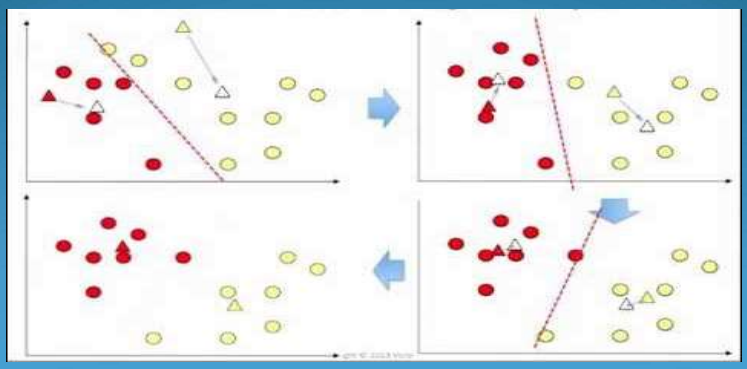
K-MEANS CLUSTERING



What is clustering?

- **Clustering** is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

Types of clustering:

1.Hierarchical algorithms:

1. Agglomerative ("bottom-up"):
2. Divisive ("top-down"):

2. Partitional clustering: Partitional algorithms determine all clusters at once. They include:

- ***K*-means and derivatives**
- Fuzzy *c*-means clustering
- QT clustering algorithm

Common Distance measures:

1. The Euclidean distance (also called 2-norm distance) is given by:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

2. The Manhattan distance (also called taxicab norm or 1-norm) is given by:

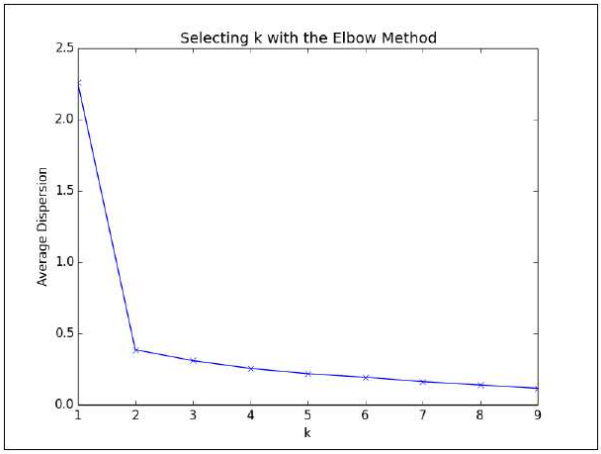
$$\sum_{i=1}^{k}|x_i - y_i|$$

K-MEANS CLUSTERING

- The **k-means algorithm** is an algorithm to cluster *n* objects based on attributes into *k* partitions, where $k < n$.
- It assumes that the object attributes form a vector space.

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$$

The elbow method



Applications of K-Mean Clustering

- It is relatively *efficient and fast.* It computes result at **O(tkn)**, where n is number of objects or points, k is number of clusters and t is number of iterations.
- k-means clustering can be applied to *machine learning or data mining*
- *Used on acoustic data in speech understanding to convert waveforms into one of k categories (known as Vector Quantization or Image Segmentation).*
- *Also used for choosing color palettes on old fashioned graphical display devices and Image Quantization.*

Difference between K means and KNN

	K Means Clustering	K Nearest neighbors
Algorithm objective	Clustering	Classification
Nature	Unsupervised	Supervised
Hyper parameter K	"k" is the number of clusters	"k" is the number of neighbors it checks
How it works	it takes a bunch of *unlabeled* points and tries to group them into clusters	it takes a bunch of *labeled* points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point