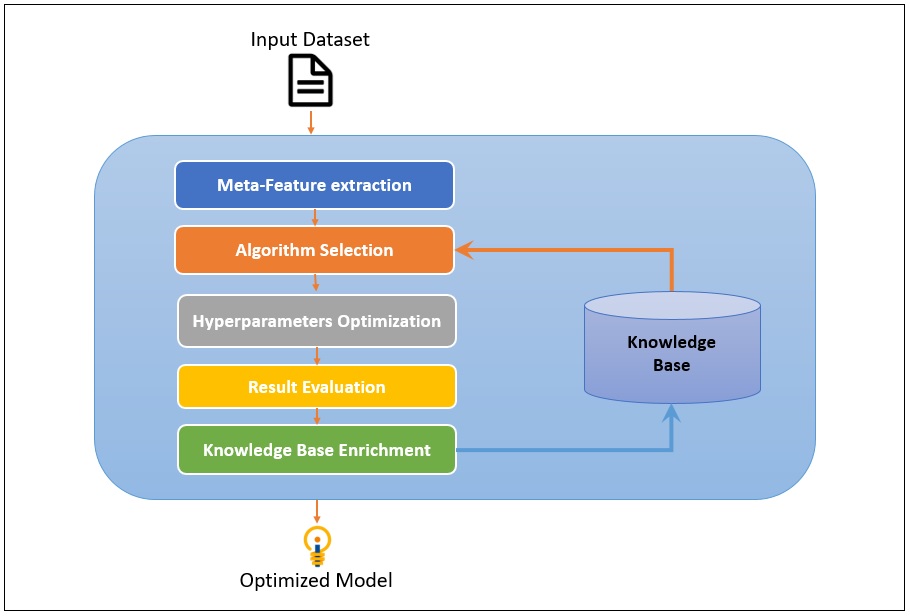
**Introduction**

* Machine learning model building is a highly iterative exploratory process where most scientists work hard to find the best model or algorithm that meets their performance requirement.
* In practice, there is no one-model-fits-all solutions, thus, there is no single model or algorithm that can handle all data set varieties and changes in data that may occur over time. Also each machine learning algorithm require user defined inputs to achieve a balance between accuracy and generalizability (which called hyperparameters).
* So the process of selecting the Algorithm and its hyperparameters usually require a lot of trails until we find it. In case of big datasets this process will need a lot of resources (Time, Hardware, effort, …)

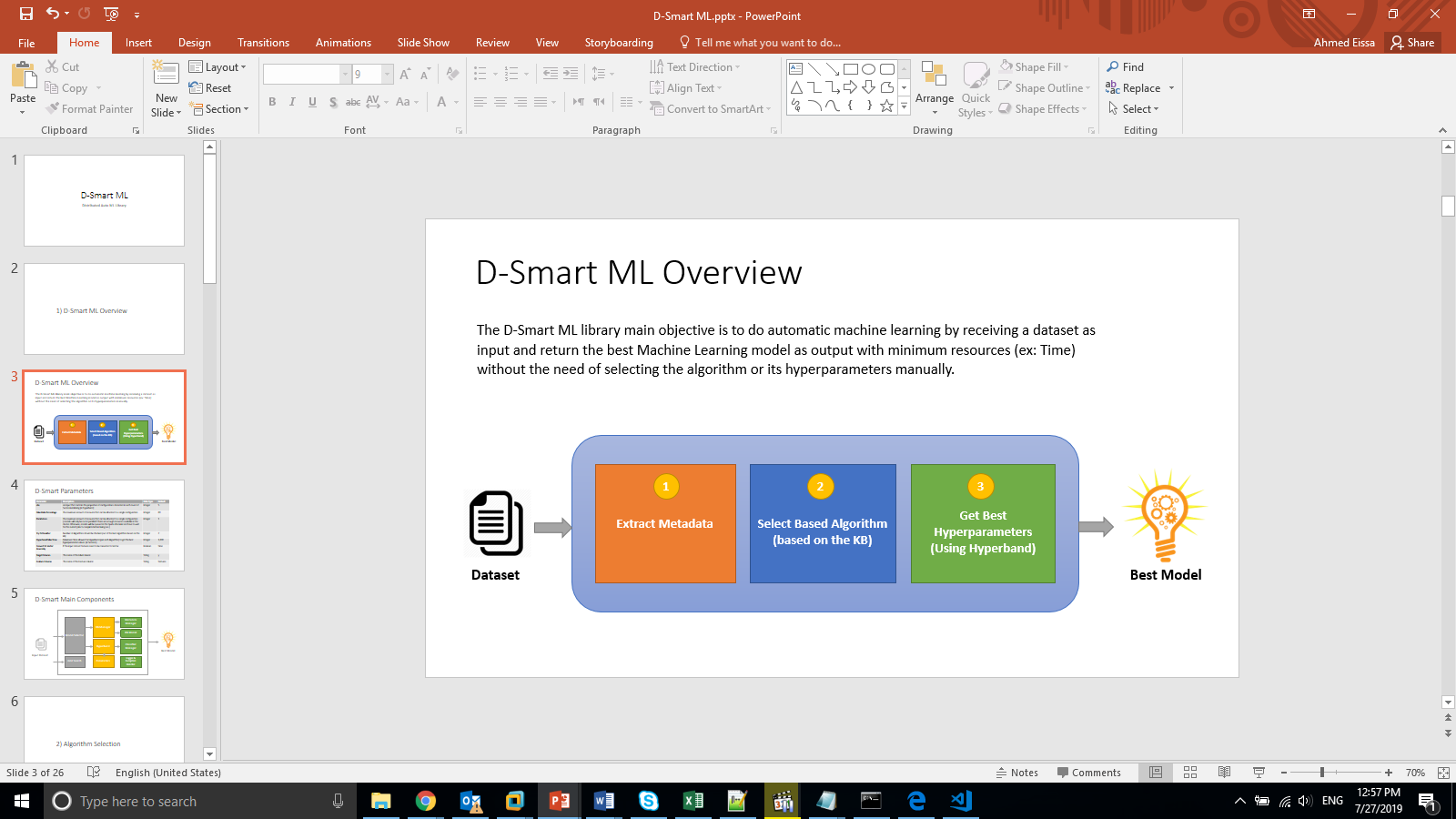
**Distributed Smart-ML Library**

* **The Main Idea**

****

The library receive dataset as an input and produce an optimized model as an output.

The library extracts some characteristics of the datasets and use an internal knowledgebase to determine the best algorithm, then use a hyperband method to find the best hyper parameters for the selected model.



The process consists of three steps as shown in the figure.

1. Metadata extraction
2. Algorithm Selection based on internal Knowledge base
3. Hyperparameter optimization using Hyperband algorithm

* **Datasets Meta Extraction & our Knowledge base**

For each dataset, a set of meta-data extracted that represent statistics and description for this dataset, the extracted meta data can be grouped as following:

* + **For All the Dataset**
    - Number of instances
    - Log (Number of instances)
    - Number of Features
    - Log (Number of Features)
    - Number of Classes
    - Number of numerical Features
    - Number of Categorical Features
    - Ratio between Categorical & Numerical
    - Number of Instances to Number of Features ratio
    - Number of Missing Value
    - Ratio of missing value
  + **For Classes**
    - Class Entropy
    - Classes Probabilities:
      * Minimum
      * Maximum
      * Standard deviation
      * Mean
  + **For Categorical Features**
    - Sum of all symbols
    - Mean of all symbols
    - Standard Deviation for all symbols
  + **For Numerical Features**
    - Skewness for all Numerical features
      * Max
      * Min
      * Standard Deviation
      * Mean
    - Kurtosis for all Numerical features
      * Max
      * Min
      * Standard Deviation
      * Mean

An important characteristic adds to the knowledge base is the accuracy against machine learning classifier.

The used set of classifiers are:

* + Random Forest
  + Logistic Regression
  + Decision Tree
  + Multilayer Perceptron
  + Linear SVC
  + Naïve Bayes
  + GBT
  + LDA
  + QDA
* **Algorithm Selection**

Since we have knowledge base contains the characteristics & the best classifier for a group of datasets, we can use machine learning to get the closest dataset in the knowledge base for an input dataset, then determine the expected best classifiers.

* **Hyper parameter Optimization**

We are using Hyperband in order to determine best hyperparameters quickly, hyperband based on Successive Halving algorithm which allocates exponentially more resources to more promising configurations, the algorithm do the following:

1. Uniformly allocate a budget to a set of hyperparameter configurations
2. Evaluate the performance of all configurations
3. Throw out the worst half
4. Repeat until one configuration remains

Hyperband consist of two loops:

* + the inner loop invokes SuccessiveHalving for fixed values of (hyperparameters configurations) and resources
  + the outer loop iterates over different values of (hyperparameters configurations) and resources

we have used number of instances (data sampling) as the hyperband resource, so the maximum resources that can be allocated is 100% of data.

* **Spark Implementation**

We have used Apache Spark to distribute the process (Feature Extraction, Algorithm Selection and hyper parameter optimization).

We have also added two classifiers (LDA & QDA) to Spark ML to increase the number of the available classifiers.