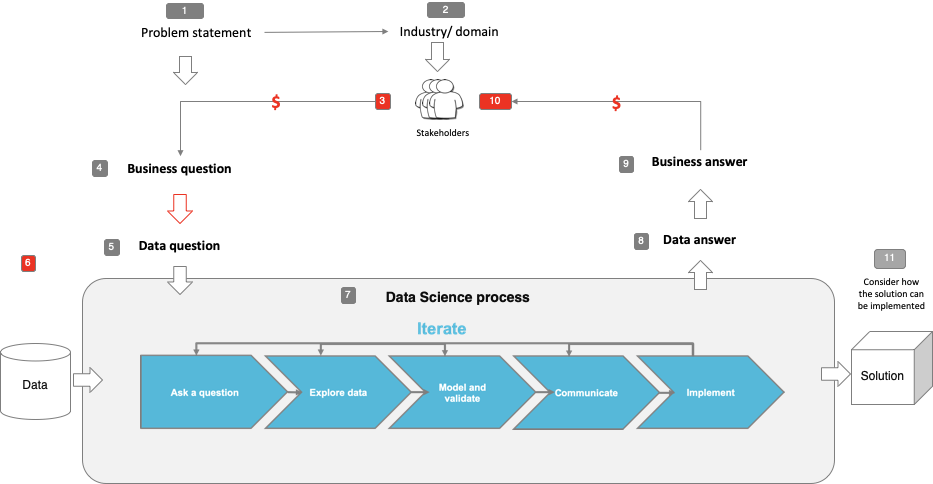
**Capstone Project**

**Document**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On the other hand, additional information may be needed.

**Identified problem statements:**

| **Problem statement** | **Dataset source** |
| --- | --- |
| Primary problem statement: Misdiagnosis of foliar diseases | https://www.kaggle.com/competitions/plant-pathology-2020-fgvc7/overview |
| Backup problem 1: Santander Customer Satisfaction | https://www.kaggle.com/competitions/santander-customer-satisfaction/overview |
| Backup problem 2: Home Credit Default Risk | https://www.kaggle.com/competitions/home-credit-default-risk/overview |

# Problem statement

* What is the problem or the opportunity that the project is investigating?
  + The U.S. apple industry is annually worth $15 billion USD (~10% of total agriculture export revenue)
  + During the growing season, the apple orchards are under constant threats from insects, fungal/bacterial/viral pathogens
  + These can results in loss of fruit bearing trees which incurs monetary losses of millions of dollars
  + Therefore, early pest and disease detection are critical for prompt deployment of pest and disease management
  + Moreover, incorrectly identifying a disease and/or misuse of treatment can even aid in the spreading of the disease, further incurring remedial costs and efforts
* Why is this problem valuable to address?
  + Machine learning value: Fast and accurate diagnosis
    - Mitigate the additional costs and efforts from remedial treatments
    - Mitigate the loss of fruit bearing trees
    - Reduce cost and effort to run diagnosis
* What is the current state (e.g. unsatisfied customers, lost revenue)?
  + Traditionally, the diagnosis of the diseases is done manually by an expert
    - This process is effort intensive as an orchard can contain thousands of trees
    - Not all experts are trained equally (discrepancies in diagnosis)
  + With help from machine learning
    - Thousands of samples can be diagnosed in minutes
    - Thereafter, the pathology experts can analyze and either accept or reject the results
* What is the desired state?
  + Develop and implement disease classification solution
* Has this problem been addressed by other research projects? What were the outcomes?
  + https://bsapubs.onlinelibrary.wiley.com/doi/10.1002/aps3.11390

# 

# Industry/ domain

* What is the industry/ domain?
  + Main industry - U.S. agriculture industry
  + Sub industry - U.S. apple industry
* What is the current state of this industry? (e.g. challenges from startups)
  + No data has been released stating the application usage of such a solution
* What is the overall industry value-chain?
  + Company source experts on plant pathology > Company uses solution to diagnose plant pathogens fast and accurately > Experts give recommendations/annotations on predicted diagnosis > Company deploy treatment
* What are the key concepts in the industry?
  + NA
* Is the project relevant to other industries?
  + This study can be extended to different industry which data involves imaging/image processing
  + Similar sub industry can make use of this solution as well (Different crop other than apples)

# Stakeholders

* Who are the stakeholders? (be as specific as possible)
  + Stakeholders can be business owners of food production company or plant pathology experts themselves
* Why do they care about this problem?
  + This problem can lead:
    - loss of fruit bearing trees
    - additional costs and efforts from remedial treatments on disease ridden crops
* What are the stakeholders’ expectations?
  + Solution to help mitigate the losses

# 

# Business question

* What is the main business question that needs to be answered?
  + Can we reduce the loss of fruit bearing trees?
  + Can we reduce the cost and effort incurred while handling spread of resistant pathogens?
  + Can we reduce cost and effort for the diagnosis process?
* What is the business value of answering this question? (quantify value and make necessary assumptions)
  + Primarily saving of cost (Monetary value)
* What is the required accuracy?
  + Accurate diagnosis is critical, as misclassification can lead to misuse of treatment
* What are the implications of false positives or false negatives?
  + NA

# Data question

* What is the data question that needs to be answered?
  + Can we build a solution that predicts accurate diagnosis of plant diseases
* What is the data required to answer the question?
  + As we are dealing with diagnosis of plant diseases, we have to look into 2D imaging of plants

# Data

* Where was the data sourced?
  + https://www.kaggle.com/competitions/plant-pathology-2020-fgvc7/data
* What is the volume and attributes of the data?
  + 1820 sample plant images (Only interested in the training dataset. Testing data set is used for competition)
  + 1820 rows with 4 one-hot encoded categorical variables
* How reliable is the data?
  + Dataset is reliable (taken from Kaggle, a data science competition platform and online community of data scientists and machine learning practitioners)
* What is the quality of the raw data?
  + Generally clean
  + High quality images
* How was this data generated?
  + Manually captured images and annotated by experts denoting the 4 categorical variables
* Is this data available on an ongoing basis?
  + Still publicly available in Kaggle

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?
  + Resizing is done to reduce image dimensions
  + Image augmentation is also done to help prevent overfitting
* What are the highlights of the Exploratory Data Analysis (EDA)?
  + Categorical distribution was balanced except for multiple diseases
  + Healthy samples are devoid of any pigmentation/holes
  + Scab samples are found to have dull yellow/greyish pigmentation indicating the fungal infection
  + Rust samples are found to have bright yellow pigmentation indicating the fungal infection
  + Multiple diseases samples are found to have symptoms from scab and rust samples as well as holes indicating presence of insects as well as infections
* Is the pipeline reusable? (for example, to process future data?)
  + Yes and it has to be used due to initialised model inputs
* What are the intermediary data structures used (if any)?
  + NA

## Modelling

* What are the main features used?
  + Features: Entire image dimensions after data wrangling
  + Dependent variables: 1820 rows, 4 columns of one-hot encoded categorical variables
* Did you find any interesting interactions between features?
  + NA (RGB channels can be explored)
* Is there a subset of features that would get a significant portion of your final performance? Which features?
  + NA
* How did you select features?
  + NA
* What feature engineering techniques are used?
  + Resizing
  + Image augmentation (https://www.tensorflow.org/tutorials/images/data\_augmentation)
* What are the models used?
  + RestNet50
  + DenseNet121
  + EfficientNetB7
  + Custom model
* How long does it take to train your model?
  + 00h:04m:36s
  + 00h:06m:20s
  + 00h:12m:24s
  + 00h:03m:40s (Respectively)
* What are the tools used? (cloud platform, for example)
  + Kaggle TPU resource
* What are the model performance metrics?
  + Categorical accuracy
  + Validation categorical accuracy
  + Loss = categorical\_crossentropy
  + Validation Loss
* Which model was selected?
  + Any of the modern pre-trained models, however if we go by ranking, DenseNet121 > EfficientNetB7 > RestNet50

## Outcomes

* What are the main findings and conclusions of the data science process?
  + API:
    - Can be extended as a mobile phone application to be used on the go (Average diagnosis time ~1-2 seconds)
  + Objective:
    - Machine learning can add immense value in terms of cost and effort savings by achieving fast and accurate diagnosis of diseases
    - Early diagnosis will prevent loss of fruit bearing trees
    - Accurate diagnosis will prevent rapid spread of resistant pathogens
    - Machine learning can also be applied to other crops and not just apple trees
  + Machine Learning Models:
    - Computational resources as a metric (Memory and Time)
    - SOTA model architectures (ResNet,DenseNet,EfficientNet) are able to classify images with high accuracy
    - Improve ResNet evaluation scores

## Implementation

* What are the considerations for implementing the model in production?
  + Scope and project deliverables should be define well first
  + Currently the solution’s functionality would just be predicting plant diseases, additional features could be added, UI/UX could be improved on the platform as well.

# 

# Data answer

* Was the data question answered satisfactorily? What is the confidence level in the data answer?
  + Yes, solution is able to predict plant diseases quickly and accurately (~1-2 seconds per prediction with ~99% categorical accuracy)

# Business answer

* Was the business question answered satisfactorily? What is the confidence level in the business answer?
  + No, there was no quantifiable value gain showcased (e.g. cost & effort savings) This would be part of future works and improvements

# Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?
  + Machine learning can add immense value in terms of cost and effort savings by achieving fast and accurate diagnosis of diseases
  + Early diagnosis will prevent loss of fruit bearing trees
  + Accurate diagnosis will prevent rapid spread of resistant pathogens
  + Machine learning can also be applied to other crops and not just apple trees

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?
  + A mobile phone application that can run predictions on the go

# 

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
  + Jupyter notebook: IOD\_Capstone\_Project.ipynb
  + Dataset: <https://www.kaggle.com/competitions/plant-pathology-2020-fgvc7/data>
  + Python code: app.py
  + Keras exported models: denseNet\_plant\_pathology.keras, EfficientNet\_plant\_pathology.keras, ResNet50\_plant\_pathology.keras
  + HTML template: index.html
  + CSS Style sheet: style.css
* What are the resources used in the project? (libraries, algorithms, etc)
  + <https://www.kaggle.com/competitions/plant-pathology-2020-fgvc7/overview>
  + <https://bsapubs.onlinelibrary.wiley.com/doi/10.1002/aps3.11390>
  + <https://www.tensorflow.org/api_docs/python/tf>
  + <https://www.kaggle.com/code/tarunpaparaju/plant-pathology-2020-eda-models/notebook>
  + <https://www.kaggle.com/code/mgornergoogle/five-flowers-with-keras-and-xception-on-tpu/notebook>
  + <https://www.kaggle.com/code/ateplyuk/fork-of-plant-2020-tpu-915e9c/notebook>
  + <https://www.kaggle.com/code/shawon10/plant-pathology-eda-and-deep-cnn>