VISUAL QUESTION-ANSWERING

AKSHAY RAJNITH
HARSHITHA KOLUKULURU
JANHAVI KASHYAP
MANISHA R
PADMAJA BHAGWAT

PROJECT OUTLINE

PROBLEM DEFINITION

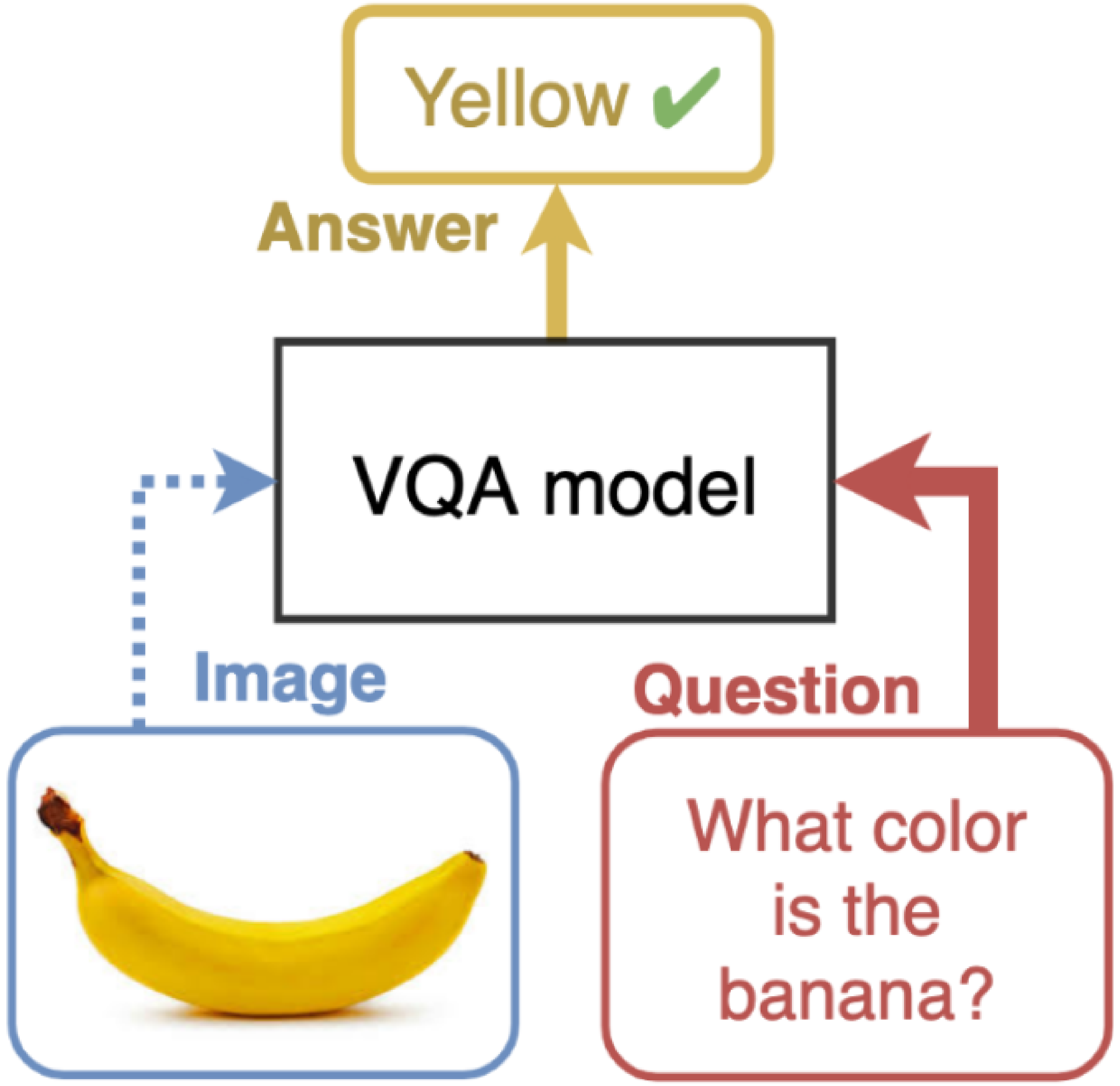
Deep learning models are generally designed to perform one task from one source of data (e.g. image classification or object detection from images in computer vision tasks; or text classification or question answering from the text in natural language processing tasks). There are times when our dataset consists of both images and text that are related and we will want to combine computer vision and natural language tasks.

VQA, lying at the intersection of Computer Vision and Natural Language Processing, is a system that takes in a natural language question about an image. We aim to build and deploy a VQA system as our project.

PROPOSED SOLUTION

We can build a multimodal model that can take both images and text questions as input and predict the answer.

A high-level flow of the minimum required tasks for this project are given in the diagram



PROJECT SCOPE

I PROOF OF CONCEPT

- EDA
- Baseline models
- Verify how the models do on unseen data
- Visualize model activations to analyse what the model is seeing

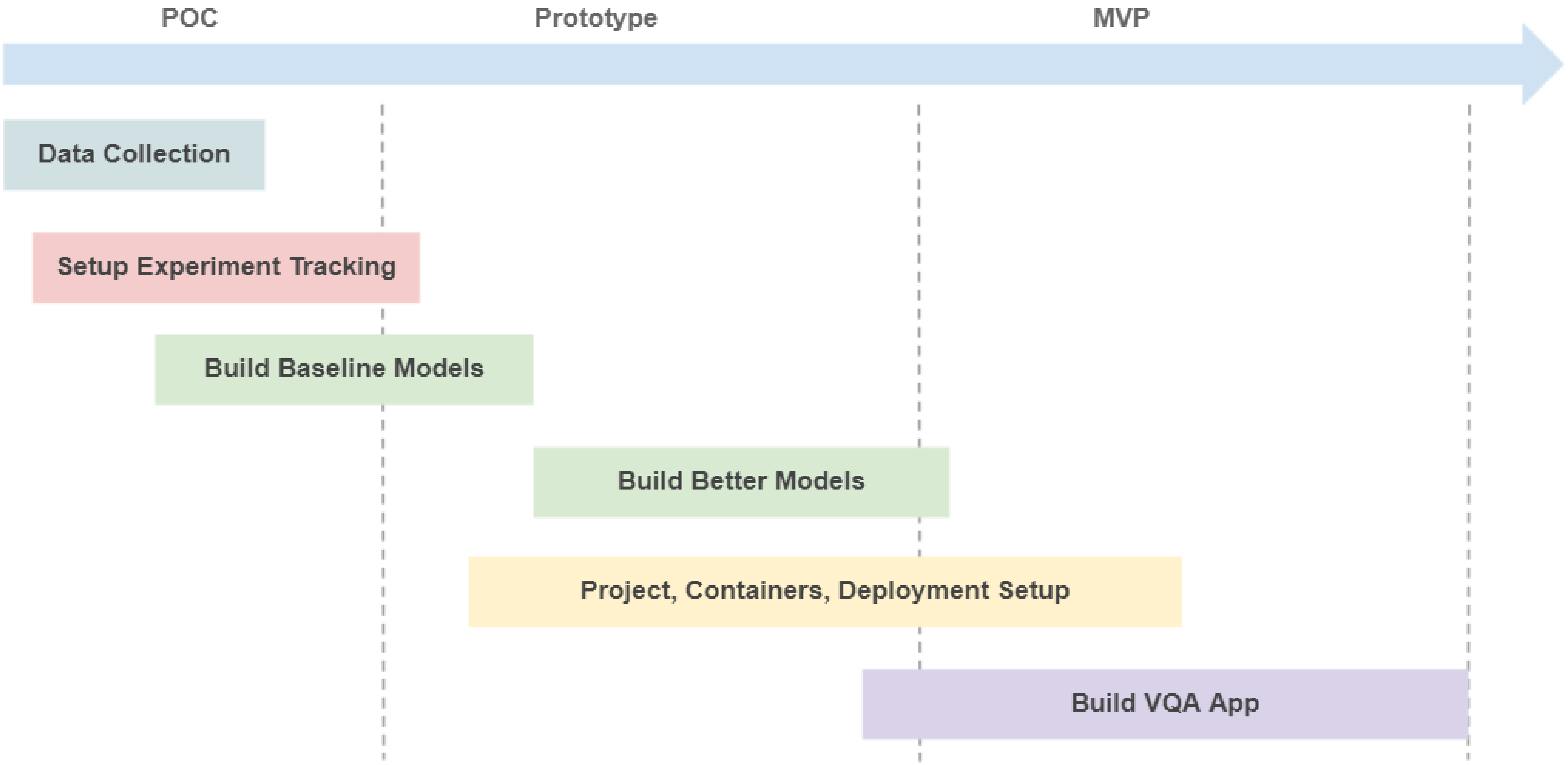
II PROTOTYPE

- Create a mock-up of screens to see what the app would look like
- Deploy one model to Fast API to service model predictions as an API

III MINIMUM VIABLE PRODUCT

- Create an app that performs VQA
- API Server for uploading images and answering questions

PROJECT WORKFLOW

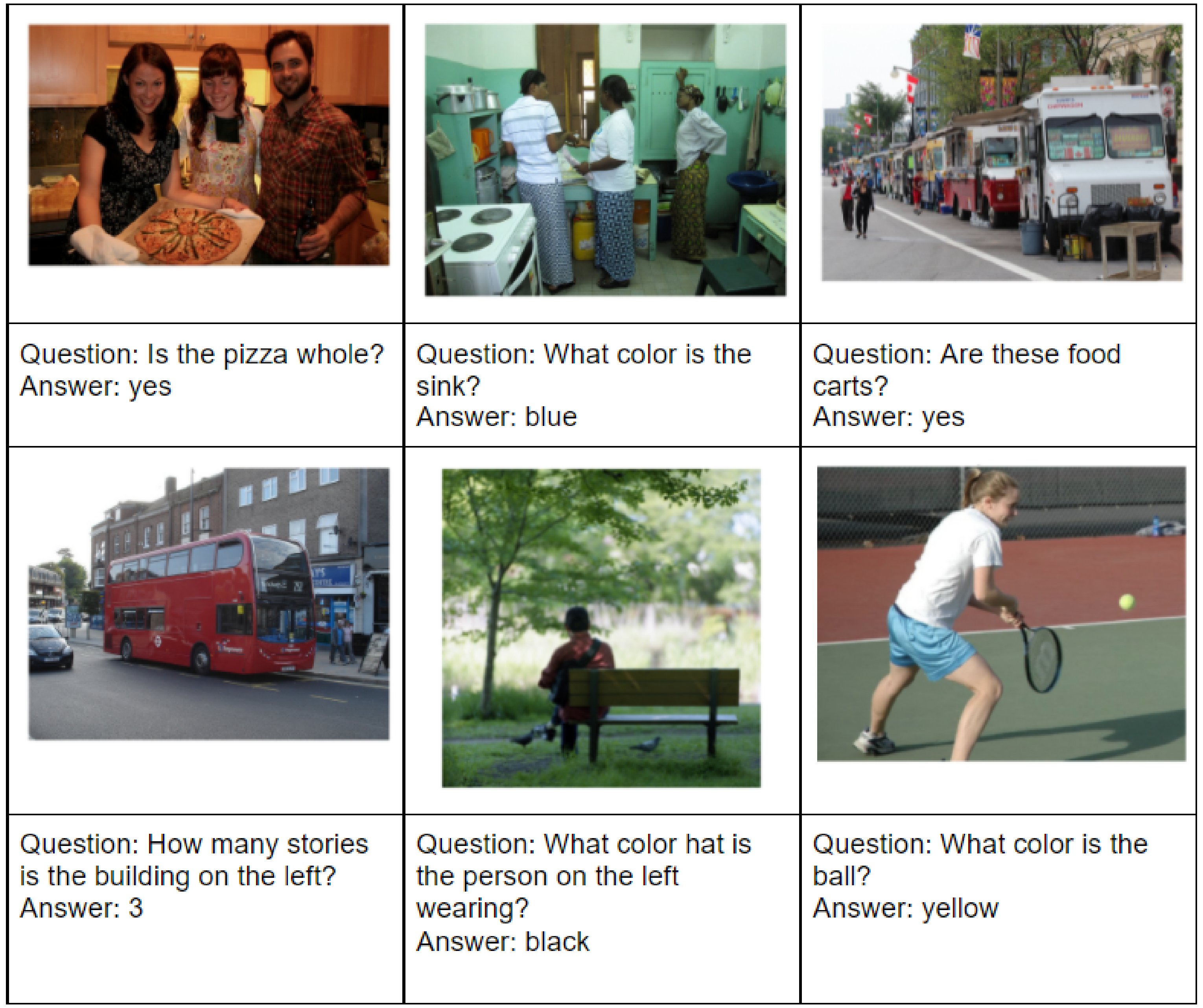


THE DATA

The data is from the Visual Question Answering dataset.

VQA is a dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer. The dataset is a combination of the Common Objects in Context (COCO) dataset with questions and answers added to them:

- Over 120,000 images (COCO)
- At least 3 questions per image with a total of over 600,000 questions
- 10 ground-truth answers per question
- 1 given answer for each question



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

MODELS

VISION

Pretrained Image models as feature extractors like region proposal networks, CNNs

LANGUAGE

- Use of Transformer based models to extract text like BERT