**Group: Skynet  
Machine Learning Proof of Concept (POC):**

* ***Start with a Problem:***
  + - Use machine learning to help identify spoiled fruit vs non-spoiled fruit to be used on a fruit sorting system that will automatically identify and discard spoiled fruit.
  + **Narrow down the problem**
    - Focus the system on identifying spoiled bananas vs. non-spoiled bananas.
  + **Build a business case after defining the problem and confirming with the team it is worth solving.**
    - **Is it an impactful problem?**
      * The system will assist farmers and vendors in automating the process of sorting through bananas to quickly identify those that are ready for market and those that should be discarded.
    - **Can you quantify the business outcome?**
      * We will be able to calculate how many more bananas we can put out into the market with the automated systems vs. without.
    - **Does the problem have a substantial volume of associated data?**
      * The USA alone has approximately 10.5 billion pounds of bananas imported yearly.
  + **Validate feasibility.**
    - **How much data is needed?**
      * The more data the more accurate we can get the Machine learning algorithm. We can start with 100 datasets, and test accuracy, and increase in multiples of 2 until desirable results are achieved.
    - **Does the dataset effectively address problem?**
      * Yes, the data set included spoiled bananas with annotations for the Machine learning to gain insight on what is a spoiled banana, while also introducing non-spoiled bananas in the data set for the Machine learning algorithm to learn from.
    - **Is the dataset complete and properly annotated?**
      * The data set will be deemed complete when desirable results are reached. Will confirm the images are properly annotated by manually marking what the Machine learning should look for on the bananas to identify if they are spoiled.
* **Prototype and test with real data**
  + **Dataset preparation**
    - **To train our model, we need a dataset consisting of pairs of input and actual labels. Data must be capable of answering the questions we want our model to address.**
      * **Does the dataset align with a problem?**
        + Yes, our data sets have the details needed to teach the machine learning algorithm the difference between spoiled and non-spoiled bananas.
      * **Is the dataset complete?**
        + To be determined, as we need to run accuracy tests to see if we are achieving desired results.
      * **How can we annotate and ensure quality dataset overtime?**
        + For initial testing, we can annotate manually within our team, once we need to to annotate at higher volume, we can outsource by providing examples of annotations along with unannotated data to an external source that can handle higher volume of data for manual annotations.

**Things to consider.**

* + - *Data availability* – Ensure there is an available dataset for your potential solution.
    - *Data Size* – Assure the datasets imputed allow proper learning.
    - *Data Distribution and Patterns* – To avoid bias, you need to take steps to address the imbalance in the datasets.
    - *Data annotation* – Convert unlabeled data into labeled data to prepare out datasets.
  + **Model training**
    - Here is where we input the data to train the Machine learning algorithm. Which Machine learning algorithm will be implemented?
    - **Understand & diversify your data.**
      * + The quality of your training data determines the behavior and performance of your model. Poor data leads to a poor model.
        + For supervised learning models, the model can only learn what is present in the data. So, make sure all the necessary information is included.
        + Your data should cover all likely scenarios. In our case, we cannot expect the model to identify spoiled bananas where the spoiled sections of the bananas are hidden behind the skin.
        + The provided data should represent real-world situations, not just artificially staged images.
        + Ensure that your model receives an equal number of different types of data to avoid bias.
  + Model evaluation
    - Compare predicted labels against withheld actual labels.
      * **Accuracy = (True Positives + True Negatives) / (Total Samples)**
      * **Precision = True Positives / (True Positives + False Positives)**
      * **Recall = True Positives / (True Positives + False Negatives)**
      * [F1score](https://www.linkedin.com/feed/hashtag/f1score) is a combined metric that considers both precision and recall providing a balanced evaluation of a classification model's performance.  
          
        **F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall)**
* Release. Learn, iterate.
  + AB Testing: Only 20% of the bananas go through the machine learning process, and those bananas are inspected manually to confirm no significant issues.
  + Iteration: If new insight is gathered, create updated versions with updated annotations and labels. For example, introduce physically damaged bananas into the data set.