

# Lecture #1: Introduction to CS1090A

aka STAT109A, AC209A, CSCIE-109A

## CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Gumb



# Lecture Outline

---

- What is data science?
- Why data science?
- How to learn and why take CS109A?
- What is this class: who, how, what?
- Demo

# Lecture Outline

---

- **What is data science?**
- Why data science?
- How to learn and why take CS109A?
- What is this class: who, how, what?
- Demo

# A little bit of history

# History: The Evolution of Data Science: **Early Methods**

---

In ancient times, scientific knowledge was largely based on empirical observations. People would gather data through direct experience, such as counting stars in the sky or measuring crop yields.



# History: The Evolution of Data Science: **Early Methods**

In ancient times, scientific knowledge was largely based on empirical observations. People would gather data through direct experience, such as counting stars in the sky or measuring crop yields.





# The Evolution of Data Science: **From Observation to Innovation**

Thousands of years ago, science was primarily empirical in nature. Individuals would observe and count entities like stars and crops. This collected data was then used to construct devices that helped explain these phenomena.



# The Evolution of Data Science: The Age of Equations

A few centuries ago, the approach to science shifted significantly. Researchers began using mathematical equations, often in the form of differential equations, to describe relationships and phenomena.

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

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

$$\begin{aligned} \nabla \cdot E &= 0 & \nabla \times E &= -\frac{1}{c} \frac{\partial H}{\partial t} \\ \nabla \cdot H &= 0 & \nabla \times H &= \frac{1}{c} \frac{\partial E}{\partial t} \end{aligned}$$

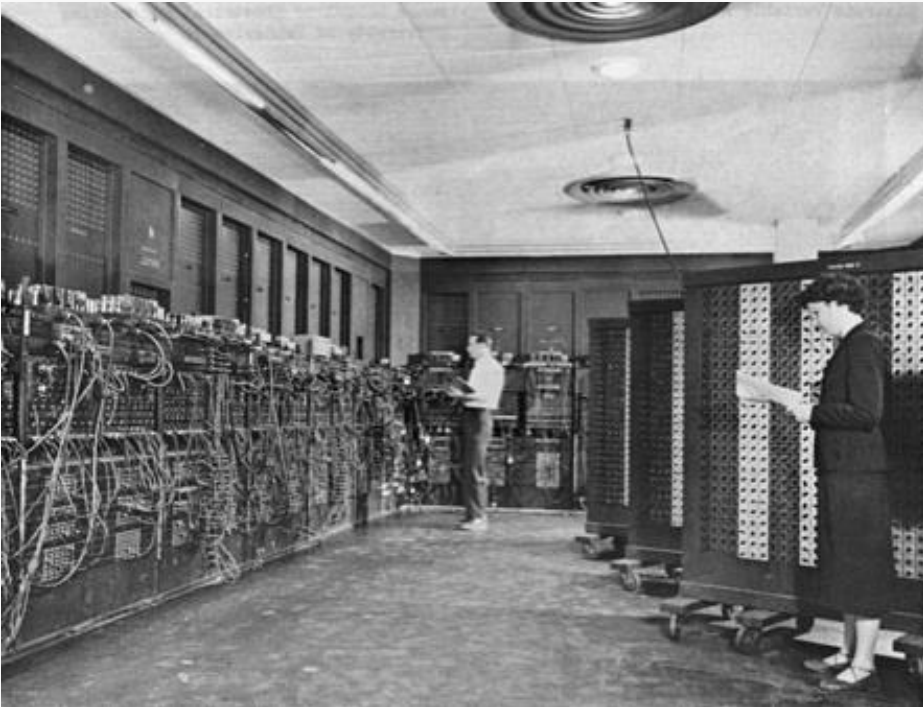
$$E = mc^2$$

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



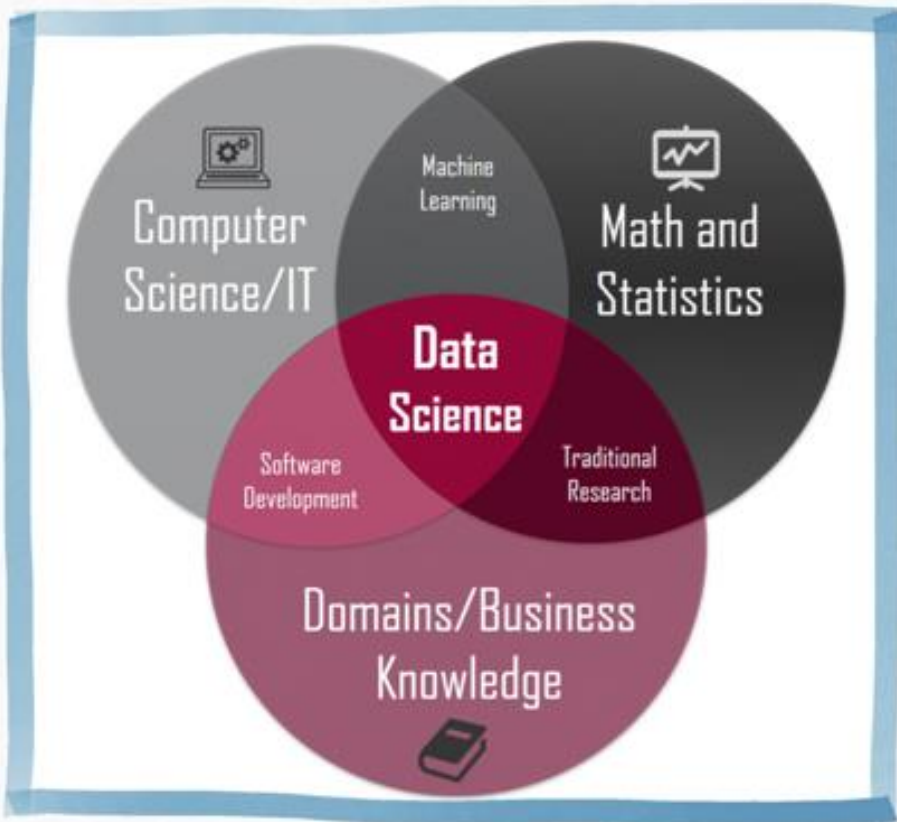
# The Evolution of Data Science: **The Computational Era**

Approximately a century ago, another paradigm shift occurred in science with the emergence of computational approaches. This allowed for complex simulations and analyses that were previously unimaginable.



# The Rise of Data Science and Machine Learning

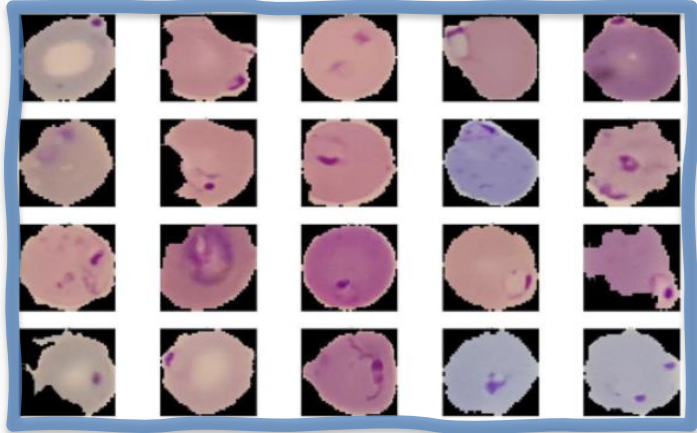
In more recent times, the focus has shifted yet again to data science and machine learning. These disciplines specialize in extracting patterns and insights from large sets of data, revolutionizing how we understand and interact with the world.



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

# The Potential of Data Science

## Disease Diagnosis



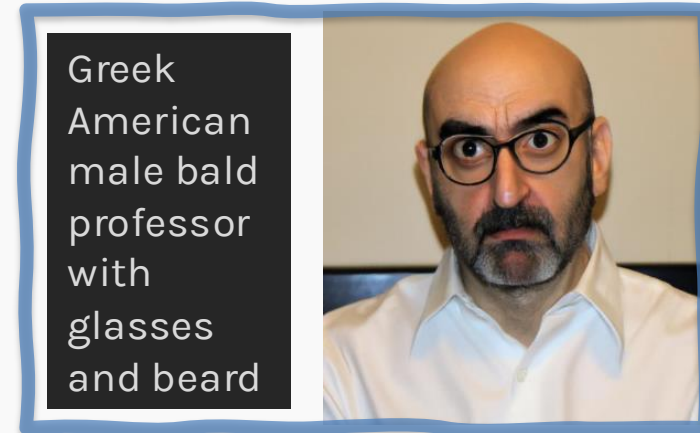
Detecting malaria from blood smears

## Drug Discovery



Discovering new drug combinations  
using language models

## Generative AI



Creating images from text prompts

## Transportation

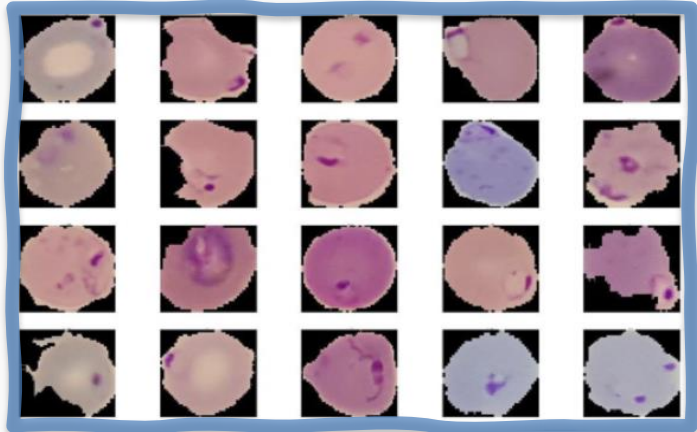


Self driving trucks for safe night shipping



# The Potential of Data Science

## Disease Diagnosis



Detecting malaria from blood smears

## Drug Discovery



Discovering new drug combinations  
using language models

## Generative AI



Creating images from text prompts

## Transportation



Self driving trucks for safe night shipping

# The Potential of Data Science

## Gender Bias



Some DS models for evaluating job applications in some fields show bias in favor of male candidates

## Racial Bias



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

# 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 do you want to predict or estimate?

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

# 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?

# Lecture Outline

---

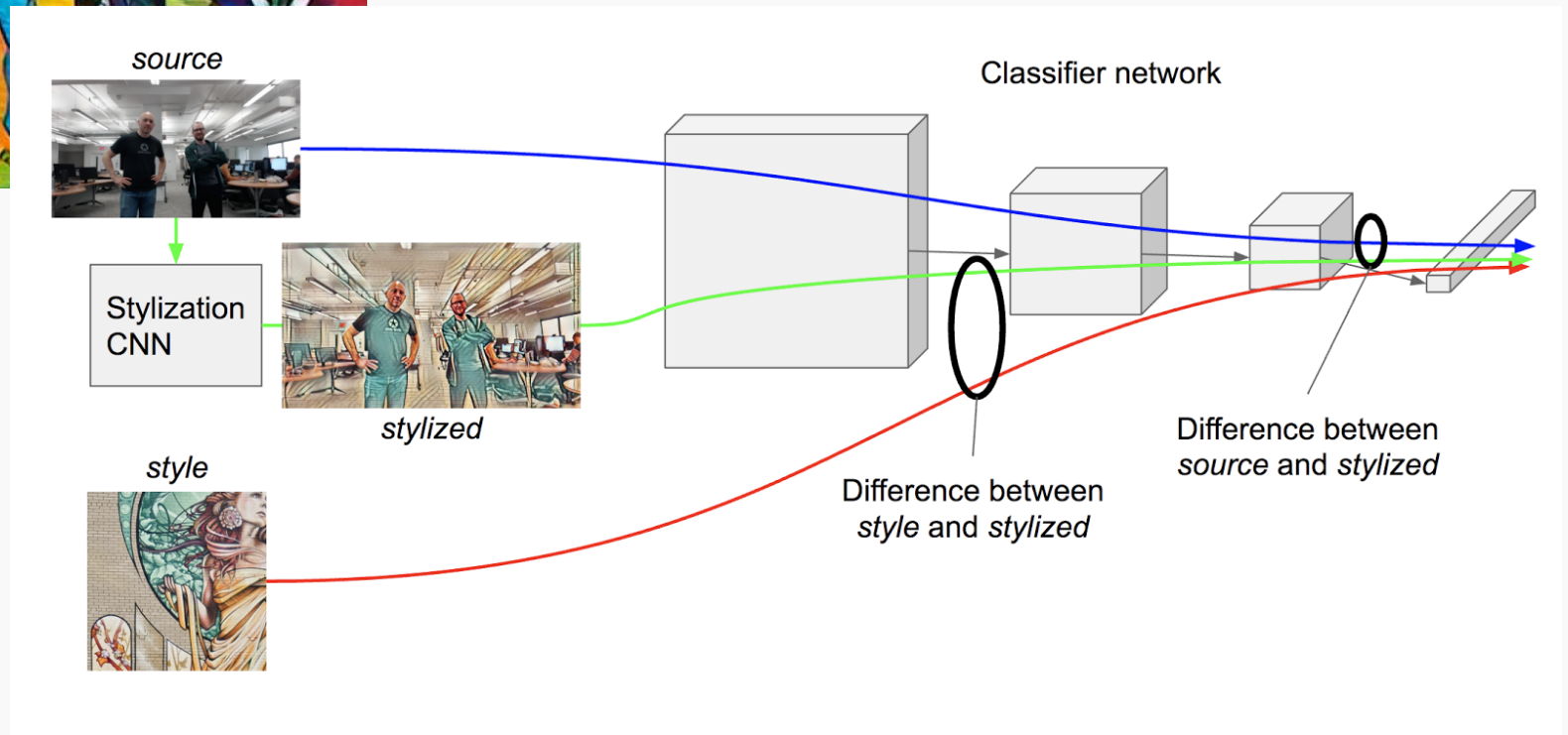
- What is data science?
- **Why data science?**
- How to learn and why CS109A?
- What is this class: who, how, what?
- Demo







## Minimise Loss



# But if you decide to do it...

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





# 50 Best Jobs in America for 2022

Job Title		Median Base Salary	Job Satisfaction	Job Openings	
#1	Enterprise Architect	\$144,997	4.1/5	14,021	<a href="#">View Jobs</a>
#2	Full Stack Engineer	\$101,794	4.3/5	11,252	<a href="#">View Jobs</a>
#3	Data Scientist	\$120,000	4.1/5	10,071	<a href="#">View Jobs</a>
#4	Devops Engineer	\$120,095	4.2/5	8,548	<a href="#">View Jobs</a>
#5	Strategy Manager	\$140,000	4.2/5	6,977	<a href="#">View Jobs</a>
#6	Machine Learning Engineer	\$130,489	4.3/5	6,801	<a href="#">View Jobs</a>
					<a href="#">View Jobs</a>

# 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

12k Shares

f

t

in

e

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

#### 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



**I want to do it because ....**



# Lecture #22: Generative Model

CS109B, STAT109B, AC209B, CSCIE-109B

## CS109B Introduction to Data Science

Pavlos Protopapas, Alex Young



# Lecture #22: Generative Model

CS109B, STAT109B, AC209B, CSCIE-109B

## CS109B Introduction to Data Science

Pavlos Protopapas, Alex Young



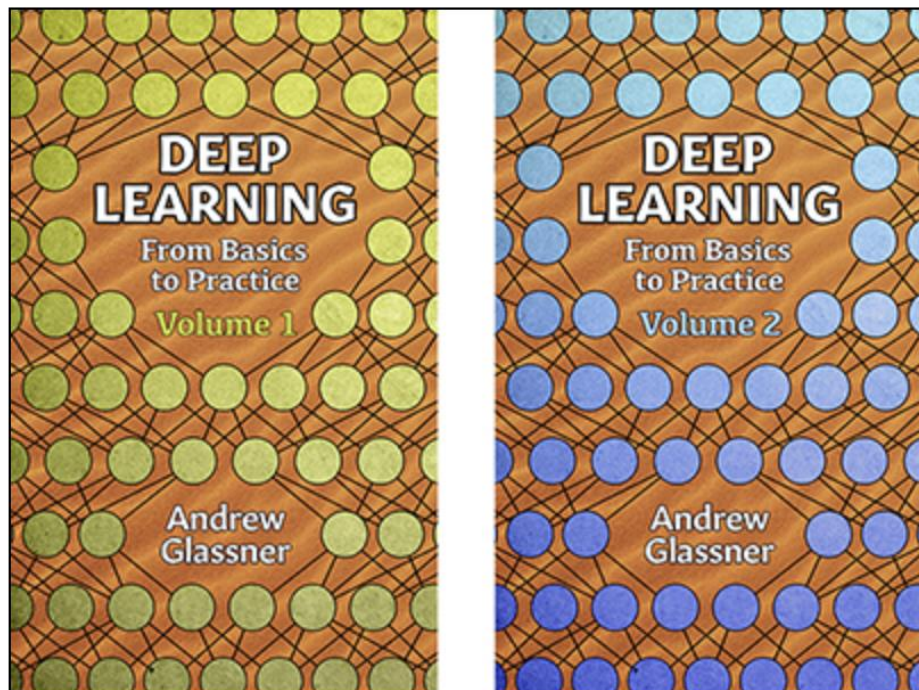
Pavlos Protopapas



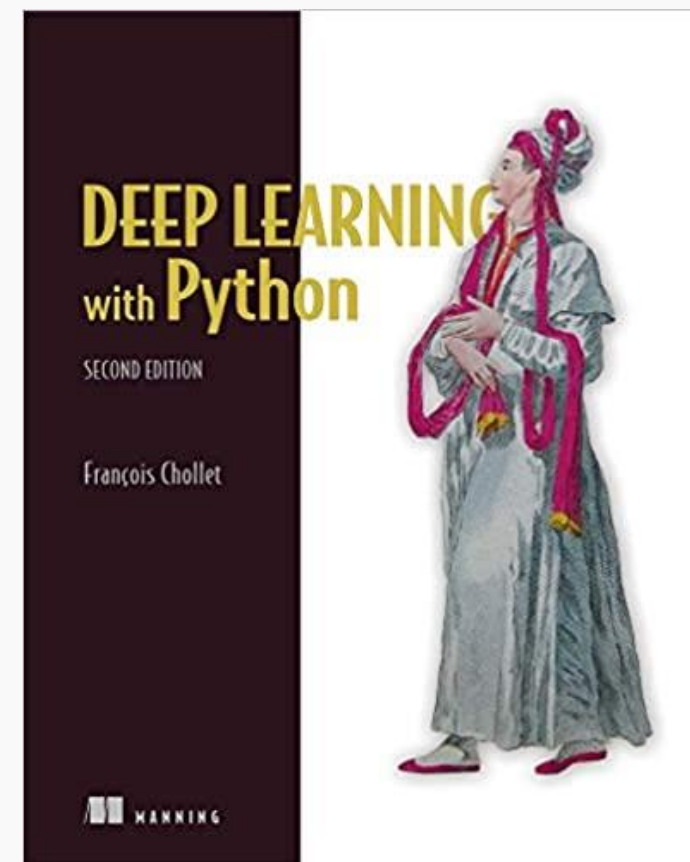
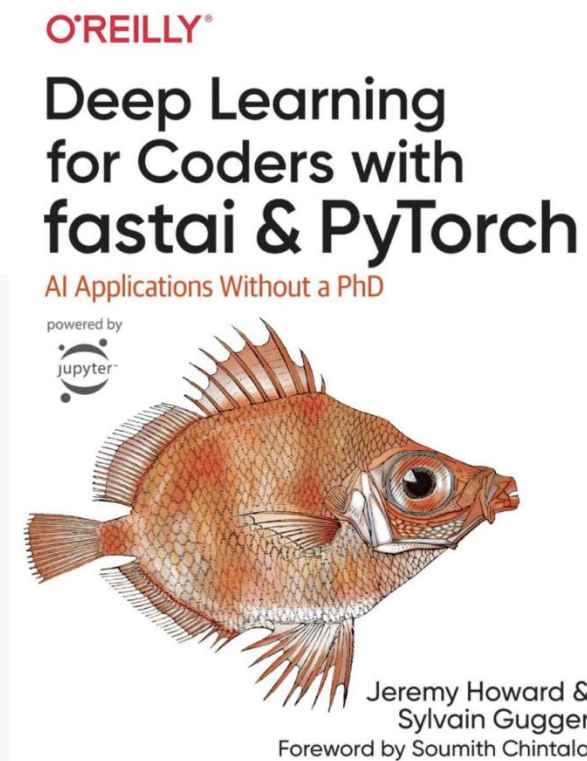
# Lecture Outline

---

- What is data science?
- Why data science?
- **How to learn and why CS109A?**
- What is this class: who, how, what?
- Demo



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



# 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	<a href="https://r/NYU_DeepLearning">r/NYU_DeepLearning</a>
DISCORD	<a href="#">NYU DL</a>
MATERIAL	<a href="#">2021 repo</a>


## 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

# Learn by Watching

 Full Stack Deep Learning

 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🌟](#).

### 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

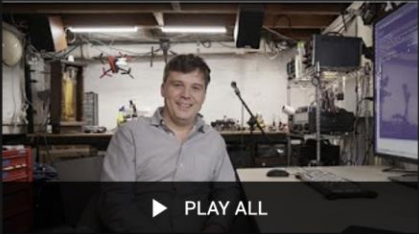
🌟Week 14-16: Projects🌟

Other Resources

### Week 1: Fundamentals

We do a blitz review of the fundamentals of deep learning, and introduce the codebase we will





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

PLAY ALL

# 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

1:51

2

0. What is machine learning?

Weights & Biases

19:59

3

1. Build Your First Machine Learning Model

Weights & Biases

21:09

4

2. Multi-Layer Perceptrons

Weights & Biases


18:58

5

3. Convolutional Neural Network

Weights & Biases

12:36




**Yannic Kilcher**  
94.3K subscribers

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 | ...

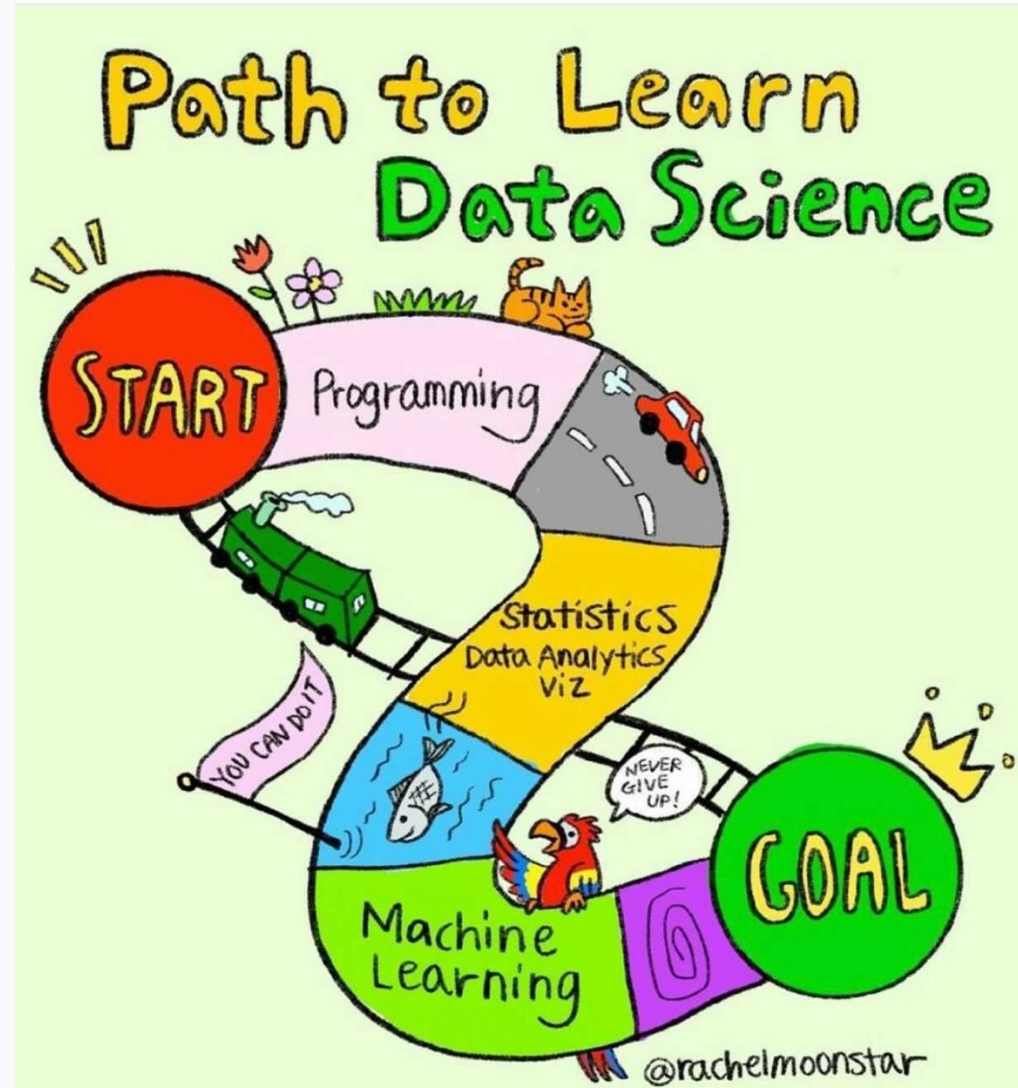


# Lecture Outline

---

- What is data science?
- Why data science?
- **How to learn and why CS109A?**
- What is this class: who, how, what?
- Demo

# Memes!





# Digestion Time

# Lecture Outline

---

- What is data science?
- Why data science?
- How to learn and why CS109A?
- **What is this class: who, how, what?**
- Demo



# Who? Instructors



**Pavlos Protopapas**

Scientific Director  
For DS and CSE  
masters programs

Principle Investigator of StellarDNN, a research lab within IACS/SEAS. Research in the intersection of [astronomy](#), ML and statistics. He uses Neural Networks to solve problems in astronomy and physics and applying NLP techniques in astronomical time series analysis.

He loves classical music and opera, and he often visits the Boston Symphony Orchestra.

A certified cook from *Le Cordon Bleu* but loves [eating](#) more than cooking.

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

tiktok: @pavlosprotopapas

# Who? Instructors



**Kevin Rader**

Senior Preceptor and  
Associate-DUS in  
Statistics

Primary role is undergraduate education, teaching several 100-level Stat courses from intro stat (104) to biostatistical methods (116).

Advises undergraduate research that applies data science and statistical analysis techniques in the domains of sports, medicine, and policy.

Loves all things Philadelphia, especially those concerning the national bird of the US



Go Birds!

Has a passion for growing and cooking his own food (mostly to help supply Pavlos' eating habits).



Fun fact: coaching 4<sup>th</sup> grade girls' soccer. Soccer analytics have not helped so far. 0 - 1 as a coach so far.

is not on tiktok...yet



# Who? Preceptor



**Chris Gumb**  
Preceptor  
SEAS

Chris has been a member of the CS109A & B teaching staff for the past 8 years.

As preceptor, he teaches some lectures, coordinates the TF team, develops course materials, and handles logistics.

When not answering your Ed posts and emails he enjoys making music and seeing films with friends.

Frequently spotted at the local independent movie theaters, he's basically made of popcorn 🍿

# Who? ~30 Teaching Fellows!

---

Wenqi Shi

Christian Rodrigo Cruz Flores

Carol (Xuan) Long

Joshua Price

Hao Shen

Paula Rodriguez Diaz

Leslie (Shixuan) Gu

Antony Tan

Michelle (Mingxiao) Song

Alessia Di Maria

Junyang Deng

Diksha Chugh

Bowen Xu

Steven Liu

Songhan Hu

Omar Abdel Haq

Haoran Zhang

Victoria (Xu) Tang

Yuan Tang

Rashmi Banthia

Victoria Okereke

Daniel More Torres

Dhati Oommen

Josh Kaplan

Teodor Malchev

Kasim Domac

Maitri Shah

Aseel Rawashdeh

Michel Arab

... and more!

# 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 and statistical learning.**
5. **Communication:** summarizing results through visualization, stories, and interpretable summaries.

# Goals of the course

## Theory/Intuition

1. Key concepts in statistical analysis & machine learning
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  
Data Viz and EDA

## Weeks 3-5: Regression

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

## Week 6: Bayesian Modeling

Bayesian Inferential Framework  
Bayesian Linear Regression

## Weeks 7-9: Classification

kNN Classification  
Logistic Regression  
Hierarchical Modeling

## Week 8: Midterm

Midterm (October 22-24 in Section)

## Week 10: Data Issues

Missingness  
Causal Inference  
Biases and Ethical Considerations

## Weeks 11-14: Tree-Based Models

Decision Trees  
Bagging  
Random Forest  
Boosting Methods  
Mixture of Experts

# After CS109A

## CS109B

### A. Neural Networks:

- MLP
- CNNs
- RNNs
- Generative models
- Deep RL

### B. Unsupervised Learning

### C. Dimension Reduction

## AC215 Next Fall

### 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



# Other related courses: not an exclusive list

---

- CS 171/271 (Visualization)
- CS 181 (ML)
- CS 182 (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).
- CS 208 (Privacy)
- CS 282R (ML: Generative Models)
- CS 282BR (Sequential Learning)
- AC 295/CS 287 (DL for NLP)

# Why?

---

Why are you here?

# Why?

---

Why am I here?

To provide the statistical/inferential/interpretive perspective:

1. What can our results tell us about the **relationships** between variables in the data? Do these relationships **vary** across individuals/observations?
2. How much **uncertainty** is there in the predictions and the estimates?
3. Are there any **data issues**? Think about: biases, missing data, ethical considerations, missing features/variables to collect or engineer, etc.



# Digestion Time

# Course Components

# Lectures, Sections, and Office Hours

---

In lecture we'll [cover the material](#) that you will need to complete the [homework](#) and to survive the rest of your life in CS109A.

We will use a mix of slides and optional exercises via *edstem*.

1. Lecture slides and associated notebooks will be posted before lecture on *edstem*.
2. Lectures will be video taped (and live streamed for the extension school students) and are usually posted on Canvas within 24 hours.

Mon & Wed 10:30-11:45am [in person](#) @Lowell Lecture Hall and @Zoom for Extension School Students (zoom link is on canvas under zoom).



# Lecture format

ASYNCHRONOUS

- Reading
- Optional Quiz/Concept questions

SYNCHRONOUS

Questions from asynchronous material and review

Live Lecture

Q&A

ASYNCHRONOUS

- Optional Coding Exercises

⋮

Repeat

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

---

*Sections* will be a mix of review material, tutorials on how to practically solve problems with Python libraries, and some hands-on exercises.

Section attendance is required!

Quizzes and the Midterm will be administered in Section.

**DCE:** Proctorio will be used for quizzes and the midterm.

You are assigned to a specific section. The full section schedule will be posted on Canvas.

# Attendance

---

Attending class isn't just required; it's something I look at closely when deciding on academic and professional recommendations.

Please understand that consistent presence and engagement in the classroom are highly valued in this course.



# Attendance

---

## From the Syllabus:

- Attendance is required for all on-campus students and will be recorded throughout the semester.
- To be eligible for certain letter grades, students must meet the following minimum attendance requirements (lectures and sections combined):
  - A requires at least 66% attendance
  - A- requires at least 50% attendance
  - B+ requires at least 33% attendance



# Attendance

---

All lectures are videotaped, so you can watch them later if you can't attend.

For every 4 sessions attended (lecture or section), students earn 1 late day, which can be used on homework (up to 2 per assignment)\*



\*DCE students are automatically granted 4 late days

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

---

Office hours will be posted before next week.

There will be a Google calendar made available through Canvas with all course components and OHs.



# Assignments

# Five Graded Components

## Homework: 30%

Homework 0: 1%  
Homeworks 1-5: 29%

Students are encouraged to work in pairs on HW assignments.

## Section Quizzes: 10%

Two 30-minute in-section quizzes (not cumulative)

Quiz 1: Sept 28-Oct 2  
Quiz 2: Nov 16-20

You will be allowed 1 page of reference notes.

## Midterm: 18%

1 in-section Midterm, with a mix of multiple choice, short answer, and coding questions.

Multiple choice and short answer will be in-person with 2 reference sheets allowed, coding questions will be a timed take-home exam.

## Final Exam: 22%

3-hour seated exam during exams period (tentatively scheduled for Dec. 11) with 4 sheets of reference notes. Roughly 90+ minutes of conceptual questions and 80+ minutes of coding

## Project: 20%

Milestone dates and details to be announced soon.

# Homework(s)

---

**There will be 5 homeworks (not including Homework 0):**

- Homework 0 (due Sept 9<sup>th</sup>; all honest attempts get full credit)
- Homework 1 - Web scraping, BeautifulSoup, Pandas, Plotting & EDA
- Homework 2 - Regression: kNN, and LinReg (Multi- & polynomial)
- Homework 3 - Regularization, Inference, and Bayesian Linear Regression
- Homework 4 - Classification: Logistic Regression and Hierarchical Models
- Homework 5 - Trees, Forests, and Boosting.

# Final Project

---

There will be a final group project (3-5 students) due during reading period.

- You can propose to use a (public) data set of your choice and your own project definition (to be approved by the instructors).
- Project proposal process starts in late September.

# Homework(s)

---

You are encouraged but not required to submit **in pairs** on HWs 1-5

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

HWs 1-5 are **due 10pm on Tuesdays**, and homework will be released on Tuesdays.

**Late submission policy:** students can earn late days based on attendance -- at most 2 late days can be applied to any single homework. Outside of these allotted late days, late homework will **not be accepted**.



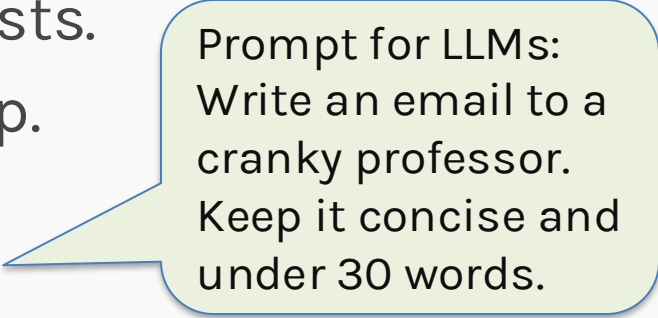
# Digestion Time

# Help

---

The process to get help is:

1. **Post** the question on *Edstem*, and hopefully, your peers will answer. The teaching staff will also monitor and respond to posts.
2. Attend **Office Hours**; this is the best way to get help.
3. For private matters, send an email to the Helpline: [cs1090a2025@gmail.com](mailto:cs1090a2025@gmail.com).
4. For personal matters, send an email to Pavlos and Kevin.



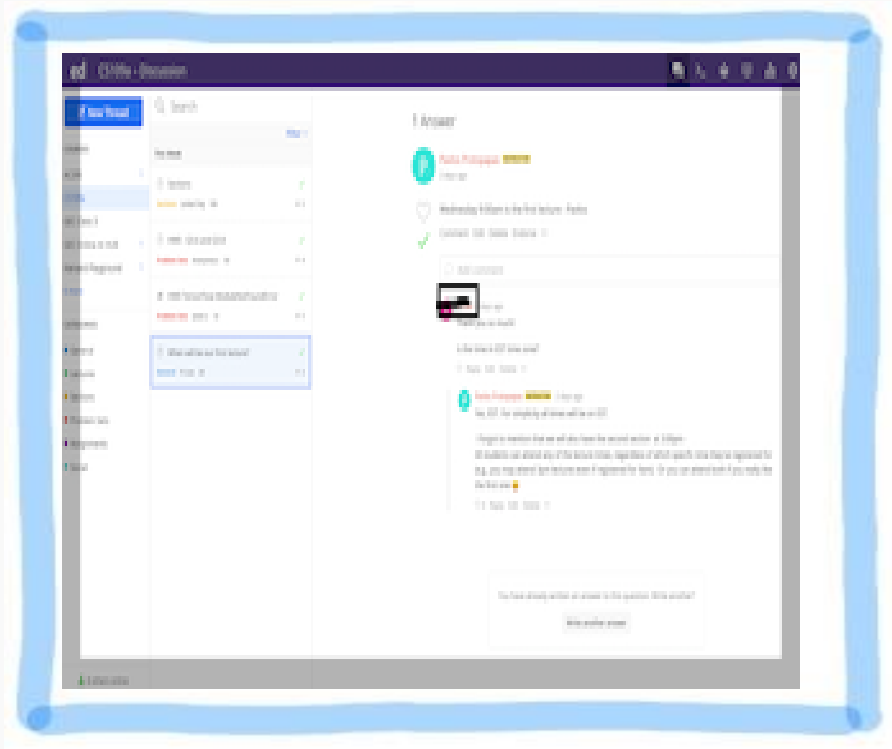
Prompt for LLMs:  
Write an email to a cranky professor.  
Keep it concise and under 30 words.

[Weekends will be slow days, so please be patient!](#)



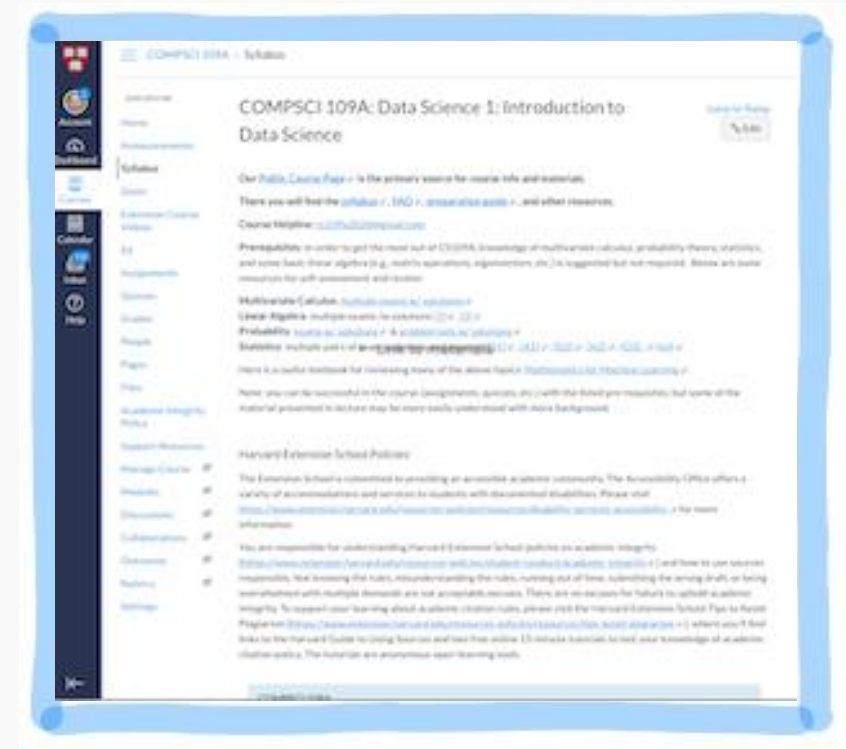
# Tools for the course

## edstem



- Discussion Forum
- Reading assignments
- Lecture slides
- Section material
- Hands on exercises

## Canvas



- Syllabus
- Schedule
- Homework Assignments
- Video Recordings
- Grades

## Can I audit this class?

Yes, CS109A does accept auditors, but all auditors must agree to abide by the rules described in the syllabus

Can I take this class asynchronously?

**College students:** This is not allowed.

**Graduate students:** This is not ideal. Attending classes is very important and part of being a student here. The decision is yours and your program academic coordinator. We feel you should attend at least 50% of the classes.

Am I prepared for this class?

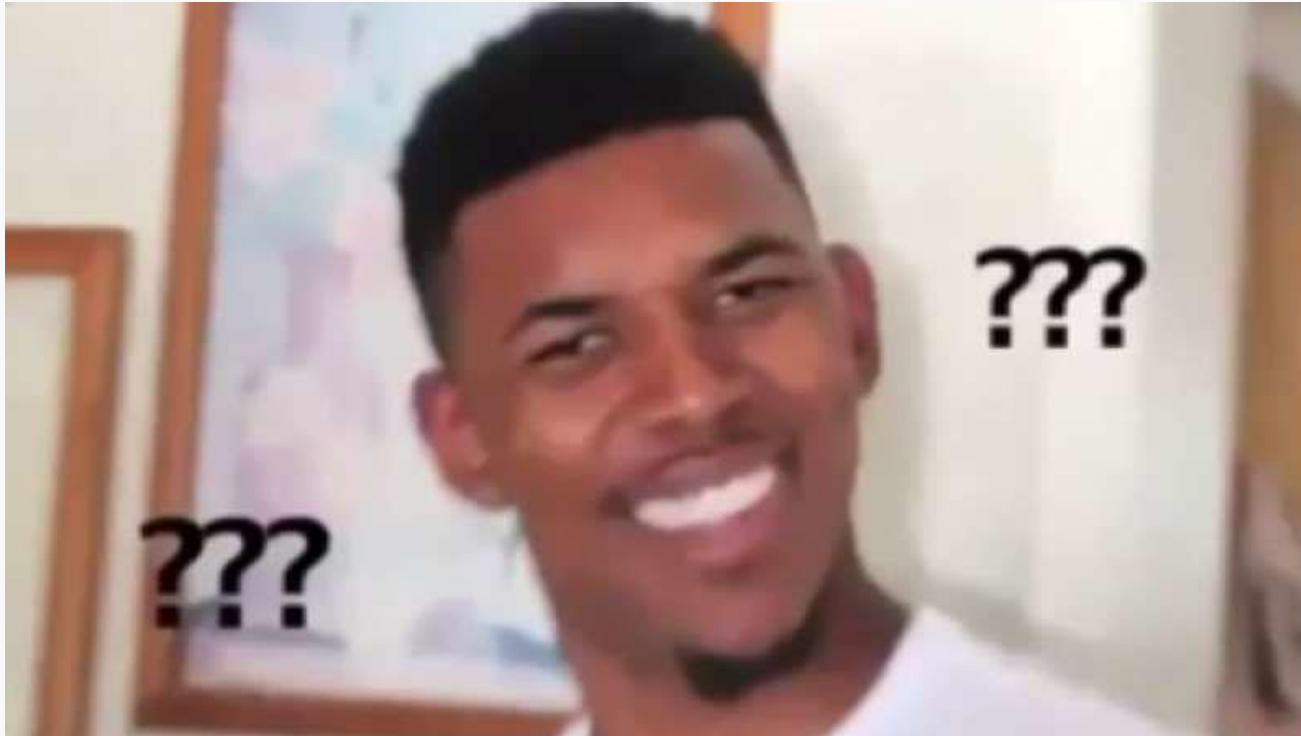
Proficiency in Python, basic math (calculus), basic stats/probability are expected.

HW0 will give you a sense of whether you have the pre-reqs.

# FAQ

---

If I miss a class, will it affect my grade?



# FAQ

---

I have a trip planned during the midterm. Can I take the midterm earlier or later?

Midterm is administered in section the week of 10/20-10/24.

Final Exam is a 3-hour exam, tentatively scheduled for 12/11.

\*DCE will be administered via Proctorio.

Make sure these are on your calendar!

I have a project in mind. Can I use it for the course?

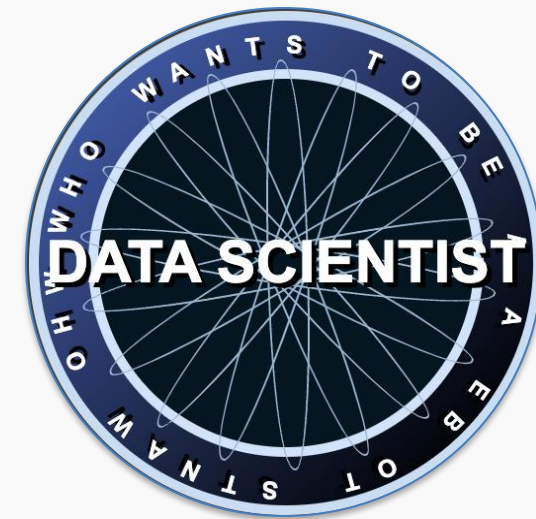
Yes, as long as the data are public and you're willing to work with other students.



# Lecture Outline

---

- What is data science?
- Why data science?
- How to learn and why take CS109A?
- What is this class: who, how, what?
- **Demo**



# CS109A

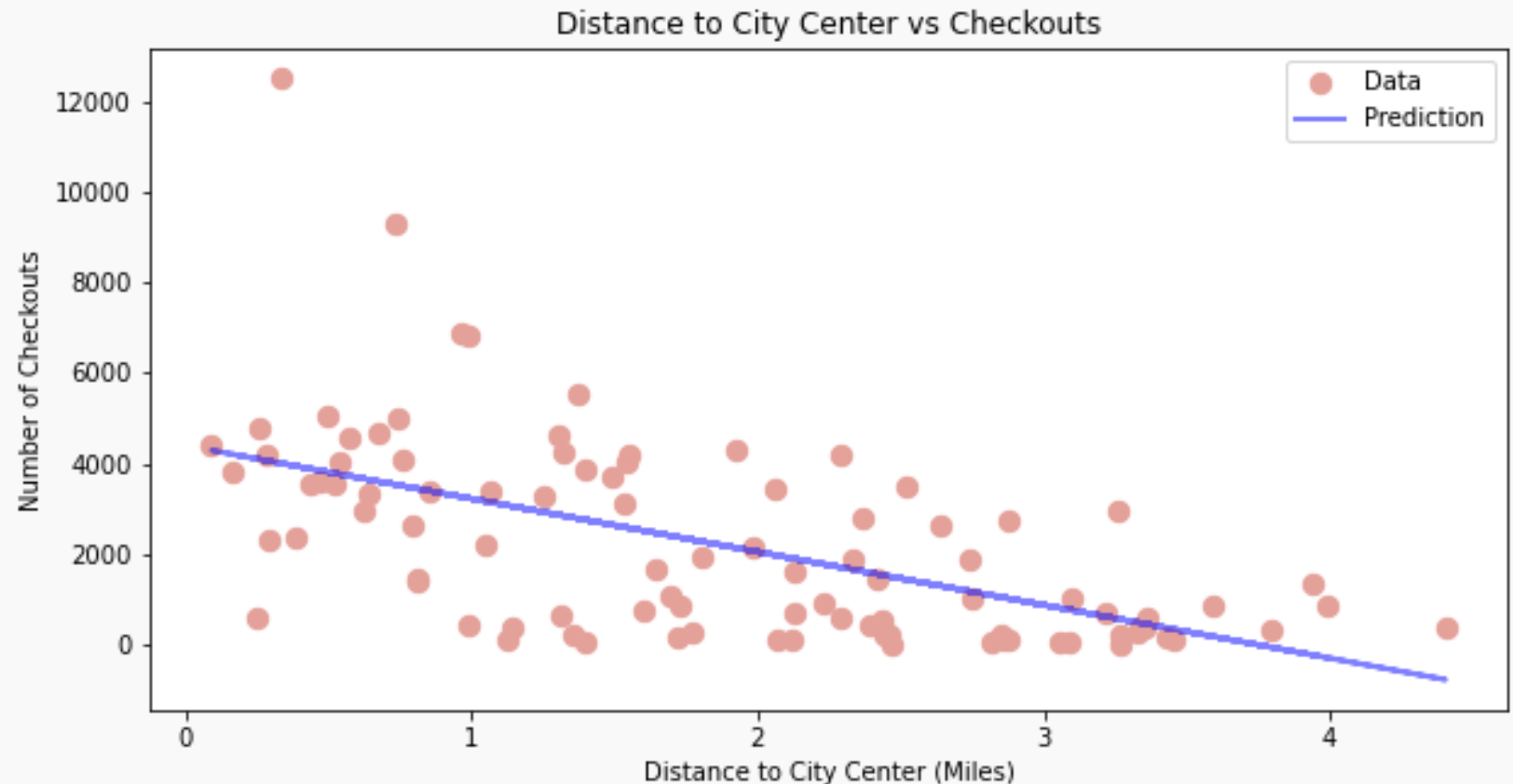
## GAME Time

Based on our "linear" model, what would most likely be the number of checkouts for a distance of 2.5 miles from the city center?



## Options

- A. 45000
- B. 12530
- C. 1450
- D. 650





What is the goal of CS109A (from the teaching staff's perspective)?

## Options

- A. To teach you practical data science.
- B. To make your life difficult and painful.
- C. To predict the next stock price crash.
- D. To enable computers to talk.

THANK YOU

Course staff available to answer questions after class  
today in:

Lowell Lecture Hall  
from  
11:45 AM – 1:00 PM