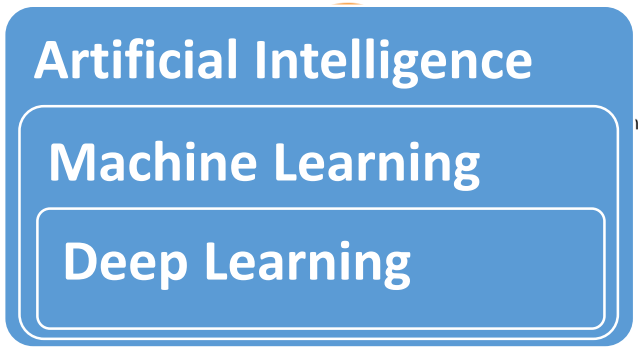
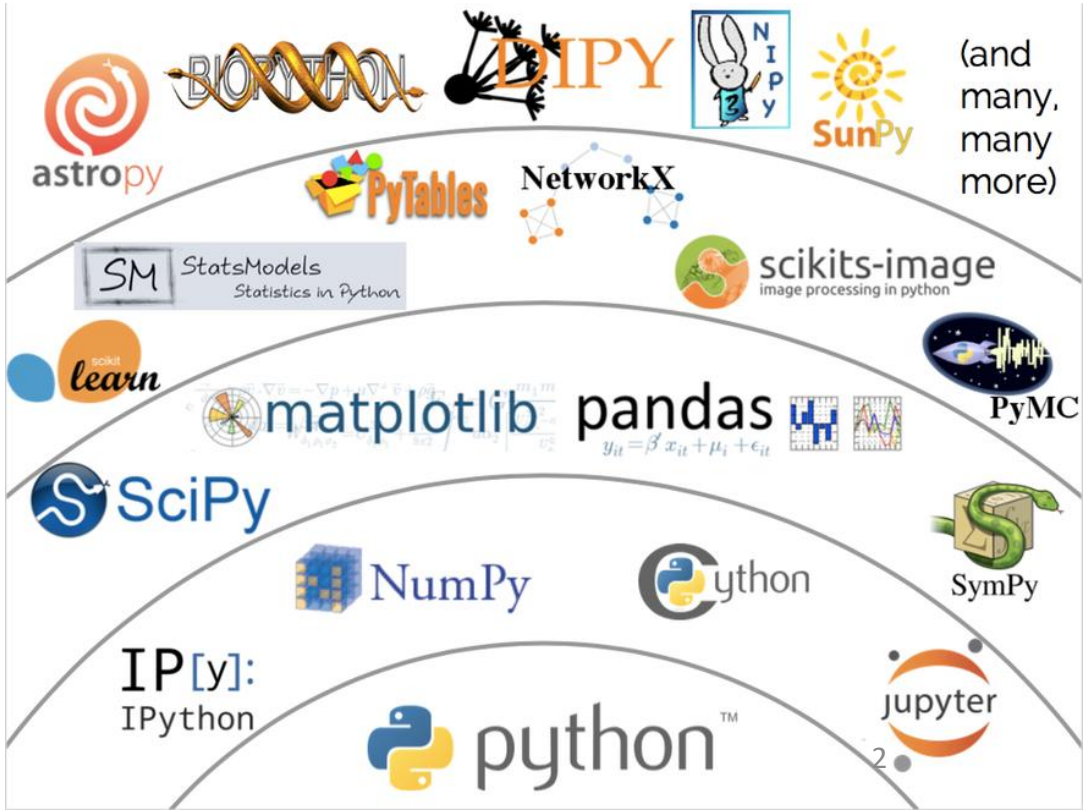
Python-Machine Learning

using Scikit-Learn package

Agenda

- Introduction
- Why Machine Learning is Needed
- Type of ML-Supervised vs unsupervised
- Classification, Regression, Clustering
- Cheat sheet
- Machine learning flow





Introduction

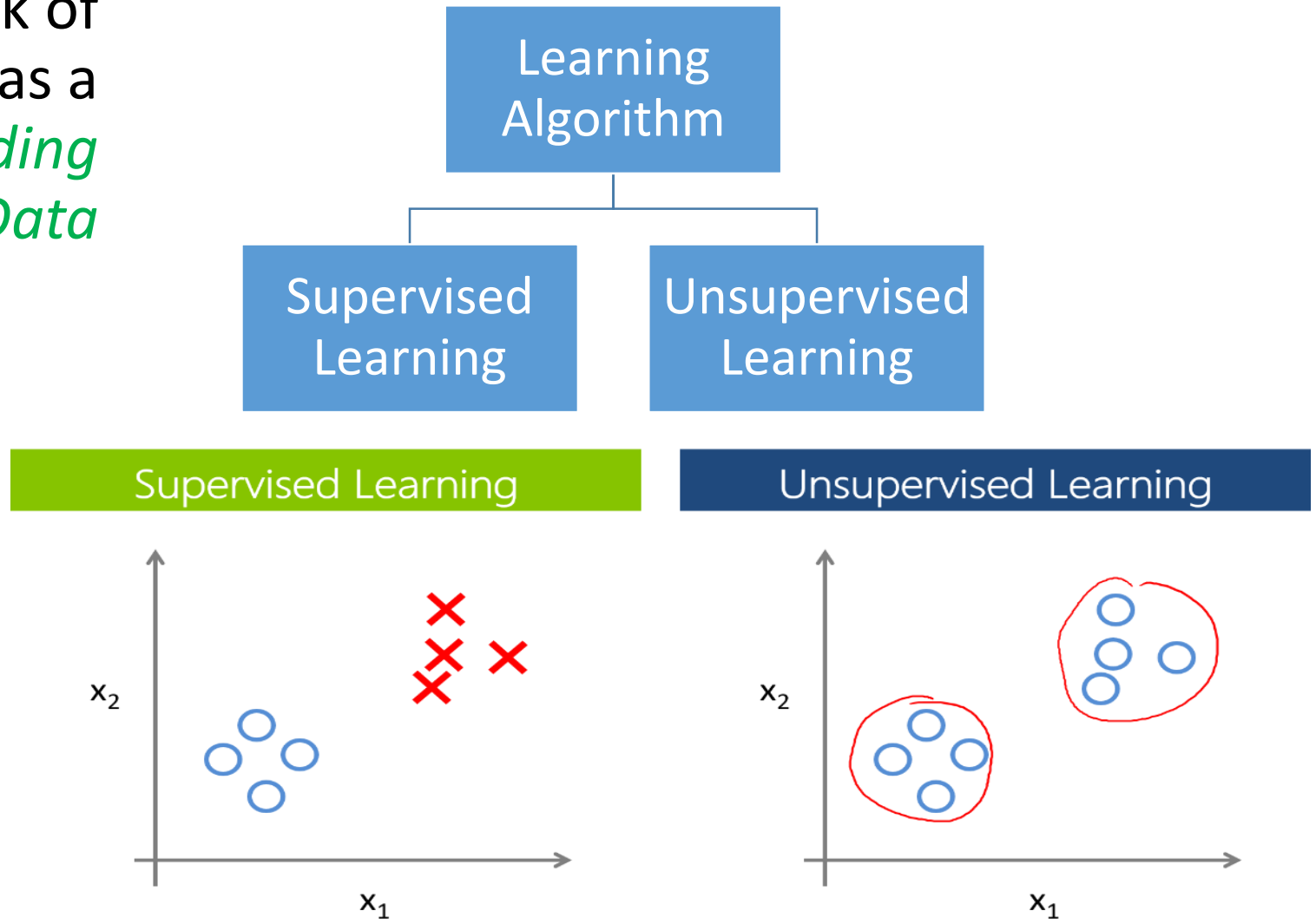
- Machine learning is where computational and algorithmic skills of data science meet the statistical thinking of data science,
- The result is a collection of approaches to inference and data exploration that are *not about effective theory* so much as *effective computation*.
- Better to think of machine learning as a *means of building models of Data*
- Machine learning along with entire Data Science ecosystem is trying to make this mathematical, model-based "learning" as same as "learning" exhibited by the human brain.

Why Machine Learning is needed

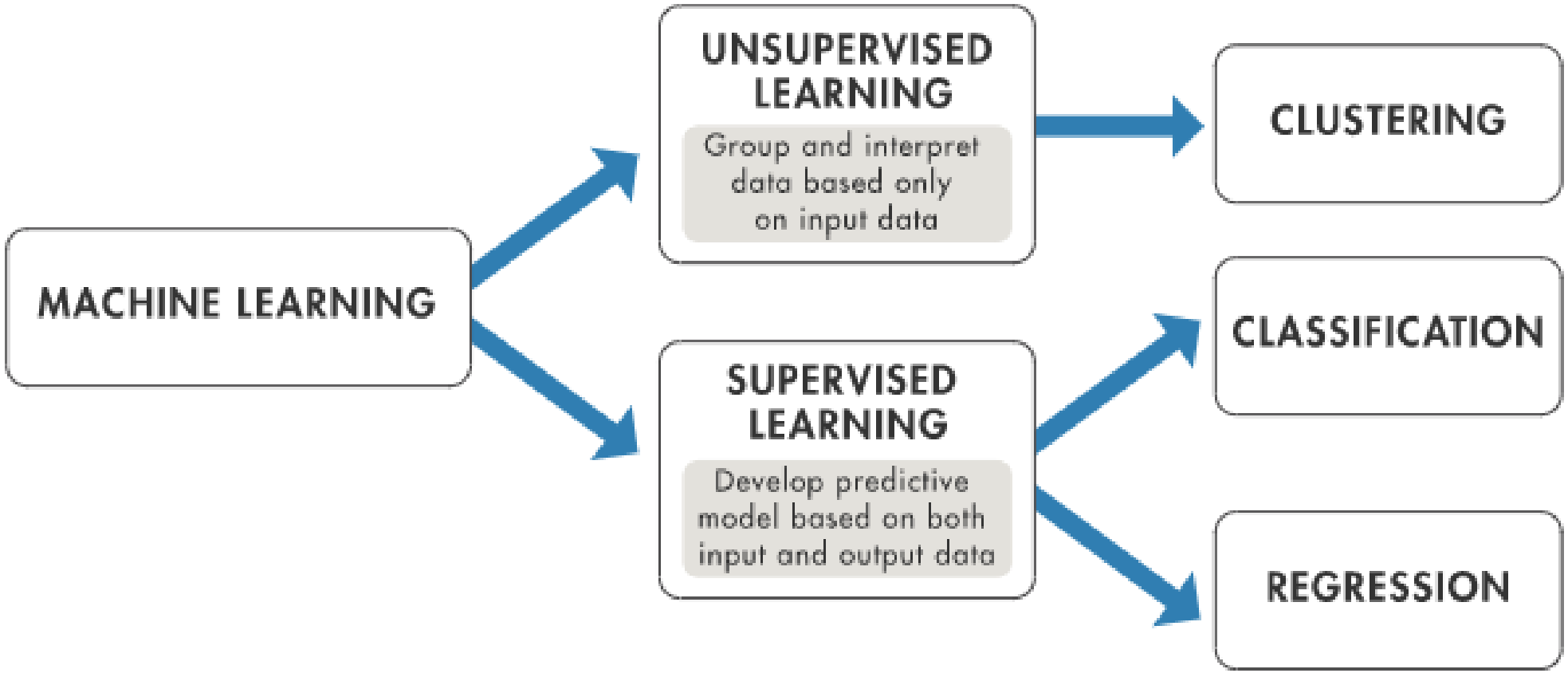
- If Programmer start making use cases / rules for complex system, then it will result in a large number of rules and exceptions.
- Machine Learning is needed in cases where humans cannot directly write a program to handle each and every case.
- So it's better to have a machine (~~rather than human~~) that learns from a large training set.

Major Classes of Learning Algorithms

Better to think of machine learning as a *means of building models of Data*



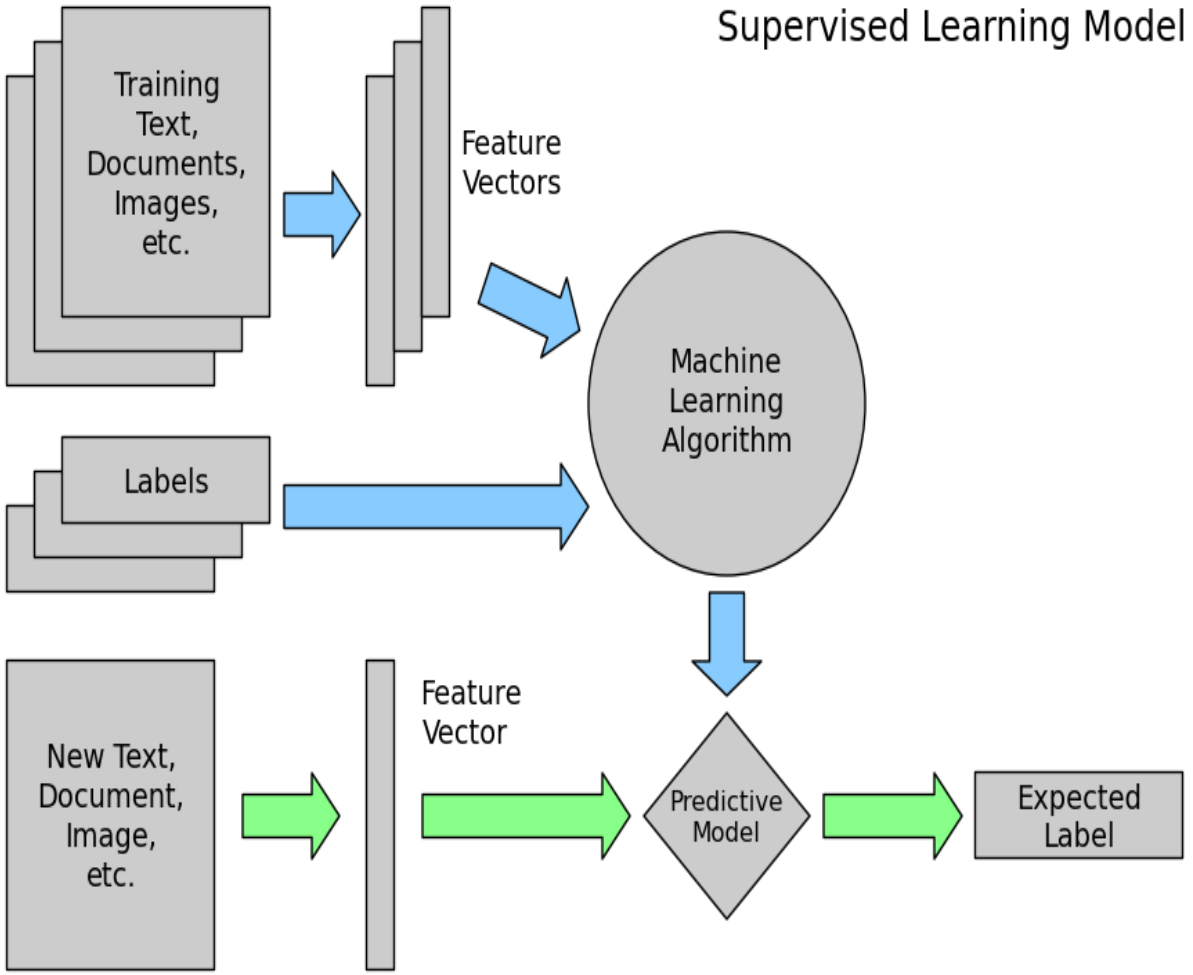
Categories of Machine Learning



Source : wikipedia

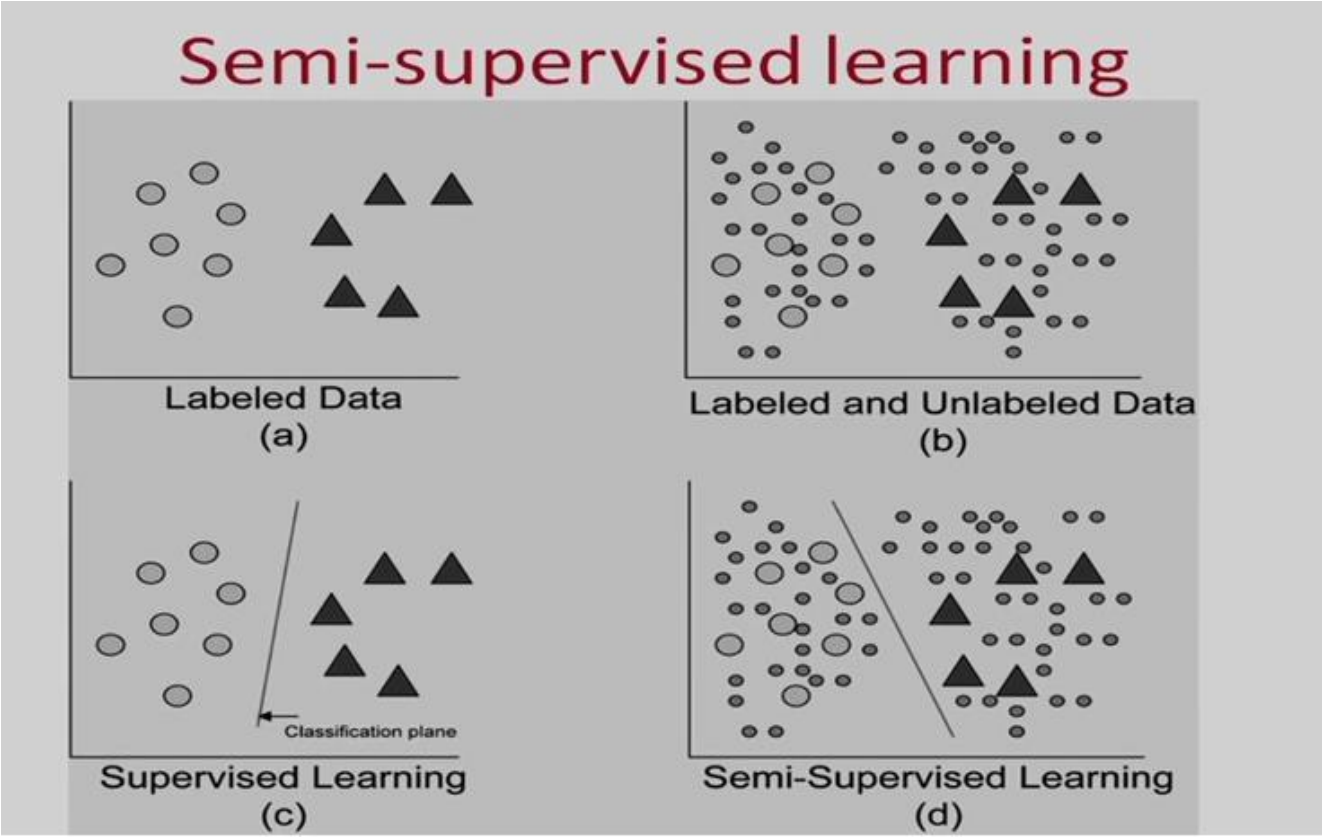
Supervised Learning

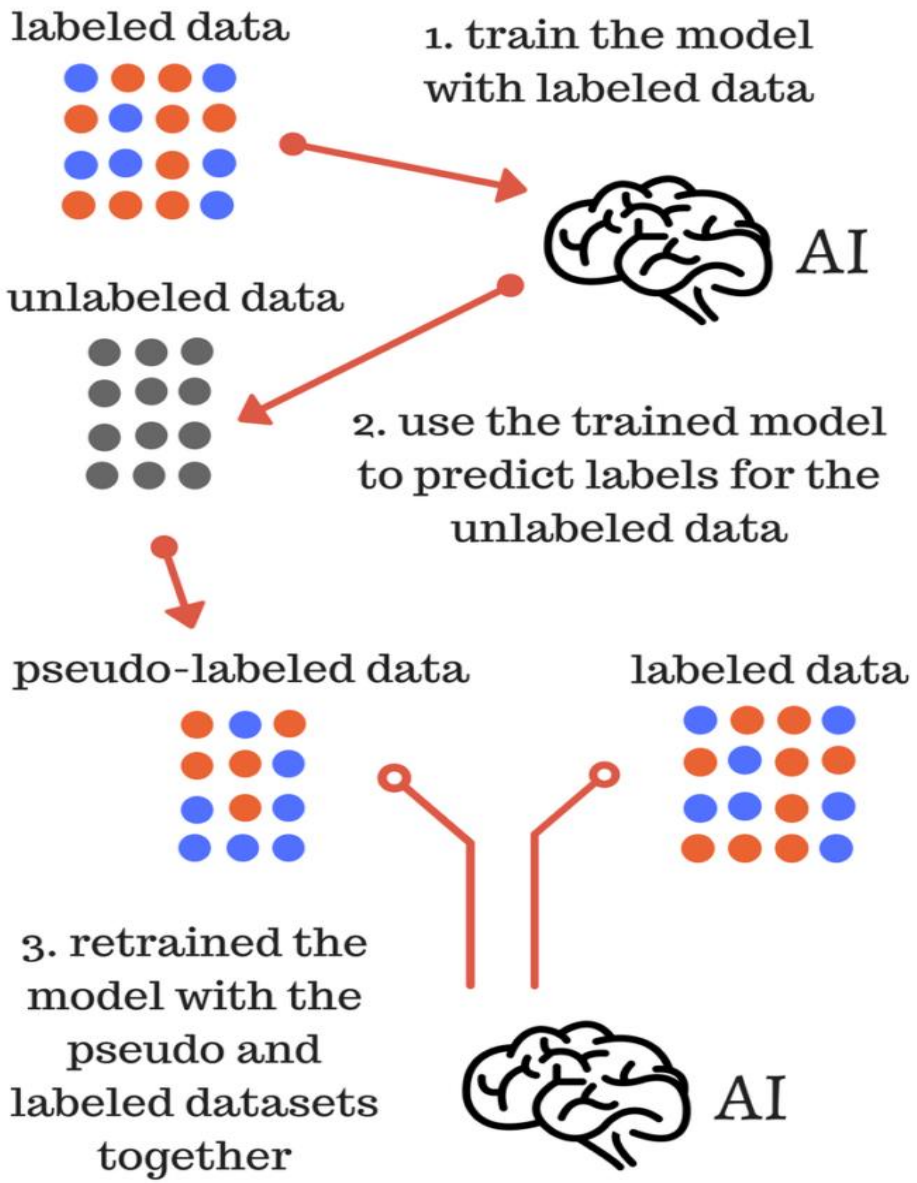
- The computer is presented with example inputs and their desired outputs and the goal is to learn a general rule that maps inputs to outputs.
- The training process continues until the model achieves a desired level of accuracy on the training data.
- Once this model is determined, it can be used to apply labels to new, unknown data.



Semi-Supervised Learning

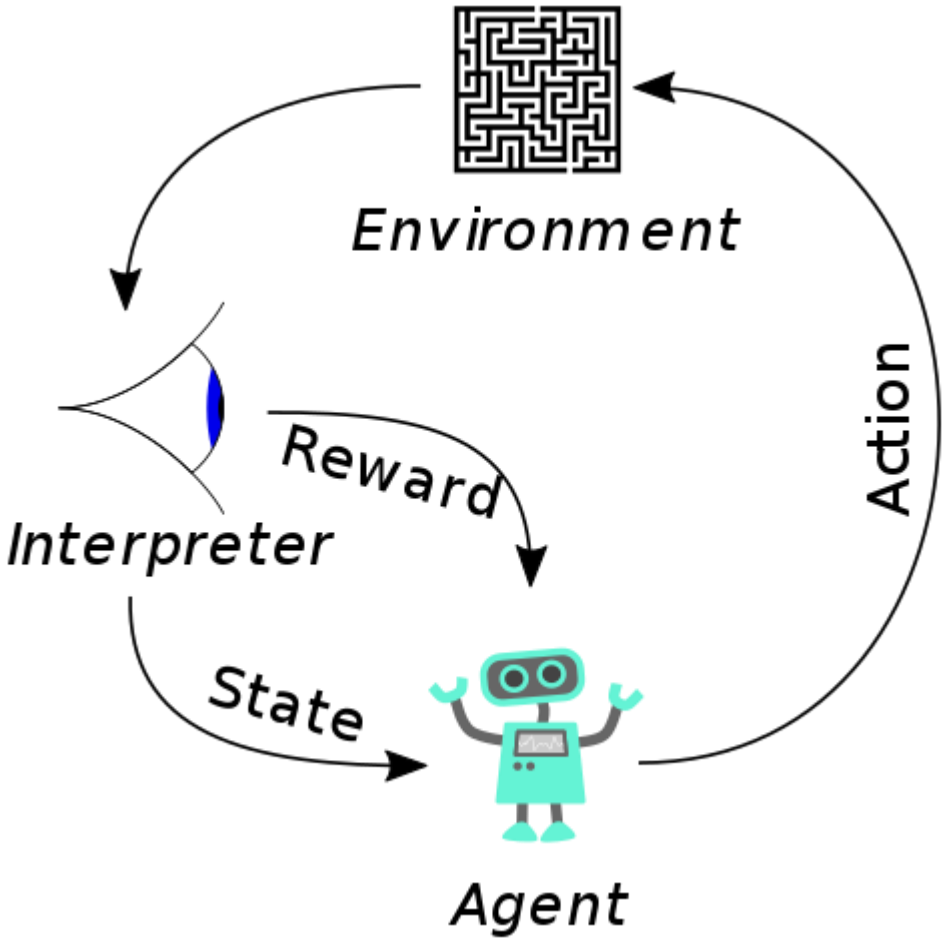
The computer is given only an incomplete training signal. The goal of a semi-supervised model is to classify some of the unlabeled data using the labeled information set.



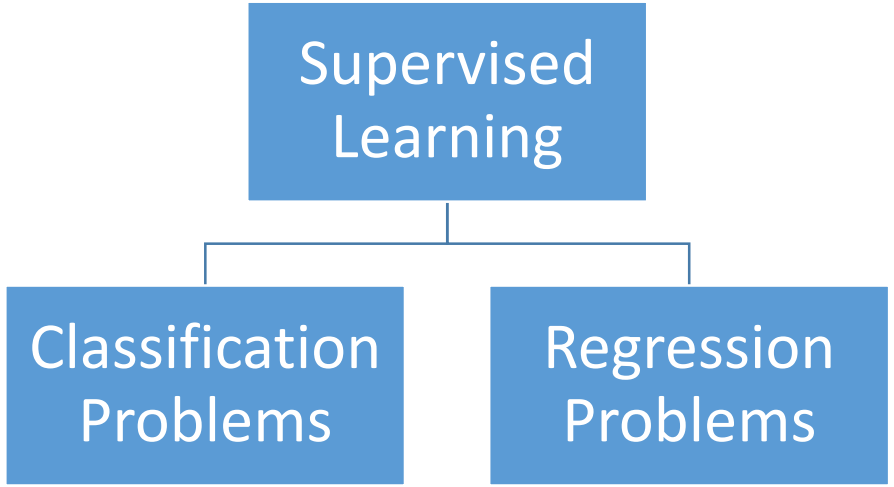


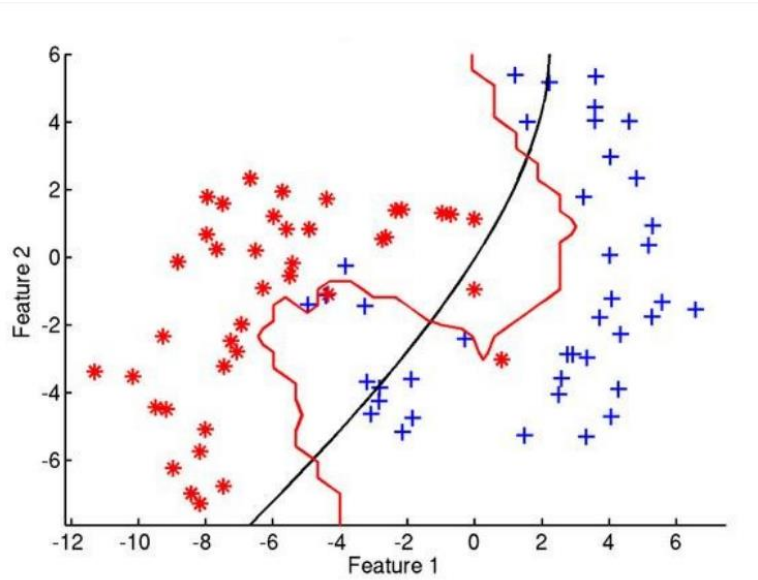
Reinforcement learning

- Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize some notion of cumulative reward.
- In the absence of training dataset, it is bound to learn from its experience. Eg a robot could be the agent and the goal for it would be to find the best way to move from one place in the house to the other without hitting into obstacles. So it is important to define a score, hit an obstacle and get a negative score (punishment), avoid an obstacle and get a positive score (reward).
- Eg The dog is the agent, the living room the environment, you are the source of the reward signal (tasty snacks).



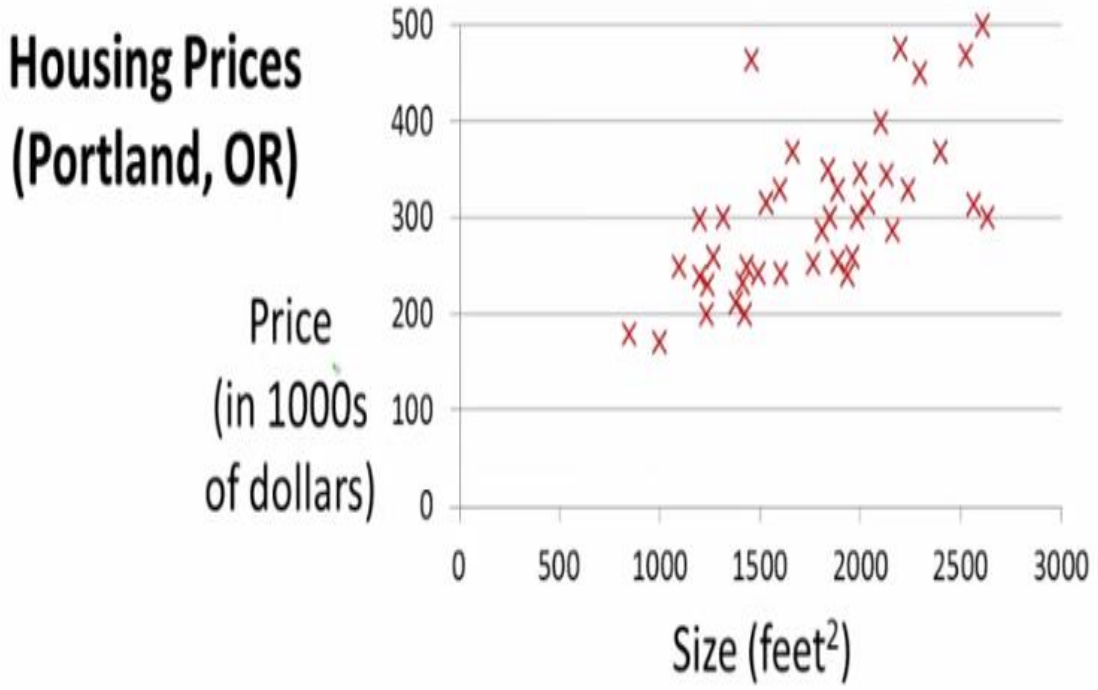
Supervised Learning: Classification Problems





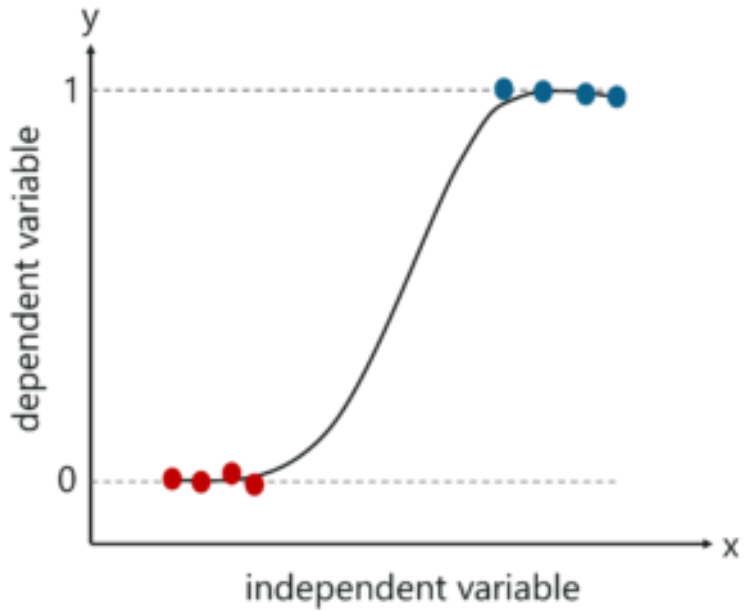
Supervised Learning: Regression Problems

- Regression is the kind of Supervised Learning that learns from the Labelled Datasets and is then able to **predict a continuous-valued output** for the new data given to the algorithm.
- It is used whenever the output required is a number such as money or height etc.



Logistic Regression

- This algorithm predicts discrete values for the set of Independent variables that have been passed to it. It does the prediction by mapping the unseen data to the **logit function** that has been programmed into it. The algorithm predicts the probability of the new data and so it's output lies between the range of 0 and 1.



Unsupervised learning

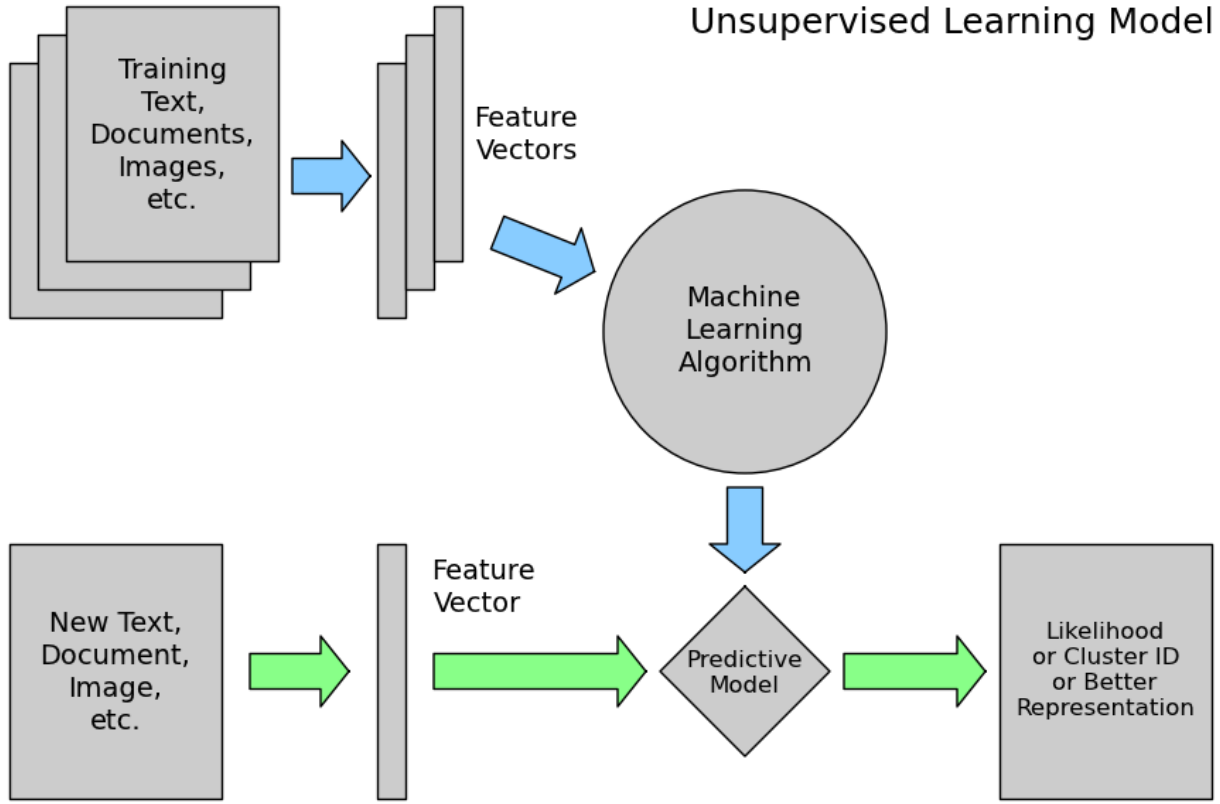
“ letting the dataset speak for itself ”

- No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or

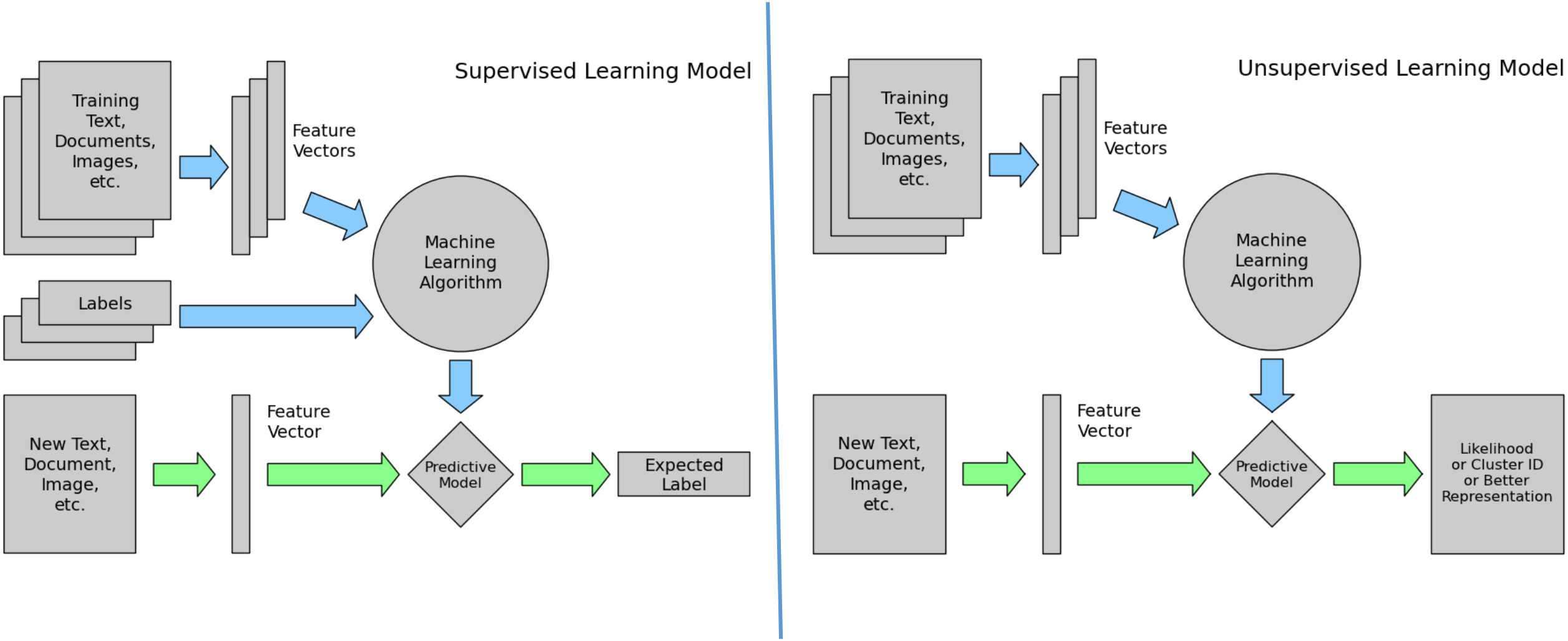
  a means towards an end (feature learning).
- It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.
- These models include tasks such as clustering and dimensionality reduction.
  - Clustering algorithms identify distinct groups of data, while
  - Dimensionality reduction algorithms search for more succinct representations of the data.

Unsupervised learning - Clustering

- The aim of unsupervised learning is to find clusters of similar inputs in the data without being explicitly told that some datapoints belong to one class and the other in other classes.
- The algorithm has to discover this similarity by itself



Supervised vs Unsupervised learning



Supervised vs Unsupervised learning

Example 1:

•You get a bunch of photos **with information about what is on them**.

•Then you train a model to recognize new photos.

Example 2:

- You have a bunch of molecules and information about which are drugs.

- Then you train a model to answer whether a new molecule is also a drug.

Example 1:

•You have a bunch of photos of 6 people

•**No information about who is on which one.**

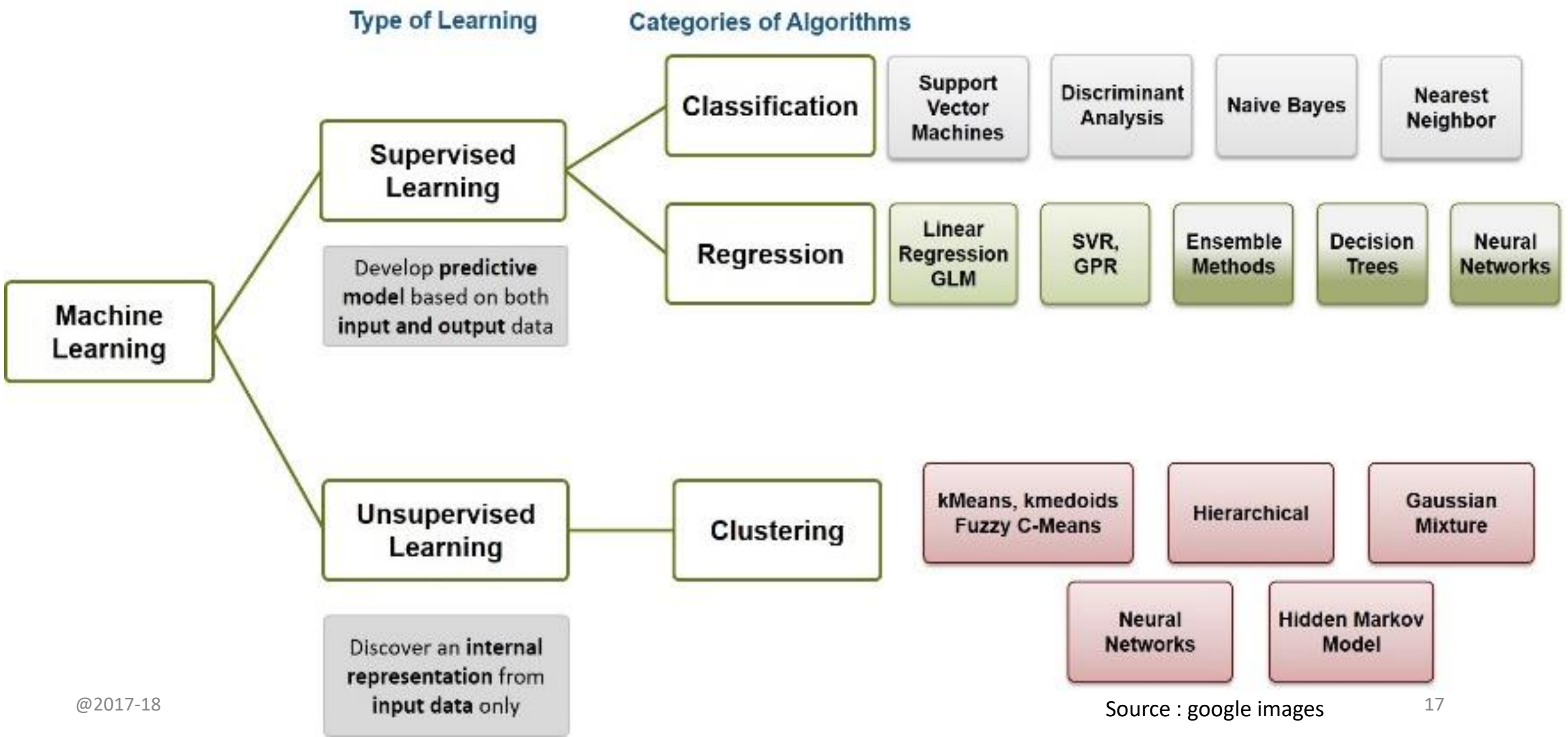
•You want to **divide** this dataset into 6 piles, each with the photos of one individual.

Example 2:

- You have molecules, part of them are drugs and part are not.
- But you do not know which are which.

- You want the algorithm to discover the drugs.

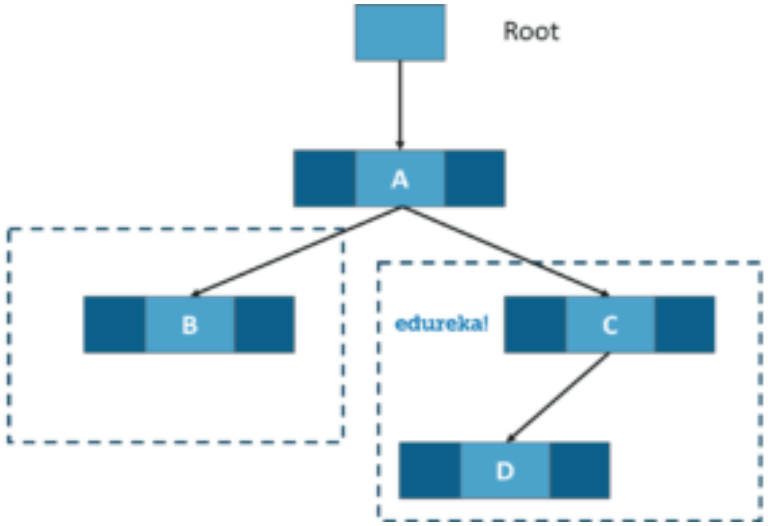
Category of Algorithms



Source : google images

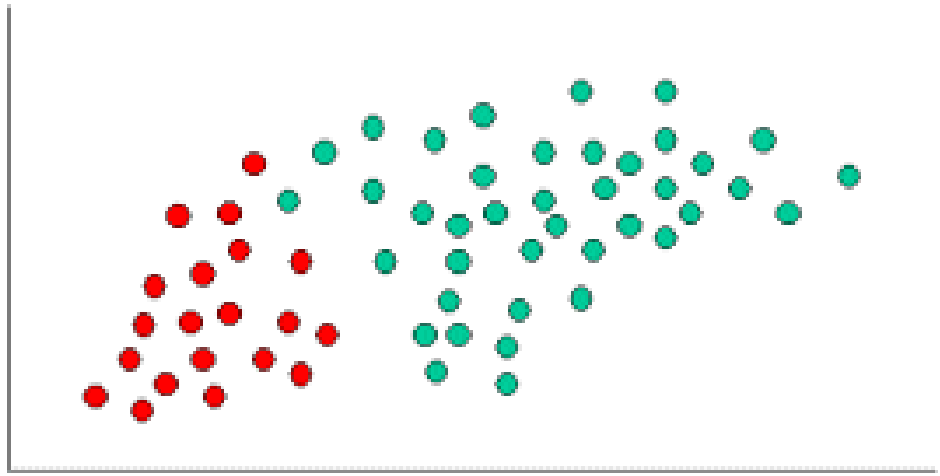
Decision Tree

- Decision Trees classify based on the feature values. They use the method of ***Information Gain*** and find out which feature of the dataset gives the best of information, make that as the root node and so on till they are able to classify each instance of the dataset. Every branch in the Decision Tree represents a feature of the dataset. They are one of the most widely used algorithms.



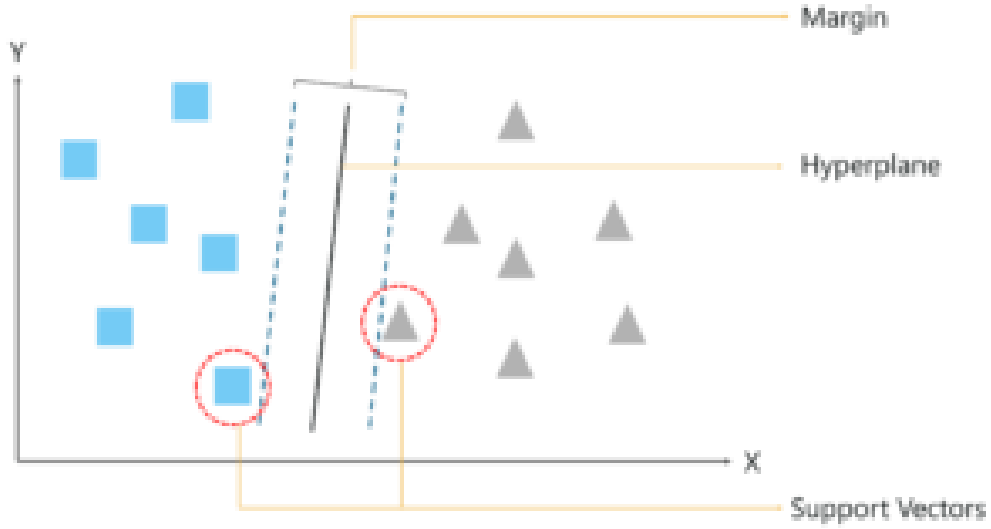
Naive Bayes Classifier

- Naive Bayes algorithms assume that the features of the dataset are all independent of each other. They work great on large datasets. Directed Acyclic Graphs (DAG) is used for the purpose of classification.



Support Vector Machines (SVM)

- SVM algorithms are based on the statistical learning theory of Vap Nik. They use Kernal functions which are a central concept for most of the learning tasks. These algorithms create a hyper-plane that is used to classify the two classes from each other.



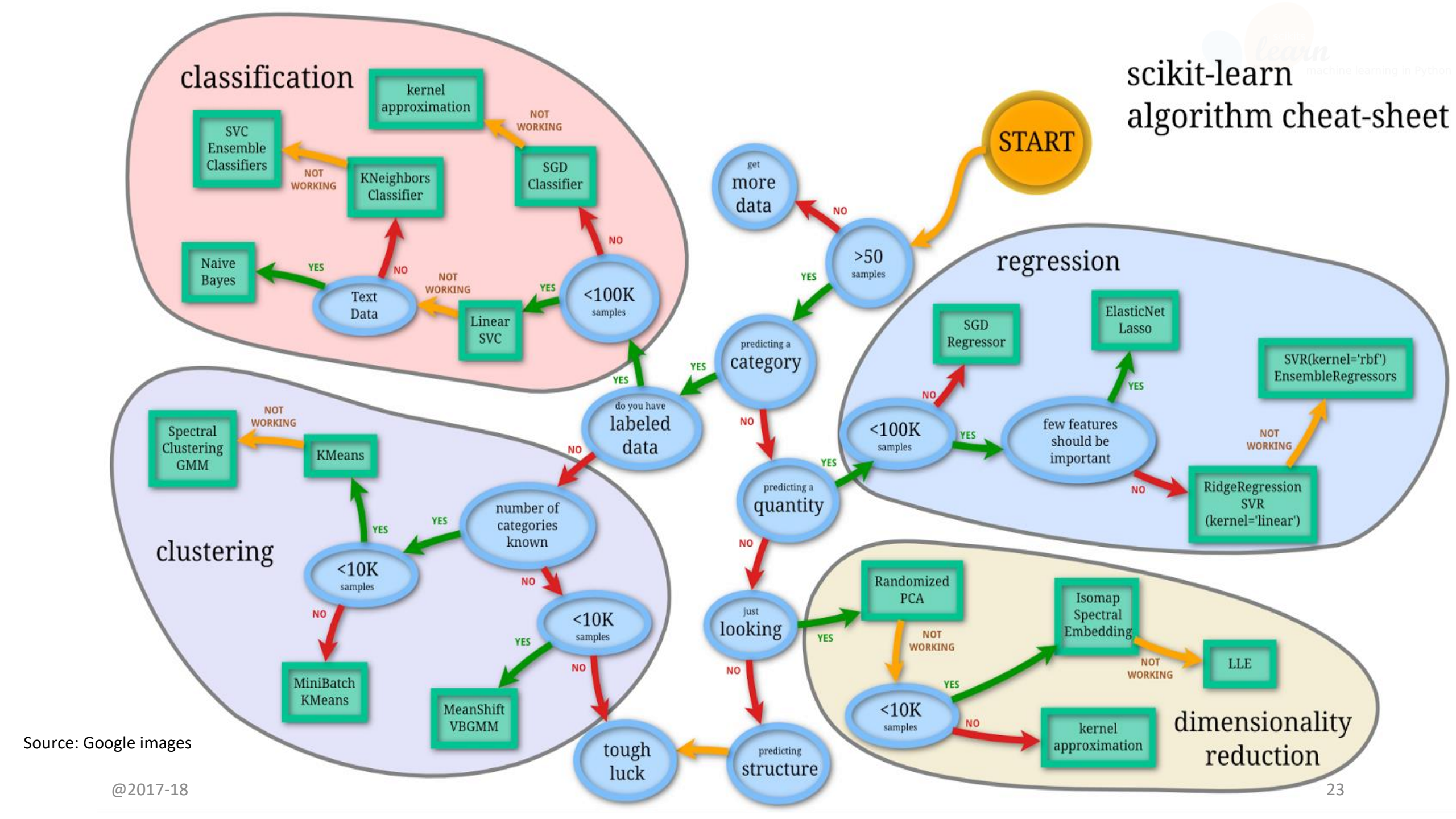
Challenges of Supervised Learning

- You could overfit your algorithm easily
- Good examples need to be used to train the data
- Computation time is very large for Supervised Learning
- Unwanted data could reduce the accuracy
- Pre-Processing of data is always a challenge
- If the dataset is incorrect, you make your algorithm learn incorrectly which can bring losses

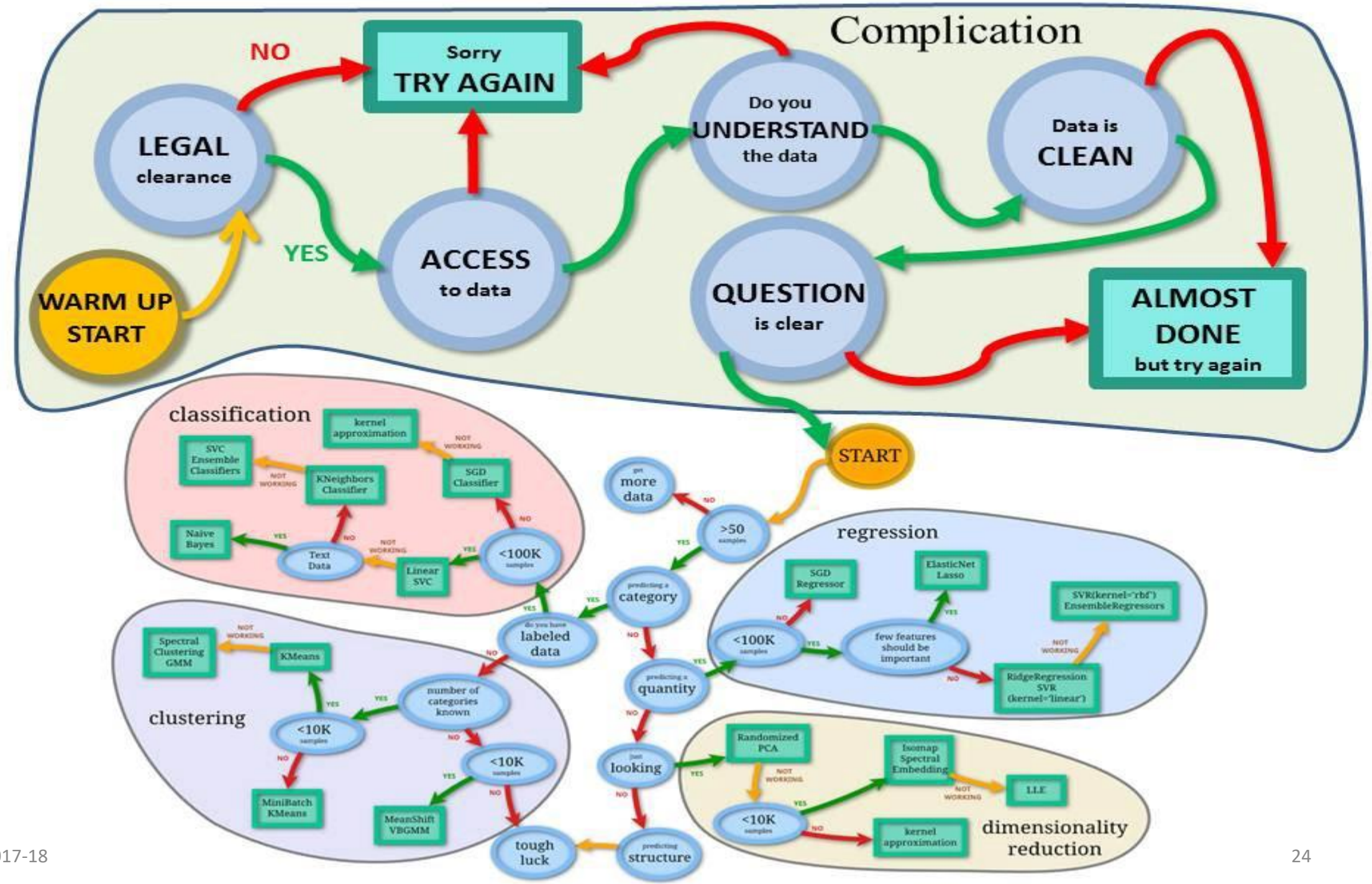
Scikit-learn

- Scikit-learn is probably the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

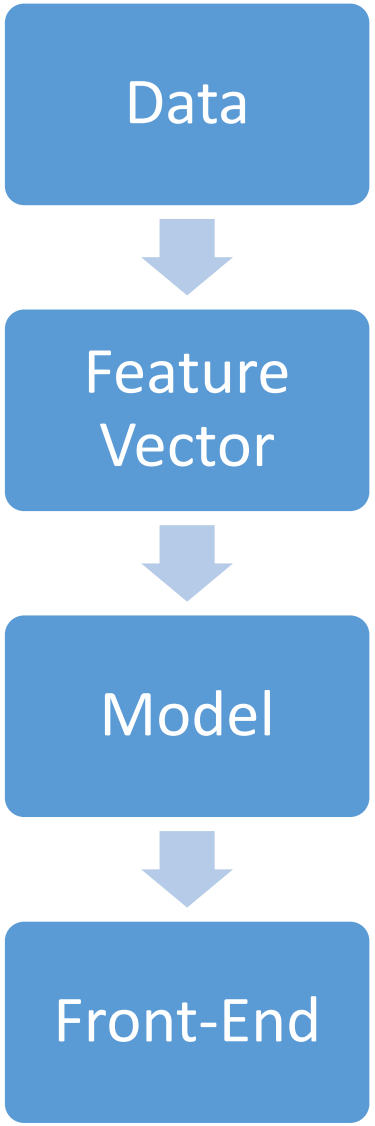
- Note that *scikit-learn is used to build models*. It should not be used for reading the data, manipulating and summarizing it. There are better libraries for that (e.g. NumPy, Pandas etc.)



Source: Google images


Complication
NO
Sorry
TRY AGAIN
LEGAL
clearance
YES
ACCESS
to data
WARM UP
START
Do you
UNDERSTAND
the data
Data is
CLEAN
QUESTION
is clear
ALMOST
DONE
but try again
START
classification
kernel approximation
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K
NO
more data
NO
>50
YES
predicting a
category
YES
NO
do you have
labeled data
NO
YES
predicting a
quantity
NO
[just]
looking
NO
predicting
structure
tough luck
regression
SGD Regressor
NO
ElasticNet Lasso
<100K
YES
few features should be important
YES
NO
SVR(kernel='rbf') EnsembleRegressors
NOT WORKING
RidgeRegression SVR (kernel='linear')
clustering
Spectral Clustering GMM
NOT WORKING
KMeans
number of categories known
YES
<10K
NO
YES
NO
<10K
YES
NO
MiniBatch KMeans
MeanShift VBGMM
dimensionality reduction
Randomized PCA
YES
NOT WORKING
Isomap Spectral Embedding
NOT WORKING
LLE
<10K
YES
NO
kernel approximation

Machine Learning Workflow



Featurize

Train

Deploy

Examples of Supervised Learning

Visual Recognition

- An AI that is learning to identify pedestrians on a street is trained with 2 million short videos of street scenes from self-driving cars. Some of the videos contain no pedestrians at all while others have up to 25. A variety of learning algorithms are trained on the data with each having access to the correct answers. Each algorithm develops a variety of models to identify pedestrians in fast moving scenes. The algorithms are then tested against another set of data to evaluate accuracy and precision.

Examples of Supervised Learning

- A robot is learning to sort garbage using visual identification. It sits all day picking out recyclable items from garbage as it passes on a conveyor belt.
- It places items such as glass, plastic and metal into 12 bins.
- Each item is labeled with an identification number on a sticker.
- Once a day, human experts examine the bins and inform the robot which items were improperly sorted.
- The robot uses this feedback to improve.

Examples of Supervised Learning

Decision Support

- An AI is learning to estimate investing risk. It is fed a large number of trades that real investors made and asked to estimate a risk/reward ratio for each trade based on company fundamentals, price and other factors such as volume.
- The estimated risk/reward ratio is then compared to the historical results of the trade at a variety of time intervals such as a day or a year.

Examples of Unsupervised Learning

Visual Recognition

- An unsupervised learner processes 10 million videos together with related textual data such as descriptions and comments.
- The learner models images in the videos using statistical analysis that allows it to identify visual patterns. These patterns can then be correlated with text to develop theories about the visual traits of various things.
- For example, such a learner might be able to build a solid model that can identify skateboards in videos. The learner is never given the right answer but can gain confidence based on a large number of samples. Likewise, the learner will discard a large number of models that don't appear to be correct.

Examples of Unsupervised Learning

- Human Behavior
- A learner that possesses visual highly developed visual and speech recognition capabilities could watch a large number of television shows to learn about human behavior. For example, a learner might be able to build a model that detects when people are smiling based on correlation of facial patterns and words such as "what are you smiling about?"

Examples of Unsupervised Learning

- Robotics
- A highly developed AI that serves as a housekeeping robot develops a theory that there is usually dust under a sofa. Each week, the theory is confirmed as the robot often finds dust under sofas. Nobody explicitly tells the robot the theory is correct but it is able to develop confidence in it nonetheless.

Example of Deep Learning

Speech Recognition

- An AI learns to tell the difference between languages. It decides a person is speaking English and invokes an AI that is learning to tell the difference between different regional accents of English. The AI decides the person is speaking Cardiff English and invokes an AI that is learning to speak Cardiff English. In this way, each conversation can be interpreted by a highly specialized AI that has learned their dialect.

Self-Driving Car

- The street in front of a moving vehicle is interpreted by a large number of specialized AI. For example, one learner is only training to recognize pedestrians, another is learning to recognize street signs.
- There might be hundreds of such specialized visual recognition AI that all feed their opinions into an AI that interprets driving events.
- In theory, a single car could use the opinions of thousands or even millions of individual AI as it navigates a street.

Robotics

- A housekeeping robot might use the opinions of a large number of AI in order to complete everyday tasks.
- For example, the robot might have a few AI devoted to dog psychology that help it deal with the household pet over the course of its day.