Machine Learning

- "Learning is any processby which a system improves performance from experience." (**Herbert Alexander Simon)**

- "Machine Learning is concernedwith computer programs that through automatically improve their performance experience."

- Machine Learning is concernedwith the development, the analysis, and the application of algorithms that allow computers to learn

- Learning:

  A computer learns if it improves its performance at some task with experience (i.e. by collecting data)

- Extracting a model of a system from the sole observation (or the simulation) of this system in some situations.

- Learning is useful when:

Human expertise does not exist (navigating on Mars),

Humans are unable to explain their expertise (speech recognition)

Solution changes in time (routing on a computer network)

Solution needs to be adapted to particular cases (user biometrics)

What Is Machine Learning?

Machine Learning is the science (and art) of programming computers so they can *learn from data*.

Here is a slightly more general definition:

> [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.
> —Arthur Samuel, *1959*

And a more engineering-oriented one:

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

—Tom Mitchell, *1997*

- Learning = Improving with experience at some task
  - Improve over task *T*,
  - With respect to performance measure, *P*
  - Based on experience, *E*.

Example: Spam Filtering

Spam - is all email the user does not want to receive and has not asked to receive

T: Identify Spam Emails

P: % of spam emails that were filtered
% of ham/ (non-spam) emails that were incorrectly filtered-out

E: a database of emails that were labelled by users

For example,

- Your spam filter is a Machine Learning program that can learn to flag spam given examples of spam emails (e.g., flagged by users) and examples of regular (nonspam, also called “ham”) emails.

- The examples that the system uses to learn are called the *training set*.

- Each training example is called a *training instance* (or *sample*).

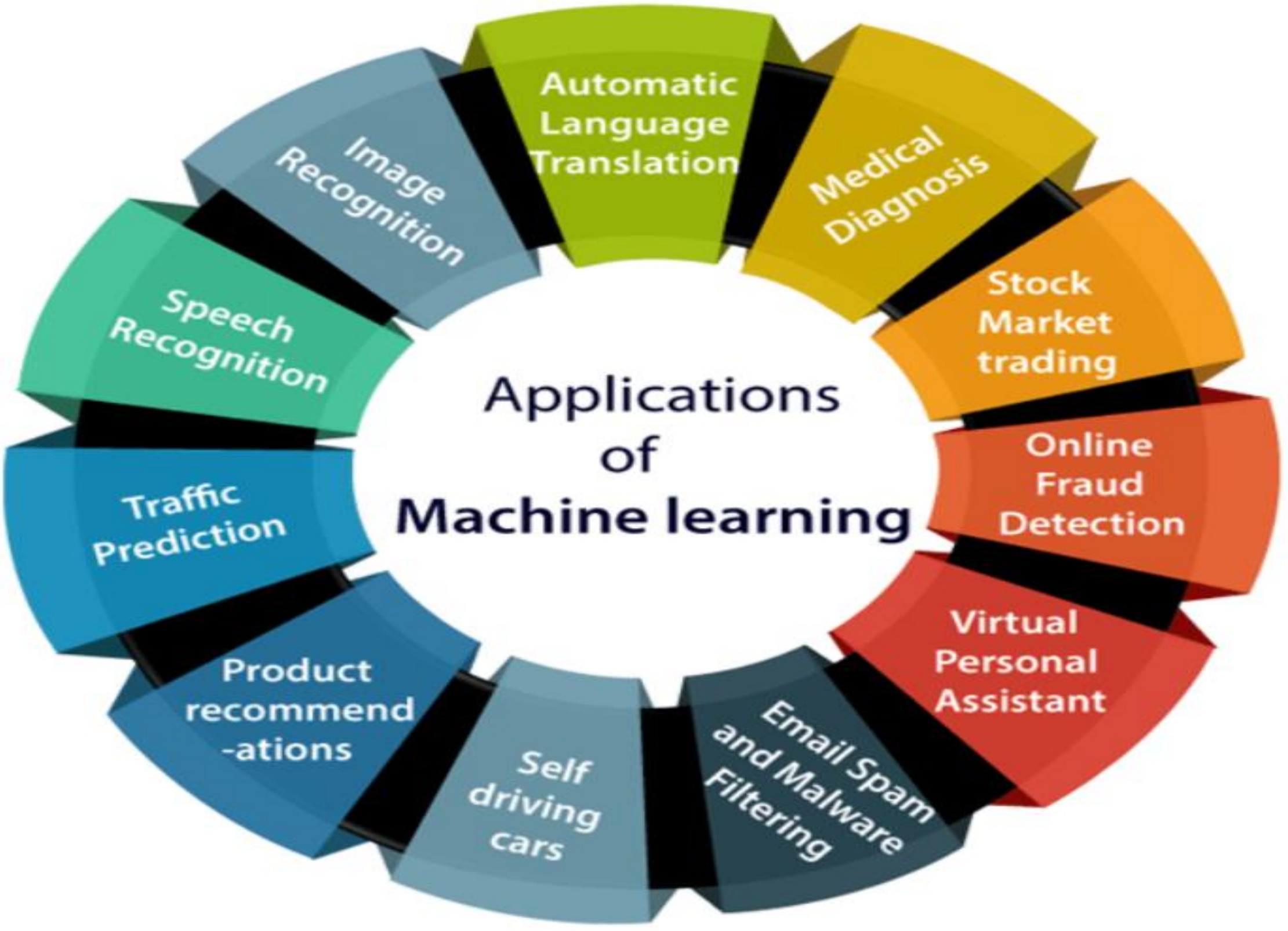
- In this case, the task T is to flag spam for new emails, the experience E is the *training data*, and the performance measure P needs to be defined; for example, you can use the ratio of correctly classified emails.

- This particular performance measure is called *accuracy* and it is often used in classification tasks.

Let's consider another example: predicting house prices.
Imagine you're a real estate agent and you want to help your clients estimate the selling price of their house. Instead of relying solely on your intuition or general market trends, you can leverage machine learning to build a predictive model.

The size of the house (in square feet), number of bedrooms and bathrooms, location (e.g., neighborhood or ZIP code), proximity to amenities like schools and parks, and other relevant factors.

Applications of Machine learning



Types of Machine Learning Algorithm

- **Supervised learning:-** labels are present for all the observations

- **Unsupervised Learning-** label is not present for any observations

- **Semi-supervised Learning:-** Semi-supervised learning falls in between above two

- **Reinforcement Learning:-** aims at using observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk. Reinforcement learning algorithm (called the agent) continuously learns from the environment in an iterative fashion.

Supervised learning

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output.

The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly.

It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model.

The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

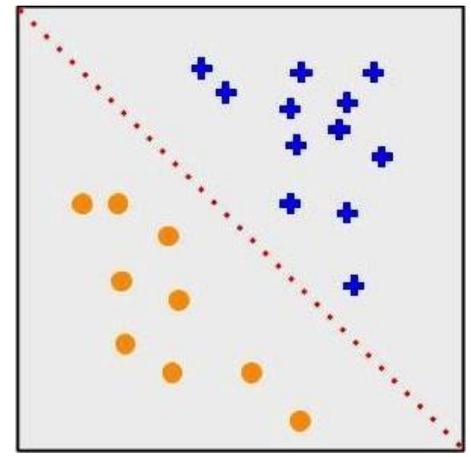
In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, medical diagnosis etc.

Supervised ML Problems

Supervised

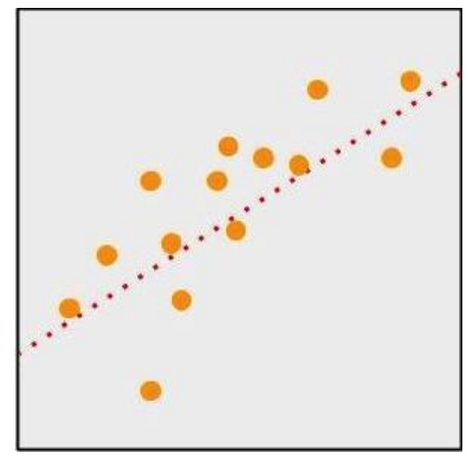
Classification

Output is a discrete variable (e.g., cat/dog)



Regression

Output is continuous (e.g., price, temperature)



Here are some of the most important supervised learning algorithms:

- Regression Techniques
- Linear Regression
- Logistic Regress ion

- Classification Techniques
- k-Nearest Neighbors
- Support Vector Machines (SVMs)
- Naive Bayes
- Decision Trees
- Neural networks[2]

CLASSIFICATION

- Given: A collection of records *(training set)*, where each record contains a set of *attributes*, and a *class*.
- Find: A *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

ML Classification Techniques

- Given: A collection of records *(training set)*, where each record contains a set of *attributes*, and a *class*.

- Find: A *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
- A *test set* is used to determine the accuracy of the model.

- Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

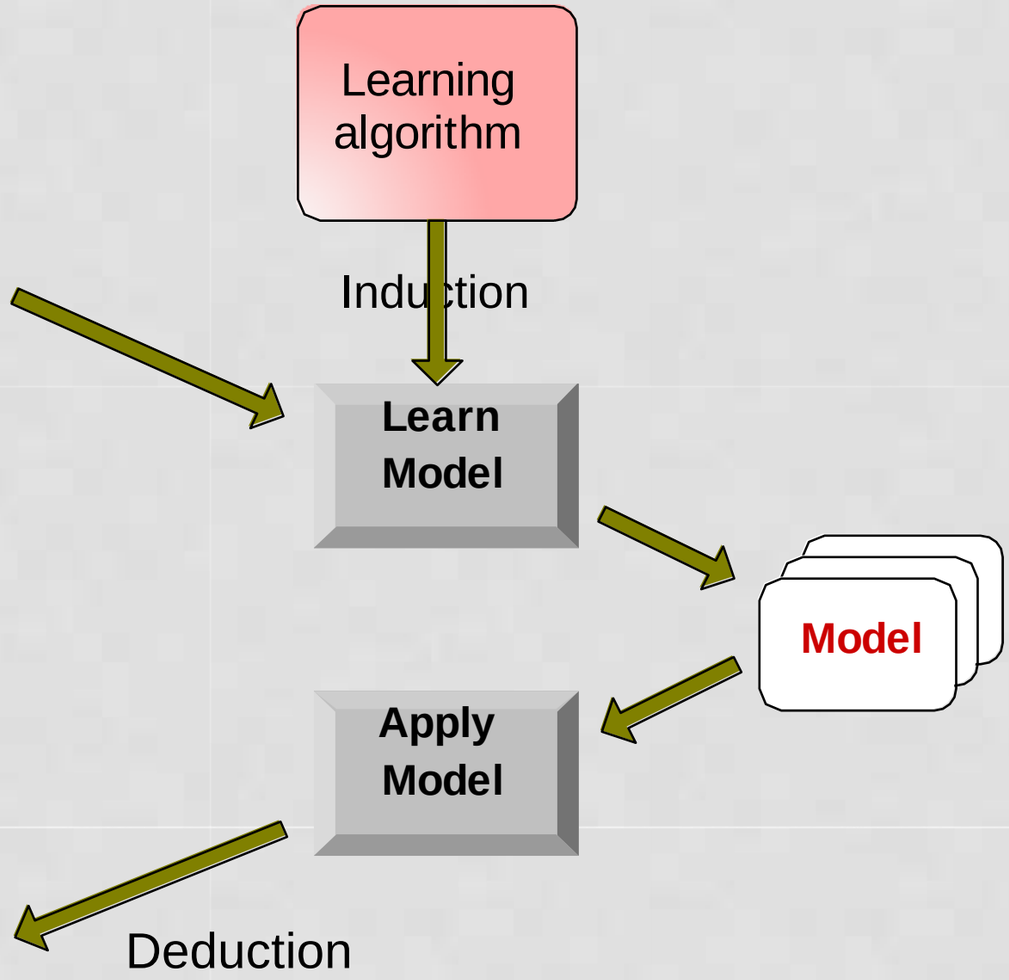
ILLUSTRATING CLASSIFICATION TASK

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



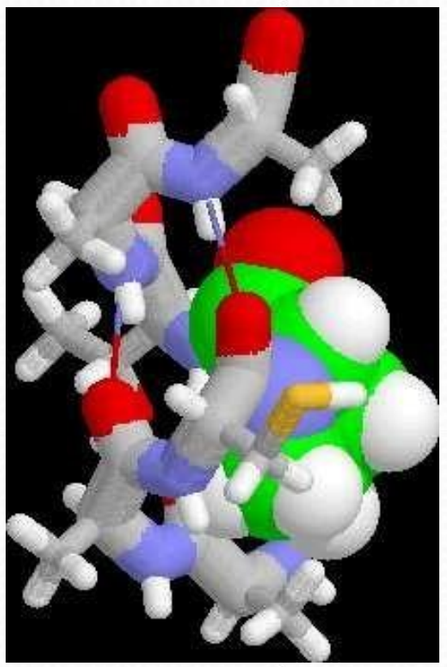
EXAMPLES OF CLASSIFICATION TASK

- Predict tax returns as “clean” or “need an audit”

- Predicting tumor cells as benign or malignant

- Classifying credit card transactions as legitimate or fraudulent

- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



- Categorizing news stories as finance, weather, entertainment, sports, etc

Semi-supervised learning is a machine learning paradigm where the dataset contains both labeled and unlabeled data. This approach is particularly useful when labeled data is scarce or expensive to obtain, but there is an abundance of unlabeled data available.

Example of Semi-supervised Learning

Imagine you're working on a sentiment analysis task where you want to determine the sentiment (positive, negative, or neutral) of customer reviews for a product.

However, you only have a small dataset of labeled reviews, but there's a vast amount of unlabeled reviews available online.

Reinforcement learning (RL) is a type of machine learning paradigm where an agent learns to make decisions by interacting with an environment in order to maximize some notion of cumulative reward.

Unlike supervised learning, where the model is trained on labeled data, and unsupervised learning, where the model learns patterns from unlabeled data, reinforcement learning focuses on learning through trial and error.