Lecture #1: Introduction to CS1090A

aka STAT109A, AC209A, CSCIE-109A

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai 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.

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

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

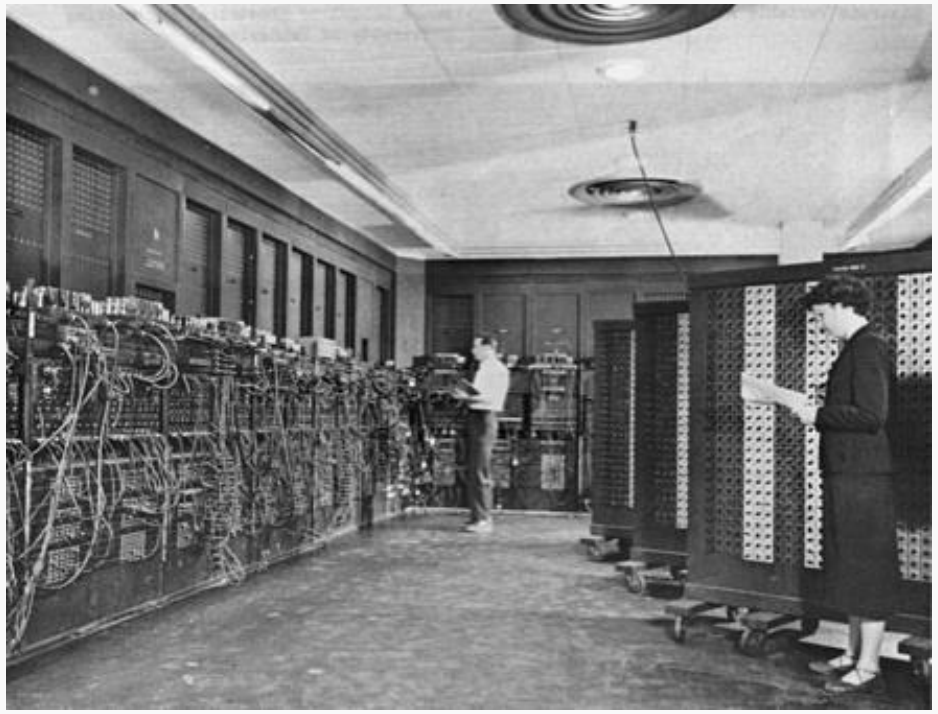
$$\nabla \cdot E = 0 \quad \nabla \times E = -\frac{1}{c}\frac{\partial H}{\partial t}$$
$$\nabla \cdot H = 0 \quad \nabla \times H = \frac{1}{c}\frac{\partial E}{\partial t}$$

$$E = mc^2$$

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

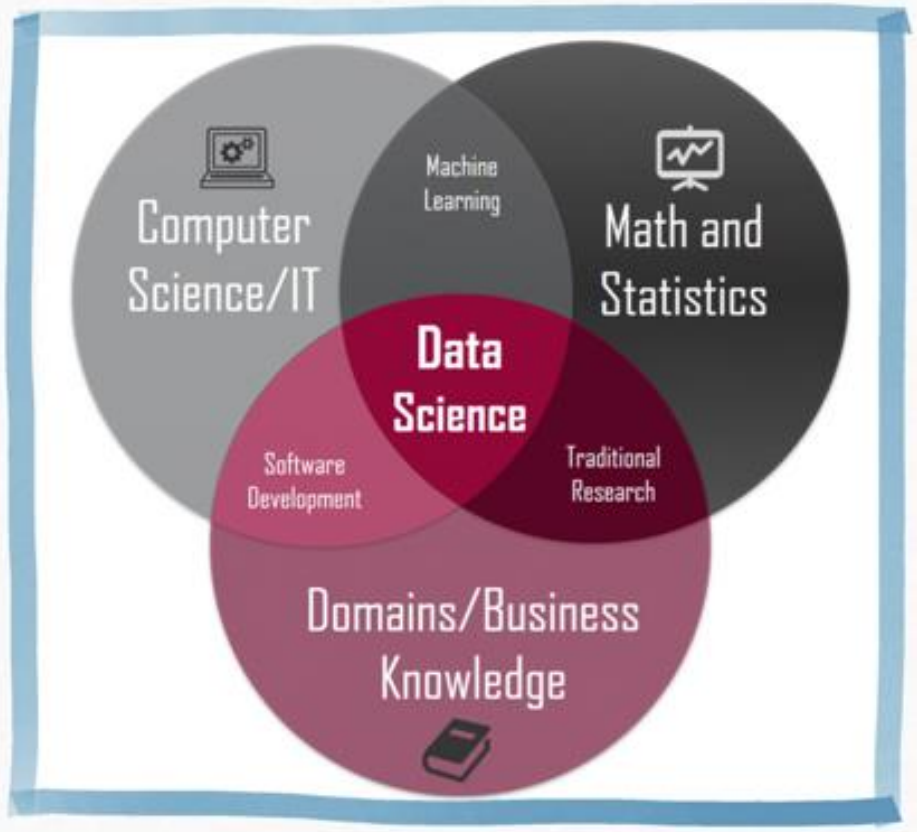
The Evolution of Data Science: **The Computational Era**

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

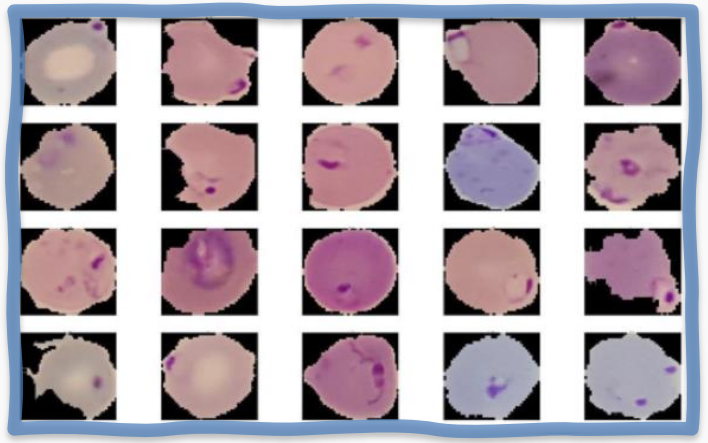
In more recent times, the focus has shifted yet again to data science and machine learning. These disciplines specialize in extracting patterns and insights from large sets of data, revolutionizing how we understand and interact with the world.



- 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

Generative AI

Creating images from text prompts

Drug Discovery

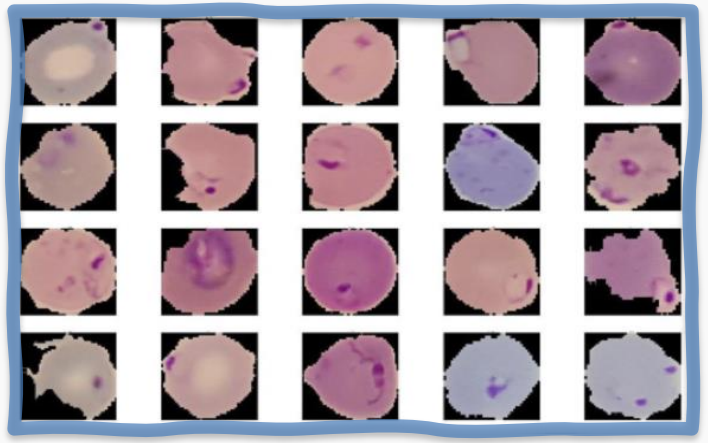
Discovering new drug combinations using language models

Transportation

Self driving trucks for safe night shipping

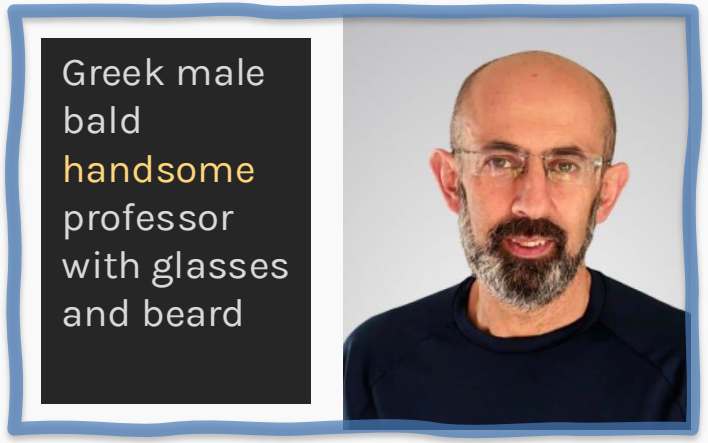
The Potential of Data Science

Disease Diagnosis



Detecting malaria from blood smears

Generative AI



Creating images from text prompts

Drug Discovery

Discovering new drug combinations using language models

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?

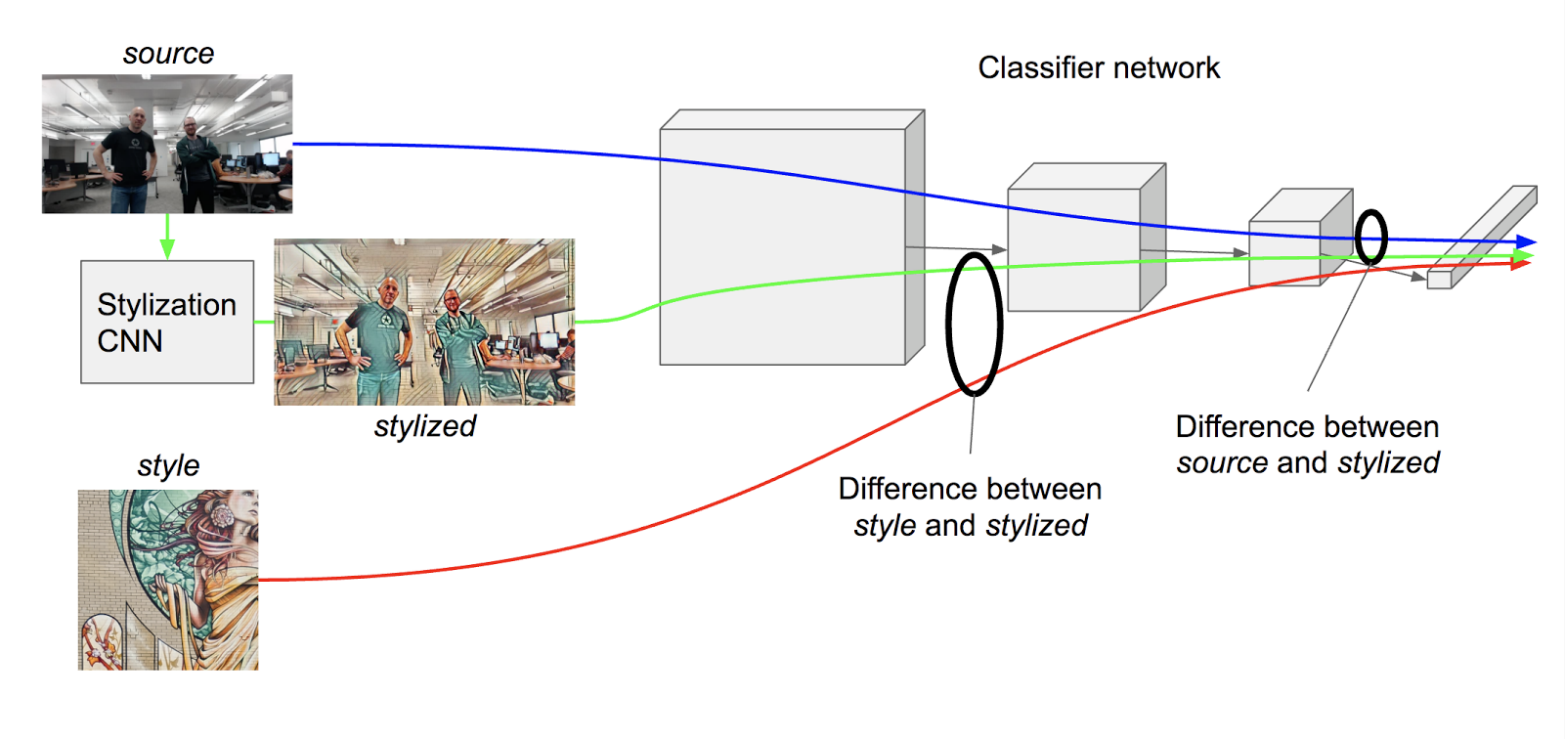
Do the results make sense?

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…

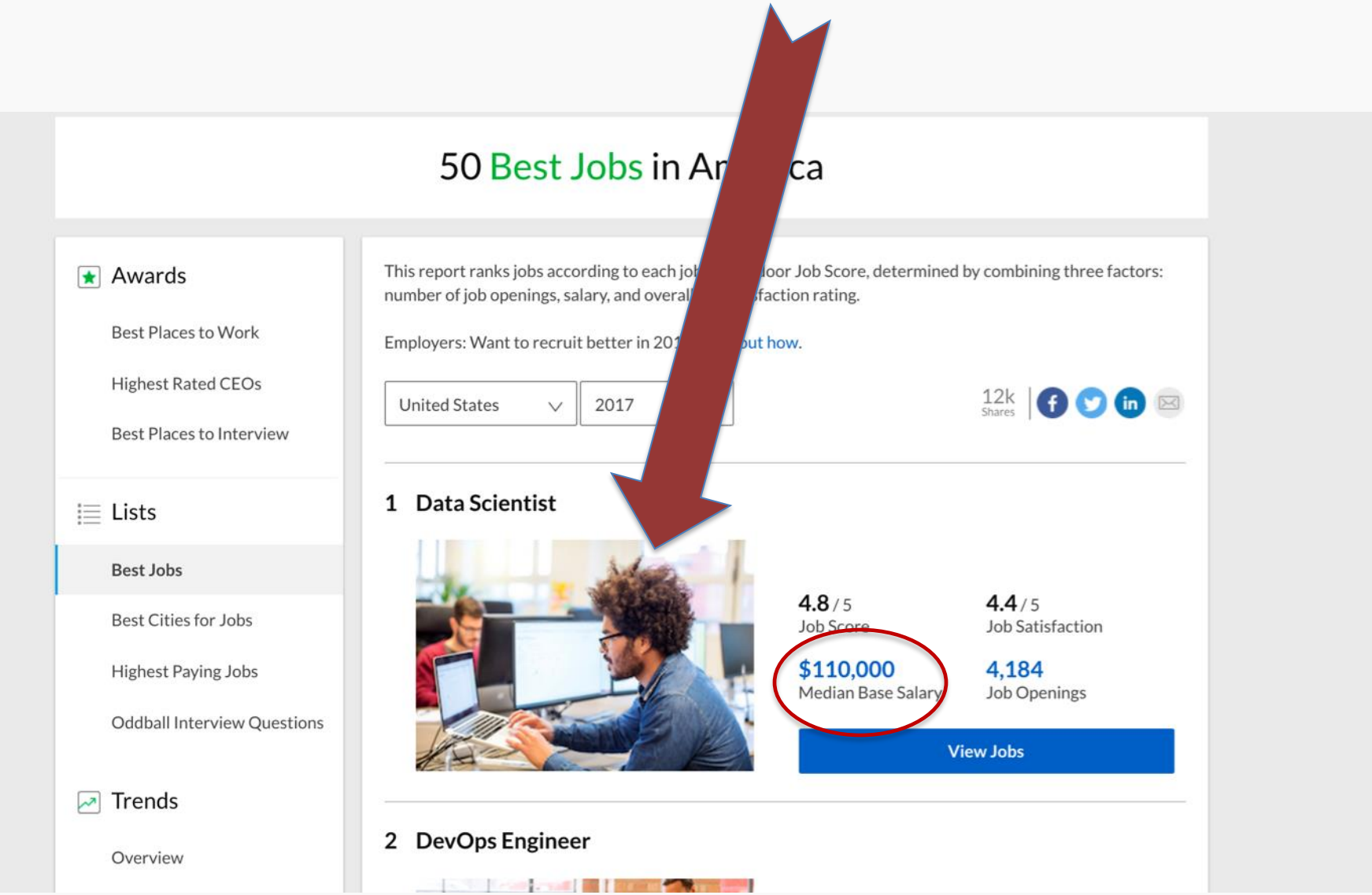
- It’s a lot of fun!
- You will be on the cutting edge of research and industry
- You'll make lots of money doing something you'll enjoy
- It’s not that hard to start and do!

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	View Jobs
#2	Full Stack Engineer	$101,794	4.3/5	11,252	View Jobs
#3	Data Scientist	$120,000	4.1/5	10,071	View Jobs
#4	Devops Engineer	$120,095	4.2/5	8,548	View Jobs
#5	Strategy Manager	$140,000	4.2/5	6,977	View Jobs
#6	Machine Learning Engineer	$130,489	4.3/5	6,801	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 jo oor Job Score, determined by combining three factors: number of job openings, salary, and overal 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

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

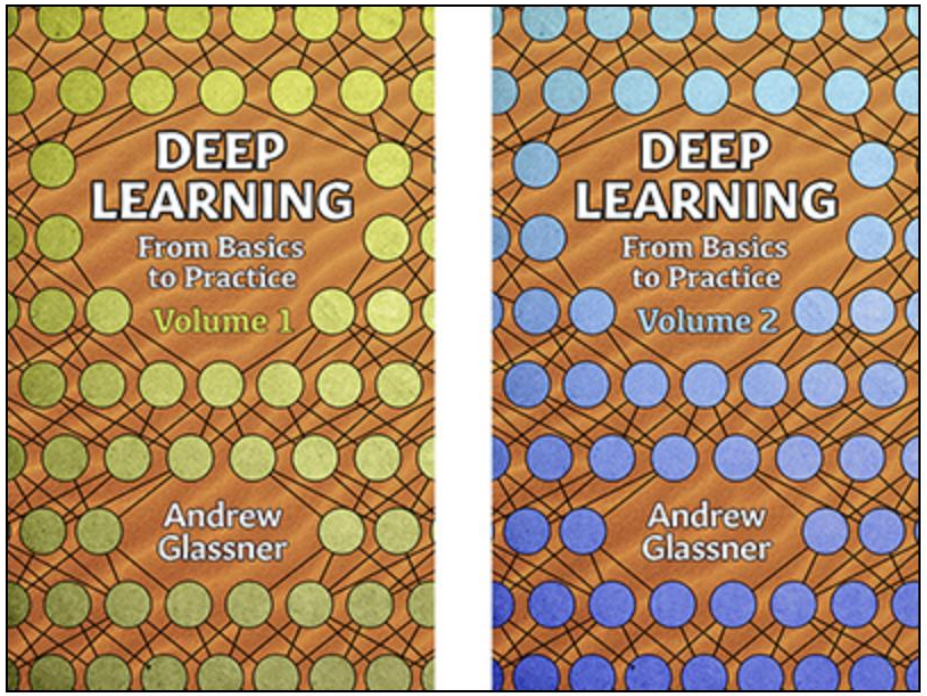
CS109B, STAT109B, AC209B, CSCIE-109B

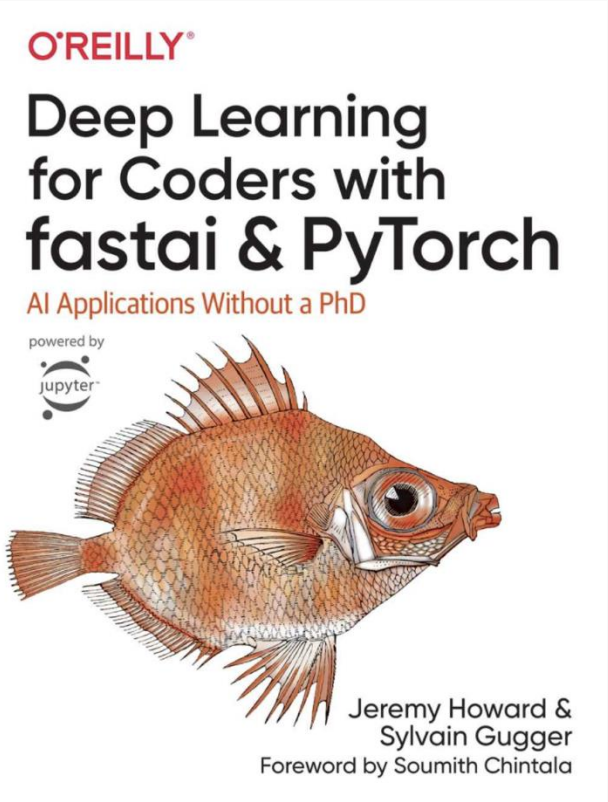
CS109B Introduction to Data Science

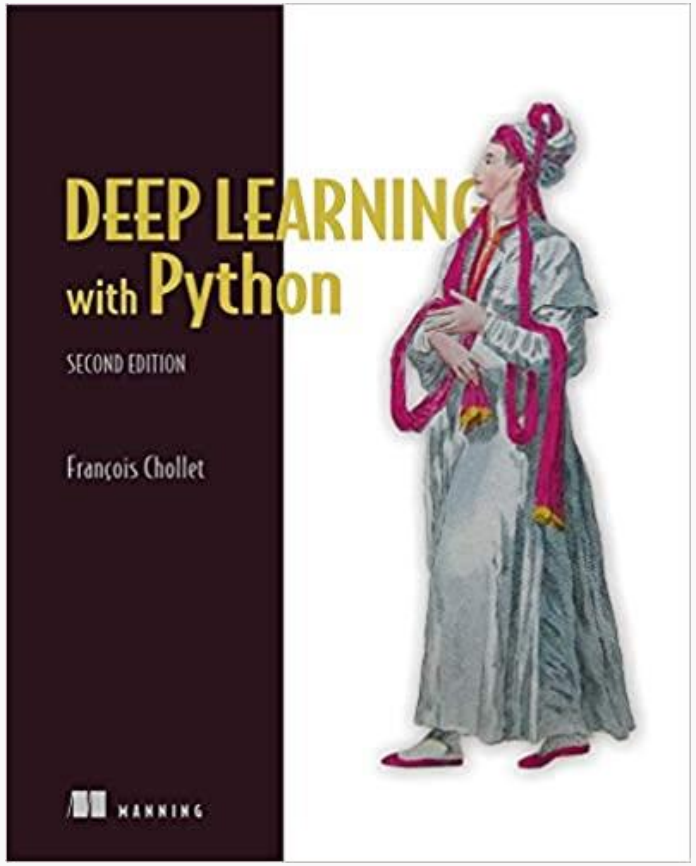
Pavlos Protopapas, Alex Young

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

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.

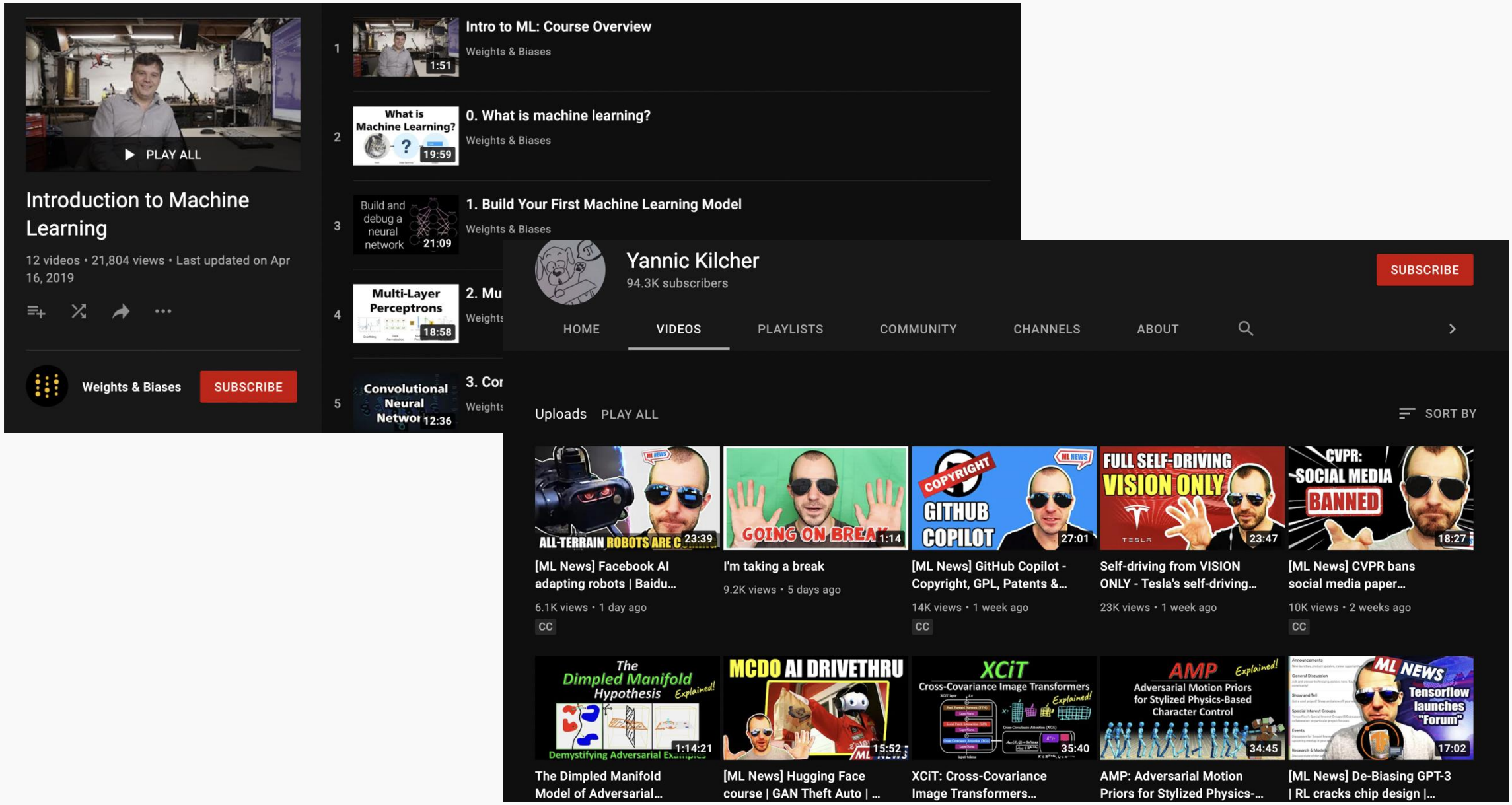
And check out the ✨**course projects showcase**✨.

email address Subscribe

Table of contents

Week 1: Fundamentals

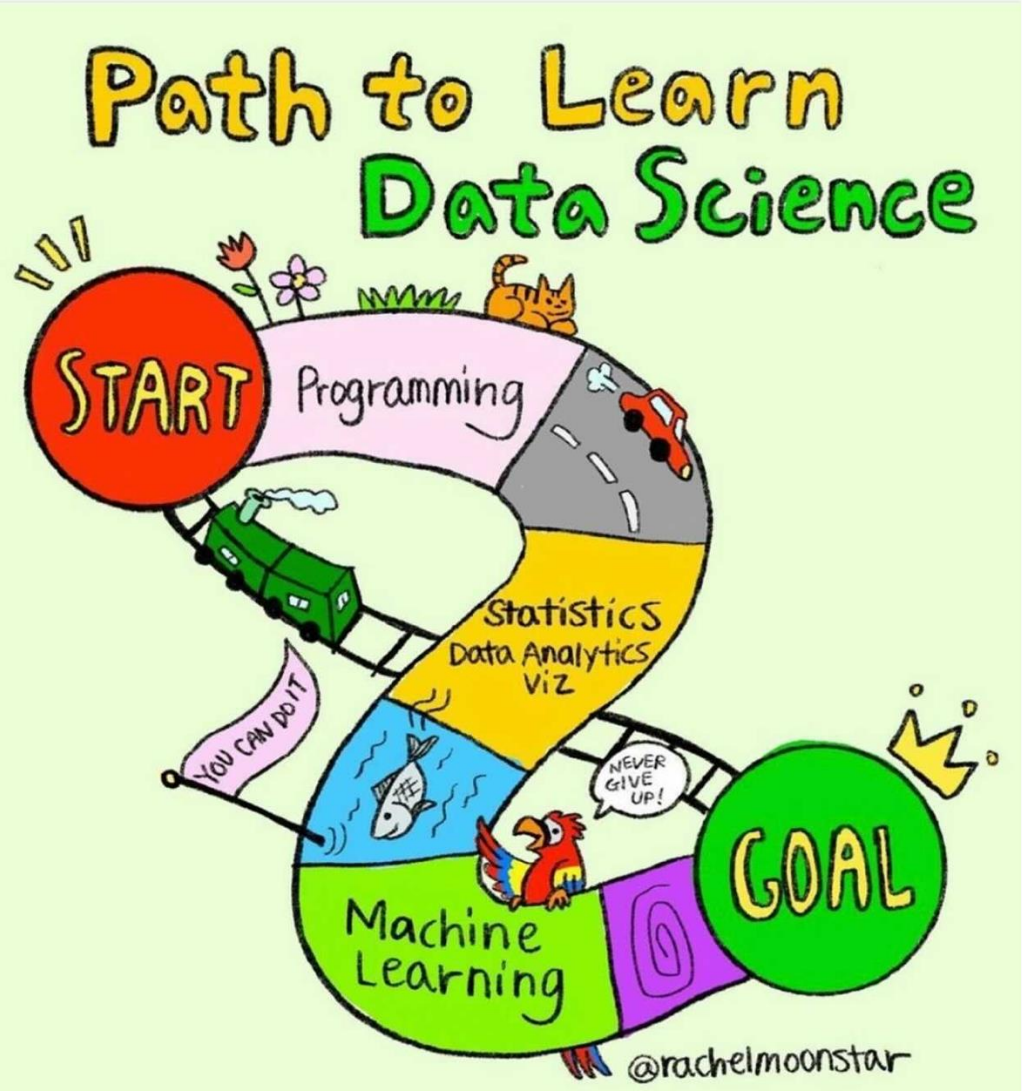
We do a blitz review of the fundamentals of deep learning, and introduce the codebase we will


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
1:51
Weights & Biases
2 What is Machine Learning?
19:59
0. What is machine learning?
Weights & Biases
3 Build and debug a neural network
21:09
1. Build Your First Machine Learning Model
Weights & Biases
4 Multi-Layer Perceptrons
18:58
5 Convolutional Neural Networ
12:36
Yannic Kilcher
94.3K subscribers
SUBSCRIBE
HOME
VIDEOS
PLAYLISTS
COMMUNITY
CHANNELS
ABOUT
Uploads
PLAY ALL
SORT BY
ALL-TERRAIN ROBOTS ARE C
23:39
[ML News] Facebook AI adapting robots | Baidu...
6.1K views • 1 day ago
CC
GOING ON BREA
1:14
I'm taking a break
9.2K views • 5 days ago
COPYRIGHT
GITHUB COPILOT
27:01
[ML News] GitHub Copilot - Copyright, GPL, Patents &...
14K views • 1 week ago
CC
FULL SELF-DRIVING VISION ONLY
23:47
Self-driving from VISION ONLY - Tesla's self-driving...
23K views • 1 week ago
CVPR: SOCIAL MEDIA BANNED
18:27
[ML News] CVPR bans social media paper...
10K views • 2 weeks ago
CC
The Dimpled Manifold Hypothesis Explained!
Demystifying Adversarial Examples
1:14:21
The Dimpled Manifold Model of Adversarial...
MCDO AI DRIVETHRU
15:52
[ML News] Hugging Face course | GAN Theft Auto | ...
XCiT
Cross-Covariance Image Transformers
Explained!
35:40
XCiT: Cross-Covariance Image Transformers...
AMP Explained!
Adversarial Motion Priors for Stylized Physics-Based Character Control
34:45
AMP: Adversarial Motion Priors for Stylized Physics-...
ML NEWS
Tensorflow launches "Forum"
17:02
[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!


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

Lecture Outline

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

Why?

Why are you here?

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/Intuition

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

PCA
Missingness

Weeks 7: Data Issues

Midterm 1

Weeks 8: Classification

Logistic Regression

Week 9: Causal Reasoning

Causal Inference

Weeks 10-13: Decision Trees

Decision Trees
Bagging
Random Forest
Boosting Methods
Mixture of Experts

Weeks 14

Ethics


MOVIECLIPS.com

After CS109A

CS109B

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

B. Unsupervised Clustering

C. Bayesian Modeling

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

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
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

⏳ 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 a distinguished engineer at LinkedIn working on responsible AI. 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? Preceptor

Chris Gumb
Preceptor
SEAS

Chris has been a member of the CS109A & B teaching staff for the past 7 years.

As preceptor, he teaches labs, 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? **~40 Teaching Fellows!**

Omar Abdel Haq
Bailey Bai
Kushagra Chitkara
Labdhi Gandhi
Leslie Gu
Panthon Imemkamon
Ziqing Luo
Megan Luu
Shiyu Ma
Tanner Marsh
Siona Prasad

Robert Roessler
Elaine Swanson
Yuan Tang
Xu (Victoria) Tang
Xinjie Yi
Jacob Yu
Haoran Zhang
Rama Edlabadkar
Aalto Lin
Li Yao
Eunice Liu

Matthew Andrews
Tina Gong
Dhati Oommen
Pranav Ramesh
Aseel Rawashdeh
Josh Rosenblum
Omar Mohammad Siddiqui
Dhrubhagat Singh
Eric Tang
Alice Wu
Matthew Andrews

… and more!

Course Components

Lectures, Labs 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 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 9:00-10:15am in person @Science Center Hall B and @Zoom for Extension School Students (zoom link is on canvas under zoom).

Lecture format

ASYNCHRONOUS

- Quiz
- Finish exercises from previous lecture
- Reading

SYNCHRONOUS

Questions from asynchronous material and review

Live Lecture

Q&A

Hands-on exercises in breakout rooms

Discussion about the exercises

Repeat ⋮

Summary and conclusions

Lectures, **Labs**, and Office Hours

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

Friday 9:00*-10:15am in person @Science Center, Hall B @Zoom for Extension School

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

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

BUT

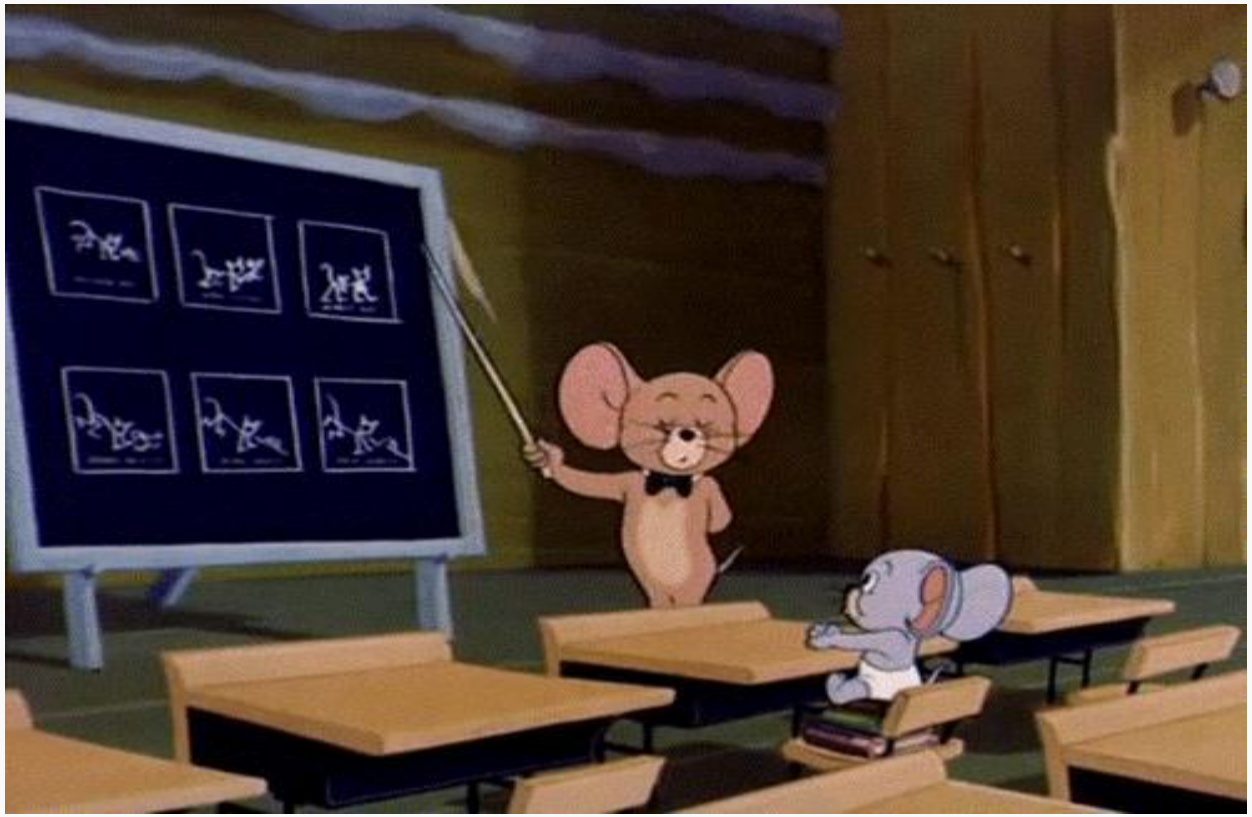
You will earn 1 extra late day for every 8 lectures/labs you attend!

Lectures, Labs, 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: 35%

Homework 0: 1%
Homeworks 1-6: 34%

Students are encouraged to work in pairs on HW assignments.

Exercises: 2%

During lecture.
All test cases are weighted equally.
Due at the beginning of the next morning lecture or lab.

We will only count the exercises category if it helps your overall grade.

Quizzes: 8%

End of each lecture.

<u>1/3 of the quizzes will be dropped from your grade.</u>
All questions are weighted equally.

Due at the beginning of the next lecture or lab.

Midterms: 30%

2 Midterms, each a mix of multiple choice and coding questions.
Multiple choice will be in-person, coding questions will be take-home exam.

Projects: 25%

Milestone dates and details to be announced soon.

Homework(s)

There will be 6 homeworks (not including Homework 0):

- Homework 0 (due Sept 13th; all honest attempts get full credit)
- Homework 1: Web scraping, BeautifulSoup, Basic Pandas, and Plotting
- Homework 2: Regression kNN and LinReg
- Homework 3: Multi- & polynomial Regression, Regularization, Inference
- Homework 4: High Dimensional Data and PCA
- Homework 5: Logistic Regression
- Homework 6: Trees, Bagging, Random Forest, and Boosting

Homework(s)

You are encouraged but not required to submit in pairs on HWs 1-6

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

HWs 1-6 are **due 10 pm Wednesdays,** and homework will be released on Wednesdays.

Late submission policy: Each student is allowed up to 4 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**.

⏳ Digestion Time

Final Project

There will be a final group project (3-5 students) due during exams 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 September 27th.

Help

The process to get help is:

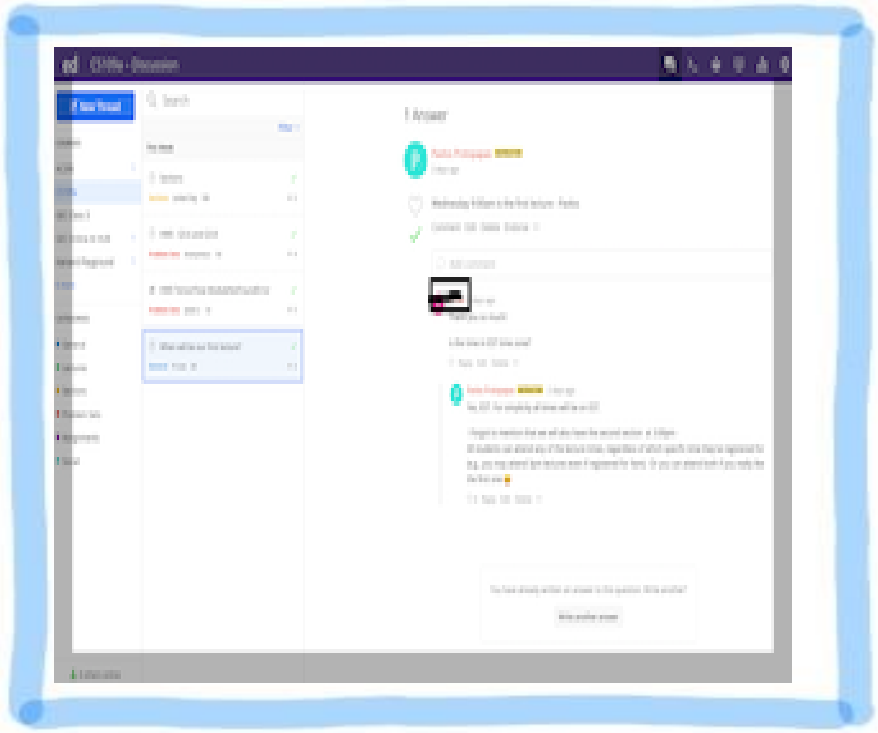
1. Post the question on *Edstem*, and hopefully, your peers will answer. The teaching staff also monitor and respond to posts.
2. Attend the Office Hours; this is the best way to get he
3. For private matters, send an email to the Helpline: **cs1090a2024@gmail.com**.
4. For personal matters, send an email to Pavlos or Natesh.

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!

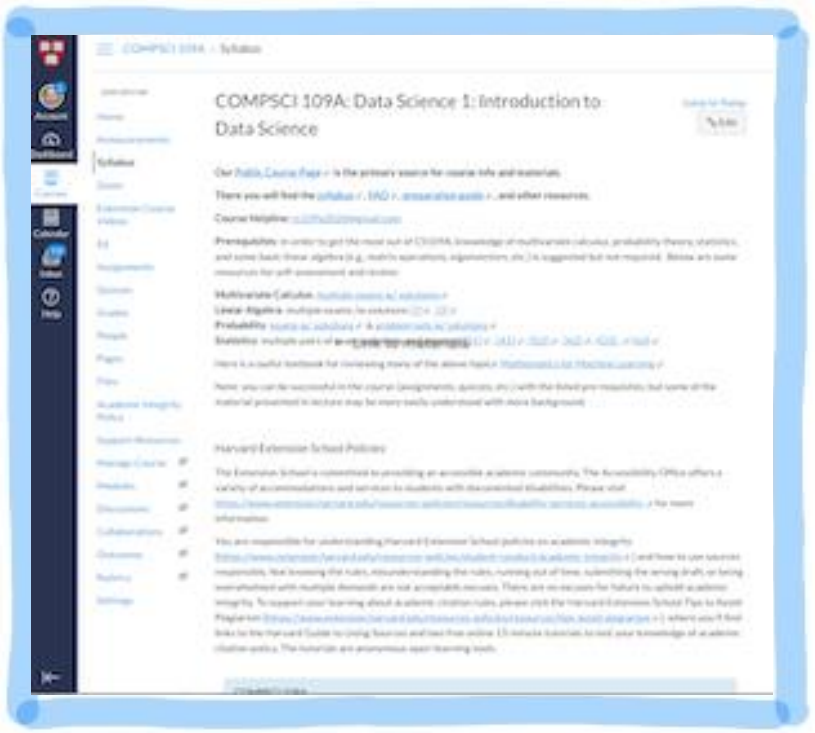
Tools for the course

edstem



- Forum
- Quizzes
- Reading assignments
- Hands on exercises
- Lecture slides

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

FAQ

Can I take this class asynchronously?

College students: This is not allowed.

Graduate students: This 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?

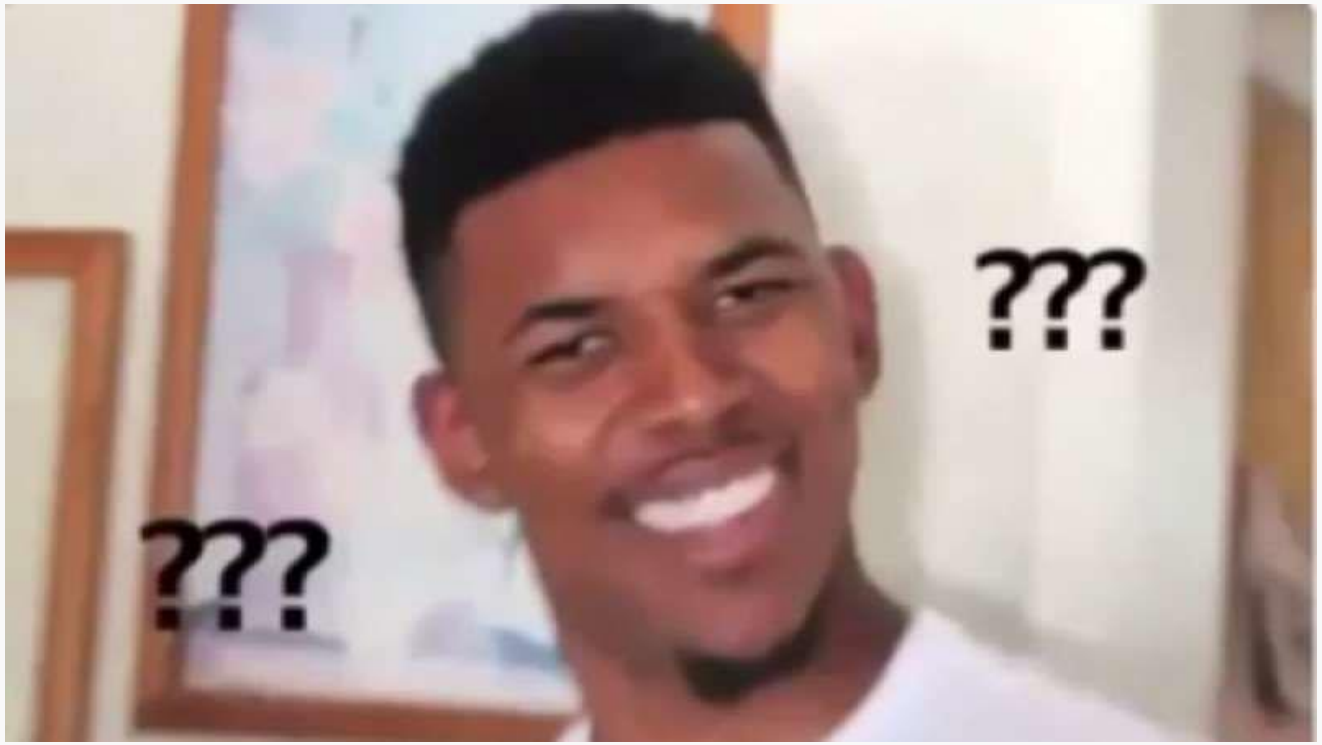
Proficiency in Python, basic math (calculus), basic stats are expected.

We offered a class called *Bedrock Data Science* this summer which helps with some of these topics.

<u>We are making these material available under resources on ED for you to brush up on your Python, linear algebra, and statistics.</u>

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 1 is on 10/18 at 9:00 am in person

Extension school have 3 Zoom time slots on 10/18 & 10/19.

Midterm 2 is on 12/11 at 9:00am in person

Extension students have 3 Zoom time slots across 12/11 & 12/12

Make sure these are on your calendar!

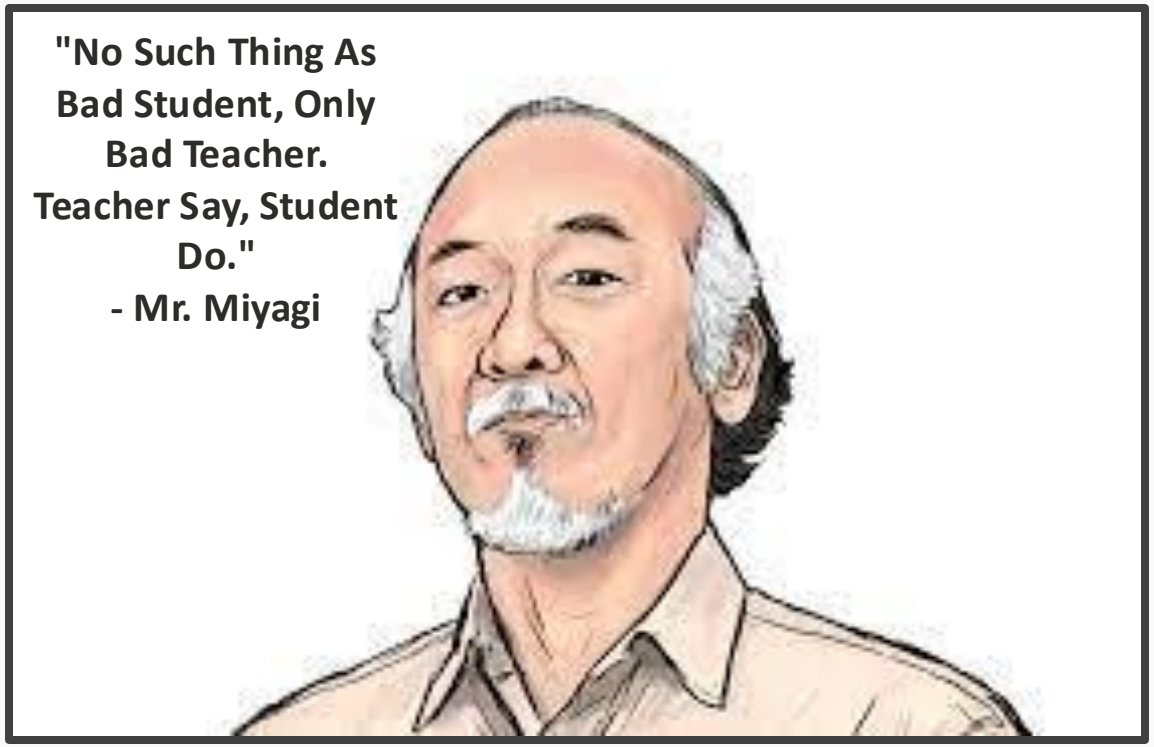
FAQ

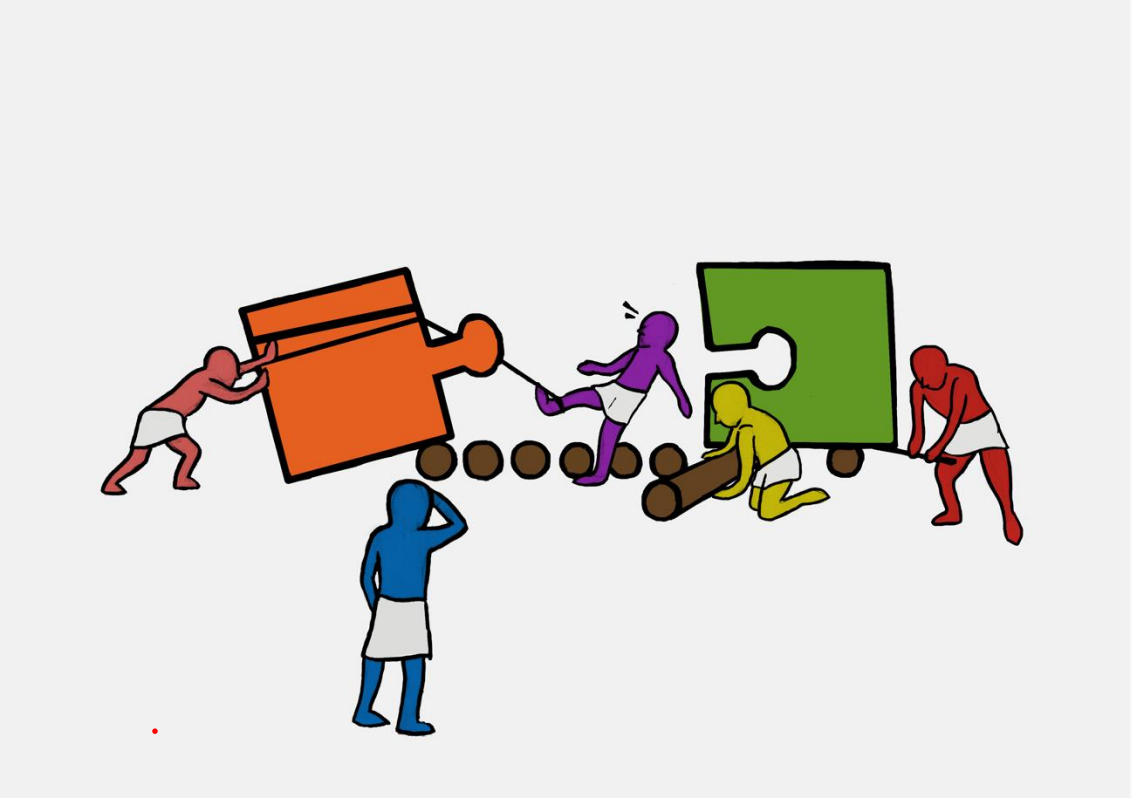
I have a project in mind. Can I use it for the course?

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**





Breakout rooms and in-class exercises

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?

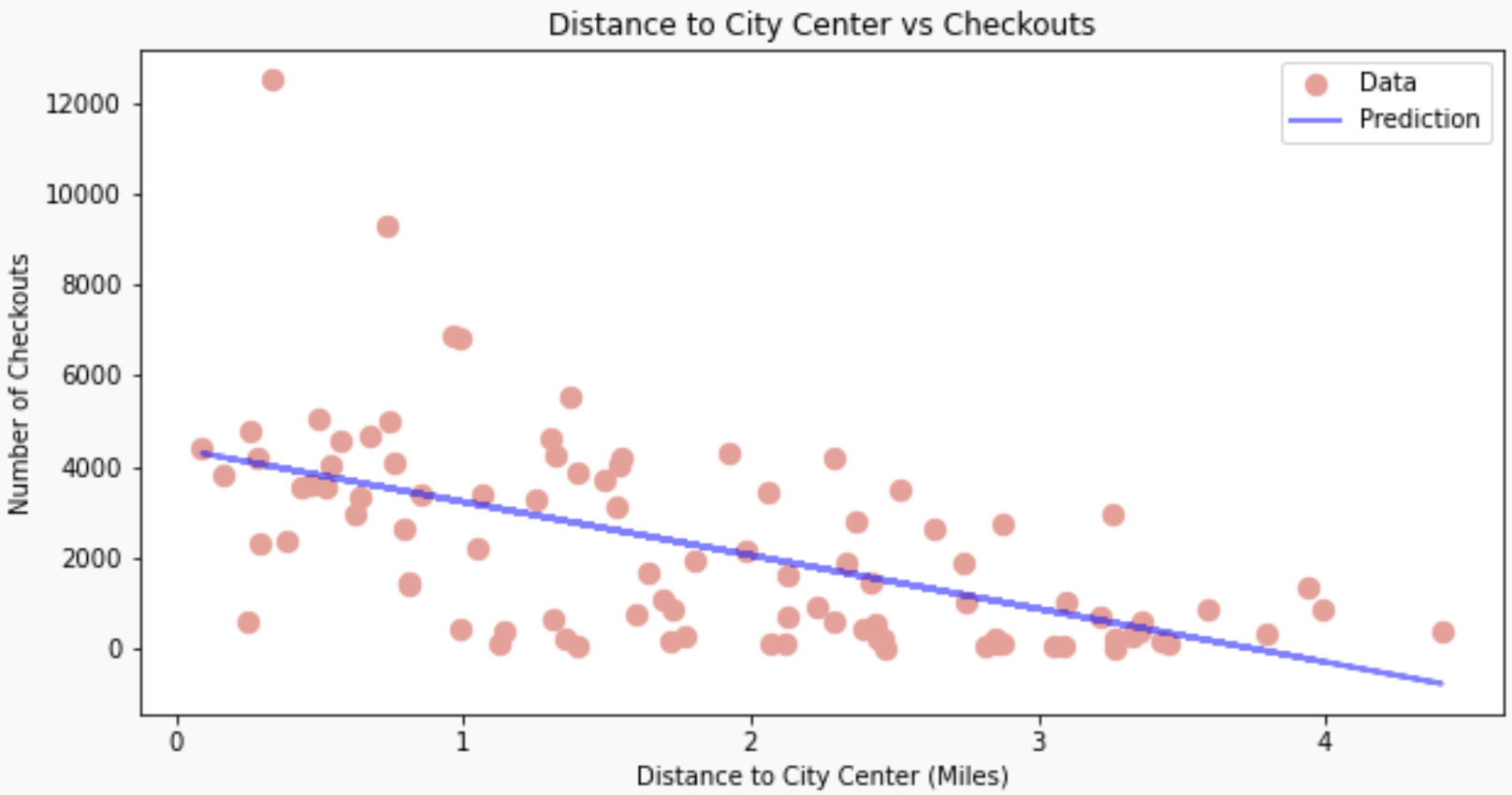
Options

A. 45000

B. 12530

C. 1450

D. 650



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?

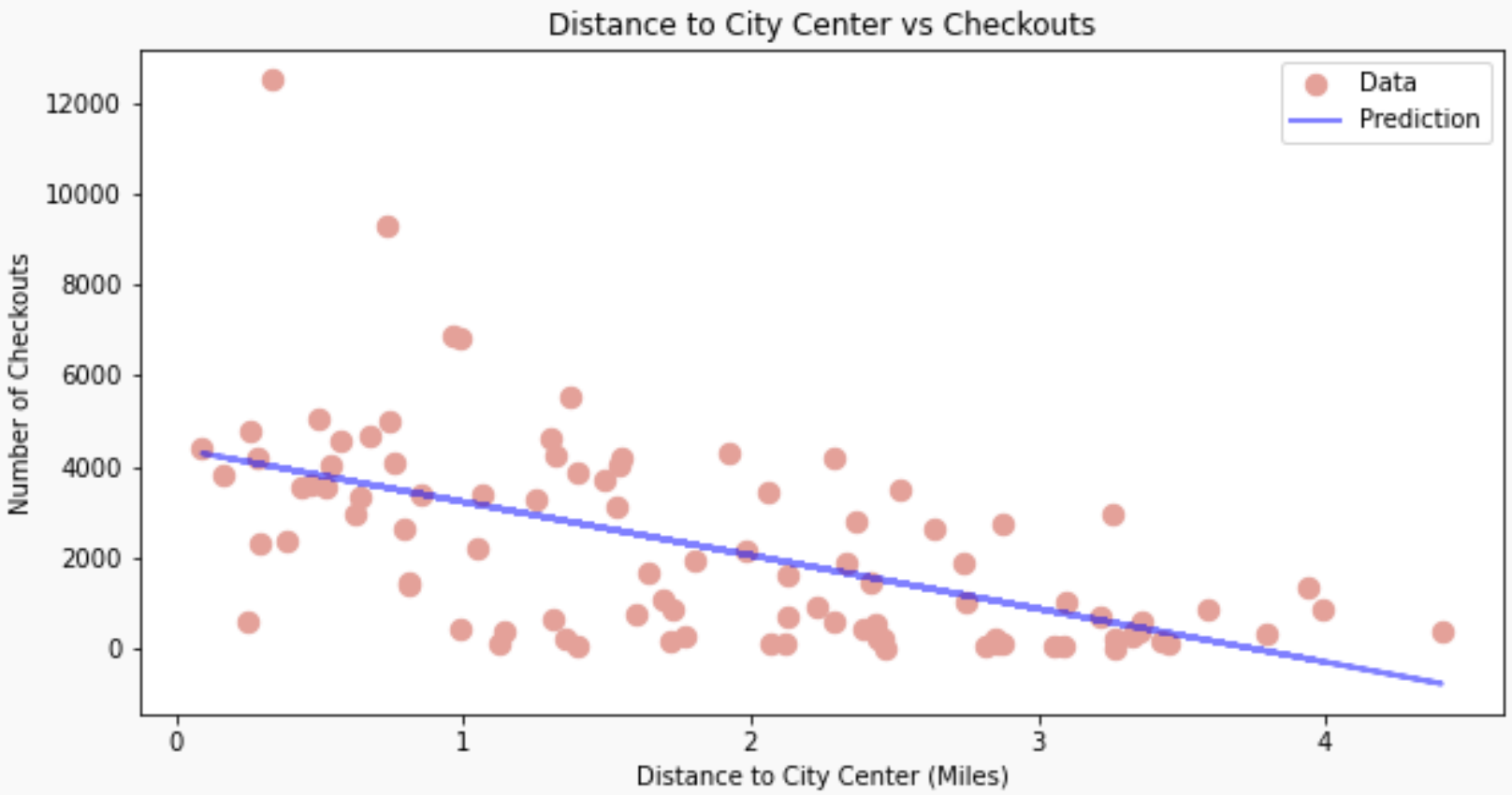
Options

A. 45000

B. 12530

C. 1450

D. 650



What is the goal of CS109A?

Options

A. To teach you data science.

B. To make your life difficult and painful.

C. To predict the next stock price crash.

D. To enable computers to talk.

What is the goal of CS109A?

Options

A. To teach you 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:

Pierce Hall Room 209
from
10:30 AM - 12:30 PM