VISUAL QUESTION-ANSWERING APP

VigNET

PROBLEM DEFINITION

VQA is a problem at the intersection of Computer vision and NLP that answers text-based questions about images. Natural language questions, given their arbitrary nature, can encompass many sub-problems including but not limited to object detection and recognition, attribute classification and counting.

A robust VQA system capable of answering a wide range of questions can help the visually impaired "see" an image. Wouldn't it be great if VigNET could lend eyes to them!

Image

Question: What colour is the parrot ?

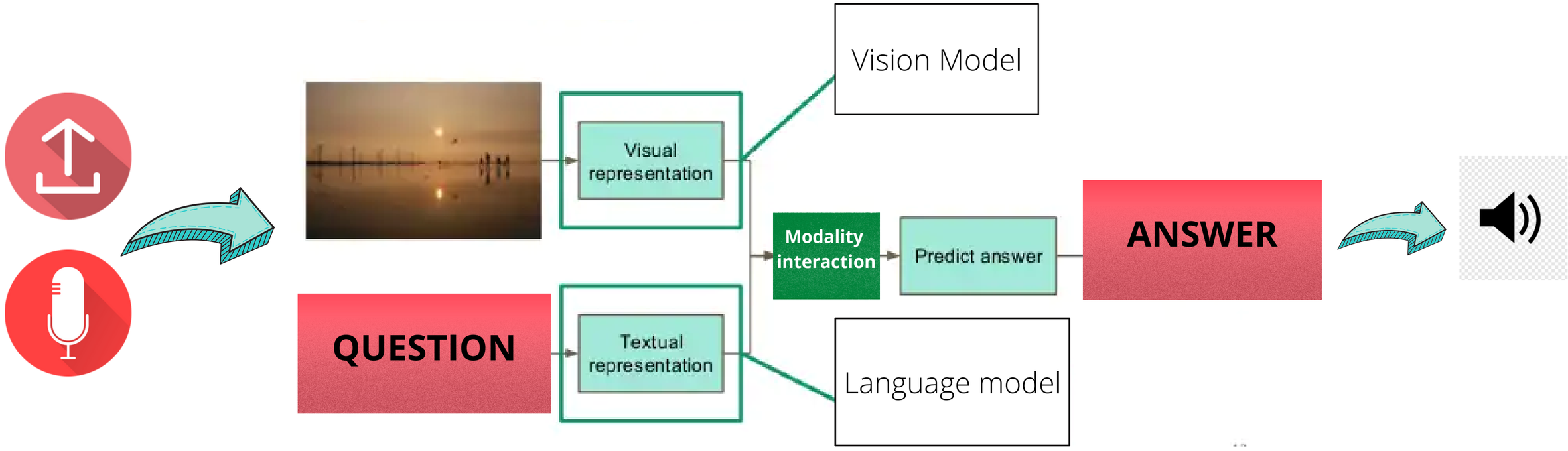
VigNET

Answer: Green

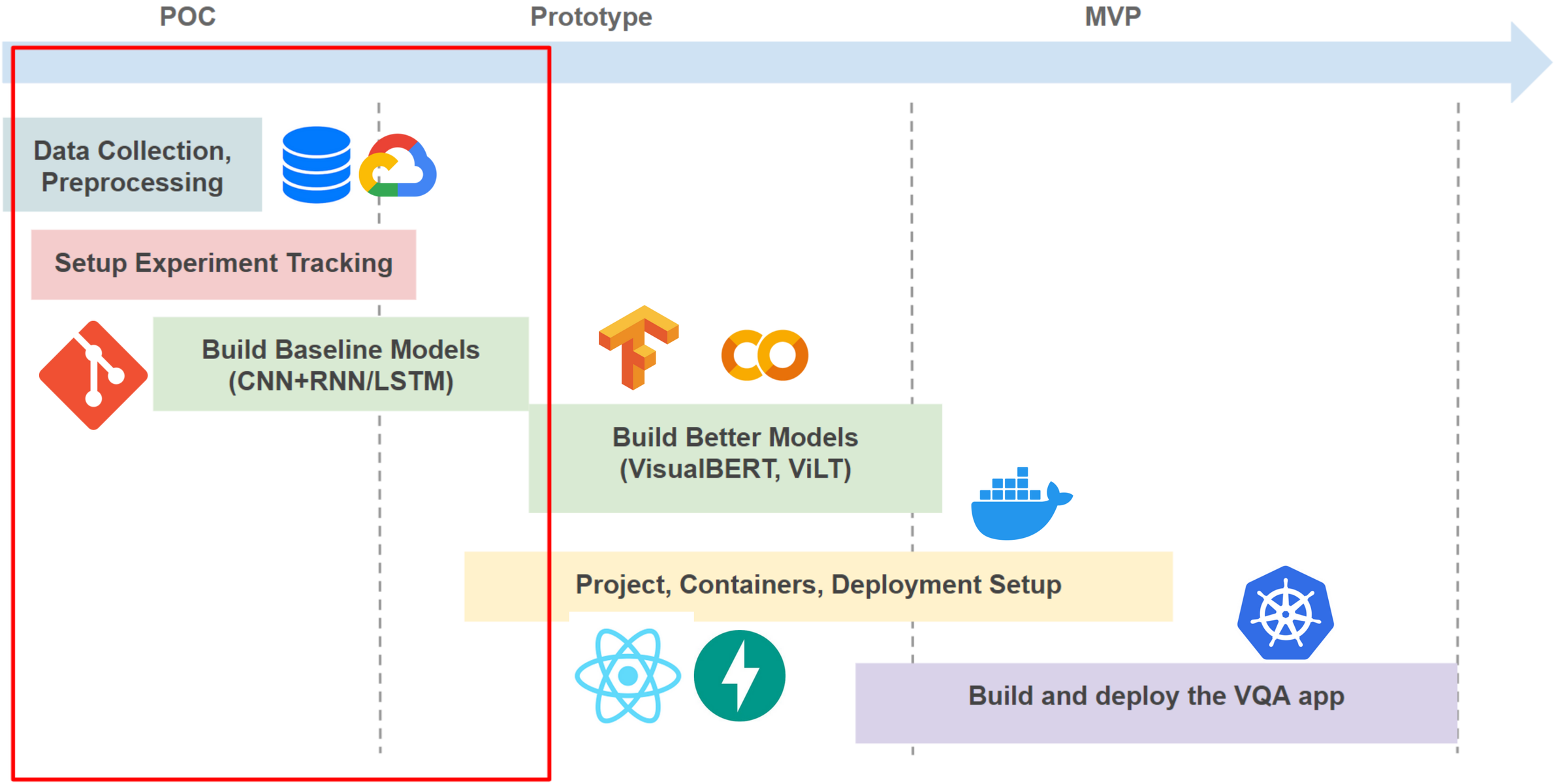
PROPOSED SOLUTION

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 WORKFLOW



PROCESS FLOW

Data Collection and Preprocessing

EDA
Training
Evaluation

Build and deploy the app

- Download data, set up storage buckets
- Build data preprocessing and augmentation pipelines

Build and evaluate models. Perform further fine-tuning to ensure robustness

- Containerise different services.
- APIs to upload images, ask questions, and predict answers.
- Build UI using React

PROJECT SCOPE

I PROOF OF CONCEPT

- Data collection
- Data Preprocessing pipelines for image and text data
- Baseline models
- Test on new images, try asking arbitrary questions
- Fine-tune SOTA models
- Inference

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