

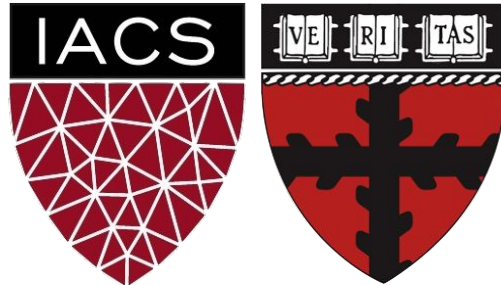
# Lecture 1: Introduction, Project Outline

Advanced Practical Data Science, MLOps

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).
- 

5. Introduction to projects
6. Project scope

# Outline

---

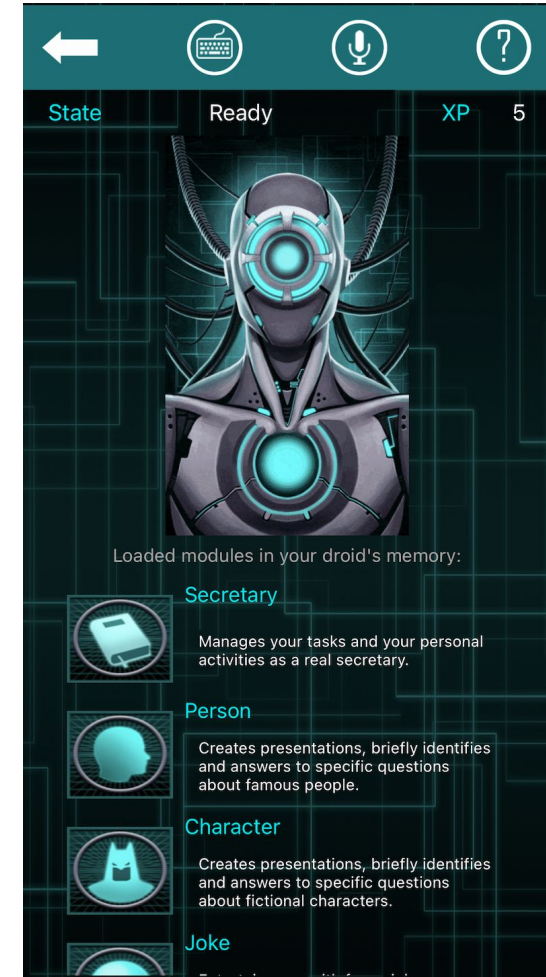
1. **Why should you take this class and why not?**
  2. Who are we?
  3. Course structure and activities?
  4. Class organization (Workload, Logistics, Grades).
- 
5. Introduction to projects
  6. Project scope

# 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 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 and CS109B
- **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<sup>[1]</sup>

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

# Motivation

---

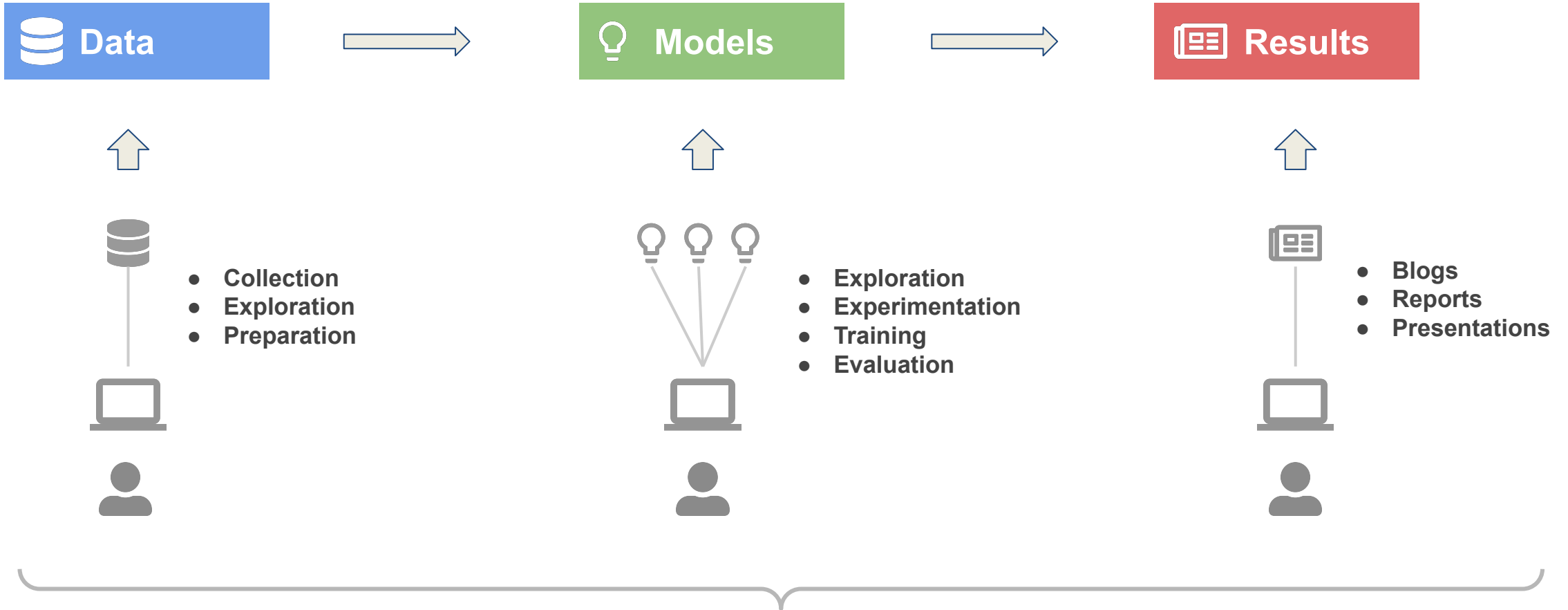
A recent International Data Corporation ([IDC](https://www.idc.com/)) survey of global organizations that are already using artificial intelligence (AI) solutions found only 25% have developed an enterprise-wide AI strategy. At the same time, half the organizations surveyed see [AI as a priority](#)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

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



# Data Science Series to Real World

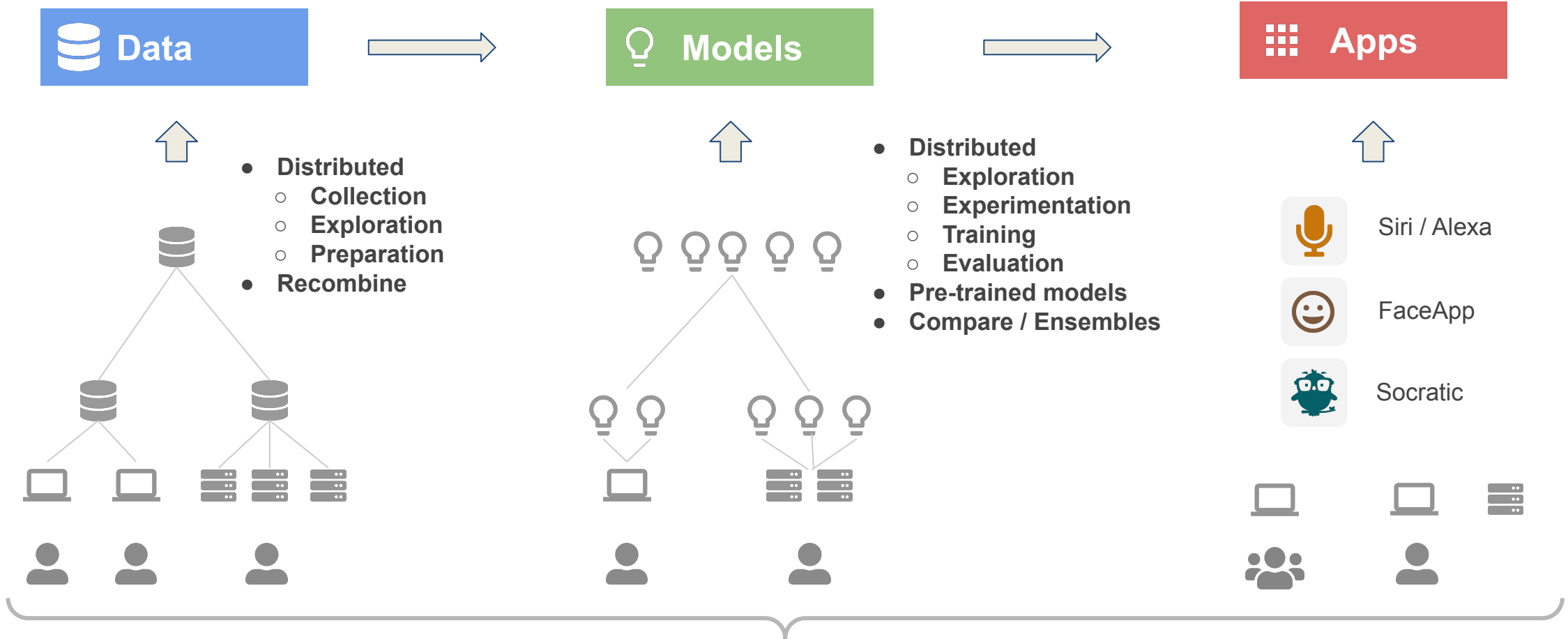
Data Science Series CS109



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

# Data Science Series to Real World

Real World



# Data Science Series to Real World (cont)

---

## Challenges:

- OS specific installations are required
- How to collaborate, sharing 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)\\_/͇

# DL Ops (MLOps for 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

## **Deep Learning Operations (DL Ops):**

DL Ops is a practice that brings together **deep learning** model development, **application** development, and **operations** together to streamline the interaction between the three and simplify the deep learning life cycle

## **Deep Learning:**

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

## **Application Development:**

- APIs/Model serving
- 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/Continuous Deployment
- Continuous data collection/model training
- Model/data monitoring








# DL Ops - Tech Stack

 **Data**

 **Models**

 **Development**

 **Operations**

-  Spark
-  Hadoop
-  NiFi
-  Kafka
-  Dask
-  Airflow
-  Pachyderm



**Data Engineers**

**Data Scientists**

**Software Engineers**

**Systems Engineers**








# DLOps - Tech Stack









 Data

 Models

 Development

 Operations

-  Spark
-  Hadoop
-  NiFi
-  Kafka
-  Dask
-  Airflow
-  Pachyderm

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

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

Data Engineers

Data Scientists

Software Engineers

Systems Engineers

# DL Ops - Tech Stack

## Data

- Spark
- Hadoop
- NiFi
- Kafka
- Dask
- Airflow
- Pachyderm

## 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



# DL Ops - Tech Stack

## Data

- Spark
- Hadoop
- NiFi
- Kafka
- Dask
- Airflow
- Pachyderm

## 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
- TensorFlow Serving
- Amazon Sagemaker Hosting
- DataRobot

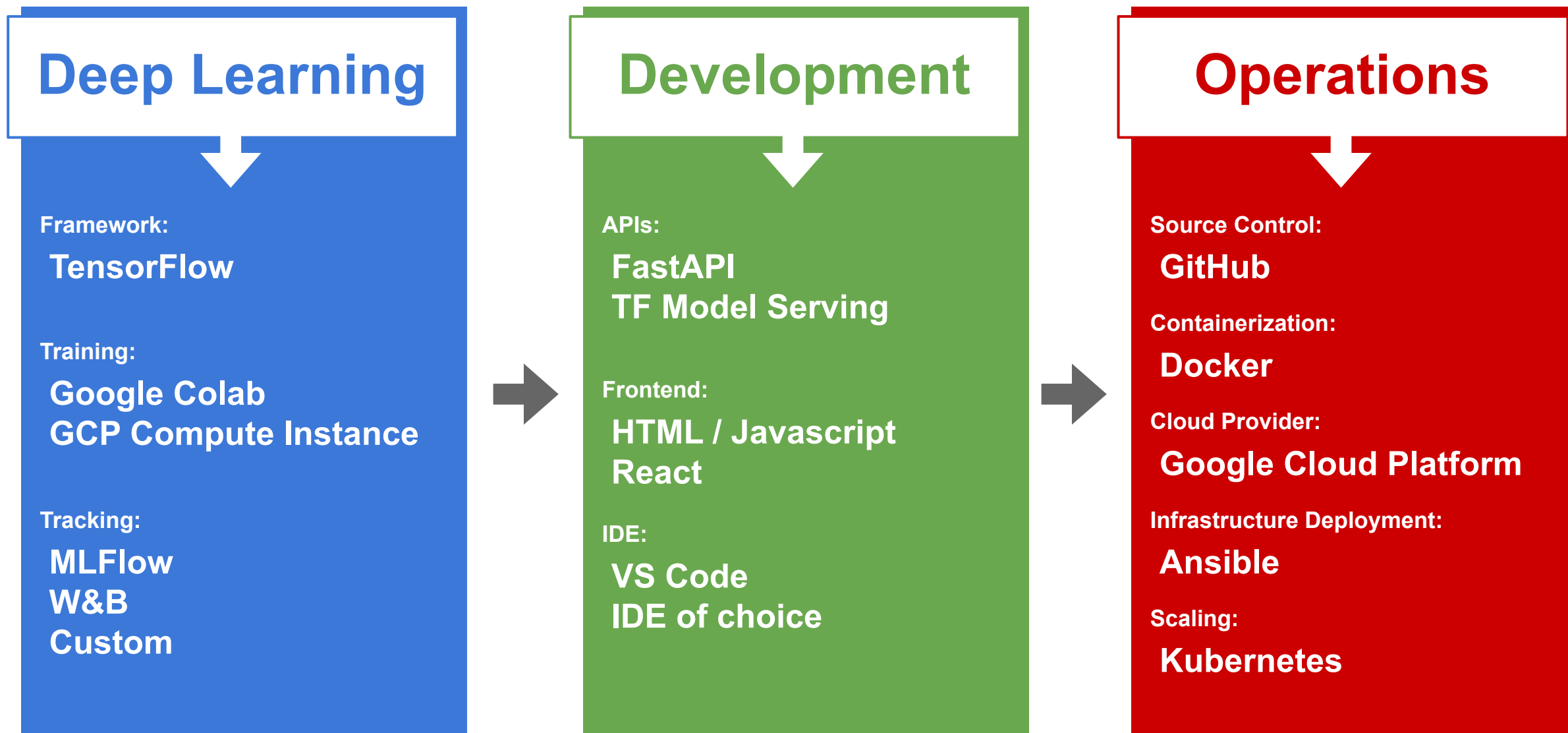
Data Engineers

Data Scientists

Software Engineers

Systems Engineers

# DL Ops - Tech Stack



# 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).
- 
5. Introduction to projects
  6. Project scope

# 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



## Gordon Hew

Financial Software Engineer

HES ALM in DS Candidate

Fun Fact: Learning how to Skateboard



## Shivas Jayaram

Deep Learning Researcher and Practitioner

TF/Teach DS/MLOps classes

Fun Fact: During covid started learning dance/yoga

# 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).
- 

5. Introduction to projects
6. Project scope

# Course Structure and Activities

---

## Modules:

- Project Outline
- Deep Learning
- Development
- Operations

## Activities:

Sessions, exercise, project, reading and quizzes

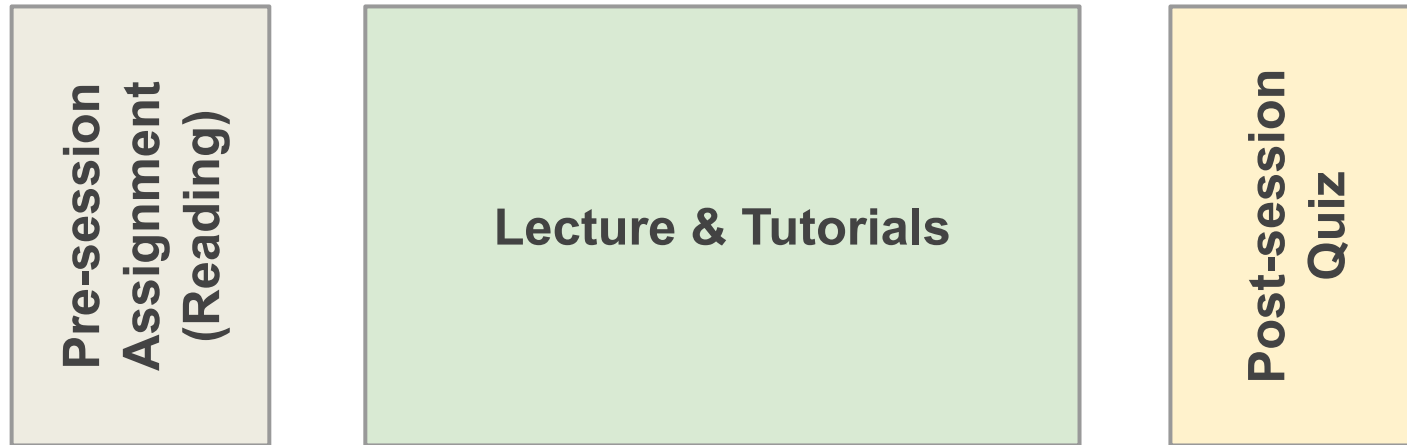
**Sessions:** Tuesdays & Thursdays 2:15 PM - 3:30 PM EST @SEC 2.224

**Office Hours:** Check <https://harvard-iacs.github.io/2021-AC215/>

# Course Structure and Activities

---

## Tuesday Session - What to expect



There will be one reading assignment per week

Quiz will include questions regarding the lecture and the reading



# Course Structure and Activities

---

## Thursday Session - What to expect

Lecture & Tutorials

Post-session  
Exercise

# Topics

31-Aug	Sessions	Process	Step	Concepts
2-Sep	1	Project Outline	Introduction to Projects	Problem Definition Proposed Solutions Project Scope
7-Sep	2,3	Deep Learning	Data	Data Pipelines Tensorflow Data Tensorflow Records Dask Cloud Storage (GCS)
14-Sep	4,5			
21-Sep	6,7		Models	Computer Vision: Classification Computer Vision: Segmentation NLP & Language Models Transfer Learning and SOTA Models Distillation and Compression
28-Sep	8,9			
5-Oct	10,11	Development	Environments	Virtual Environments & Virtual Machines
12-Oct	12,13		Containers	Containerization & Docker
19-Oct	14,15		Design & Implement	App Design Setup & Code organization APIs & Model serving App frontend
26-Oct	16,17			
2-Nov	18,19	Operations	Deployment, Scaling, & Automation	Google Cloud Platform (GCP) Kubernetes Ansible Deployment
9-Nov	20,21			
16-Nov	22,23			

## **Introduction to Projects**

- Problem Definition
- Proposed Solution
- Project Scope

## **Deep Learning - Data**

- Data Pipelines
- TensorFlow Data
- TensorFlow Records
- Dask
- Cloud Storage

## **Deep Learning - Models**

- Computer Vision: Classification
- Computer Vision: Segmentation
- NLP & Language Models
- Transfer Learning and SOTA Models
- Distillation and Compression

## Development

- Virtual Environments, Virtual Machines
- Containers & Docker
- App Design
- Setup and Code organization
- APIs and Model serving
- App frontend

## **Operations - Deployment, Scaling, & Automation**

- Google Cloud Platform (GCP)
- Kubernetes
- Ansible

# Calendar

	Sun	Mon	Tue	Wed	Thu	Fri	Sat	
<b>Week 1</b>	29	30	31	1	2 Session	3	4	September
<b>Week 2</b>	5	6	7 Session	8	9 Session	10	11	
<b>Week 3</b>	12	13	14 Session	15	16 Session	17	18	
<b>Week 4</b>	19	20	21 Session	22	23 Session	24	25	
<b>Week 5</b>	26	27	28 Session	29	30 Session	1	2	
<b>Week 6</b>	3	4	5 Session	6	7 Session	8	9	October
<b>Week 7</b>	10	11	12 Session	13	14 Session	15	16	
<b>Week 8</b>	17	18	19 Session	20	21 Session	22	23	
<b>Week 9</b>	24	25	26 Session	27	28 Session	29	30	
<b>Week 10</b>	31	1	2 Session	3	4 Session	5	6	
<b>Week 11</b>	7	8	9 Session	10	11 Session	12	13	November
<b>Week 12</b>	14	15	16 Session	17	18 Session	19	20	
<b>Week 13</b>	21	22	23 Session	24	25	26	27	
<b>Week 14</b>	28	29	30	1	2	3	4	December
<b>Week 15</b>	5	6	7	8	9	10	11	
<b>Week 16</b>	12	13	14	15	16	17	18	



# 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).**

- 
5. Introduction to projects
  6. Project scope

# Workload (per week)

---

- 1 hour *Reading*
- 2.5 hours *Sessions*
- 1.5 hour *Office Hour*
- 3 hours *Exercise/Homework*
- 3 hours *Project Milestones*
- ~ 11 hours/ week

# Expectations

---

- Readings
- Exercise/Homeworks: Continuing and finishing what we start in the session.
- Milestones
- Presentations of project progress

# Logistics

---

- [Survey](#)
- [Make project groups](#)

# Course Components

## Course web page



## Topics in Applied Computation: Advanced Practical Data Science, MLOps

Fall 2021

[Pavlos Protopapas](#)

Office Hours: By appointment

Course helpline: [ac215.fall2021@gmail.com](mailto:ac215.fall2021@gmail.com)

Welcome to AC215: Advanced practical data science, MLOps.

This course aims to review existing Deep Learning flow while applying it to a real-world problem. Then we will build and deploy an application that uses the deep learning model to understand how to productionize models. This course follows the CS109 model of balancing between concept, theory, and implementation.

Split into three parts; the course starts with the review of Deep Learning concepts for data and modeling and how to apply them to different tasks, including vision and language tasks. The next part will be Development, where you use the models you trained in part 1 and incorporate them into real-world applications. Finally, you will Deploy the application in Google Cloud Platform (GCP). The three parts will cover in detail topics such as Transfer learning, Containerization using Docker, and Scaling deployments using Kubernetes.

At the end of this module, you will build efficient deep learning models and design, build and deploy applications that scale.

## ED Stem

## Canvas

[≡](#) [APCOMP 215](#) > [Syllabus](#)

2021-2022 Fall

[Home](#)

[Announcements](#)

[Syllabus](#)

[Modules](#)

[Assignments](#)

[Quizzes](#)

[Zoom](#)

### APCOMP 215: Advanced Practical Data Science

Our [Public Course Page](#) is the primary source info, syllabus and materials.

Syllabus Fall 2021 - APCOMP 215 / CSCIE-115

Course helpline: [ac215.fall2021@gmail.com](mailto:ac215.fall2021@gmail.com)

# Grades

Assignment	Final Grade Weight
Quiz	10%
Exercises	20%
Milestone 1	5%
Milestone 2	15%
Milestone 3	20%
Final Presentation & Deliverable	30%
<b>Total</b>	<b>100%</b>

# Final Details

---

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

Submissions for project milestones and projects will be using GitHub

# 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).
- 

- 5. Introduction to projects**
6. Project scope



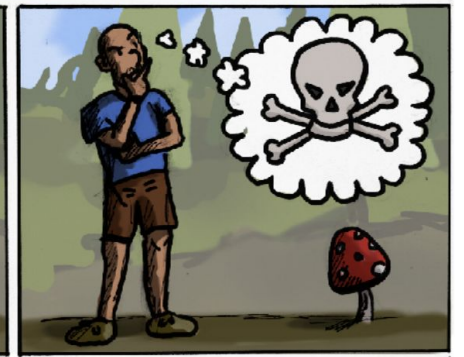
# Introduction to Projects

---

- Mushroom Identification (**in class demo**)
- Austin Pets Alive (APA)
- Visual Question Answering
- Caption this Pic

# Mushroom Identification (In class demo)

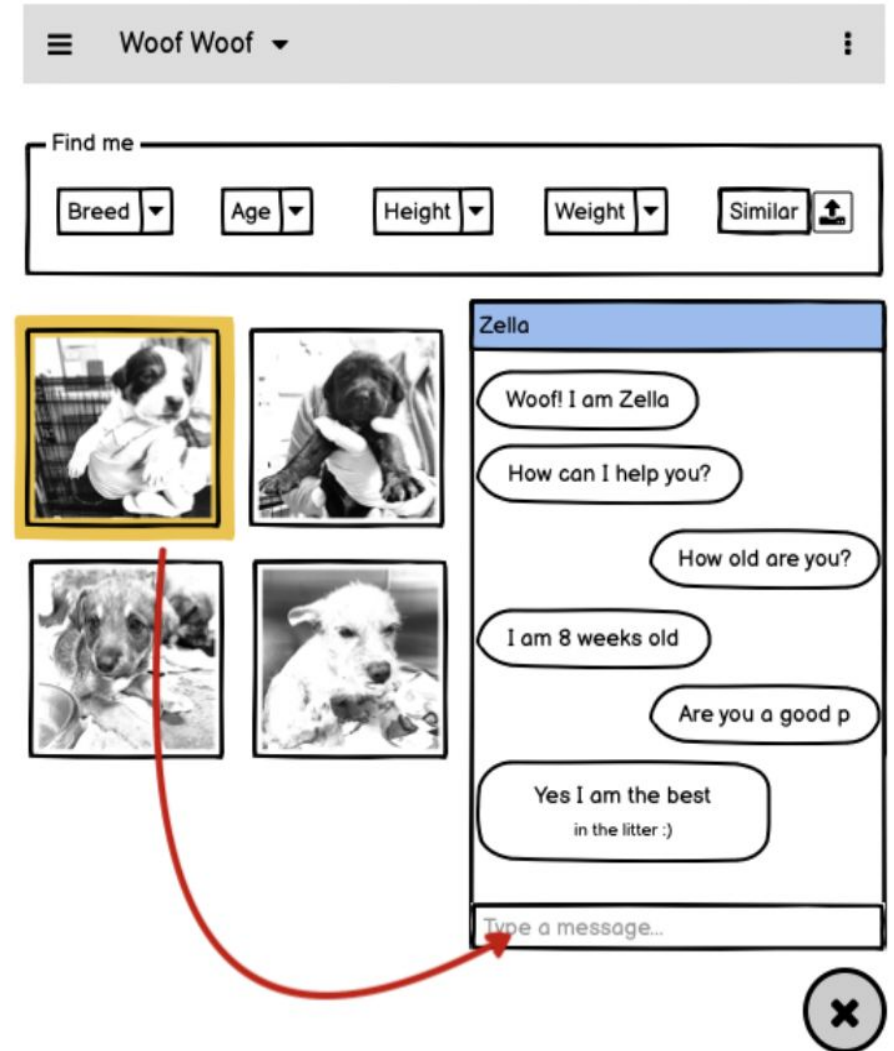
- Pavlos likes to go the forest for mushroom picking
- Some mushrooms can be poisonous
- Help build an app to identify mushroom type and if poisonous or not
- [Project Summary](#)



Credit: Nikolas Protopapas



# Austin Pets Alive (APA)

- APA is an association of pet owners
- They would like to help future dog owners find a dog who is a perfect fit for them
- Help build an app that can help owners find the right pet
- [Project Summary](#)



# Visual Question Answering

- The VQA dataset contains open-ended questions about images
- Build an app that uses a multimodal model that can take both images and text questions as input and predict the answer
- [Project Summary](#)

		
Question: Is the pizza whole? Answer: yes	Question: What color is the sink? Answer: blue	Question: Are these food carts? Answer: yes
		
Question: How many stories is the building on the left? Answer: 3	Question: What color hat is the person on the left wearing? Answer: black	Question: What color is the ball? Answer: yellow



# Caption this Pic

- The dataset consists of Flickr8k and Common Objects in Context (COCO)
- Create an app that allows users to upload images, and have them captioned
- [Project Summary](#)

a zebra walks beside a line of traffic cylinders behind a sign indicating lions  
a zebra standing on the ground next to a sign  
A zebra walks along a road next a sign that reads lions.  
A zebra stands alone in a zoo enclosure.  
A zebra walks toward the lion enclosure at a zoo..



the dog is swimming in the water



black pitbull dog is running through the dirt



downhill skier in black pants and jacket



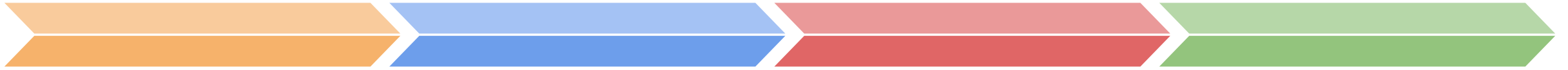
# Outline

---

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).
- 

5. Introduction to projects
6. **Project scope**

# Project Scope



## 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



# Setup & Installation

---

[Setup & Installation Details](#)

**THANK YOU**