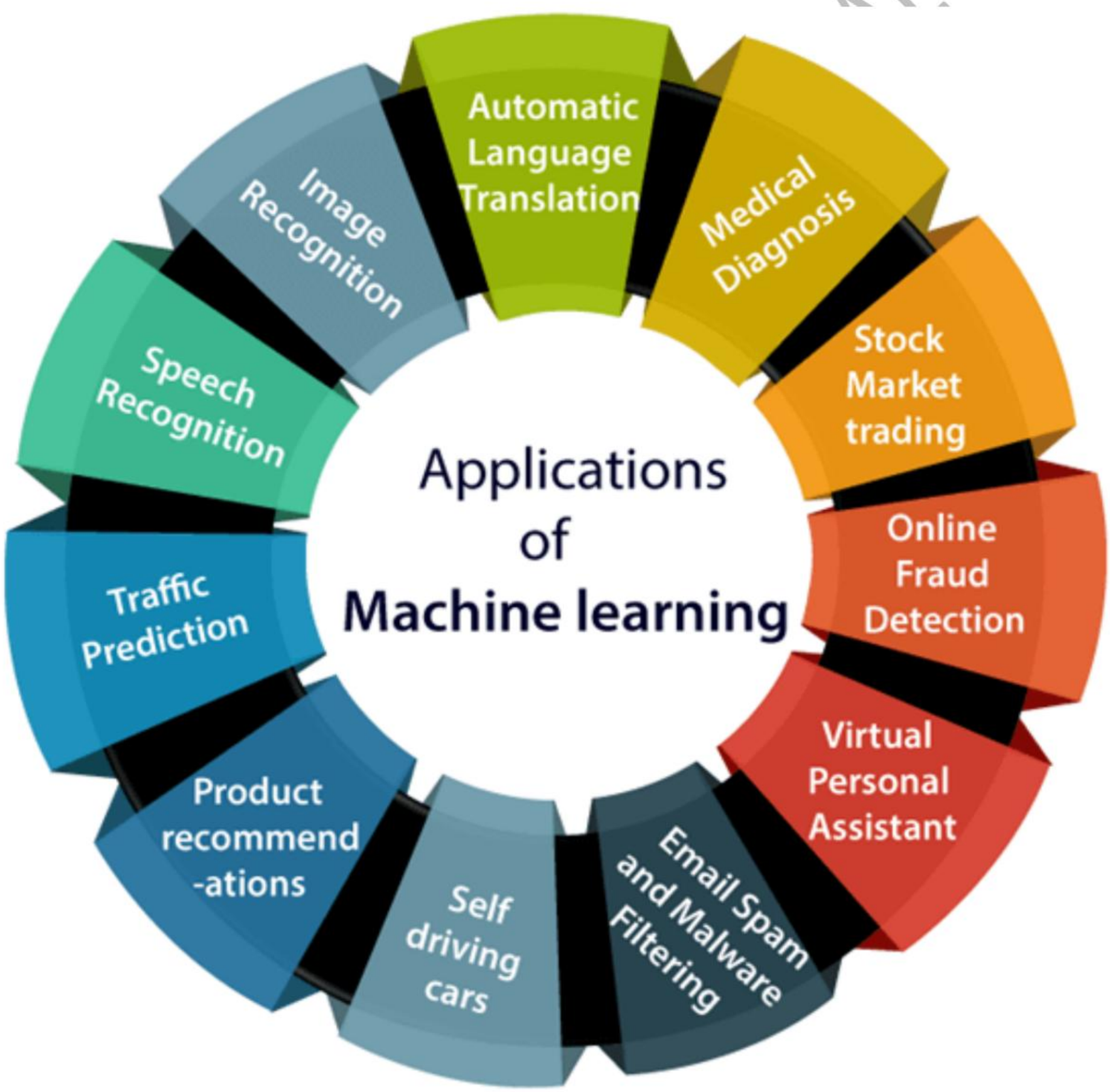
Chapter 1: Machine Learning Introduction

Introduction:

- Machine Learning also termed ML aims to grant computers the ability to learn by making use of statistical techniques.
- It deals with algorithms that can look at data to learn from it and make predictions/classifications/clustering.



ML Lifecycle:

- ML lifecycle includes the following steps:
    1. Gathering Data
    2. Understanding Data
    3. Model Selection
    4. Train & Test the Model
    5. Deploy the Model.

Step1: Gathering Data:

- Data can be collected from various sources such as files, database, internet or mobile devices.
- The quantity and quality of the data will determine the efficiency of the output.
- **Dataset:**
    - Dataset is collection of data which is arranged in some order. (mostly csv files).
- **Types of Dataset:**
    - Numerical Data:
        - Such as house prices, temp, etc
    - Categorical Data:
        - Such as Yes/No, True/False, GoOut/StayHome, etc.
- **Dataset Sources:**
    - Kaggle dataset, UCI repository, AWS, Googles Dataset Search Engine, etc.

Step2: Understanding Data:

- It is done to make the data suitable for machine learning model.
- **Handling Missing Data:**
  - Missing data can cause problem to the model and hence missing values can be handled by either:
    - By deleting row or column which consists of null values.
    - By calculating mean of row or column which contains missing values and put it in the place of missing values.
  -
- **Encoding Categorical Data:**
  - Since Machine Learning works on mathematics and numbers and if our dataset is having some categorical variable then it should be encoded by either using:
    - LabelEncoder
    - OneHotEncoder
- **Feature Scaling:**
  - It is done to ensure the features / independent variables are in the same scale so that no variable can dominate the other variable.

Step3: Model Selection:

- Model Selection depends on the type of problem we have at hand.
- **Regression algorithms (How much?)**
  - They should be used if there is a relationship between the input variable and output variable. Eg: House Price Prediction, Salary Prediction
- **Classification algorithms (Which Class?)**
  - They should be used if the output variable is categorical. Eg: yes/no, true/false, play/don't play
- **Clustering algorithms (Which group?)**
  - They should be used to group objects into clusters such that objects with most similarities remain into one group and with less or no similarities into another group.

Step 4: Train & Test Model:

- In Machine Learning we create models to predict the outcome of certain events.
- To measure if the model is good enough, we can use a method called Train/Test.
- Train/Test is a method to measure the accuracy of your model.
- It is called Train/Test because you split the data set into two sets:
  - a training set and
  - a testing set.
- 80% for training, and 20% for testing.
- You *train* the model using the training set.
- You *test* the model using the testing set.
- *Train* the model means *create* the model.
- *Test* the model means test the accuracy of the model.

Step 5: Deploy the Model:

- The final step is to deploy the model in the real world system.
- We do the deployment provided the model that we have prepared is producing accurate result or acceptable results as per our requirement.

	Supervised Learning	Unsupervised Learning
Usage	It is used for predicting the output.	It is used for finding hidden patterns
Data	Input data is provided to the model along with the output.	Only input data is provided to the model.
Categories	It can be categorized into Regression and Classification problems	It can be categorized into Clustering and Association problems
Example	LR, MLR, PR, LoR, NB, DT	Kmeans clustering

	Regression	Classification
Usage	It can be used to solve problems such as House price prediction, Weather prediction	It can be used to solve problems such as Pass/Fail, Cycle/Bike/Car
Data	It can be used with continuous data	It can be used with discrete data
Working	We try to find the best fit line which can predict the output more accurately	We try to find the decision boundary which can divide the dataset into different classes
Types	Regression algorithm can be further divided into Linear and Non-Linear Regression	Classification algorithm can be further divided into Binary classifier and Multi-class classifier