Lecture 1: Introduction

AC215

Pavlos Protopapas

Institute for Applied Computational Science, Harvard

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).

---

1. Projects

Outline

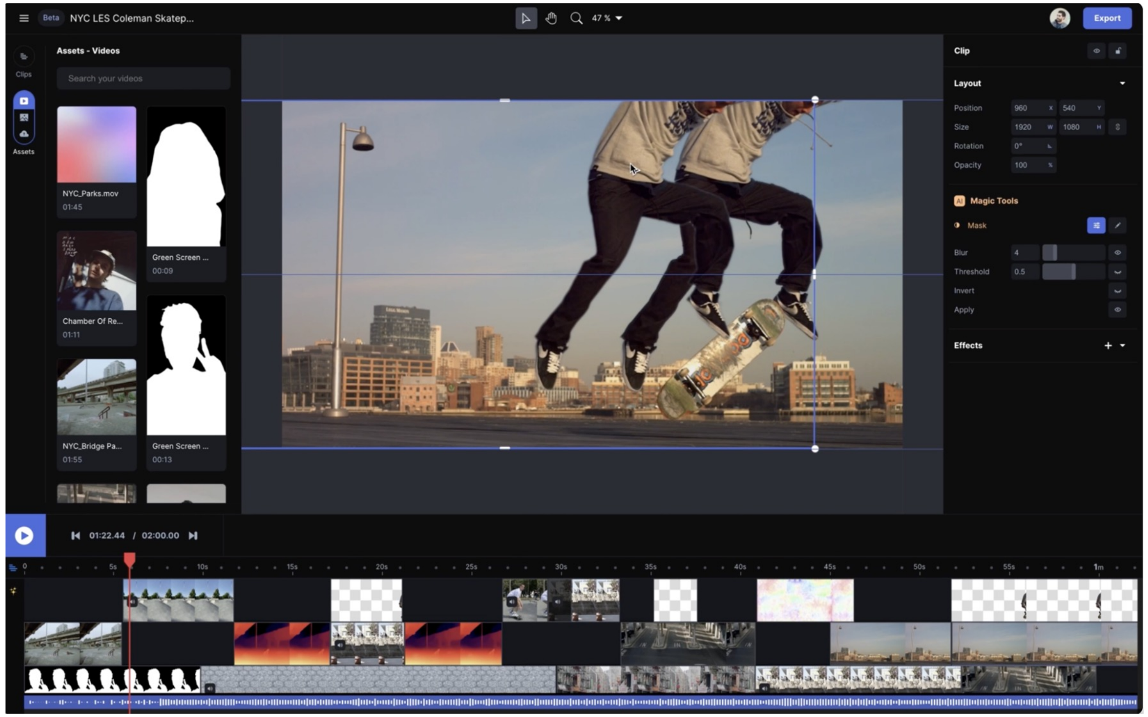
1. **Why should you take this class and why not?**
2. Who are we?
3. Course structure and activities?
4. Class organization (Workload, Logistics, Grades).

---

1. Projects

Why you should take this class

So you can build awesome apps like this:



Position 960 X 540 Y

Size 1920 W 1080 H

Rotation 0°

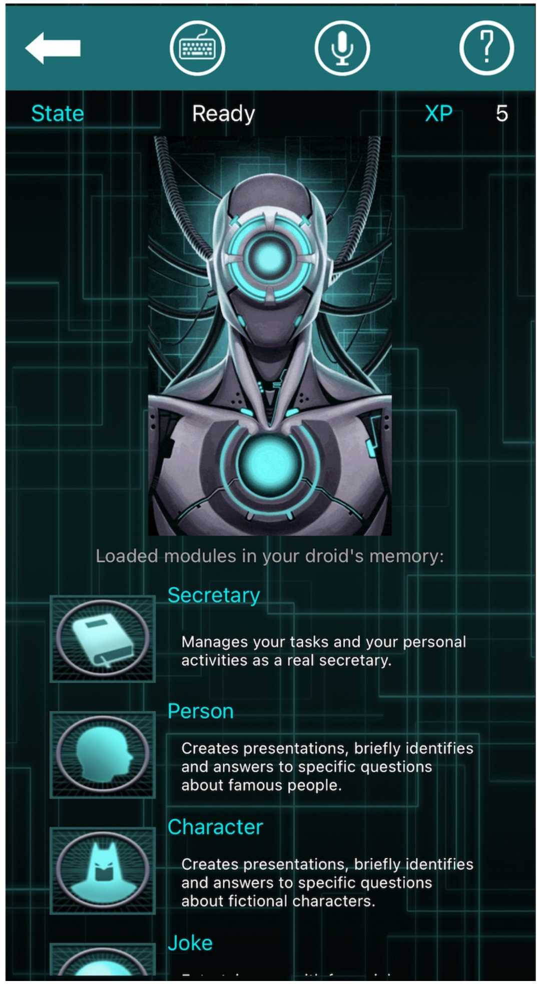
Opacity 100 %

Blur 4

Threshold 0.5

Invert

Apply



Loaded modules in your droid's memory:

Secretary

Manages your tasks and your personal activities as a real secretary.

Person

Creates presentations, briefly identifies and answers to specific questions about famous people.

Character

Creates presentations, briefly identifies and answers to specific questions about fictional characters.

Joke

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

Why you should take this class

Because you want to learn how to:

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

Why you shouldn’t take this class

- You are **not** familiar with most of the concepts covered in CS109A/B
- **For example:**
  - Basic Machine Learning
  - CNNs, RNNs, Autoencoders, {GANs, etc}.
  - Basic shell commands

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

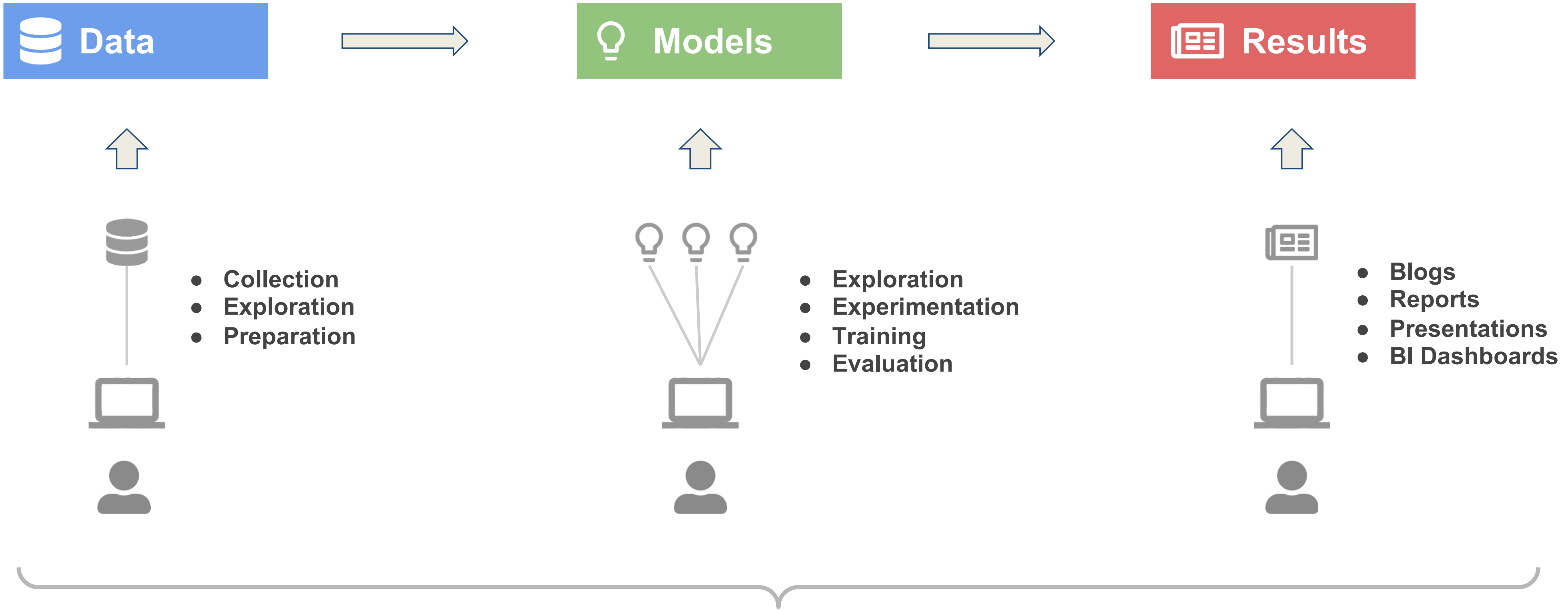
Motivation

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/

Data Science Series to Real World

Data Science Series CS109 A/B



Data

- Collection
- Exploration
- Preparation

Models

- Exploration
- Experimentation
- Training
- Evaluation

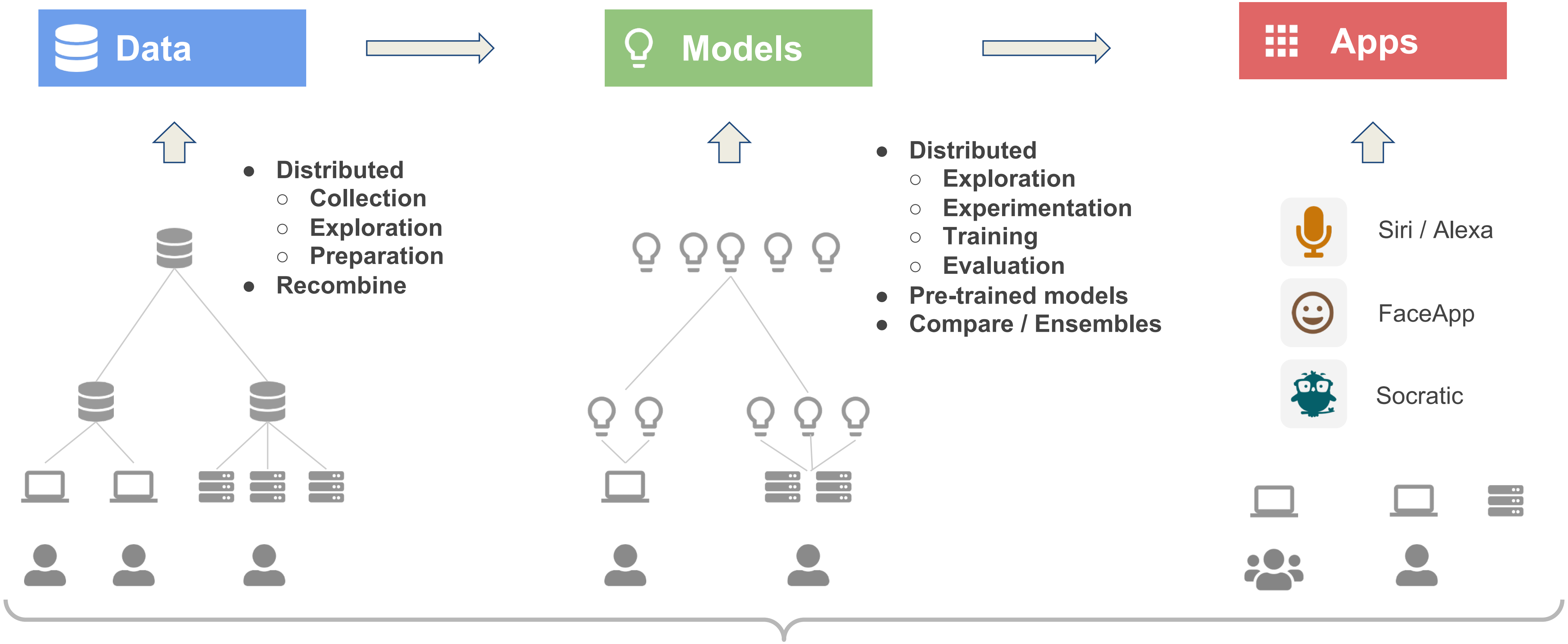
Results

- Blogs
- Reports
- Presentations
- BI Dashboards

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

Data Science Series to Real World

Real World



Data → **Models** → **Apps**

- **Distributed**
  - **Collection**
  - **Exploration**
  - **Preparation**
- **Recombine**

- **Distributed**
  - **Exploration**
  - **Experimentation**
  - **Training**
  - **Evaluation**
- **Pre-trained models**
- **Compare / Ensembles**

Siri / Alexa

FaceApp

Socratic

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:

- OS specific installations required
- How to collaborate code?
- How to share datasets & models?
- Need for multi GPUs or training for more than 12 hours
- Automate data collection / model training
- New team member onboarding
- “It works on my machine” ¯_(ツ)_/¯

Ops for Machine / Deep Learning

Development Operations (DevOps):

DevOps is a practice that brings together software development (Dev) and operations (Ops) to streamline the process for better productivity and shorten development life cycle

Machine / Deep Learning Operations (MLOps):

MLOps is a practice that brings together machine learning or deep learning model development, application development, and operations together to streamline the interaction between the three and simplify the machine learning life cycle

MLOps - Tasks

Machine / Deep Learning:

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

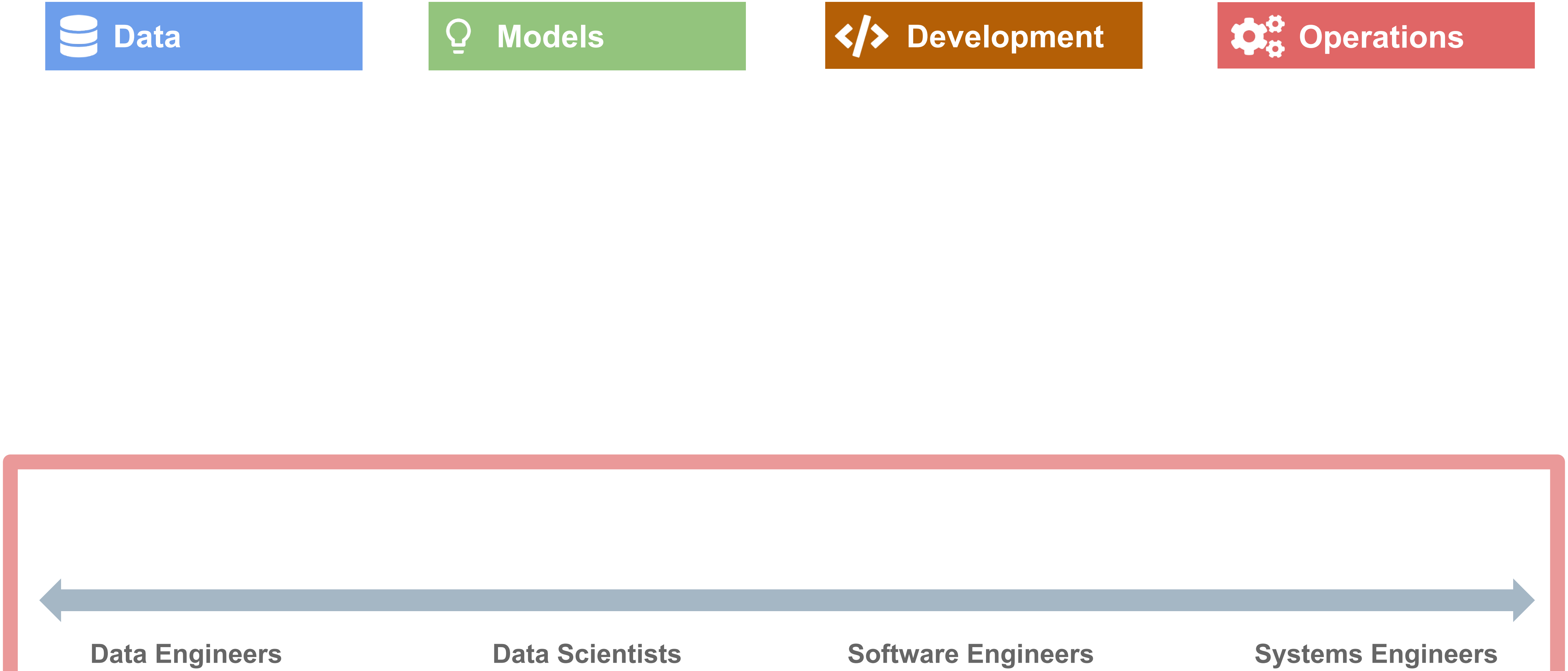
Application Development:

- 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

Models

Development

Operations

Data Engineers

Data Scientists

Software Engineers

Systems Engineers

MLOps - Tech Stack

Data

Models

Development

Operations

- Spark
- Hadoop
- Kafka
- Dask
- Airflow
- Pachyderm
- DagsHub
- Elastic
- Neo4j
- Weaviate

Data Engineers

Data Scientists

Software Engineers

Systems Engineers

MLOps - Tech Stack

Data

- Spark
- Hadoop
- Kafka
- Dask
- Airflow
- Pachyderm
- DagsHub
- Elastic
- Neo4j
- Weaviate

Models

- TensorFlow
- PyTorch
- MXNet
- JupyterLab
- Google Colab
- Deepnote
- Google AI Platform
- Amazon Sagemaker

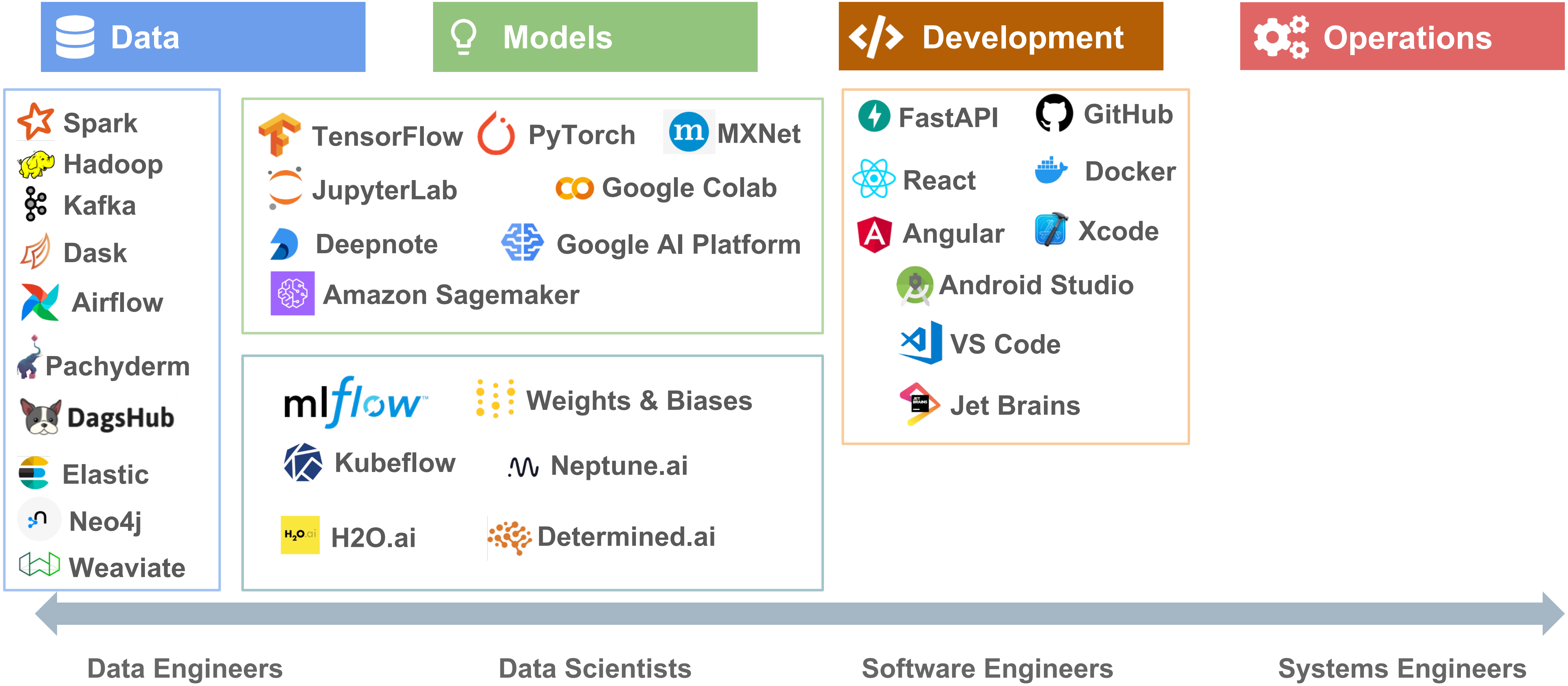
- mlflow
- Weights & Biases
- Kubeflow
- Neptune.ai
- H2O.ai
- Determined.ai

Development

Operations

Data Engineers — Data Scientists — Software Engineers — Systems Engineers

MLOps - Tech Stack



Data

Spark
Hadoop
Kafka
Dask
Airflow
Pachyderm
DagsHub
Elastic
Neo4j
Weaviate

Models

TensorFlow
PyTorch
MXNet
JupyterLab
Google Colab
Deepnote
Google AI Platform
Amazon Sagemaker

mlflow
Weights & Biases
Kubeflow
Neptune.ai
H2O.ai
Determined.ai

Development

FastAPI
GitHub
React
Docker
Angular
Xcode
Android Studio
VS Code
Jet Brains

Operations

Data Engineers
Data Scientists
Software Engineers
Systems Engineers

MLOps - Tech Stack

Data

- Spark
- Hadoop
- Kafka
- Dask
- Airflow
- Pachyderm
- DagsHub
- Elastic
- Neo4j
- Weaviate

Models

- TensorFlow
- PyTorch
- MXNet
- JupyterLab
- Google Colab
- Deepnote
- Google AI Platform
- Amazon Sagemaker

- mlflow
- Weights & Biases
- Kubeflow
- Neptune.ai
- H2O.ai
- Determined.ai

Development

- FastAPI
- GitHub
- React
- Docker
- Angular
- Xcode
- Android Studio
- VS Code
- Jet Brains

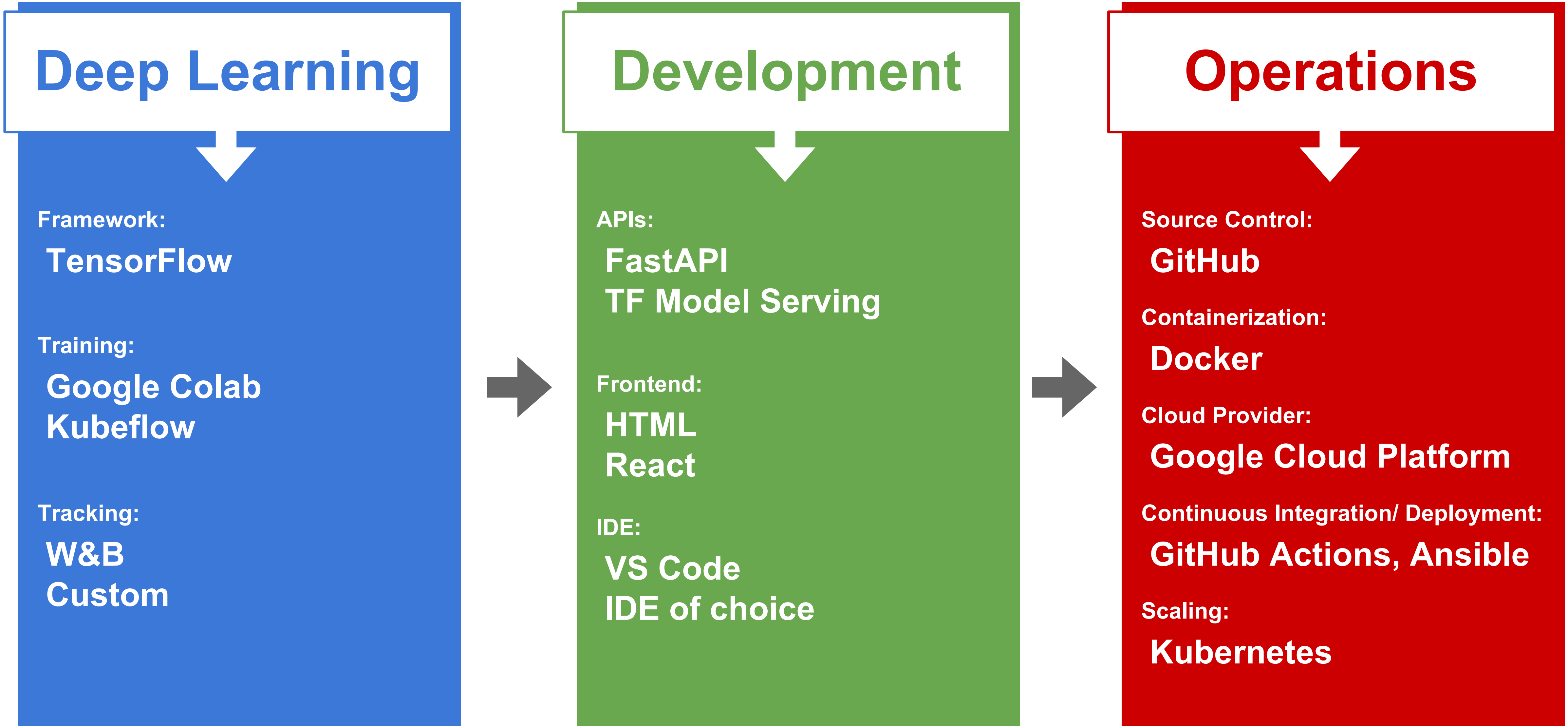
Operations

- GCP
- AWS
- Kubernetes
- Jenkins
- Ansible
- GitHub Actions
- Cloud Functions

- TensorFlow Serving
- Amazon Sagemaker Hosting
- DataRobot

Data Engineers — Data Scientists — Software Engineers — Systems Engineers

MLOps - Tech Stack



Deep Learning

Framework:

TensorFlow

Training:

Google Colab
Kubeflow

Tracking:

W&B
Custom

Development

APIs:

FastAPI
TF Model Serving

Frontend:

HTML
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?
4. Class organization (Workload, Logistics, Grades).

---

1. Projects
2. Experiment Tracking
3. Model Compression Techniques

Who?

Pavlos Protopapas

- Scientific Director of IACS.
- Teaches CS109a, CS109b and AC215.
- He is a leader in astrostatistics and he is excited about the new telescopes coming online in the next few years.
- PI of stellarDNN a research lab on 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
- Fun facts:
  - 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.
  - During a failed military service he was declared the worst soldier in NATO

Who ?

Rashmi Banthia

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

Fun Fact: Enjoys kaggle competitions

Andrew Smith

Passionate about using machines to model and assist the human creative process

Fun Fact: Has produced concerts on five different continents

Connor Capitolo

Machine Learning Engineer

Graduated from Master’s in DS program in May 2022

Fun Fact: Loves to go fly fishing

Who ?

Shivas Jayaram

Deep Learning Researcher, Educator and Practitioner

Working on medical-pharma knowledge platform startup

Fun Fact:

Tale Lokvenec

Fun Fact:

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).

---

1. Projects

Course Structure and Activities

Modules:

- Virtual Environments and Virtual Machines
- Containers
- Data
- Model
- ML Workflow Management
- App Development
- Scaling & Deployment

Activities:

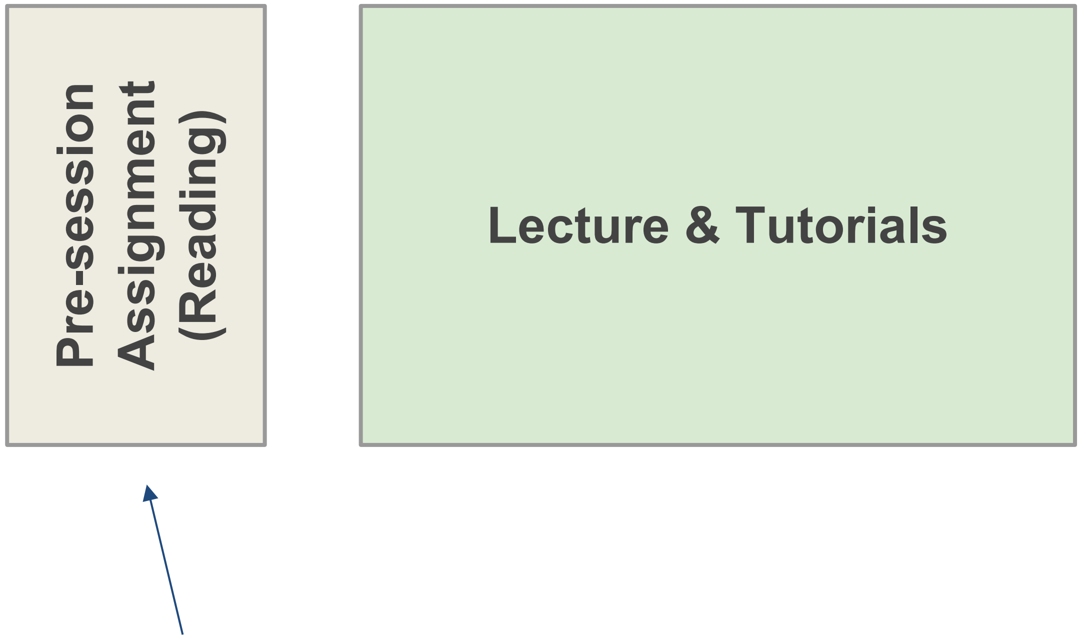
Sessions, exercise, project, reading and quizzes

Sessions: Saturdays 8:30 PM - 10:30 PM IST

Office Hours: Tuesdays 9:00 PM IST

Course Structure and Activities

Weekly Session - What to expect



There will be one reading assignment per week

Topics

- Virtual Environments and Virtual Machines
- Containers
- Data
- Model
- ML Workflow Management
- App Development
- Scaling & Deployment

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).**

---

1. Projects

Workload

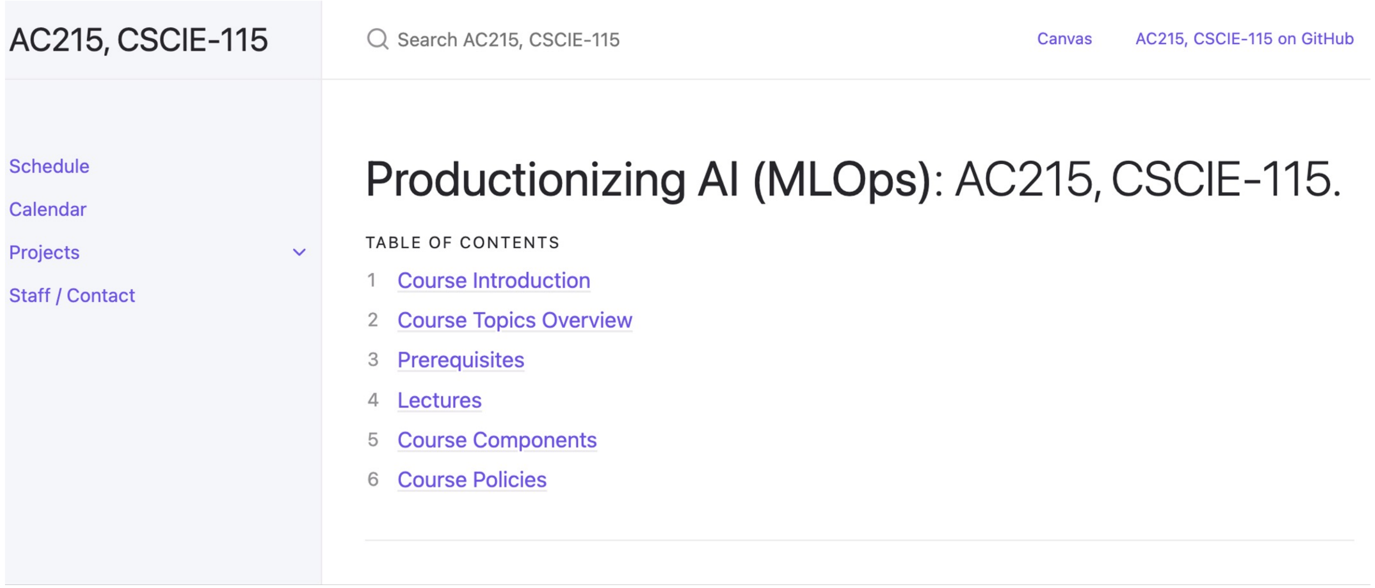
- 1 hour *Reading*
- 3 hours *Session*
- 1 hour *Office Hour*
- 5 hours *Project Milestones*
- ~ 12 hours/ week

Expectations

- Readings
- Sessions: Continuing and finish tutorials we start in the session.
- Milestones
- Presentations of project progress

Course Components

Course web page



AC215, CSCIE-115

Search AC215, CSCIE-115

Canvas

AC215, CSCIE-115 on GitHub

Schedule

Calendar

Projects

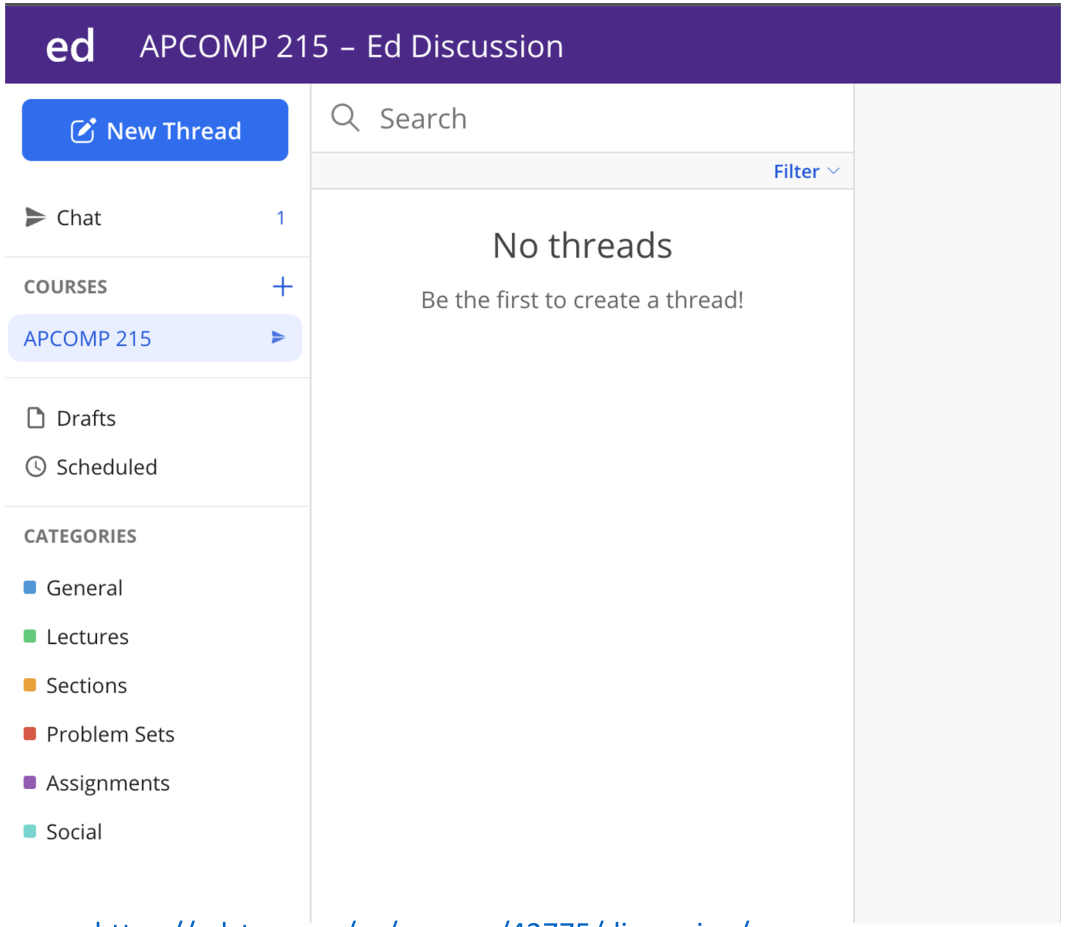
Staff / Contact

Productionizing AI (MLOps): AC215, CSCIE-115.

TABLE OF CONTENTS

https://harvard-iacs.github.io/2023-AC215/

ED Stem



ed APCOMP 215 – Ed Discussion

New Thread

Search

Filter

Chat 1

COURSES

APCOMP 215

Drafts

Scheduled

CATEGORIES

General

Lectures

Sections

Problem Sets

Assignments

Social

No threads

Be the first to create a thread!

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

Grades

Assignment	Final Grade Weight
Milestone 1	5%
Milestone 2	10%
Milestone 3	15%
Milestone 4	25%
Milestone 5	10%
Milestone 6	35%
Total	**100%**

Final Details

- We will be using ED for discussions, announcements and surveys
- Quizzes: Individual
- Exercises/Homework: Individual
- Projects: Group

Submissions for project milestones and projects will be using GitHub

Logistics

- Survey
- Make project groups

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).

---

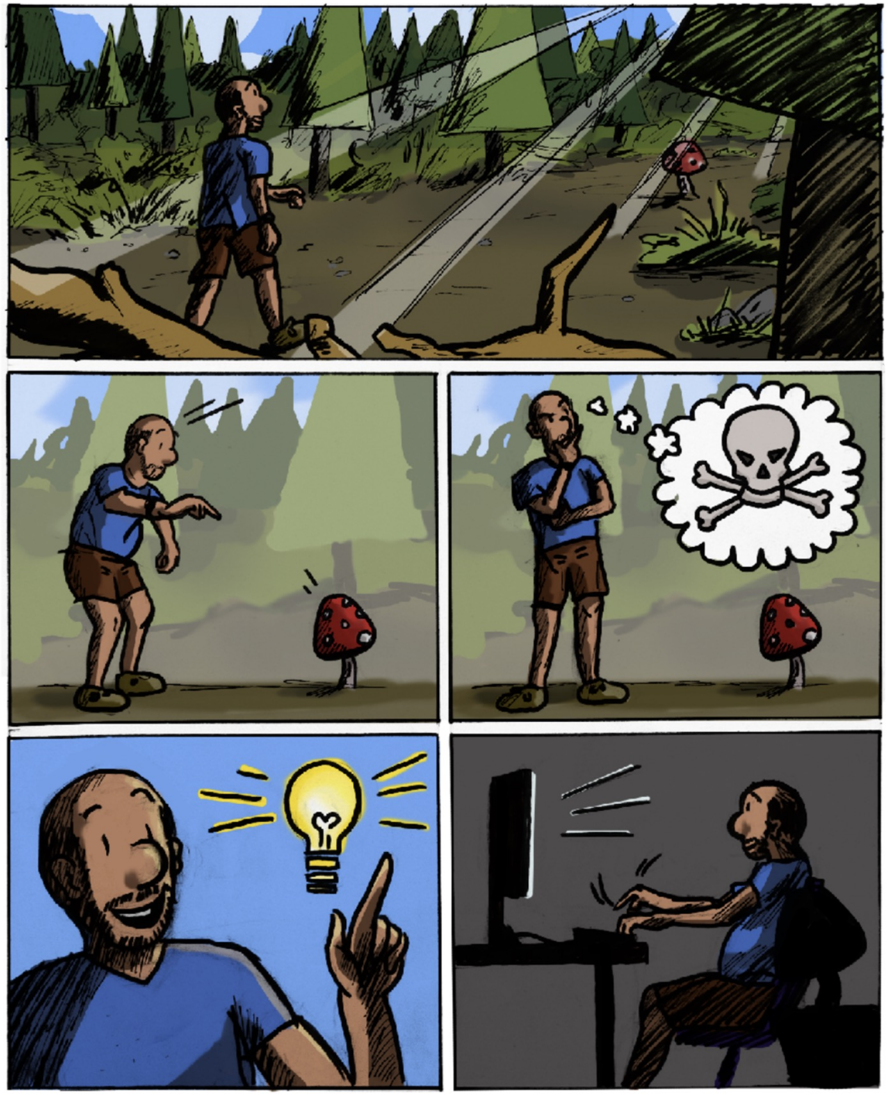
1. **Projects**

Projects

In Class Demo Mushroom Identification App

Project Idea

- Pavlos likes to go the forest for mushroom picking
- Some mushrooms can be poisonous
- Help build an app to identify mushroom type and if poisonous or not
- [Project Summary]



Credit: Nikolas Protopapas

Problem Definition

Pavlos like to go to the forest to do mushroom picking. It is a fun activity and also rewarding as some mushrooms are edible. The problem is in the forest where Pavlos goes to pick mushrooms there are many varieties of poisonous mushrooms. Some of the mushrooms are obvious but there are some which he requires help in identification.

Proposed Solution

Pavlos will have is phone with him when he is in the forest. What if he could just take a picture of the mushrooms and and app could tell him what type of mushroom it is and weather it is poisonous or not

Project Execution Steps

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

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 if a mushroom is poisonous or not
- **Users:** External
- **Duration:** Months

Project Scope (Mushroom App)

Proof Of Concept (POC)

- Scrap mushroom data
- Verify images
- Experiment on some baseline models
- Verify new unseen mushrooms are predicted by the model(s)
- Visualize model activations to analyse what the model is seeing

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 Mushrooms
- API Server for uploading images and predicting using best model

THANK YOU