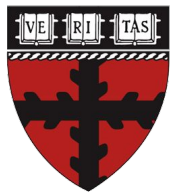
Lecture 1: Introduction

AC215

Pavlos Protopapas
SEAS/Harvard



Outline

1. Why should you take this class? And why not?
2. Who are we?
3. Course structure and activities: What can you expect?
4. Class organization: What are the workload, logistics, and grading system?

Projects

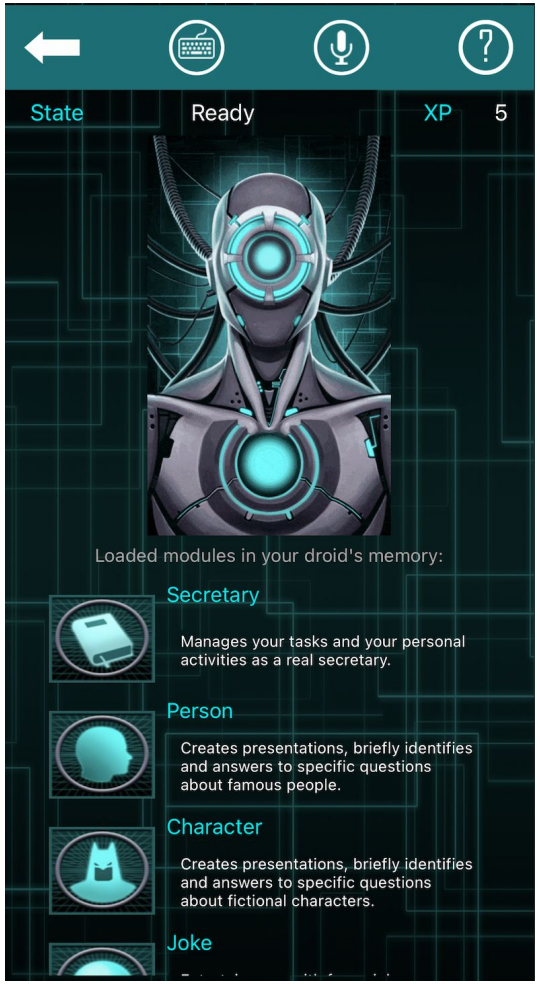
Outline

1. **Why should you take this class? And why not?**
2. Who are we?
3. Course structure and activities: What can you expect?
4. Class organization: What are the workload, logistics, and grading system?

Projects

Why You Should Take This Class

So you can build awesome apps like this:



- https://runwayml.com/
- https://www.databot-app.com/

Why You Should Take This Class

Because you want to:

- Put your models into production
- Build an application using your models
- Take advantage of available models
- Integrate and orchestrate applications
- Deploy and manage increasing amounts of data

Why You Should Take This Class

What students said in previous iterations

- *“This course helped me with kickstart a startup”*
- *“This course helped me land a job - nailed that interview”*
- “*The ‘money’ course. Truly one of the most practical course*”
- *“I knew some of these concepts but going through all gave me a good idea why and what”*
- “*It was fun! I love Pavlos*”

Reasons This Course May Not Be For You

1. Lack of Commitment: This course demands full engagement with both materials and projects.

2. Expectation of a Traditional Lecture Format: This is a project-based course, focused on hands-on learning.

3. Limited Background in Prerequisites: If you are unfamiliar with key concepts from CS109A/B such as:
   - Basic Machine Learning
   - CNNs, RNNs, Autoencoders, Language Models, GANs, etc.
   - Basic shell commands.

Reasons This Course May Not Be For You

4. Unwillingness to Receive and Apply Feedback: We provide detailed feedback on your projects and expect you to implement changes in subsequent milestones.

5. Seeking a "Easy A": While project-based courses may seem easier, we hold rigorous evaluation standards.

6. Thinking this is CS109C: This course is not about learning new models and methods. Instead, it focuses on productionizing your existing knowledge and skills.

7. You do not like cheese!

Reasons This Course May Not Be For You

What students said in previous iterations

- *"This is more of an engineering course than a traditional data science or ML course."*
- *"You won't learn many new methods here."*
- *"Why work on GCP instead of AWS, which is the leading cloud computing platform?"*

Reasons This Course May Not Be For You

- *"No homework! Because this is a team-based course, you might miss out on learning some key concepts."*

Motivation

Mckinsey Global Survey findings on Adoption of AI shows nearly 25% year over year increase in the use of AI. 50% of companies spend between 8 and 90 days deploying a single AI model, with 18% taking longer than 90 days. A report by IDC that surveyed 2,473 organizations and their experience with ML found that a significant portion of **attempted deployments fail**, quoting **lack of expertise**, as one of the key factors.[1]

[1] https://arxiv.org/pdf/2011.09926.pdf

Motivation

A recent International Data Corporation (IDC) survey of global organizations that are already using artificial intelligence (AI) solutions found only 25% have developed an enterprise-wide AI strategy. At the same time, half the organizations surveyed see AI as a priority and two thirds are emphasizing an "AI First" culture.

IDC: https://www.idc.com/

Enrollment: Cap, Pending Status and Criteria

Some of you have petitioned for the course but have not yet been approved. Why?

I typically avoid capping courses unless it impacts the quality and experience.

Currently, the class is capped at 110, with 25 students pending.

Enrollment: Cap, Pending Status and Criteria (cont)

Challenges:

- It's difficult to find teaching staff due to the specialized material covered.
- I want to monitor the progress of projects and interact with all students.
- We can't provide cloud credits for everyone.

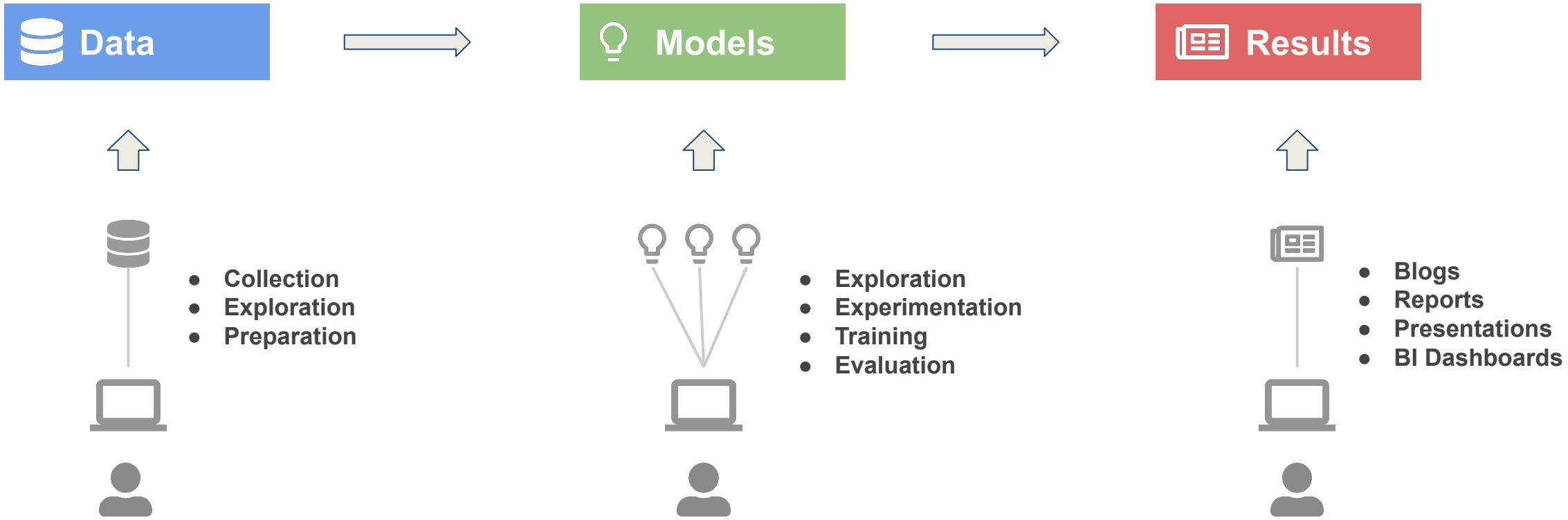
Considerations for Enrollment:

- Likelihood of the student taking the course next year.
- Whether the course is a program requirement.
- Comments in my.harvard.

I anticipate some students will drop the course, which will open up additional spots.

Data Science Series to Real World

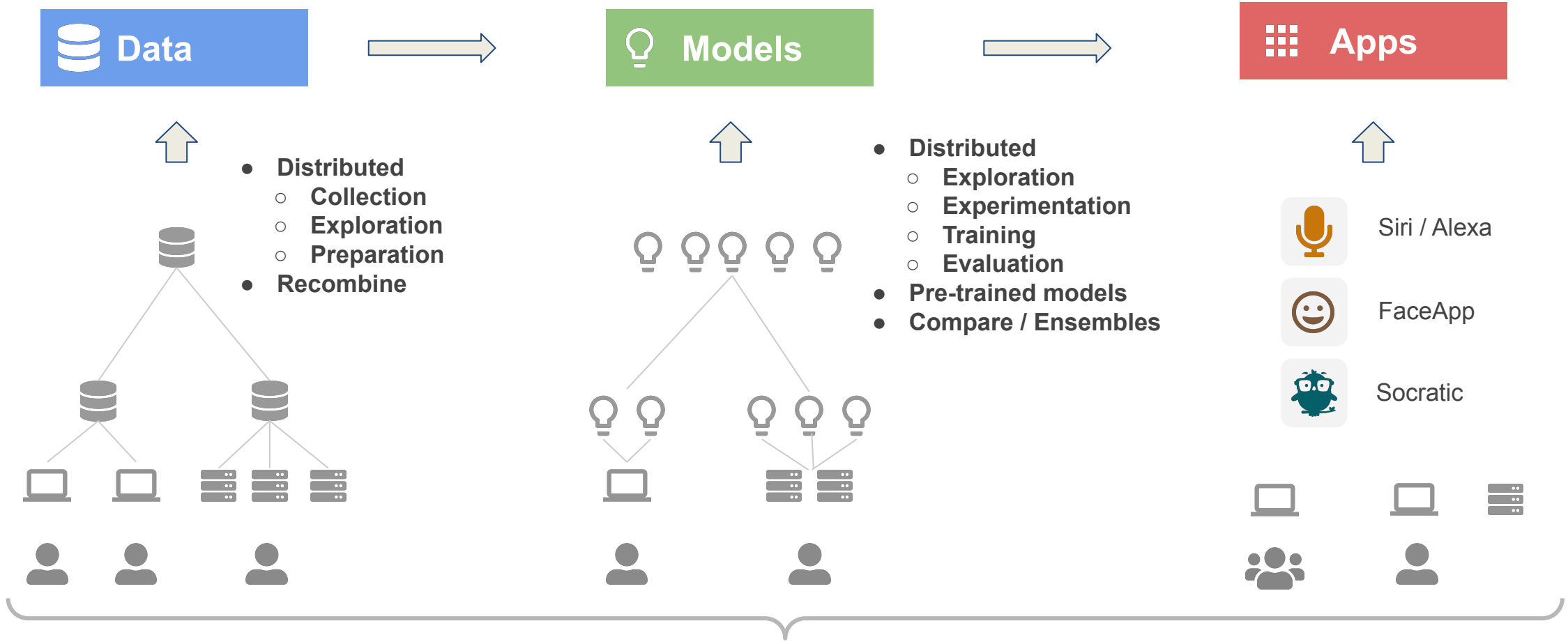
Data Science Series CS109 A/B



Single developer on one computer. Projects are individual to 2-3 member team.

Data Science Series to Real World

Real World



Team of developers on single/ multi node clusters on a Cloud Platform. Projects are 5+ member teams

Data Science Series to Real World (cont)

Challenges:

- Onboarding Procedures for New Team Members:

- Required Installations for Specific Operating Systems:

  All team members have the necessary software and tools installed on their systems.

- Guidelines for Code Collaboration:

  Version control and code reviews.

- Methods for Sharing Datasets and Models

- Automation of Data Gathering and Model Training

- **Resolving “It Works on My Machine” Issues**

Ops for Machine / Deep Learning

Development Operations (DevOps):

Unifies software development (Dev) & operations (Ops) for efficiency.

AI Operations (AI Ops):

Integrates ML/DL model development with app development & operations.

AI Ops - Tasks

Machine / Deep Learning:

- Data collection & exploration
- Model exploration & selection
- Training & evaluation
- Distillation & compression

Application Development:

- APIs / Model serving
- ML integration
- Web & mobile apps
- Edge device apps
- Automation scripts

Operations:

- Provisioning and managing deployment servers, on-demand GPU servers
- Maintain 100% uptime of app / apis
- CI/CD: Continuous Integration / Deployment
- Continuous Data Collection / Model Training
- Model/data monitoring
- Model/data versioning
- ML Workflow Management

MLOps - Tech Stack

Data Engineers | Data Scientists | Software Engineers | Systems Engineers

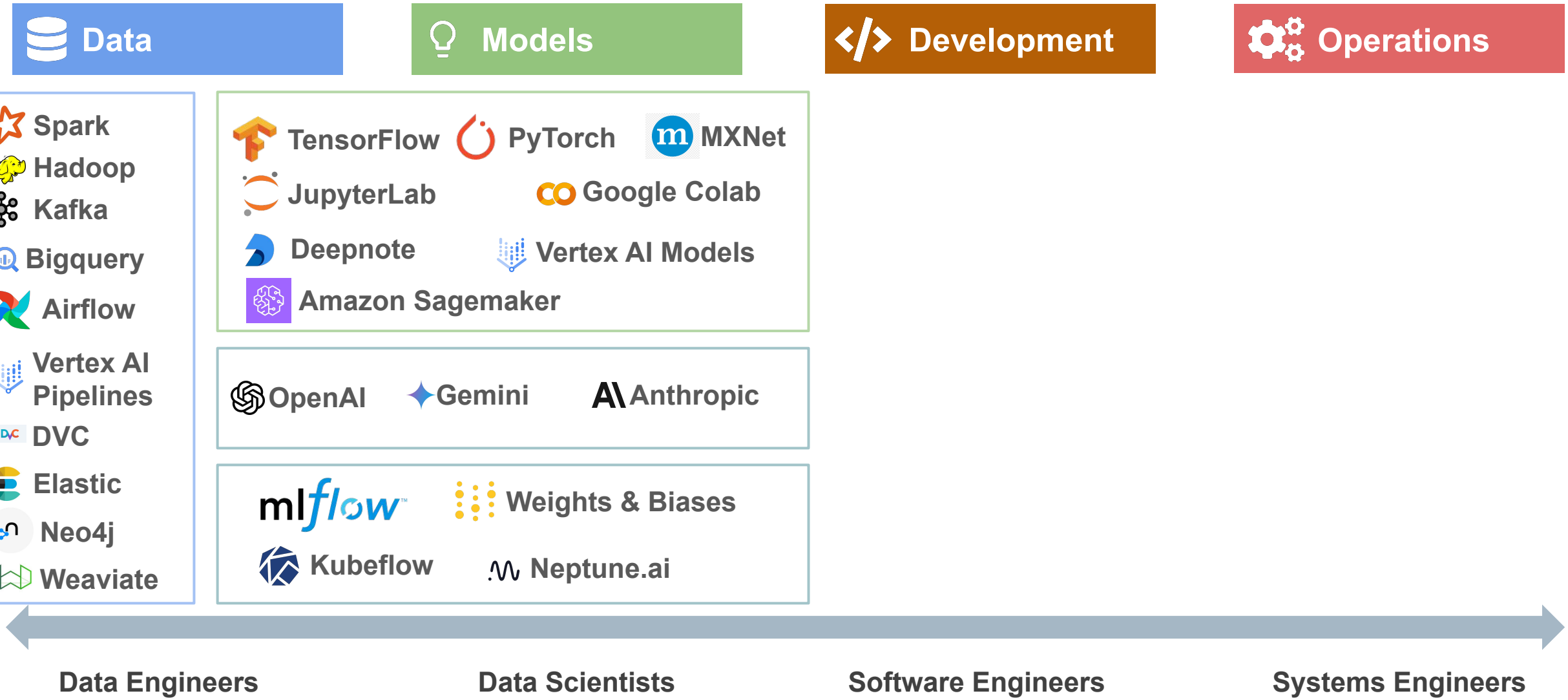
MLOps - Tech Stack

Data | Models | Development | Operations

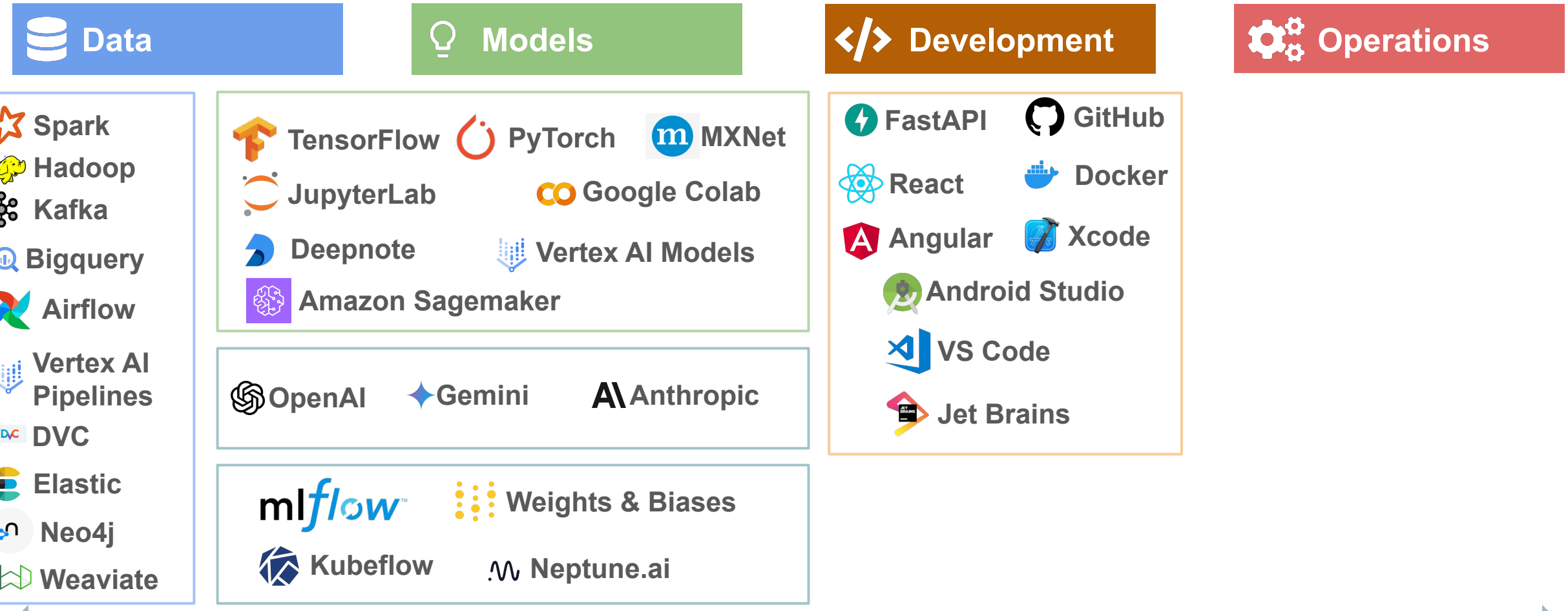
- Spark
- Hadoop
- Kafka
- Bigquery
- Airflow
- Vertex AI Pipelines
- DVC
- Elastic
- Neo4j
- Weaviate

Data Engineers | Data Scientists | Software Engineers | Systems Engineers

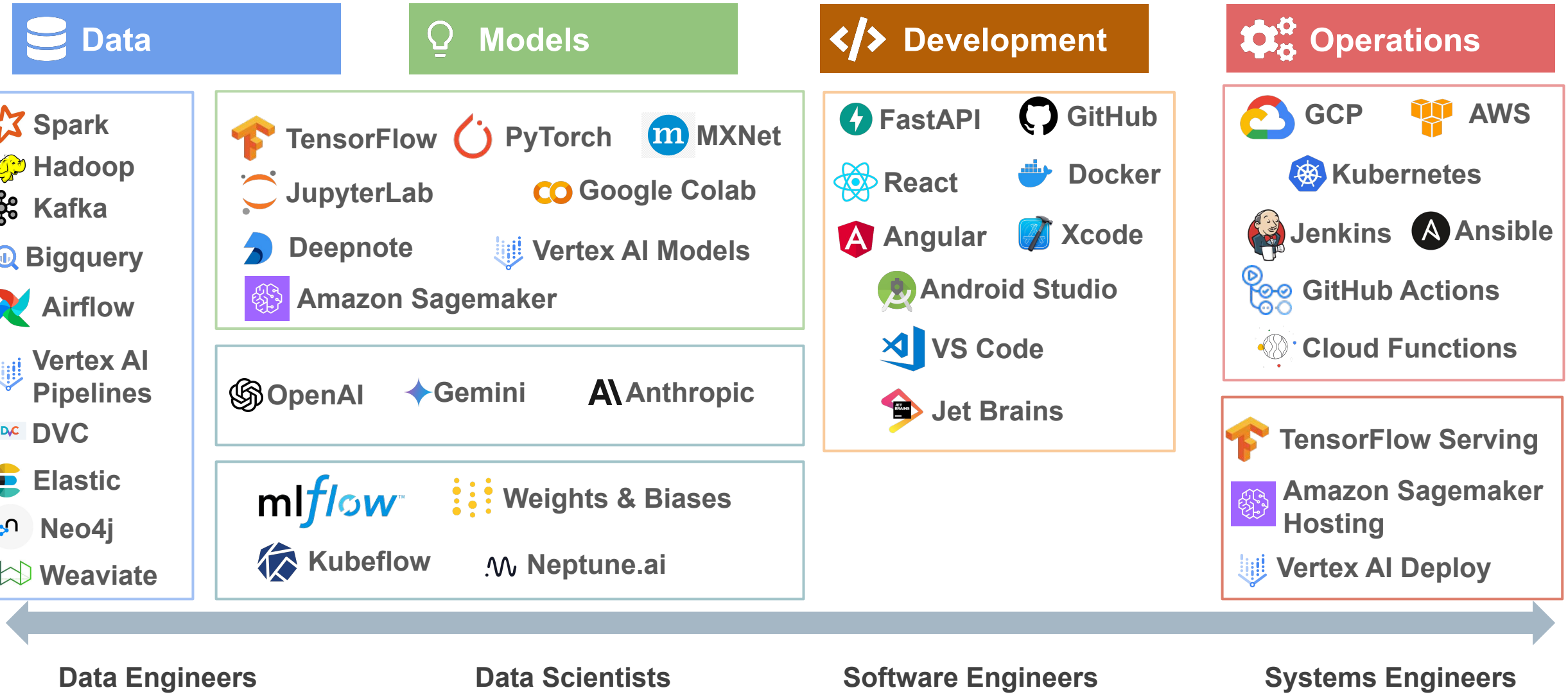
MLOps - Tech Stack



MLOps - Tech Stack



MLOps - Tech Stack



MLOps - Tech Stack

AI / ML

Data:
- Label Studio
- DVC

Framework:
- TensorFlow
- PyTorch

ML Tasks:
- Google Colab
- Vertex AI

Tracking:
- W&B
- Custom

Development

APIs:
- FastAPI
- Model Serving

LLM APIs:
- OpenAI
- Gemini

Frontend:
- React

IDE:
- VS Code
- IDE of choice

Operations

Source Control:
- GitHub

Containerization:
- Docker

Cloud Provider:
- Google Cloud Platform

Continuous Integration/ Deployment:
- GitHub Actions, Ansible

Scaling:
- Kubernetes

Outline

1. Why should you take this class? And why not?
2. **Who are we?**
3. Course structure and activities: What can you expect?
4. Class organization: What are the workload, logistics, and grading system?

Projects

Who? The Astro-AI-Statistician Who Rocks The Kitchen!

Roles:

- The Science Wizard: Scientific Director at OMPD (DS and CSE Masters).
- Course Maestro: Instructs CS109a, CS109b, and AC215 like a boss.
- Astro-Guru: Stoked about the next-gen telescopes that are about to revolutionize our view of the universe!

Research Lab:

- StellarDNN Maverick: Tackles cosmic mysteries where astronomy, ML, and stats collide!
- His Interests? Cracking open differential equations with deep neural networks, being a detective in deep neural network inference, and teaching NLP techniques to chat with stars (well, in astronomical time series analysis, but let's keep it jazzy!).

Who? The Astro-Statistician Who Rocks The Kitchen!

Fun Facts:

- Musical Soul: Classical tunes and opera are his jam. The Boston Symphony Orchestra is his second home!
- Culinary Artist: Holds a cooking badge of honor from Le Cordon Bleu and enjoys both whipping up a storm and devouring the results.
- Adventure Junkie: From biking up mountains to skiing down them, from kayaking to hooking fish mid-air—this professor is always on the go!

Who?

Shivas Jayaram

Deep Learning Researcher, Educator and Practitioner

Currently working on a medical-pharma startup

Fun Fact: Just started a new hobby - Beekeeping

Rashmi Banthia

TF for many Data Science classes here at Harvard including CS109A/B.

Fun Fact: Enjoys kaggle competitions

Chris Gumb

SEAS preceptor, assisting with courses for the DS and CSE Master's programs.

Fun Fact: Used to work at a (now defunct) comedy club in Boston.

Ignacio Becker

He has a Ph.D. in Computer Science and trained as an astronomer.

His main area of research is AI applied to astrophysical problems. Nowadays, he focuses on developing models to leverage the massive real-time data stream.

Who?

Luis Ribeiro

Machine Learning Engineer
Fun Fact: I'm considering creating a YouTube channel about AI/DS

Li Yao

TF for many Data Science classes including CS109
Fun Fact: Has a collection of rare stones

Yasmine Morrison

Venture capitalist investing in pre-seed/seed startups.

Fun Fact: Started a podcast called Venture Analytics (launching soon)

Outline

1. Why should you take this class? And why not?
2. Who are we?
3. **Course structure and activities: What can you expect?**
4. Class organization: What are the workload, logistics, and grading system?

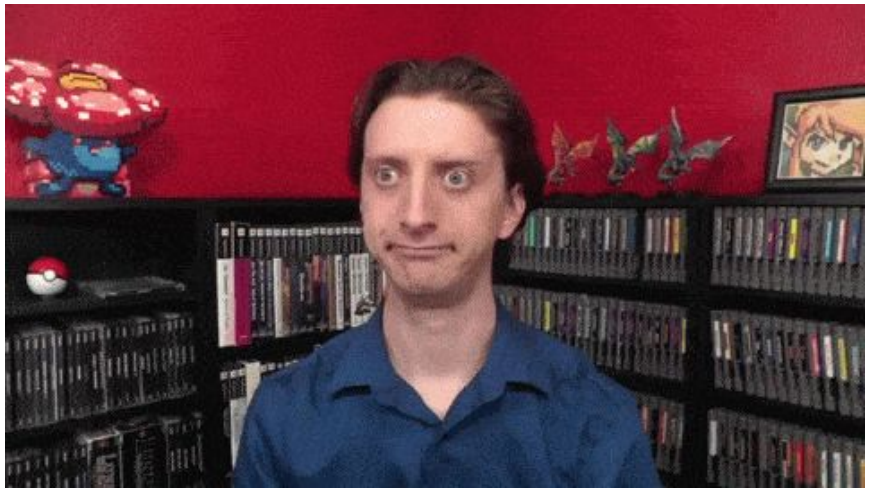
Projects

Course Structure and Activities

- Two lectures/tutorials per week
- Projects
- Assignments

Lectures

Two lectures per week. There will be a combination of lecture style and hands on tutorials.



Lectures

Where do I find the class material?

The official web page for this course is:
https://harvard-iacs.github.io/2024-AC215

All lecture PDFs and link to tutorials will be posted here:
https://harvard-iacs.github.io/2024-AC215/schedule/

Topics

- Containers
- Data Pipelines and Cloud Storage
- Data Parallelization
- Data Versioning
- Advanced Training Workflows
- Advanced Inference Workflows
- Pipeline Design
- App Design, Setup, and Code
- APIs & Frontend
- Scaling with Kubernetes (k8s)
- Additionally, we offer several tutorials on the latest LLM tools, including fine-tuning, agents, and Retrieval-Augmented Generation (RAGs).

Team Projects: Crafting Your Own AI Solutions

Objective: Hands-on project development in AI & MLOps; transform your idea into a functional app.

Guidance: Weekly demonstrations from Pavlos' project provide practical insights and a reference point.

Milestones: Assess project evolution and grasp of MLOps concepts; crucial for grading.

Creativity: Open platform for start-up ideas, research, or personal hobbies.

Assessment: Milestones are key to **grades** and holistic development.

Group Formation: Starts today, aiming for teams of 3-4.

Projects Milestones

Milestone	Description	Due Date	Points
MS1	Project Proposals & Team Formation	09/19	4
MS2	MLOps & Advanced Training	10/18	10
MS3	Midterm	10/31	25
MS4	Deployment	11/15	14
MS5	Final Presentation and Deliverables	12/11	35

For the most up-to-date and accurate information, please visit: https://harvard-iacs.github.io/2024-AC215/

Assignments/Homework

	Description	Due Date	Points
HW1	Containers / Setup quota etc		4
HW2	Create a simple api and consume it		4
HW3	Use GCP to create a VM and deploy a container		4

For the most up-to-date and accurate information, please visit: https://harvard-iacs.github.io/2024-AC215/

Outline

1. Why should you take this class and why not?
2. Who are we?
3. Course structure and activities?
4. **Class organization (Workload, Logistics, Grades).**

Projects

Workload

- 1 hour *Reading*
- 2.5 hours *Lectures*
- 1 hour *Office Hour/meet with your TF*
- 7.5 hours *Project Milestones*
- ~ 12 hours/ week

Expectations: 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.

Expectations: Attendance

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

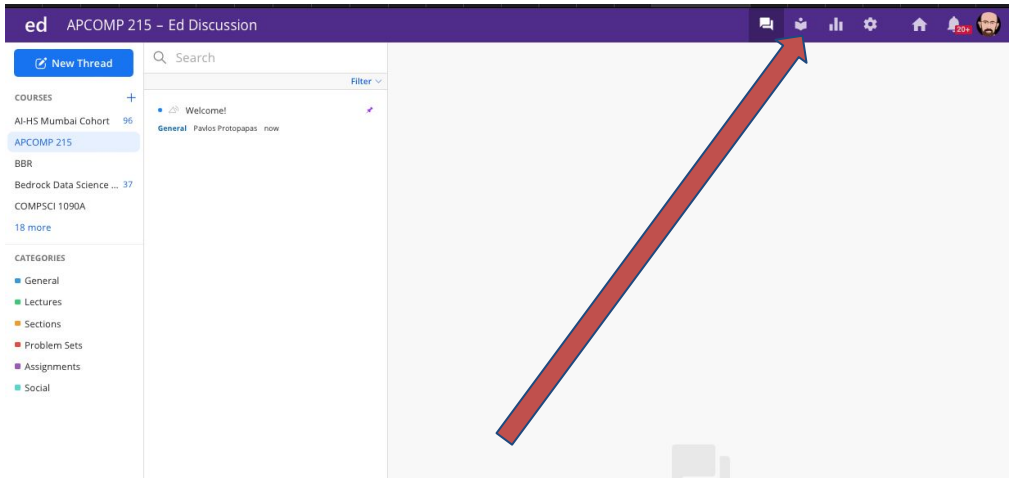
BUT

You will earn 1 extra late day for every 5 lectures you attend!

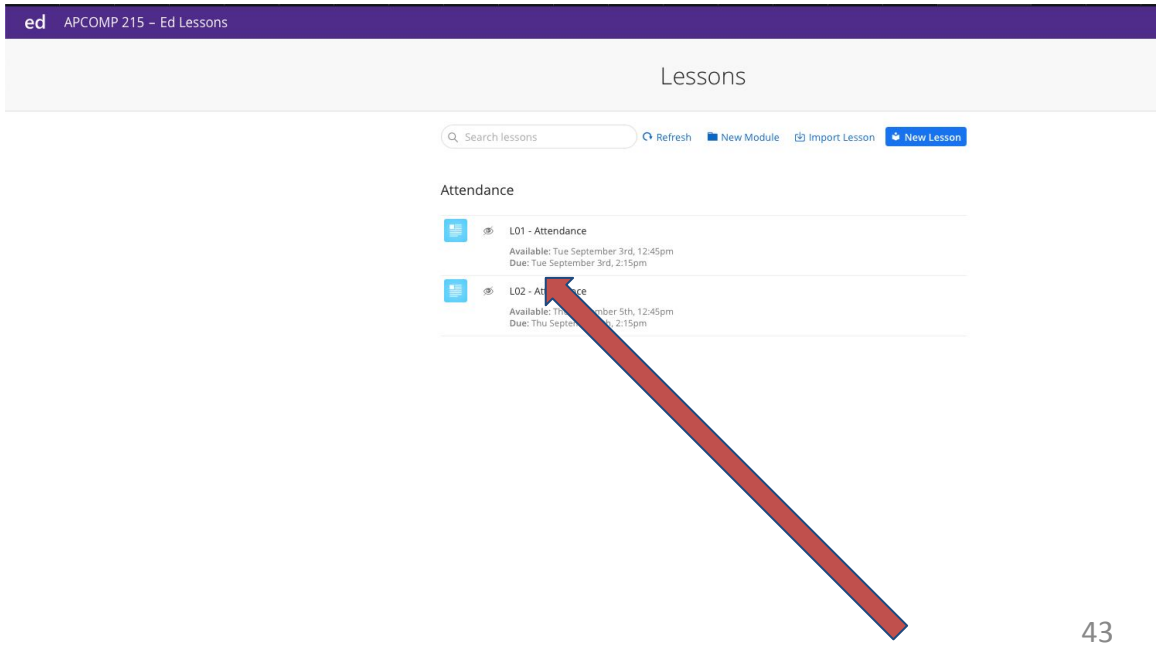
Expectations: Attendance

How to record attendance:

Step 1: go to edstem



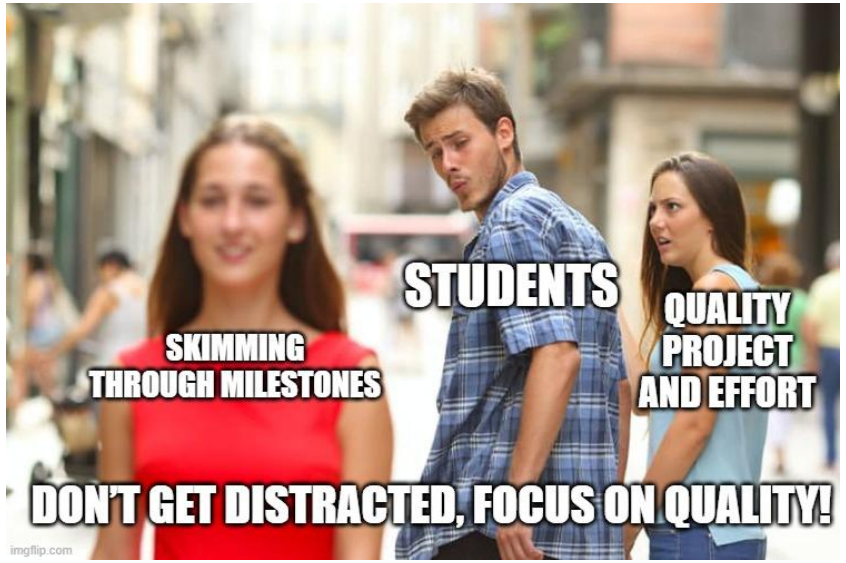
Step 2: Copy the secret word from the board



Expectations: Projects

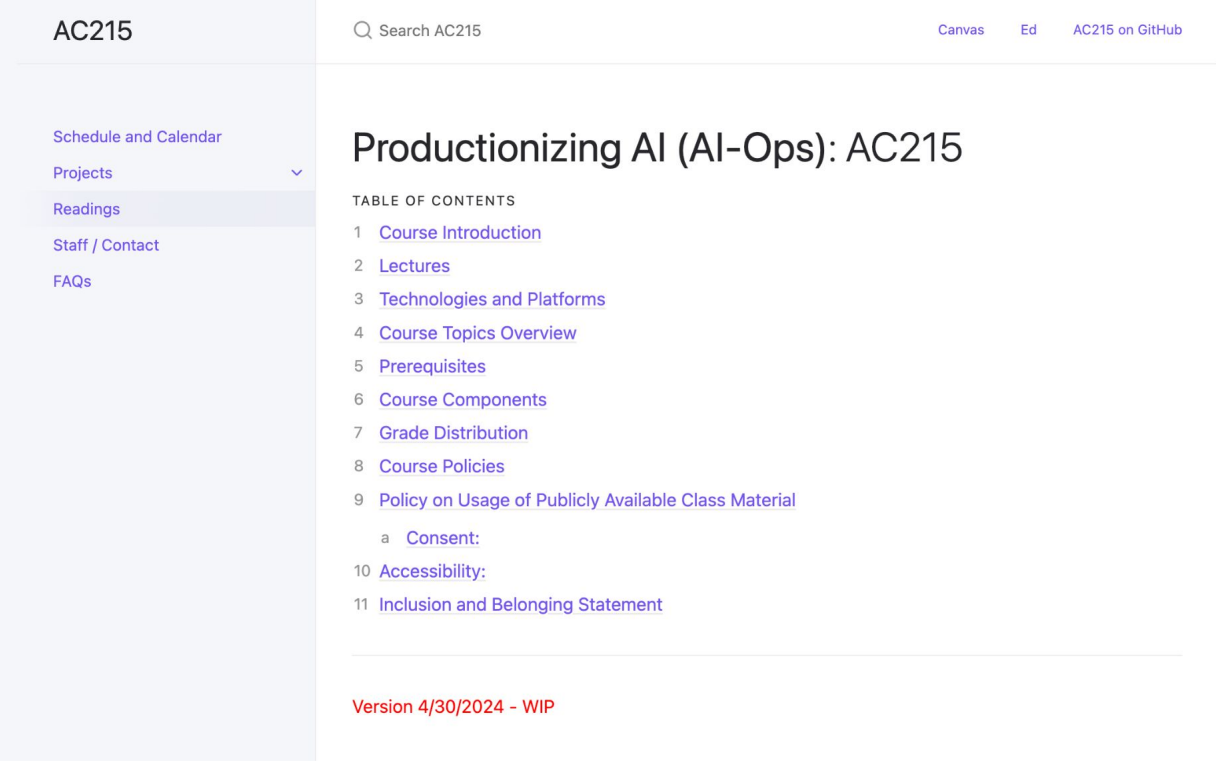
This course is project-based, and your grade will depend on the quality of your project. We will place significant emphasis on both the effort you put into the project and the quality of your completion of its milestones.

Project-based courses may offer opportunities for higher grades, but it is ultimately up to you to learn and apply the feedback provided to improve your work for future milestones.



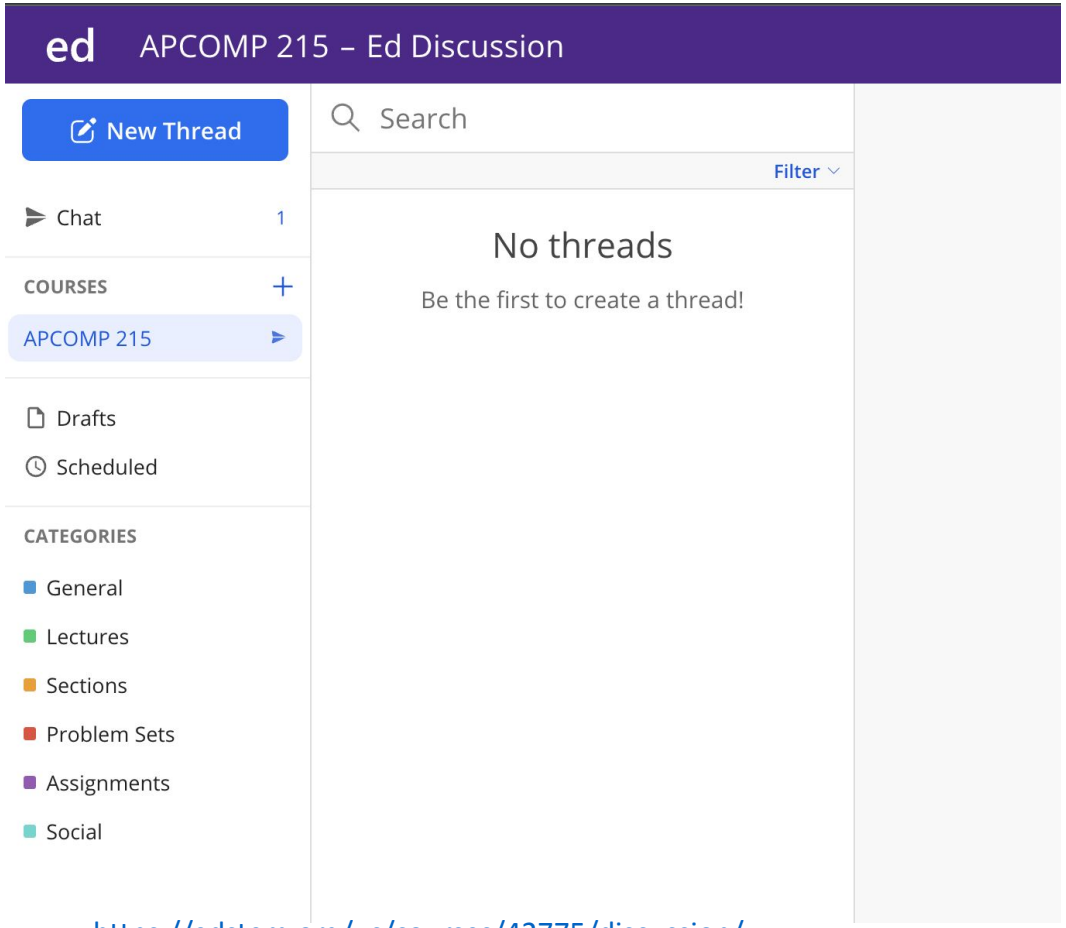
Course Components

Course web page



https://harvard-iacs.github.io/2024-AC215/

ED Stem



https://edstem.org/us/courses/42775/discussion/

Grades

Assignment	Final Grade Weight
Milestone 1	4%
Milestone 2	10%
Milestone 3	25%
Milestone 4	14%
Milestone 5	35%
Homework 1	4%
Homework 2	4%
Homework 3	4%
Total	**100%**

For the most up-to-date and accurate information, please visit: https://harvard-iacs.github.io/2024-AC215/

Final Details

- We will be using ED for discussions, announcements and surveys
- Canvas for some of the submissions and group formations

Submissions for project milestones and projects will be using GitHub – details will follow soon

Outline

1. Why should you take this class and why not?
2. Who are we?
3. Course structure and activities?
4. Class organization (Workload, Logistics, Grades).

Projects

Projects

In Class Demo: **Formaggio.me**, an AI cheese sommelier.

Project Idea

- Pavlos likes cheeses and when he throws parties he always have cheese for his guests.
- He wants to build an app to identify cheese types and learn what cheeses go with each other, how to pair it with wines, the way this particular cheese is made, a history of the cheese etc.
- Project Summary

Formaggio.me Problem Definition

Imagine being able to identify a cheese by simply taking a photo of it. Our app uses AI-powered visual recognition technology to help you identify the cheese you're looking at, and then provides you with a wealth of information about it.

Take a photo of the cheese, and our app will identify it for you. Then, dive deeper into the world of cheese with our interactive chatbot. Ask questions about the cheese's origin, production process, nutritional information, and history.

Formaggio.me Problem Definition

Want to host a cheese-tasting party? **Formaggio.me** makes it easy. Use our app to select the perfect cheeses for your gathering, and then get expert advice on pairing them with wines, crackers, and other accompaniments. Our chatbot is always available to help you plan the perfect cheese platter.

Formaggio.me is your one-stop-shop for all things cheese. With our app, you'll never be stuck wondering what that delicious cheese is or how to pair it with other foods. Whether you're a cheese aficionado or just starting to explore the world of cheese, **Formaggio.me** is the perfect companion for your culinary journey.

Proposed Solution

Key Features:

- Visual cheese identification using AI-powered technology
- Interactive chatbot for asking questions about cheese
- In-depth information on cheese origin, production process, nutritional information, and history
- Expert advice on pairing cheese with wines, crackers, and other accompaniments
- Perfect for cheese enthusiasts, party planners, and anyone looking to explore the world of cheese

Project Execution Steps

- Project Ideation / Requirements
- Data Exploration
- Model Exploration
- Prototyping
- Model Serving
- Product Development
- ML Integration
- Deployment

Resources

We will be using Google Cloud Platform (GCP). Each student receives a credit allocation, and we have extensive experience working with GCP. All tutorials will be conducted in GCP environments using TensorFlow and PyTorch.

While AWS tutorials will be provided, they are not fully supported or extensively tested.

Additionally, we expect to have GPU credits from Modal, along with a guest lecture.

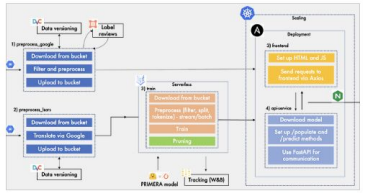
Previous Year's Projects

AC215 Fall 2023

RAG Detective: Retrieval Augmented Generation with website data

This article was produced as part of the final project for Harvard's AC215 Fall 2023 course.

Ian Kelk
Dec 10, 2023 · 16 min read



ML-powered restaurant review summarisation with cultural insights

This article was produced as part of the final project for Harvard's AC215 Fall 2023 course.

Varun Ullanat
Dec 14, 2023 · 13 min read

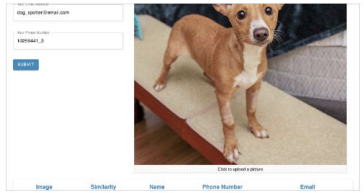
PAVVY—Learning Tools For Lecture Content

Caleb Saul
Dec 12, 2023 · 10 min read

ScienceTutor App: An Educational Application For Children

This article was produced as part of the final project for Harvard's AC215 2023 course.

Isjnancy
Dec 12, 2023 · 11 min read



Spotted! An app to help lost dogs get home.

by Olga Leushina and Alex Coward

Alex Coward
Dec 14, 2023 · 10 min read

DawgAI: Helping Dog Lovers Identify their Favorite Breeds

Final Project for Harvard's AC215 Fall 2023 Course.

Curren Iyer
Dec 12, 2023 · 9 min read

https://medium.com/institute-for-applied-computational-science/ac215-fall2023-projects/home

How to Scope your Project

Proof Of Concept (POC)

- Experiment potential ideas
- Check feasibility of the idea
- Use a subset of data to make experiments simpler to run
- E.g.: Verify if our language task can be performed by transfer learning using a transformer model
- **Users:** Internal team
- **Duration:** Days to few weeks

Prototype

- A mockup or functional product that can showcase your ideas
- E.g.: A mockup web app to show user experience and flow
- **Users:** Internal team
- **Duration:** Weeks

Pilot

- A usable and functional product of your solution
- Used to test out the product with real users and performing real use cases
- E.g,: An api endpoint of a model for prediction, a simple one page app to showcase a model's prediction capability
- **Users:** Internal / External
- **Duration:** Weeks

Minimum Viable Product (MVP)

- Expanding on the Pilot to build something that real users can use
- E.g.: Production deployed app that can predict the cheese and respond to prompts
- **Users:** External
- **Duration:** Months

Project Scope (Cheese App)

Proof Of Concept (POC)

- Scrap cheese images and documents (books etc)
- Verify images and pdfs
- Experiment on some baseline models
- Verify new unseen cheeses are predicted by the model(s)
- Verify ideas using any instruct-LLMs

Prototype

- Create a mockup of screens to see how the app could look like
- Deploy one model to Fast API to service model predictions as an API

Minimum Viable Product (MVP)

- Create App to identify Cheeses and respond appropriately to a series of prompts
- API Server for uploading images and predicting using best model
- API Server for serving the language models

Office Hours (For this week only)

Li Yao - Tuesday 3:30 - 4:30 PM

Rashmi B - Friday 9:30-10:30 AM - Online (Canvas -> Zoom)

Logistics

- Survey
- Make project groups
- Setup & Installations

THANK YOU