ML Workflow

Data Cleansing
Feature Engineering/Selection
Data Transforms

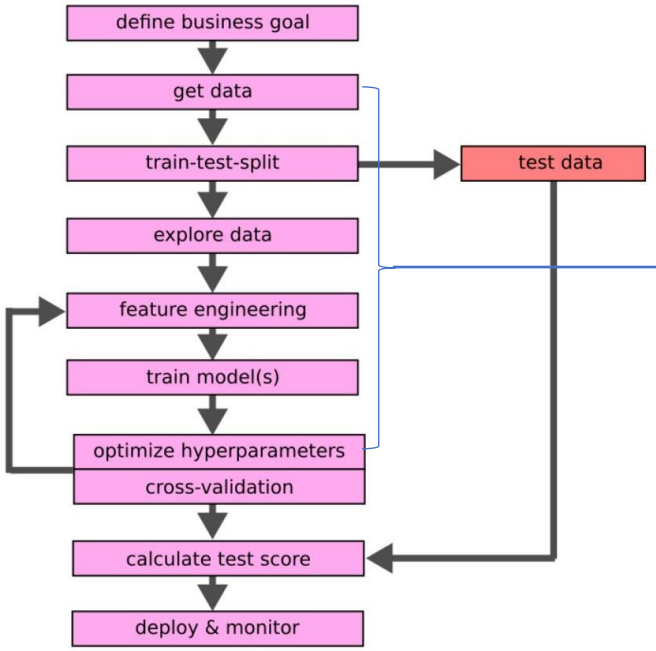
Behzad Azarhoushang

Spiced Academy – Colime cohort

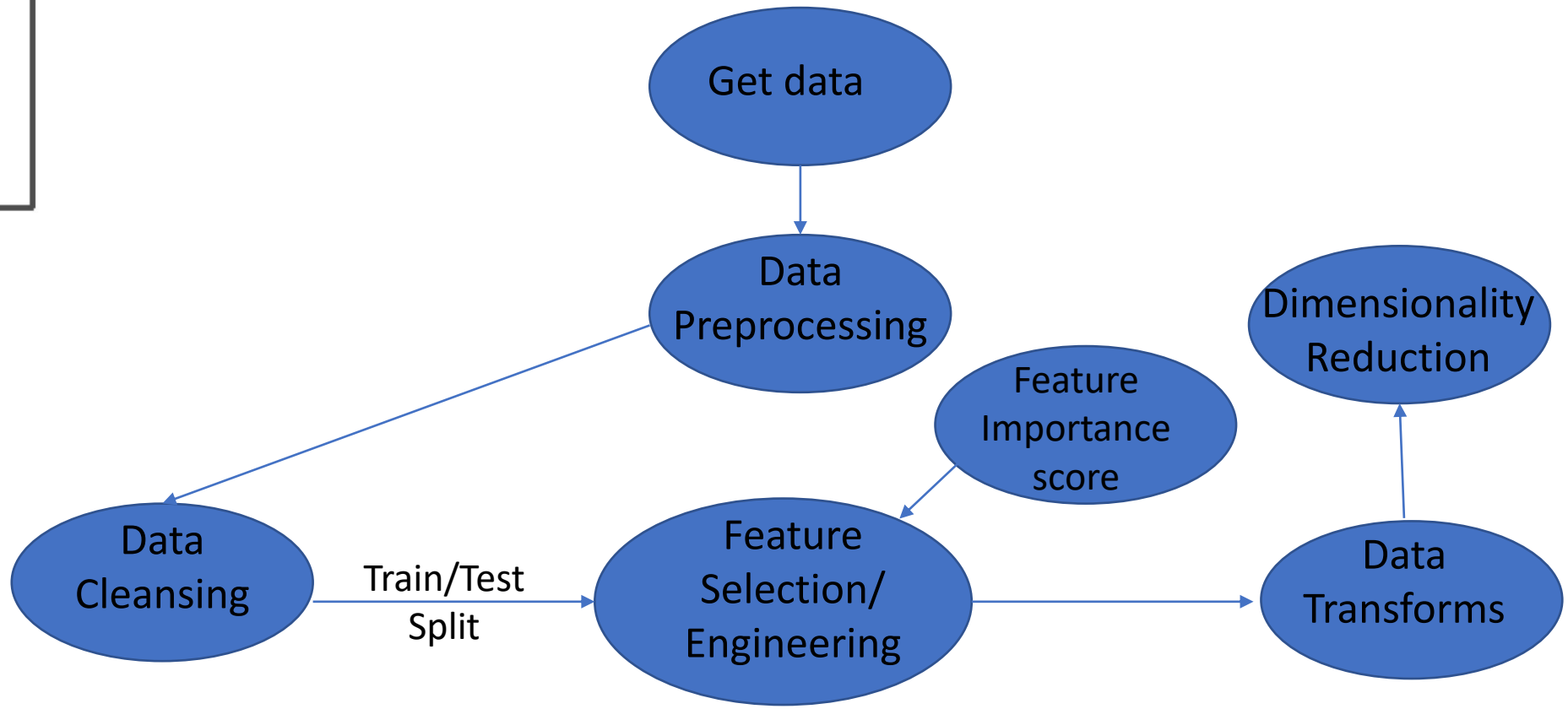
Intro

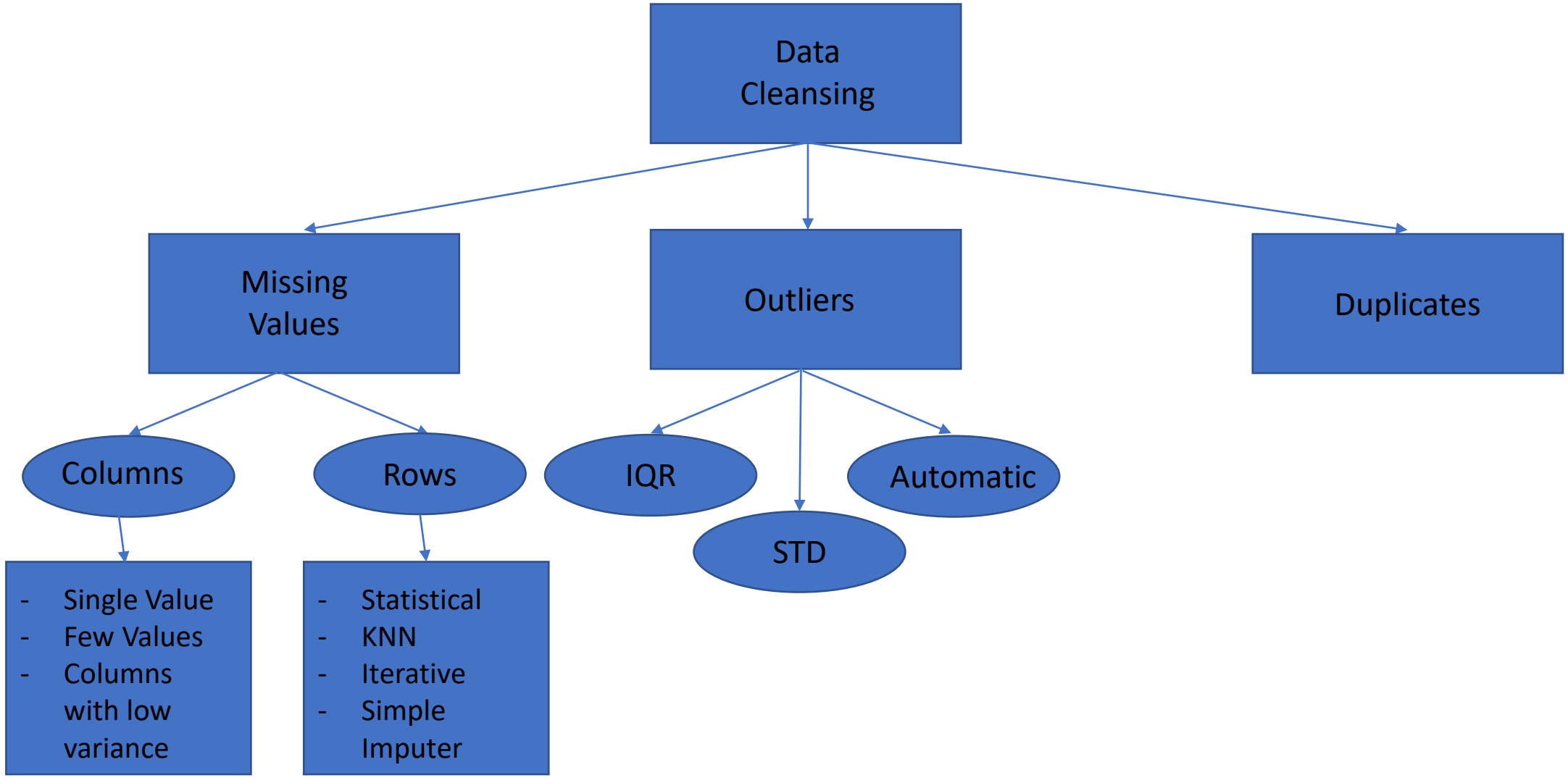
- I tried to write a summary of all common methods that being used in data processing stage by experts in the field of data science.
- As majority of expert argue: the data processing takes more than 80% of the job in any ML project (your model is as good as your data).
- I divided the Data processing into three main parts
  - Data Cleansing
  - Feature Engineering/Selection
  - Data Transforms
- I created a folder for each part which contain specific jupyter notebook for each or sets of method (only if they somehow linked together or when there is comparison involved)
  - In each folders there are some notebooks with '(Notes)' next to their name that provide summary of each method as well as their usage and their logic.
  - Also there is a 'Table of Contents' notebook that illustrates the sequence of topics
  - There are also short comments on top of each block of codes
- I got all codes from logikbot.com. In some cases, you need to download the dataset by yourself. They are all available in 'Kaggle'.

ML Workflow – Big Picture



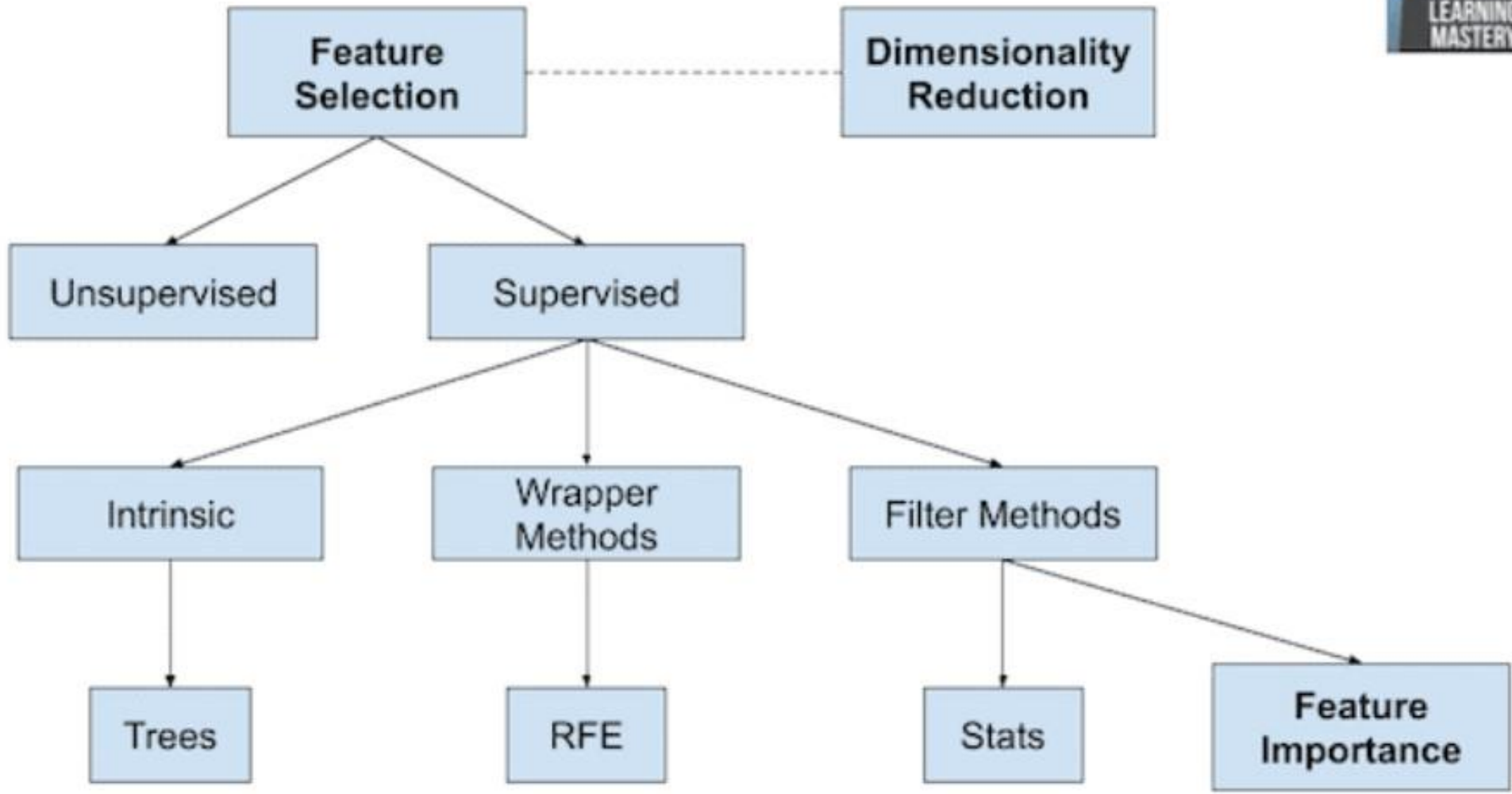
- Here my focus is on step 2 (get data) to step 7 (cross validation)



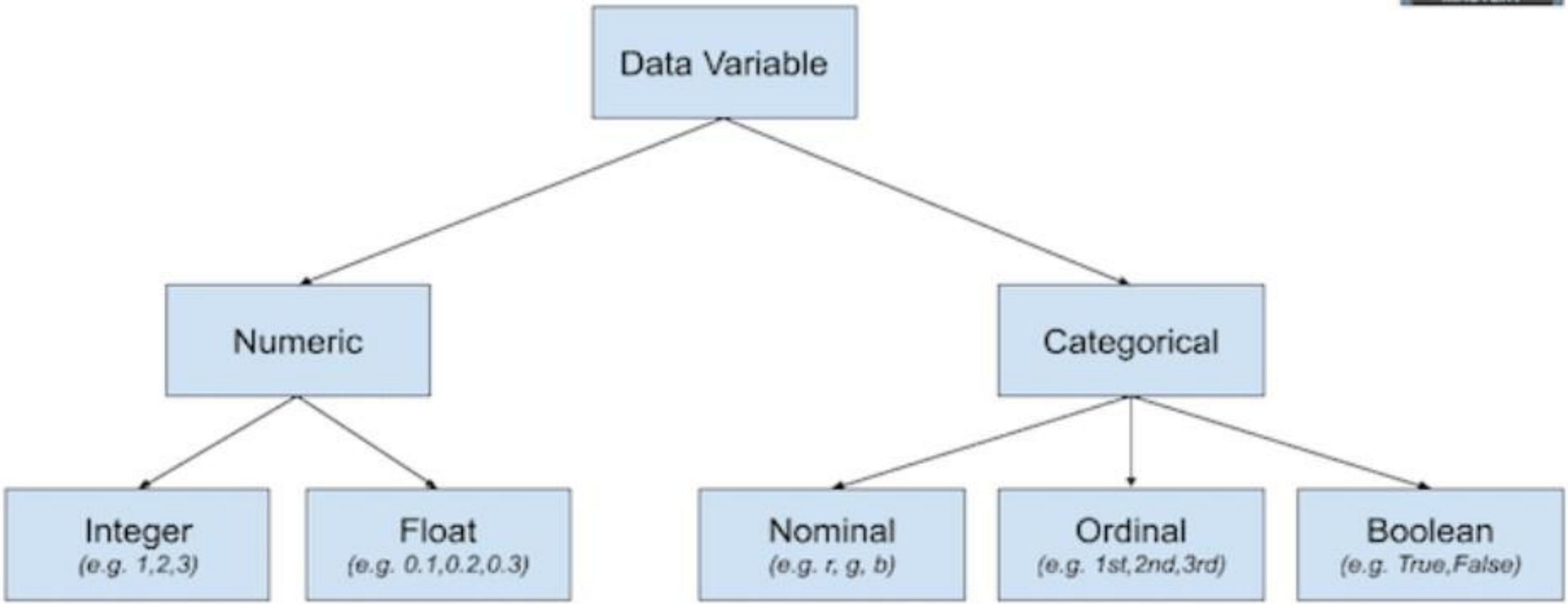


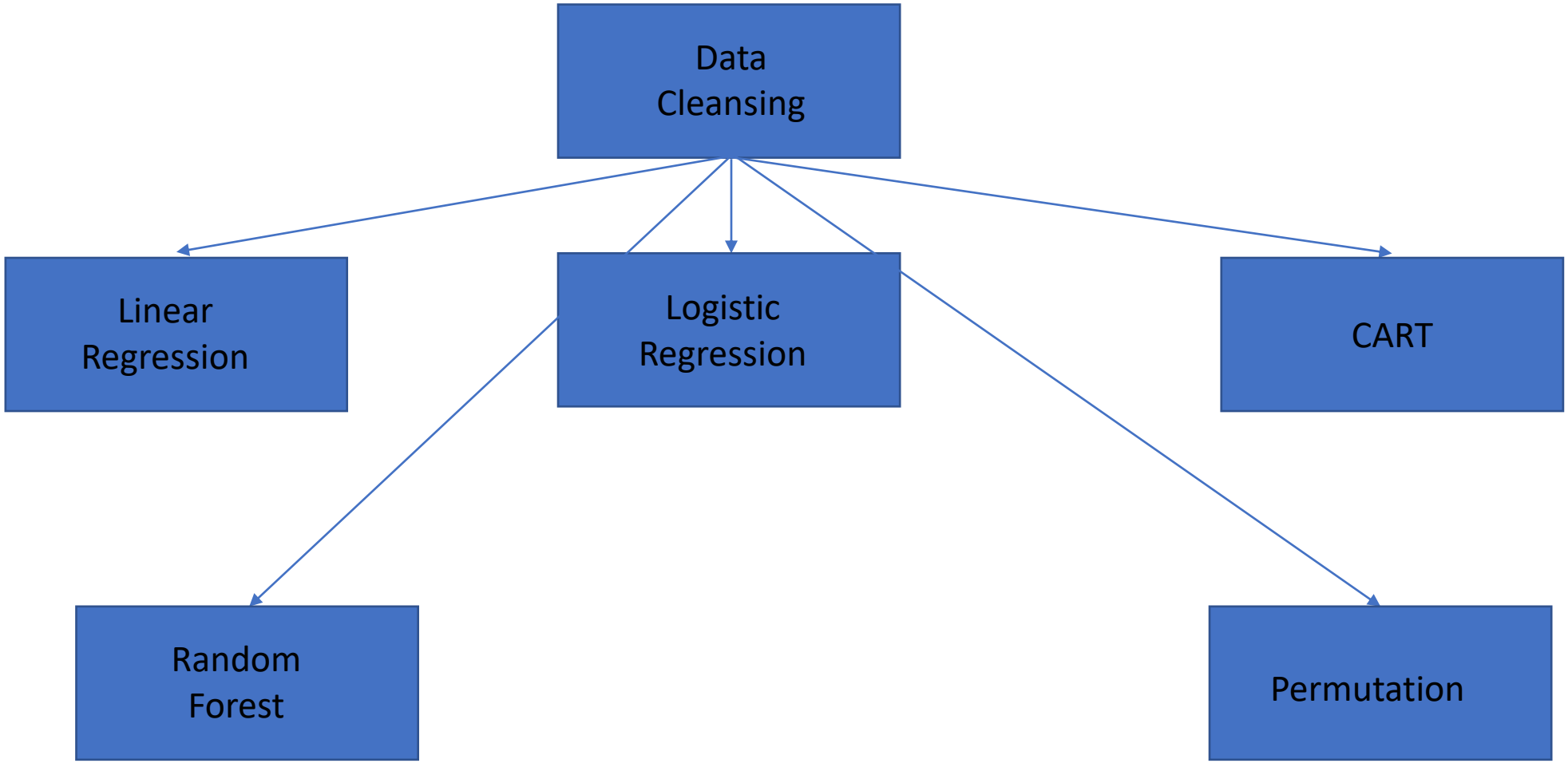
- There are lots of discussion about when we should do the train/test split and it seems there is no clear consensus among experts. But the rule of tomb is to do any above technique after split if you think it leads to data leakage. There is no consensus about what leads to data leakage. So go figure! I think at the end of the day our own judgment but more importantly experience will tell us where we should do the split

Overview of Feature Selection Techniques

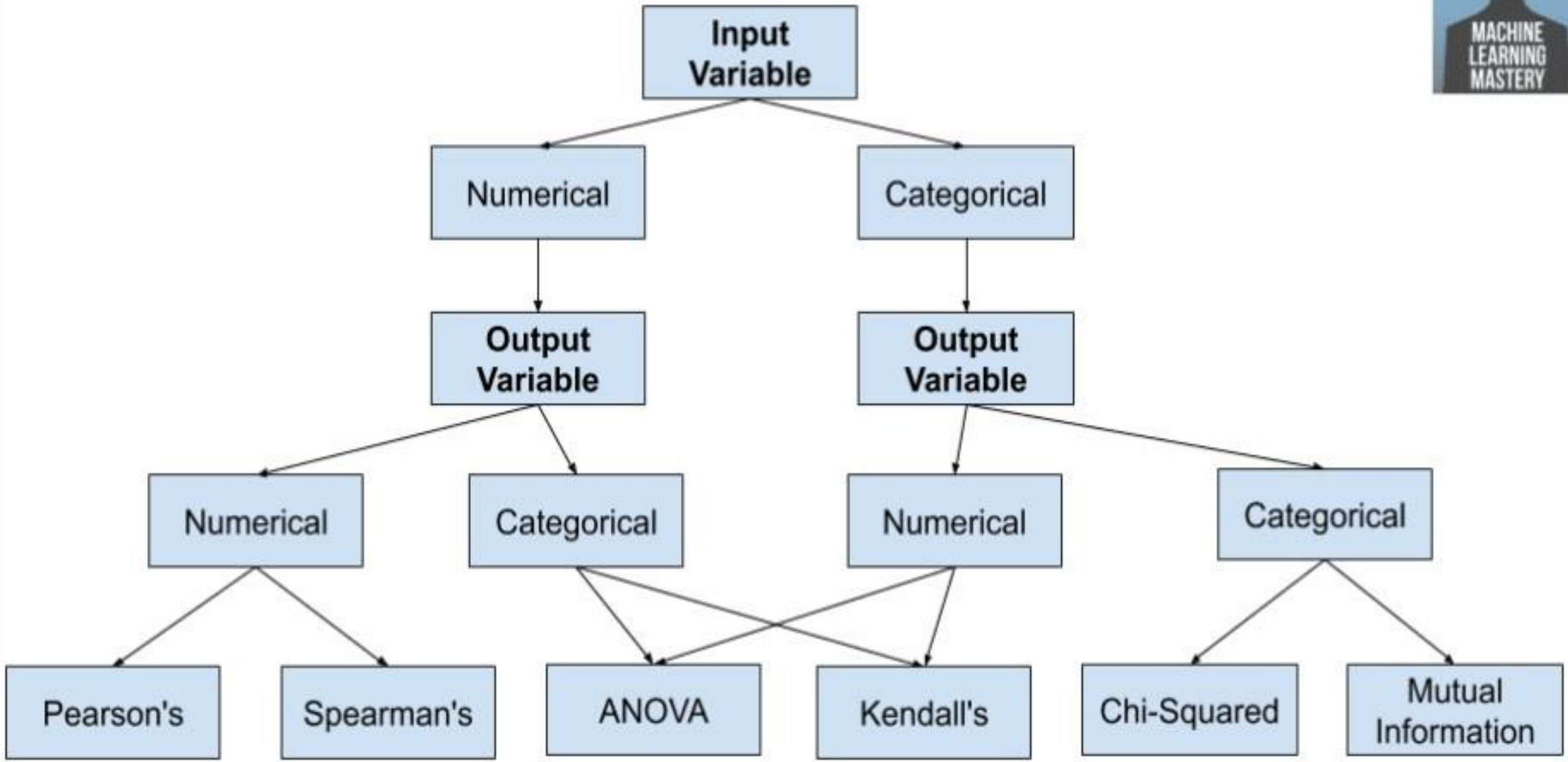


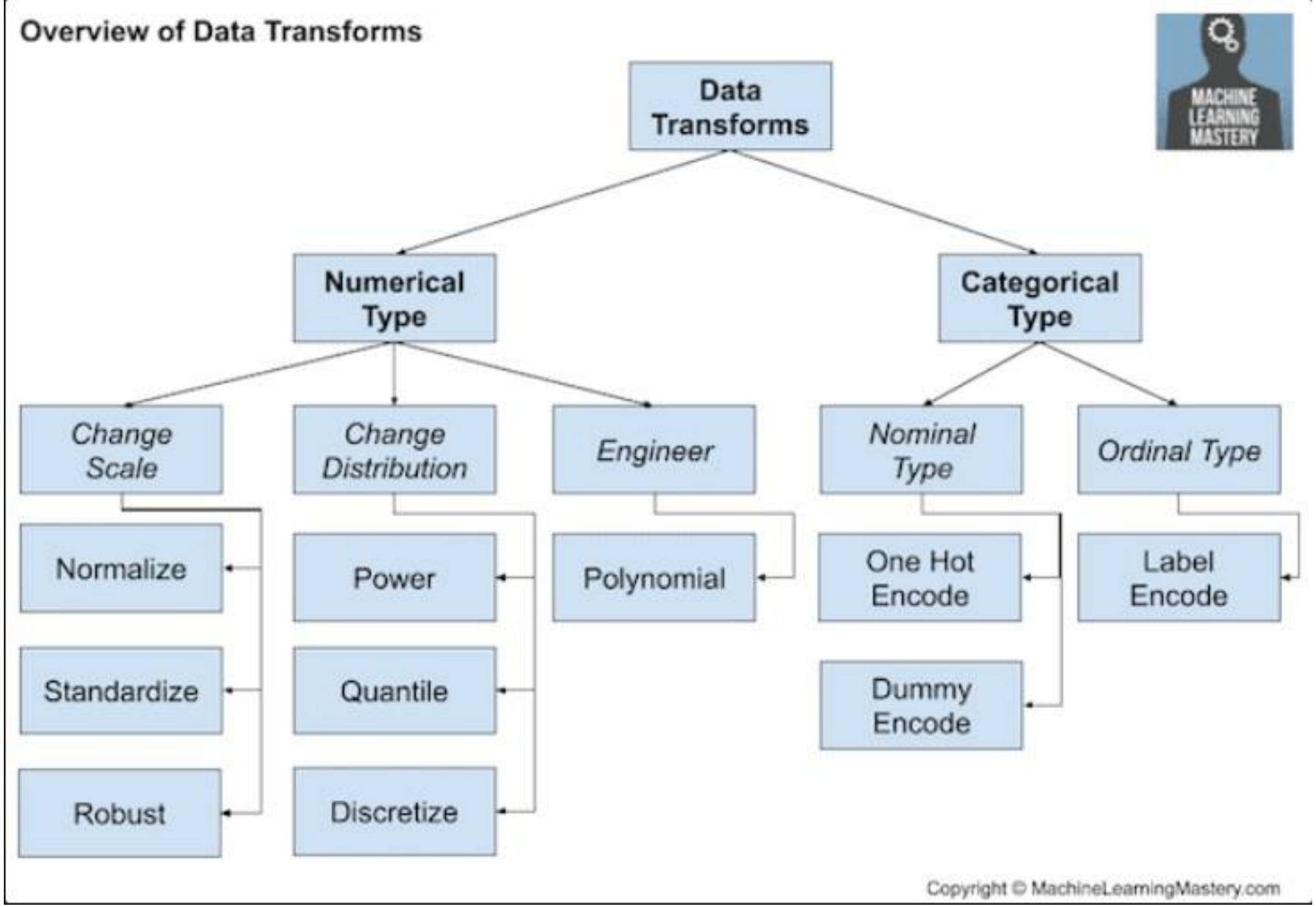
Overview of Data Variable Types



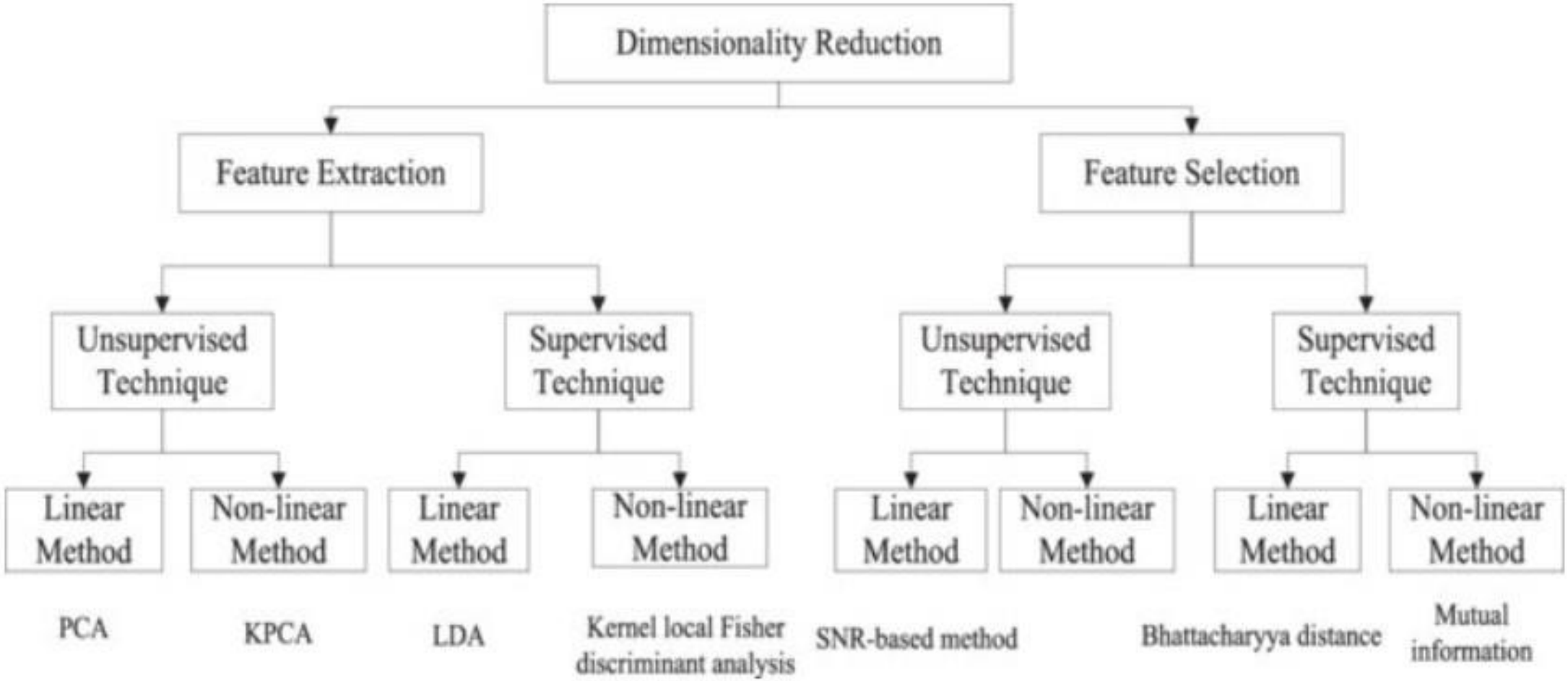

Data Cleansing
Linear Regression
Logistic Regression
CART
Random Forest
Permutation

How to Choose a Feature Selection Method





- There are also specific transforms such as column transformer and target variable transforms that are mostly related to smoothing the preparation of data for modeling
- More info about them is available -> Data Transform -> Advanced Transfer (Notes) and related notebooks



- Dimensionality Reduction could be done either during or after the Feature extraction (engineering) / Feature selection. Or you can do it after analyzing the performance of model using different sets of features to see which one works better.

Source: https://www.researchgate.net/publication/340059299_PCA-based_Feature_Reduction_for_Hyperspectral_Remote_Sensing_Image_Classification/figures?lo=1

Summary

- Obviously, there are many more techniques out there. But those that I mentioned here are the most common one.
- Don’t forget that each project/dataset is unique and need its own specific approach especially for feature selection/engineering which highly depend to our domain knowledge. But if you don’t have related domain knowledge always talk with field experts which are your collogues in your company’s department(s).
- Why I always used ‘Future Engineering/Selection’ together rather than separating them? Well, I think below link could help to clarify it.
  - https://innovation.alteryx.com/feature-engineering-vs-feature-selection/#:~:text=Feature%20engineering%20enables%20you%20to,features%20to%20a%20manageable%20number.
- I made this package for myself because of this is the way my brain organized things. Yours may work differently. However, I thought if it can help someone else, why not sharing it.
- At the end, please share with me if you find more techniques or comments so we could improve it together.