AI for Agriculture

Enter the Agri world!

Integrated Farm Assist Platform

Our Team

Manan Arora
ML Engineer and Data Engineer

Saif
Product Development

Jaya Palewar
Data Analyst

Bharatdeep Maan
DevOps Engineer

Anmol Dureha
Hustler and Tool Optimizer

Pranati Balodia
Product Development

Big Picture: Agriculture Industry in India

India has the 10th largest arable land resources in the world.

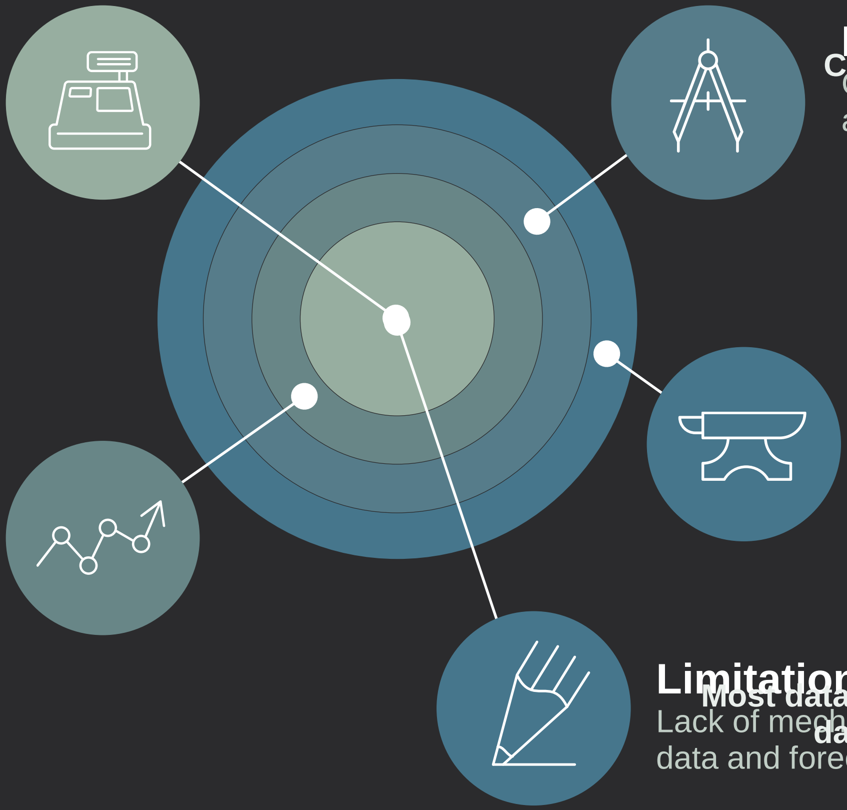
Industry growth has been volatile over the past decade, ranging from 5.8% in 2005-06 to 0.4% in 2009-10 and -0.2% in 2014-15

The average income of a farmer household at current prices is Rs 96,703 (US$ 1,505.27)

Agricultural sector employs nearly half of the total workforce, but it contributes to 17.5% of the GDP

Total agricultural exports from India grew at a CAGR of 16.45 per cent over FY10-18 to reach US$ 38.21 billion in FY18.

Problem : Information Asymmetry for the Farmers in India



Unaware of Best Practices

Unequitable access to latest technological innovations and detailed domain level information

Crop loss of 15.7% annually due to pest attacks - US$ 36 billion

Increasing Climate Uncertainty

Changing weather patterns, ecology changes and soil uncertainties disrupting yield

Climate change could reduce farm incomes by 15-18%

Resource Constraints

Due to constraints like size of land, lack of irrigation, electricity and low-cost finance

85% of operational land holdings are less than 2 hectares and account for 45% of total cropped area

Lack of Market/ Awareness

Insufficient knowledge of market economics and price movements

Incomes of all farmers declined at 1.36% per annum between 2011-12 and 2015-16

Limitations for Data Driven Decisions

Lack of mechanism for assessing decisions based on data and forecasting trends

Most data is scattered on different databases and even offline

Project Initial Plan

01a

Speech and Text Based Query Input from Farmers

01b

Supplementary Information – User Geotag, Profile Information, Weather etc.

02

Conversion of Speech to Text

Toolkits:
- Amazon Transcribe
- Azure Speech - Text
- Indic NLP

03

Context Detection, Sematic Parsing, POS Tagging, Paraphrases

Models:
- CNNs, RNNs & LSTMs

Toolkits:
- Google Diaglogflow

Training Data:
- KCC Data

04a

Connection to Knowledge Bank to Answer farmer queries on best practices

Toolkits:
- Azure QnA

Data:
- Secondary Sources

04b

Machine learning model to predict crop yield, give automatic alerts

Models:
- Decision Trees, SVMs, ANN

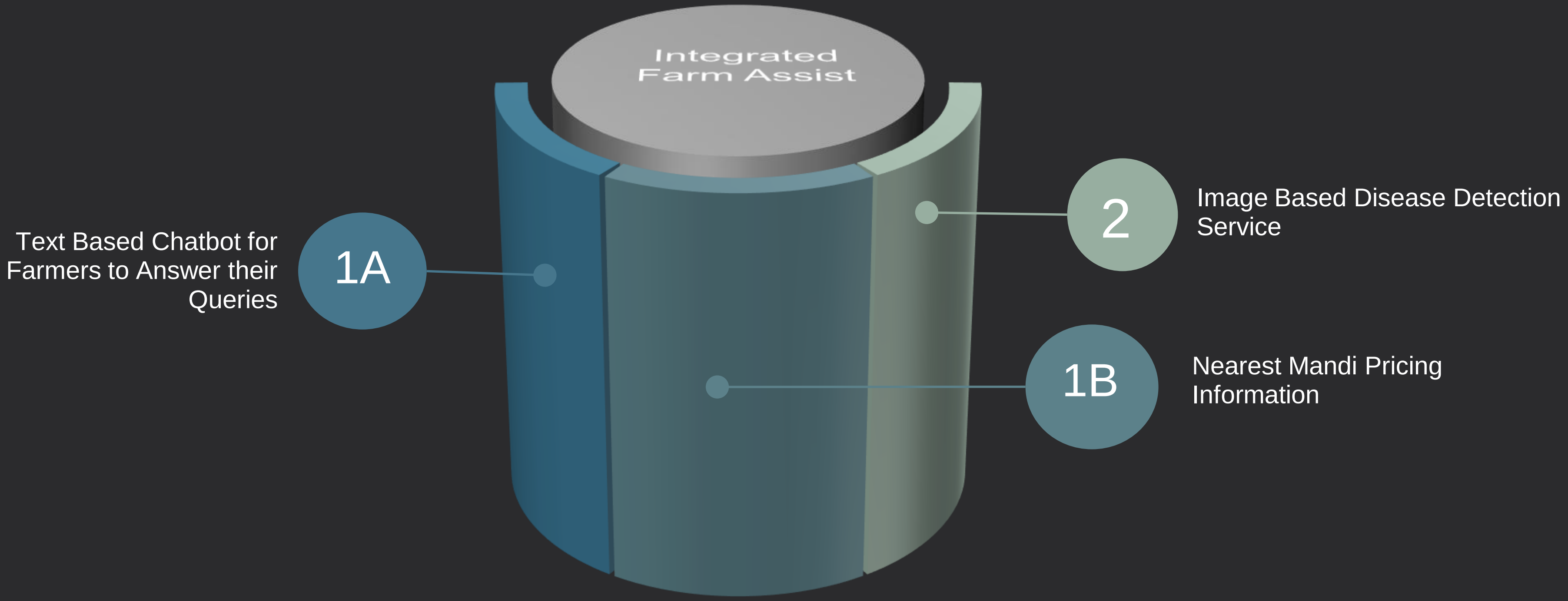
Training Data:
- Historical Data
- Weather Data, Disease Data

Tailored Alerts

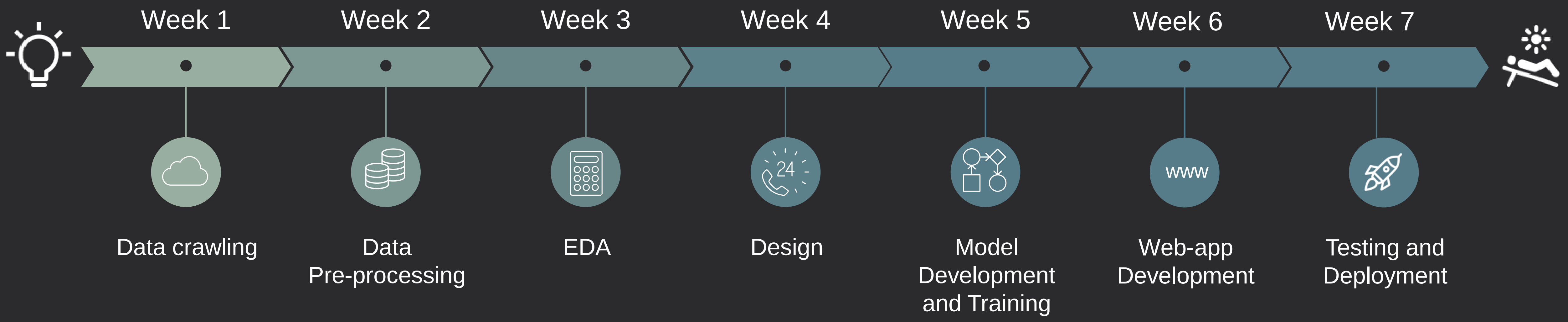
05

- Response to User
- Collection of Feedback
- Yield Predictions

Solution : Overview



Product Roadmap



Putting the Pieces Together - Datasets Used

Kissan Call Center Data
8 Largest Cotton and Wheat Producing States
Data Type: Text | Years 2015 and 2016
5.2 Million Records
Selenium Web crawling

Plant Village Data
Crowd AI Plant Disease Data
38 Classes , 14 Crops
55,261 Images

Maize Field Images
Maize images for Northern Leaf Blight diseases on Maize
18,222 Images | 105,735 Bounding Boxes

Mandi Pricing Data
All India for Mandi' Price on each Day
Data Type: Numerical | Years: 2015+
5.8 Million Records
Selenium Web crawling

Other Datasets we Crawled but were unable to Utilise:
- Soil Health Card Data – 1400 Records - Ajnala, Amritsar, Punjab
- Disease Best Practises Data – Plantix Data – 555 rows

Exploratory Data Analysis – Kisan Call Centre Data

Cotton Data – Out of the 5.2 Million Records

01

0.26M
Number of records of cotton

5 Categories of Queries within cotton
- Plant Protection: 45%
- Weather: 10%
- Fertilizer Use and Availability: 8.6%
- Market Information: 8.5%
- Varieties: 8%

02

120K
Number of Queries in Plant Protection category of cotton

03

67K
Number of unique Queries in plant protection for cotton (unique geographic column)

04

3K
Number of unique Queries in plant protection for cotton (after cleaning stop words, defined library, punctuations, stemming)

05

2.3K
Attack of White Fly Queries

Exploratory Data Analysis – Kisan Call Centre Data

Wheat Data – Out of the 5.2 Million Records

01

0.36M
Number of records of Wheat

5 Categories of Queries within cotton

- Plant Protection: 45%
- Weather: 10%
- Fertilizer Use and Availability: 8.6%
- Market Information: 8.5%
- Varieties: 8%

02

68K
Number of Queries in Plant Protection category of wheat

03

39K
Number of unique Queries in plant protection for wheat (unique geographic column)

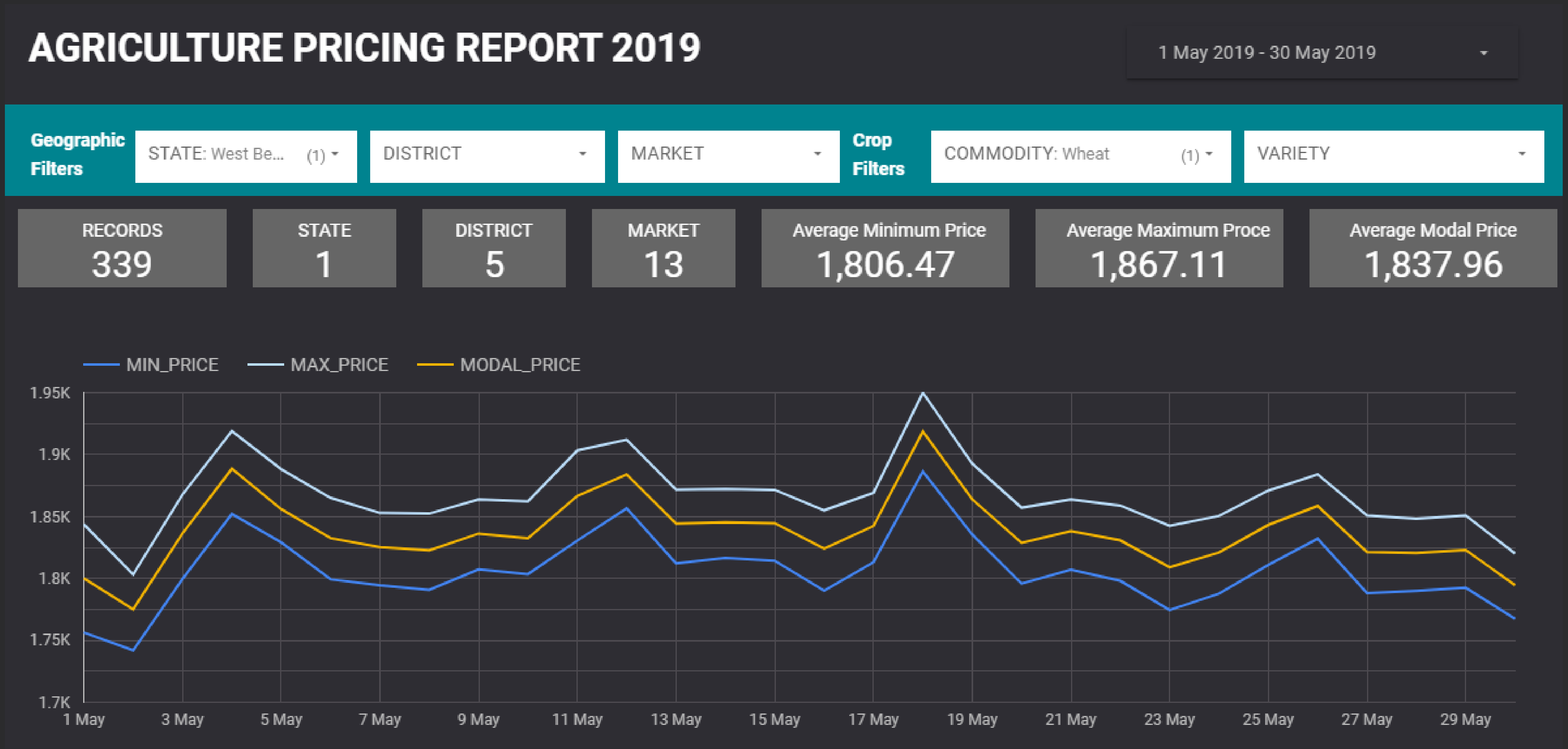
04

2.8K
Number of unique Queries in plant protection for cotton (after cleaning stop words, defined library, punctuations, stemming)

05

2K
Number of Queries on zinc deficiency in wheat

Exploratory Data Analysis – Pricing Data



Based on Google Big Query and Data Studio

Tested, Not Deployed: Chatbot with Rasa NLU

Pre-processing

Team used several methods to pre-process KCC data for Chatbot mentioned below:

- Tokenization, Lemmatization, Stemmers
- Cosine Similarities
- Using Latent Dirichlet allocation probabilistic model to find out similar questions within the dataset
- Clustering the questions using the embedding's (Word2Vec) and K-Means clustering algorithm

Implementation Steps

- Generating entity examples using Chatito - Team used Chatito to create a knowledge base for the Chatbot. This was used for training.
- Data preparation and format- For training data in rasa, we preferred markdown format over json format.
- Intent- Intent name specified with list of questions under it.
- Entity - It was specified inside the knowledge-based questions.

Success, In Production: Chatbot with Dialogflow

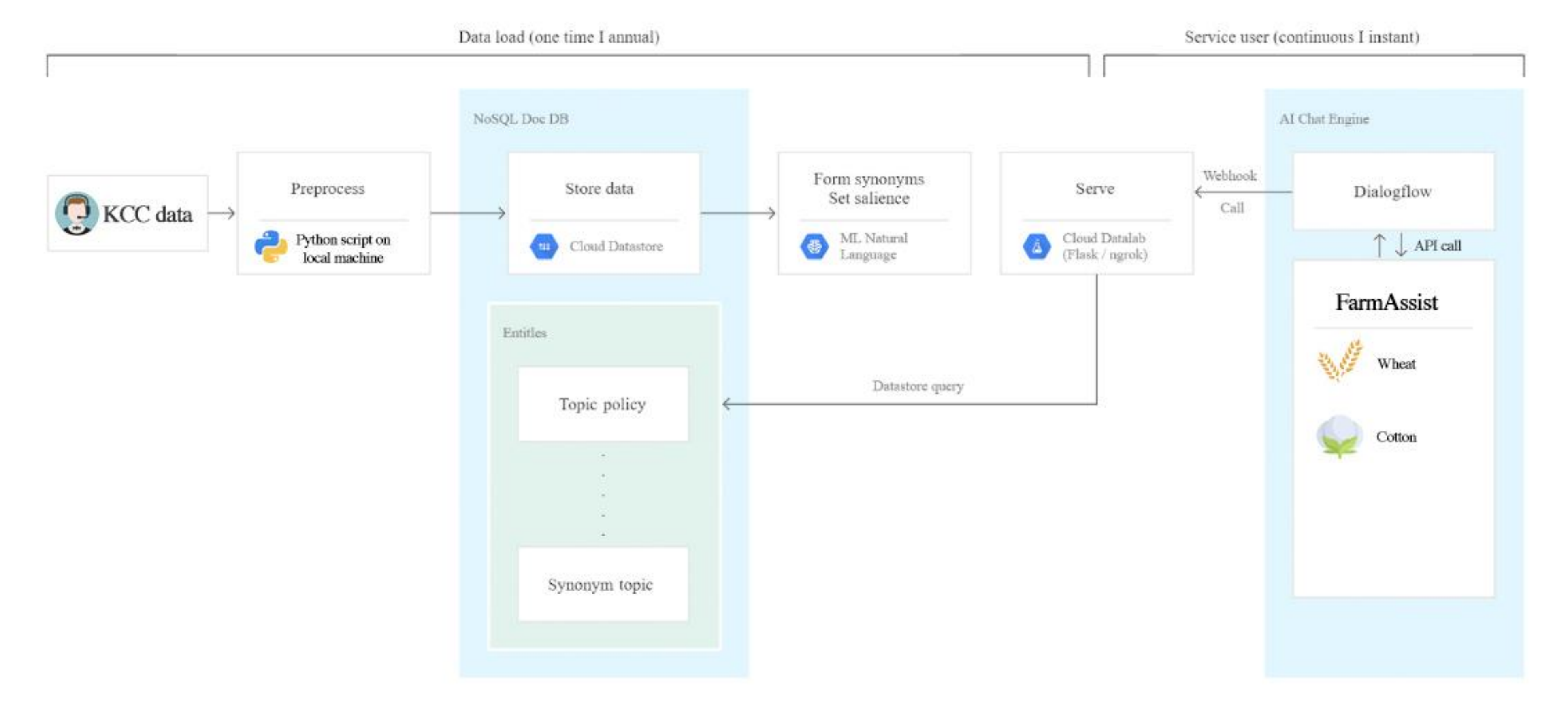
Why Dialogflow?

- Quick and easy to start building
- Built on Google infrastructure
- Easy to scale
- Strong natural language understanding (NLU) capabilities

Implementation Steps

- We used the knowledge base beta connector
- ‘Knowledge Connectors’ lets you add your content in csv or plain text format as a Knowledge Base to your Agent.
- For a given user utterance, the Bot would then try to match it to the predefined intent, or a system generated intent based on its match to the text in the Knowledge Base.

Solution Architecture



Dialogflow

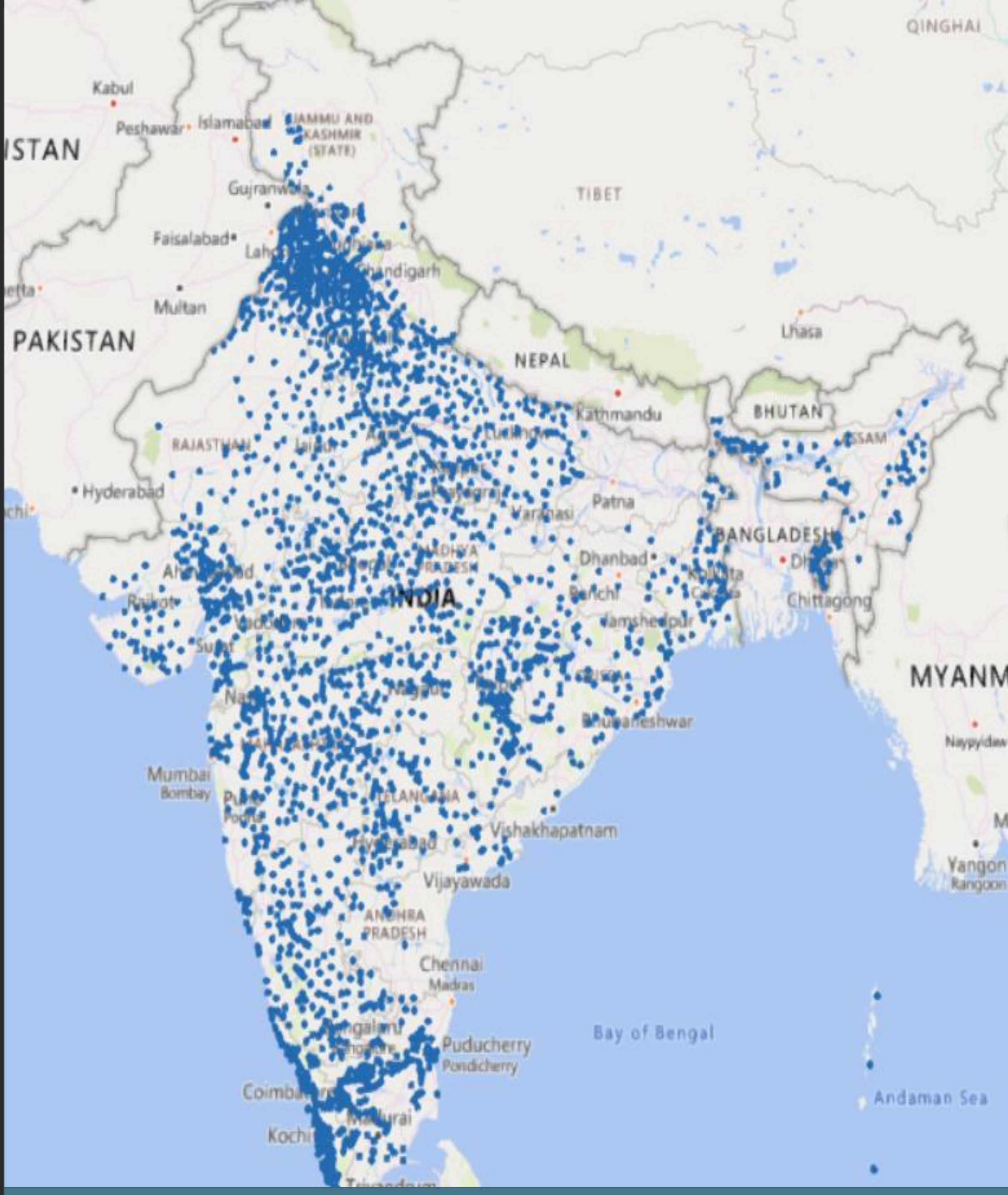
Success, In Production: Pricing Data

- Data for all the crop varieties and commodities, from all markets was combined for 2019.
- Latitude Longitude was derived using the Google Geocoding API for all the market
- K-Nearest Neighbors algorithm was trained to form clusters using the Latitude and Longitude Data
- Latitude Longitude is captured from the device and based on that the five closest Mandi price is displayed on the app.

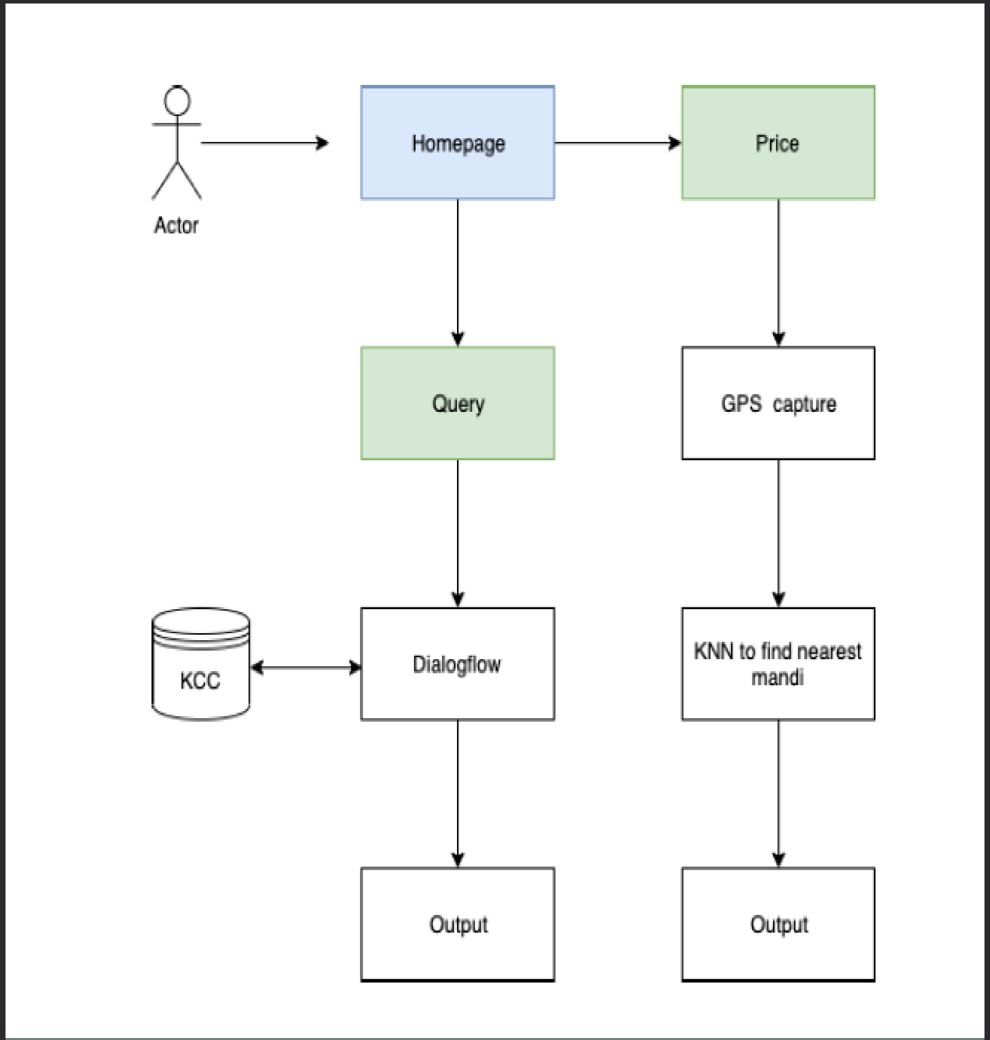
<u>Tools and Algorithms Used:</u>

Google Geocoding Services (Maps Platform) KNN, Bootstrap Studio

Data from 2300 Local Markets Across India



Solution User Architecture



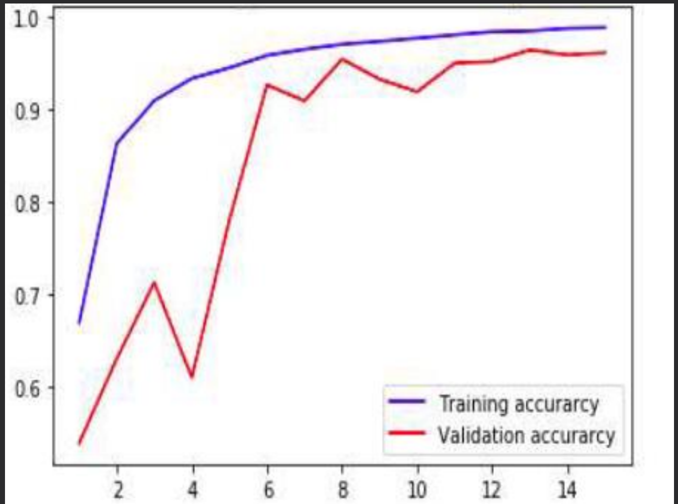
Dialogflow

Demo – Component 1

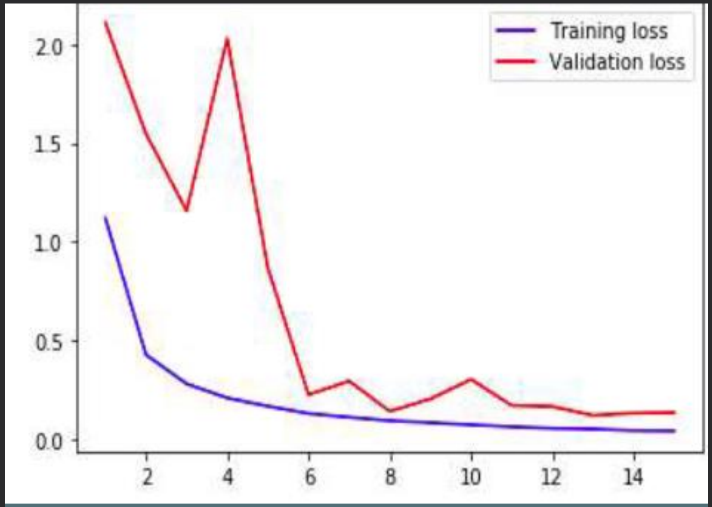
End to End – Chatbot and Pricing Platform

Testing Ground on Image Data: Plant Village Data

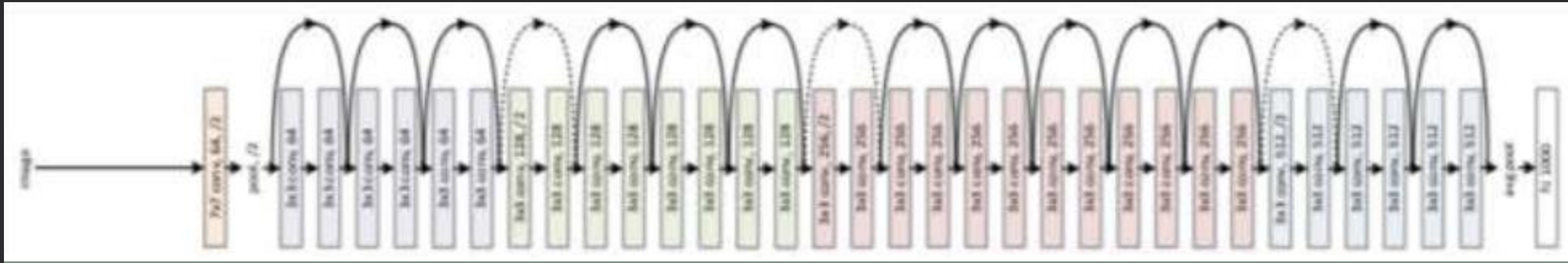
The model was trained for 15 Epochs with a validation split of 20% of the data



Training and Validation Accuracy



Training and Validation Loss



ResNet50 Architecture

Maize Field Images: Northern Light Blight

Faster RCNN Model

- **VGG Layer**: Pre- trained VGG 16 Model is applied to the image develop feature map

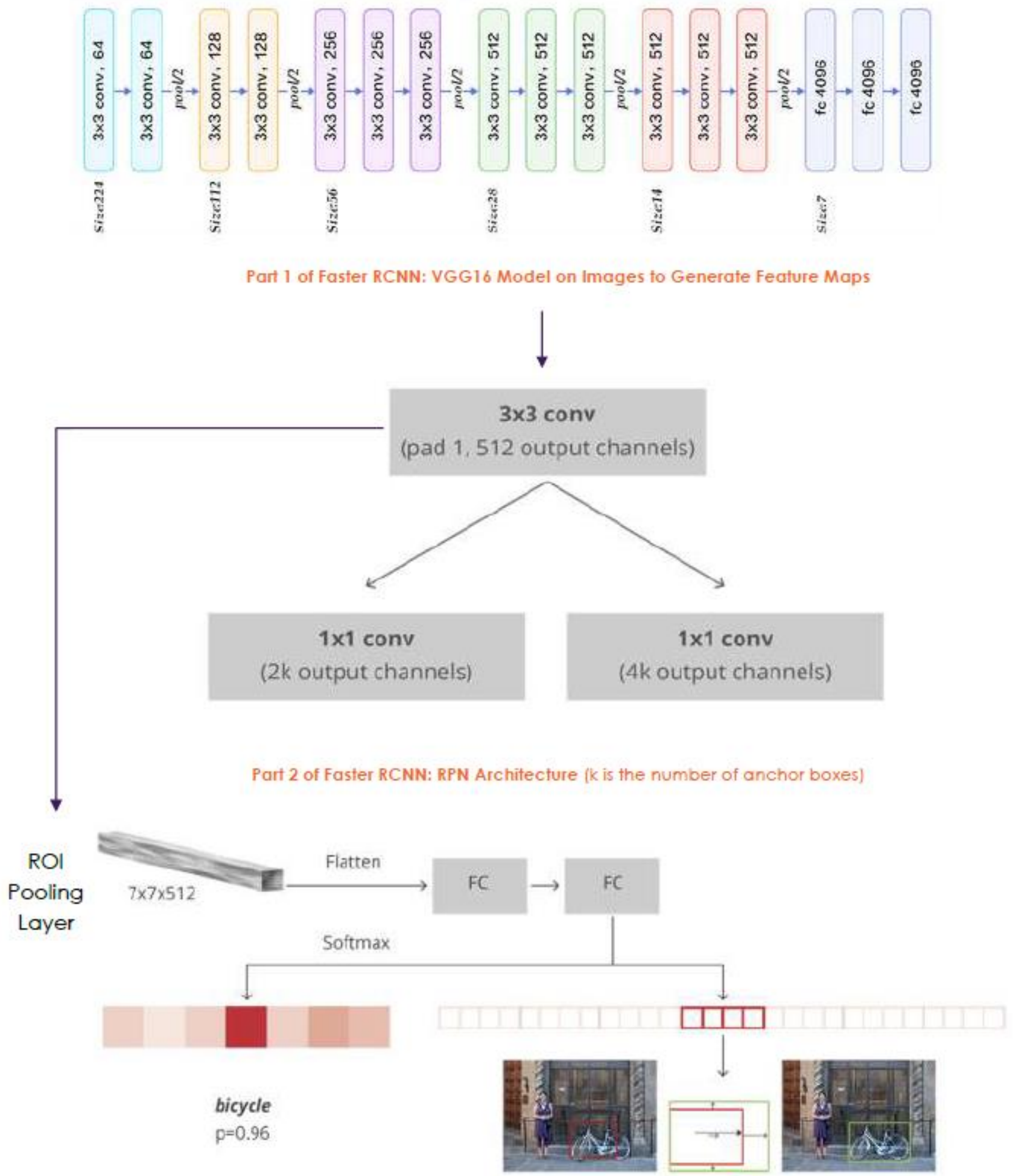
- **RPN Layer**: Each point in feature map is considered as an anchor. We need to define specific ratios and sizes for each anchor

    - Then RPN is connected to Convolution Layer with 3*3 filter, 1 padding and 512 output channels.
    - Intermediate Output: Connected to 1*1 Fully Connected Layer - box classification and box regression

- **RoI Layer:** ROI pooling is used for these proposed regions (ROIs). The output is 7x7x512.

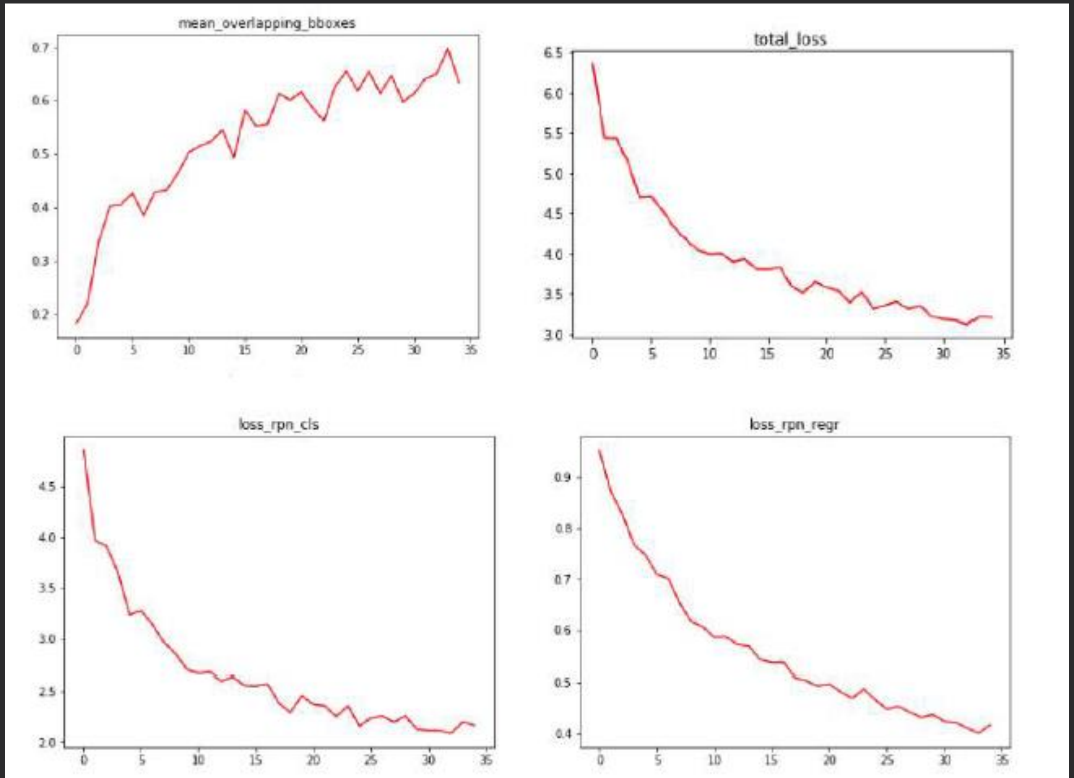
    - Then, we flatten this layer with some fully connected layers.

- **Classifier Layer:** The final step is a softmax function for classification and linear regression to fix the boxes’ location.

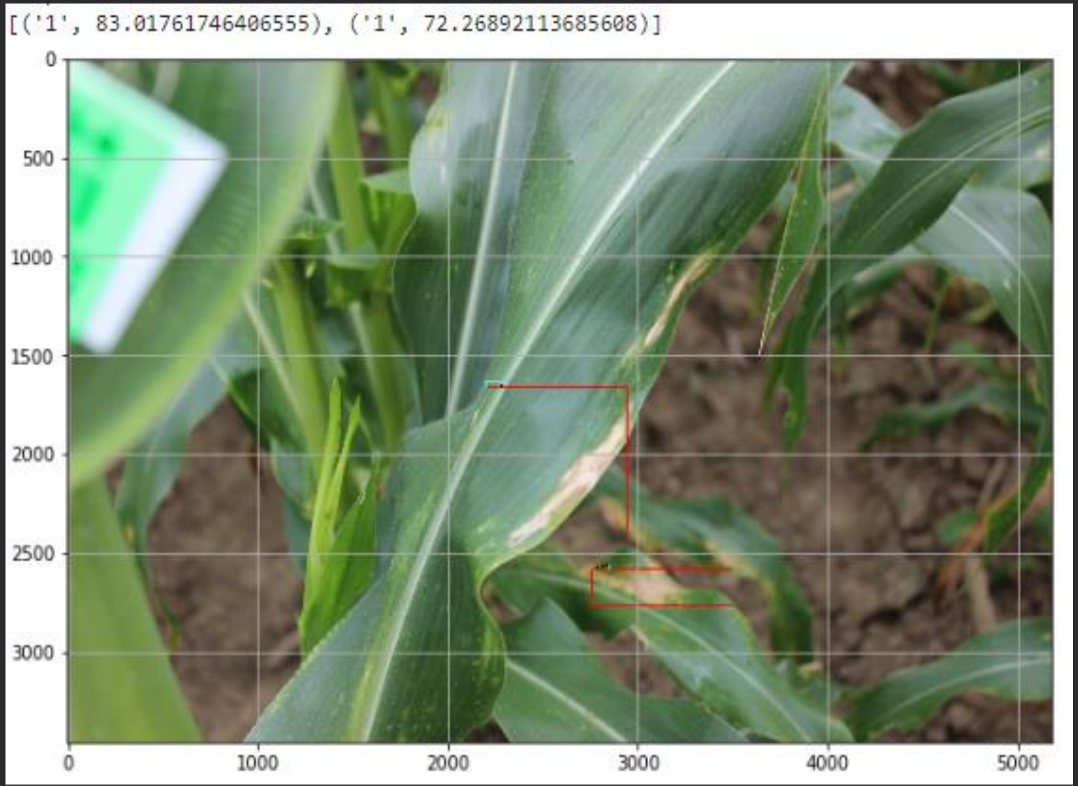


Real Model: Maize Field Images – Northern Light Blight

Train Data: 1462 Images – 6257 Bounding Boxes, 361 Images – 1442 Bounding Boxes



Performance Evaluation (35 Epochs)



Actual Images

Training Data: Mean Overlap of Bounding Boxes: 69.8% | Classification Accuracy: 79% | RPN Regression Loss: 2.2 | RPN Classification Loss: 0.401
Test Data : Mean Average Precision: 67.5%

Summarizing the Challenges

- Struggled with Government Website Structure while crawling data, frequently server down
- NLP Problems generally require a lot of manual intervention in terms of defining intents and entities
- Rasa NLU: Memory Intensive, a lot of production related issues and no out of box integration
- Dialogflow: Only high level customisations are possible.
- Inability to deploy Speech to Text on Server – Problem caused by PyAydio
- Navigating the complex ecosystem of cloud services, we chose Google Cloud Platform and struggled with multiple services it offers what to choose and what not to.
- Ended up getting a billing of $150 over and above 600$ of Free credits on GCP. Mistakenly kept the Cloud SQL Instance on which dried up a lot of credits.

Smooth Seas Do No Make Skilful Sailors

Way Forward: How this can turn into an actual On - Ground Solution?

- Add Support for Multiple India Languages in the Chatbot to ensure Usability
- Add Better Speech to Text Compatibility for Indian Languages
- Integrate Other Authentic Prescriptive sources that ca help
- Integrate Additional information like Soil, Weather
- Fetch Real Time Pricing Data from Government sources instead of the current static data
- Collate a more comprehensive dataset for different crops and disease to train image model for detection
- **Strategic Partnerships:** CABI India (Vinod Pandit, Director), Prof. Dharmendra Saraswat (Purdue University), CropIn (Richa Hukumchand, R&D)

Launch as an Application for On Ground Usage

Thank You!