# Machine Learning (Introduction)

Arthur Samuel described machine learning as: "the field of study that gives computers the ability to learn without being explicitly programmed. There are several types of machine learning based on the problem and dataset we have.

## Supervised Machine Learning

Given dataset already has correct output. The idea is to find the relationship between input and output. Supervised machine learning is categorized into regression and classification problems. In regression problem, we try to predict result within a continuous output. On the other hand, classification predicts result in a discrete output. These are the mainly used algorithm in supervised learning:

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression Support Vector Machines (SVMs)
- Decision Trees
- Random Forests
- Neural networks

These are some example of use case for regression and classification problems:

- Regression - Given a picture of a person, we have to predict their age on the basis of the given picture. Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.
- Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.

## Unsupervised Learning

When we try to predict a model for unlabeled data, it's called unsupervised learning. In these problems, we look for structure in dataset based on relationship among the variables. The mostly common algorithm used unsupervised learning are:

- Clustering (k-Means, Hierarchical Cluster Analysis (HCA), Expectation Maximization)
- Visualization and dimensionality reduction (Principal Component Analysis (PCA), Kernel PCA, Locally-Linear Embedding (LLE), t-distributed Stochastic Neighbor Embedding (t-SNE))
- Association rule learning (Apriori, Eclat)

These are few examples for this type of learning:

- Clustering: Take a collection of 1,000,000 different genes, and find a way to automatically group these genes into groups that are somehow similar or related by different variables, such as lifespan, location, roles, and so on.
- Non-clustering: The "Cocktail Party Algorithm", allows you to find structure in a chaotic environment. (i.e. identifying individual voices and music from a mesh of sounds at a cocktail party).

# Machine Learning (Introduction)

## Semi-Supervised Learning

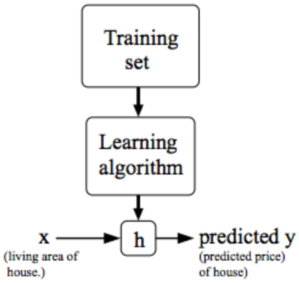
Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data. This is called semi-supervised learning. Most semi-supervised learning algorithms are combinations of unsupervised and supervised algorithms. For example, deep belief networks (DBNs) are based on unsupervised components called restricted Boltzmann machines stacked on top of one another. RBMs are trained sequentially in an unsupervised manner, and then the whole system is fine-tuned using supervised learning techniques.

## Reinforcement Learning

Reinforcement Learning is a very different beast. The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return. It must then learn by itself what is the best strategy, called a policy, to get the most reward over time. A policy defines what action the agent should choose when it is in a given situation.

## Model Representation

Imagine we have supervised learning with training set of $x^i$, by using learning algorithm, we learn a model for giving structure to dataset and predict output variable $y^i$. The learning algorithm is called hypothesis.



## Cost Function

The accuracy of hypothesis is measured by using a cost function. It is calculated by taking average difference between predicted values and actual values of output y:

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

This calculated the mean squared error, we choose θ in a way to minimize the cost function for the model.