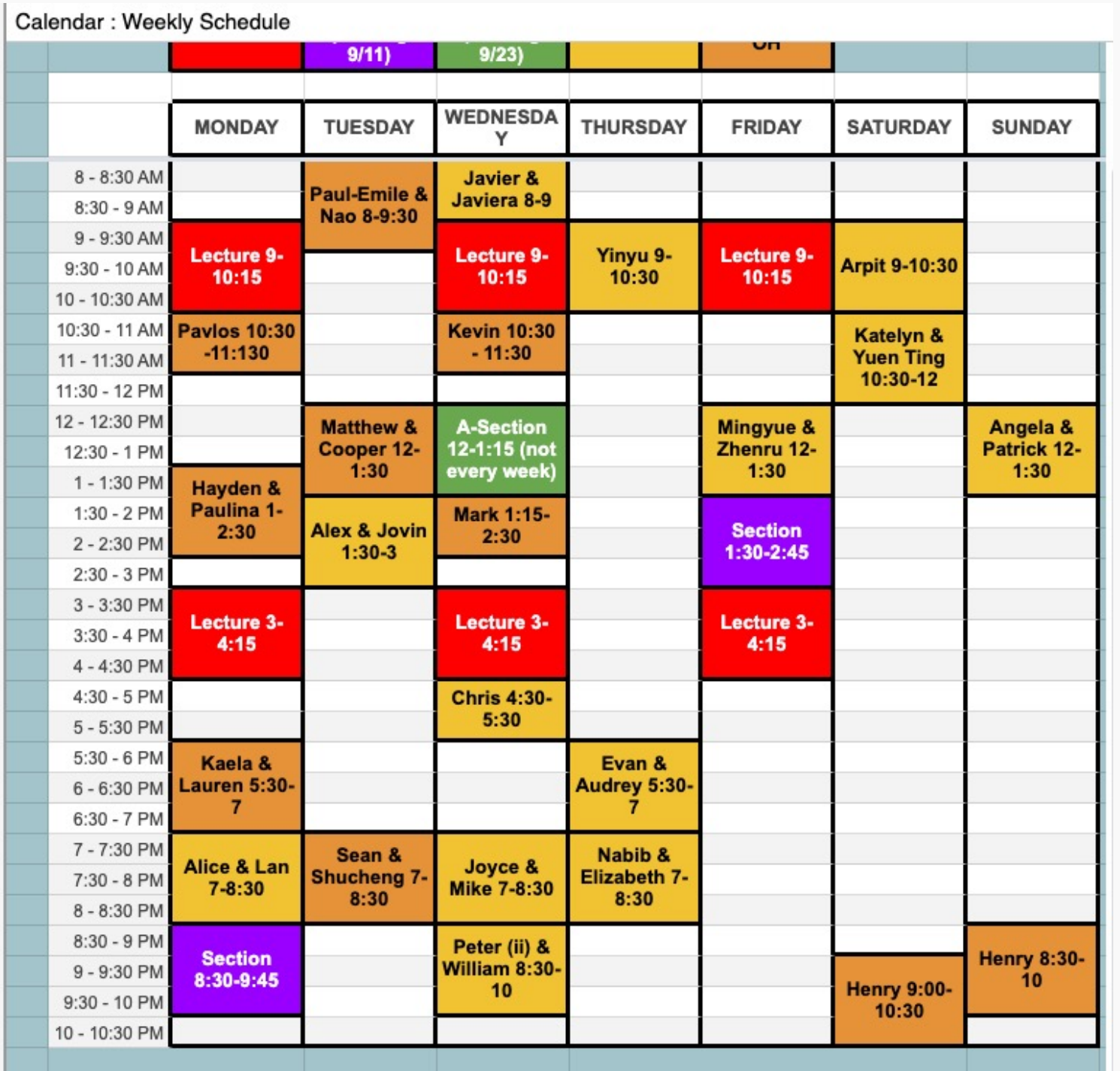
**NOTE 1**: The materials in the *Advanced Sections* are required for all AC 209A students. There will be one extra question in most homework for AC 209 students which will be based on the A-Section materials.

**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 (TBD)

Expect something like this in the next few days



# Assignments

# Five Graded Components

## Homework: 52%

Homework zero: 1%
Individual Homework (2): 16%
Paired Homework (6): 35%

HW4 and HW7 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.

## Midterm: 10%

A mix of multiple choice and coding questions.

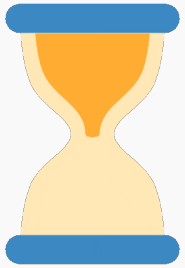
## Projects: 26%

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 9)
- Homework 1: Web scraping, Beautiful Soup
- Homework 2: Regression kNN and LinReg
- Homework 3:  Multi-regression, polynomial reg and model selection
- Homework 4*: Regularization, inference
- Homework 5: High Dimensional Data and PCA
- Homework 6: Logistic Regression
- Homework 7*: Random Forest, Boosting and Neural Networks
- Homework 8: Ethics and model interpretation



# Digestion Time

# 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 11:59.59 pm Wednesdays,** and homework will be released on Wednesdays.

Late submission policy: Each student is allowed up to 3 late days over the semester with at most 1 day applied to any single homework. Outside of these allotted late days, late homework will **not be accepted**.

# Final Project

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

- We will provide seven (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 October 10th.

# Help

The process to get help is:

1. Post the question in *Edstem*, 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: **cs109a2021@gmail.com**. All the instructors and TFs monitor the Helpline.
4. For personal matters, send an email to Pavlos, or Natesh.

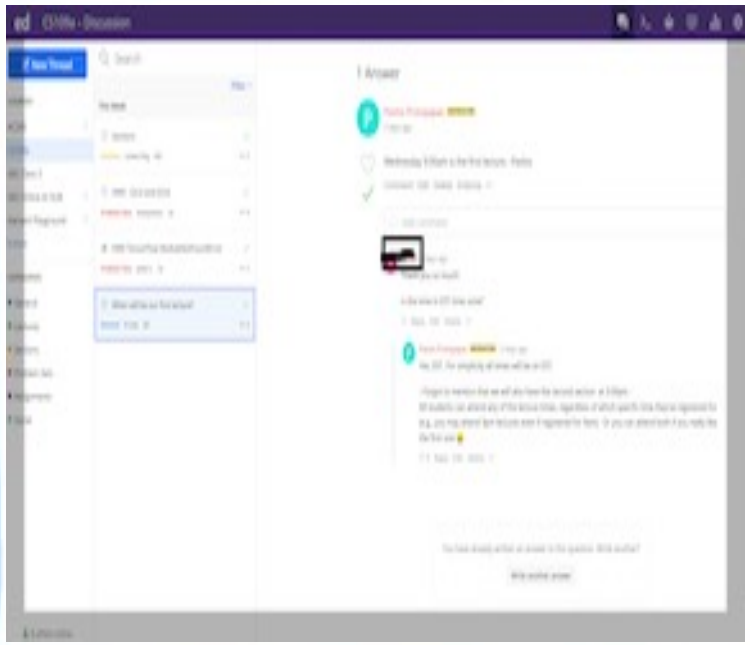
**Sundays will be slow days, so please be patient!**

# Tools for the course

## Web page

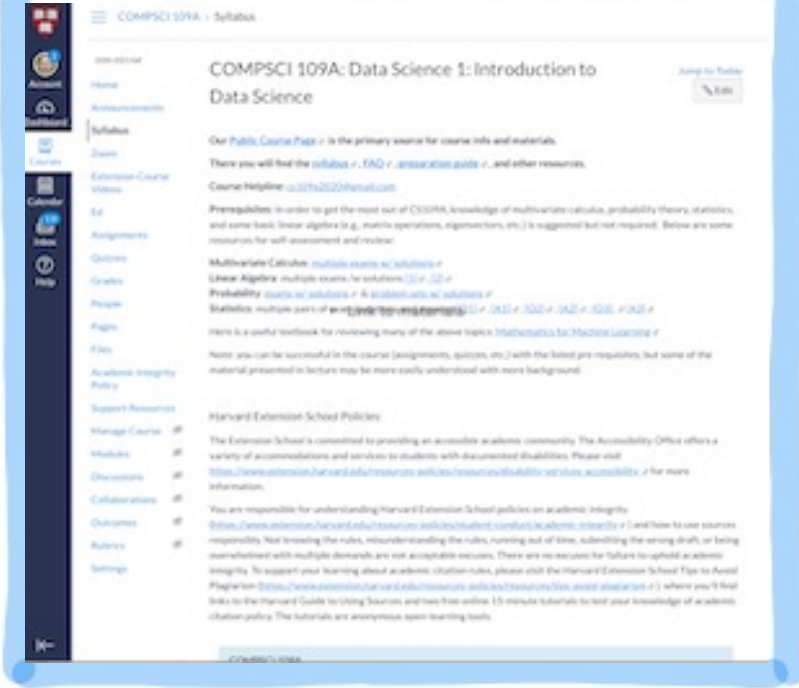

- Syllabus
- Calendar
- Link to materials

## edstem

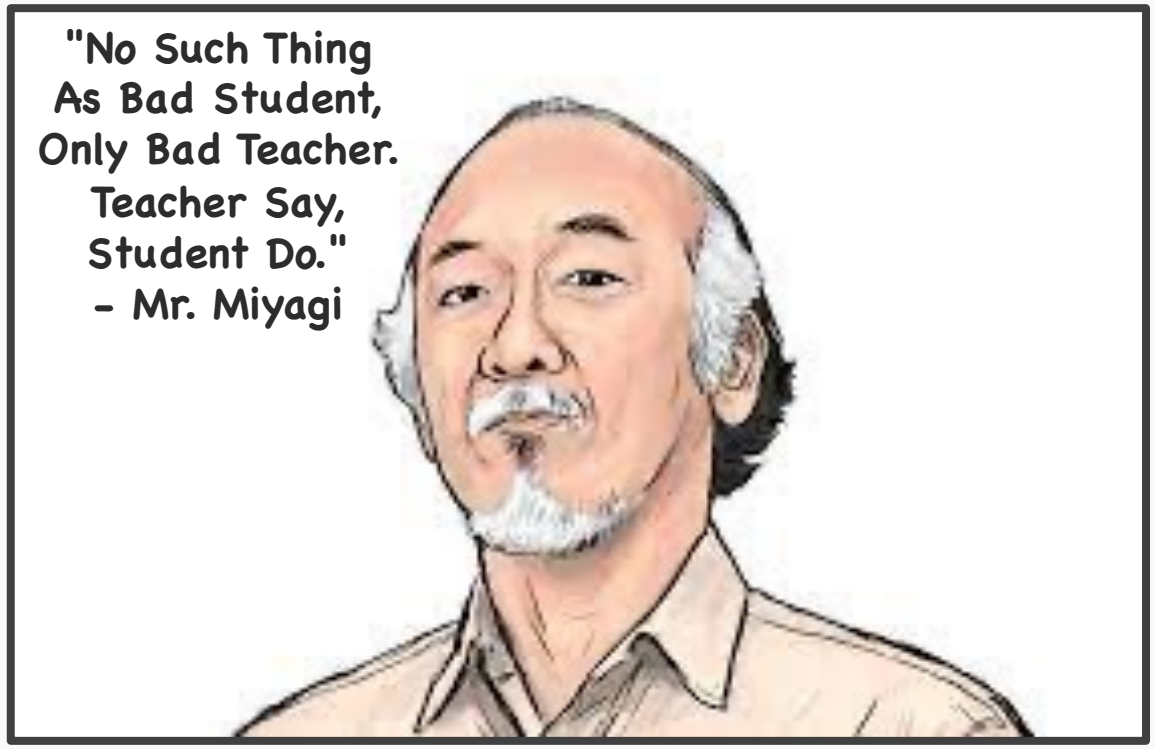


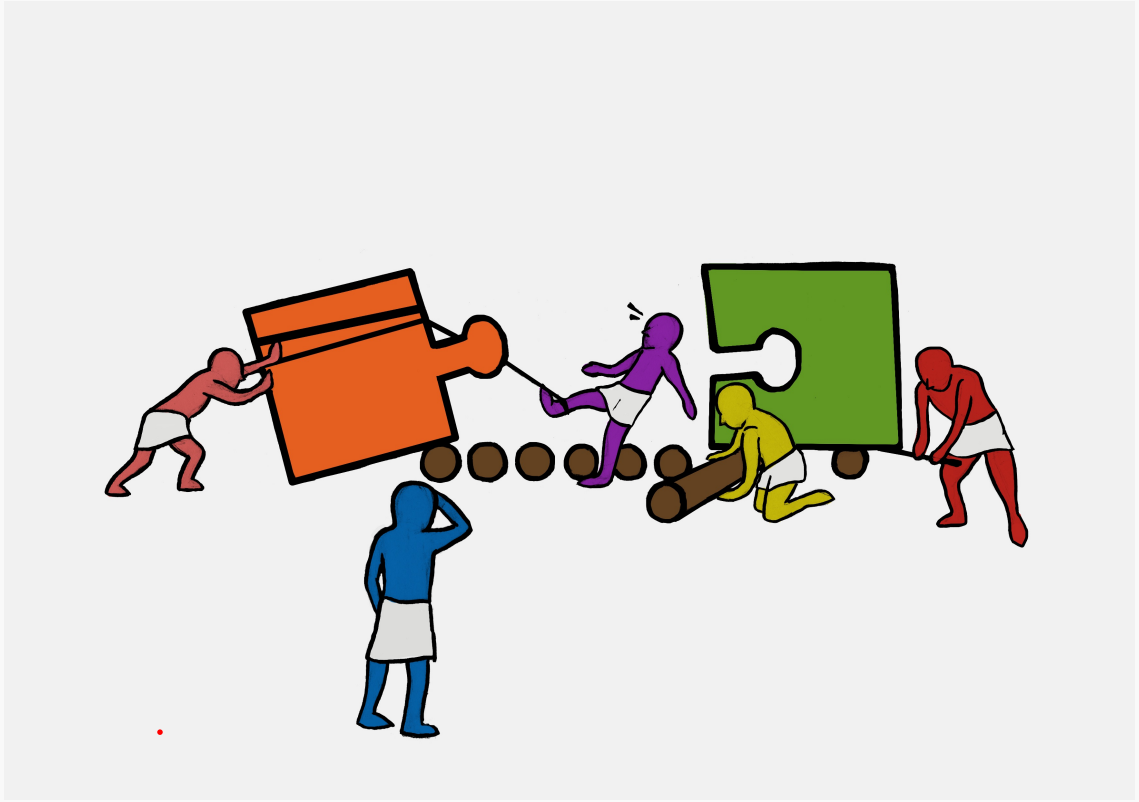
- Forum
- Quizzes
- Reading assignments
- Hands on exercises
- Links to lectures

## Canvas




- Homework
- Grades



Breakout rooms and in-class exercises