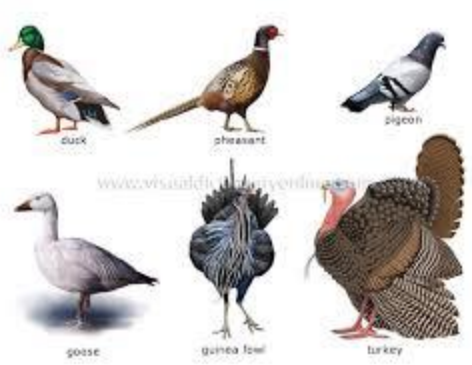
Clustering

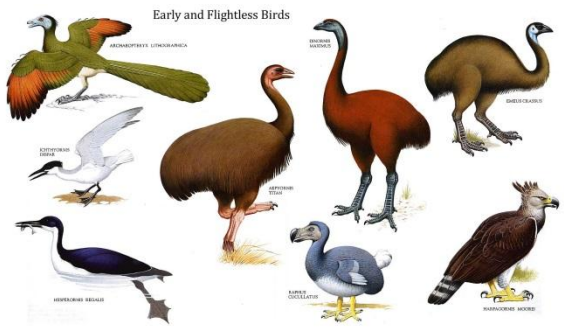
Can You Group These Gems?

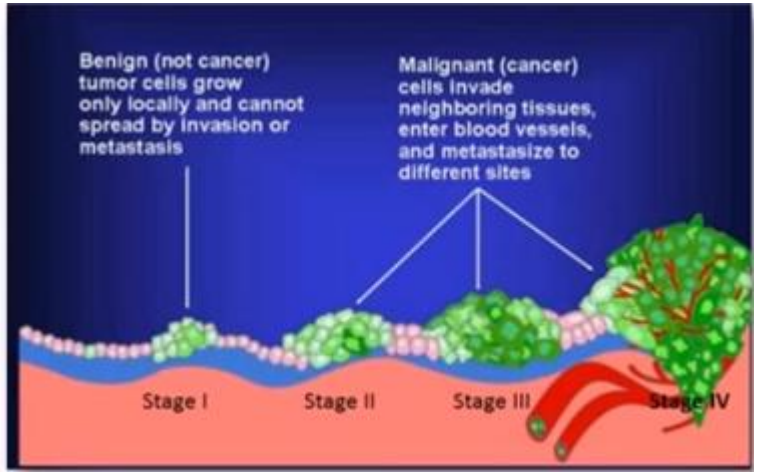
What is Clustering?

- Clustering is an algorithm which finds natural groupings inside your data when these groupings are not obvious.
- It finds the hidden variable that accurately segment your data.
- A cluster is a collection of data instances which are similar to each other and are dissimilar to data instances in other instances.

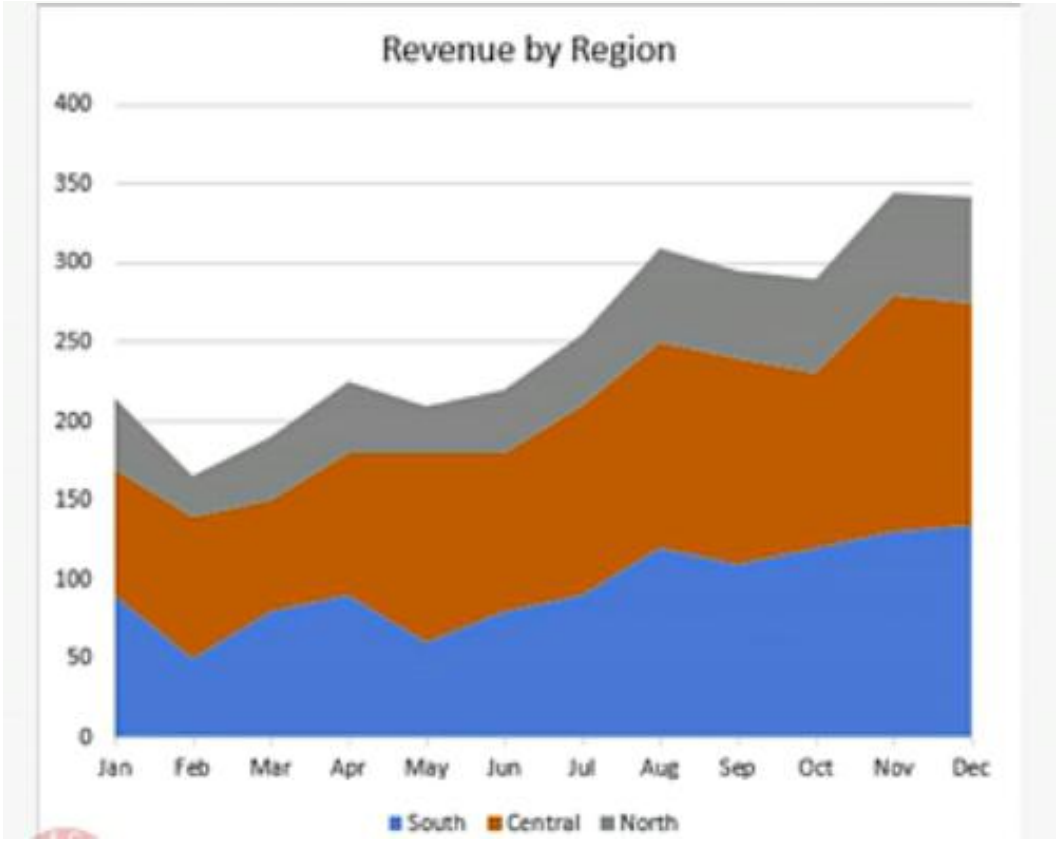
Can You Identify These Groups?

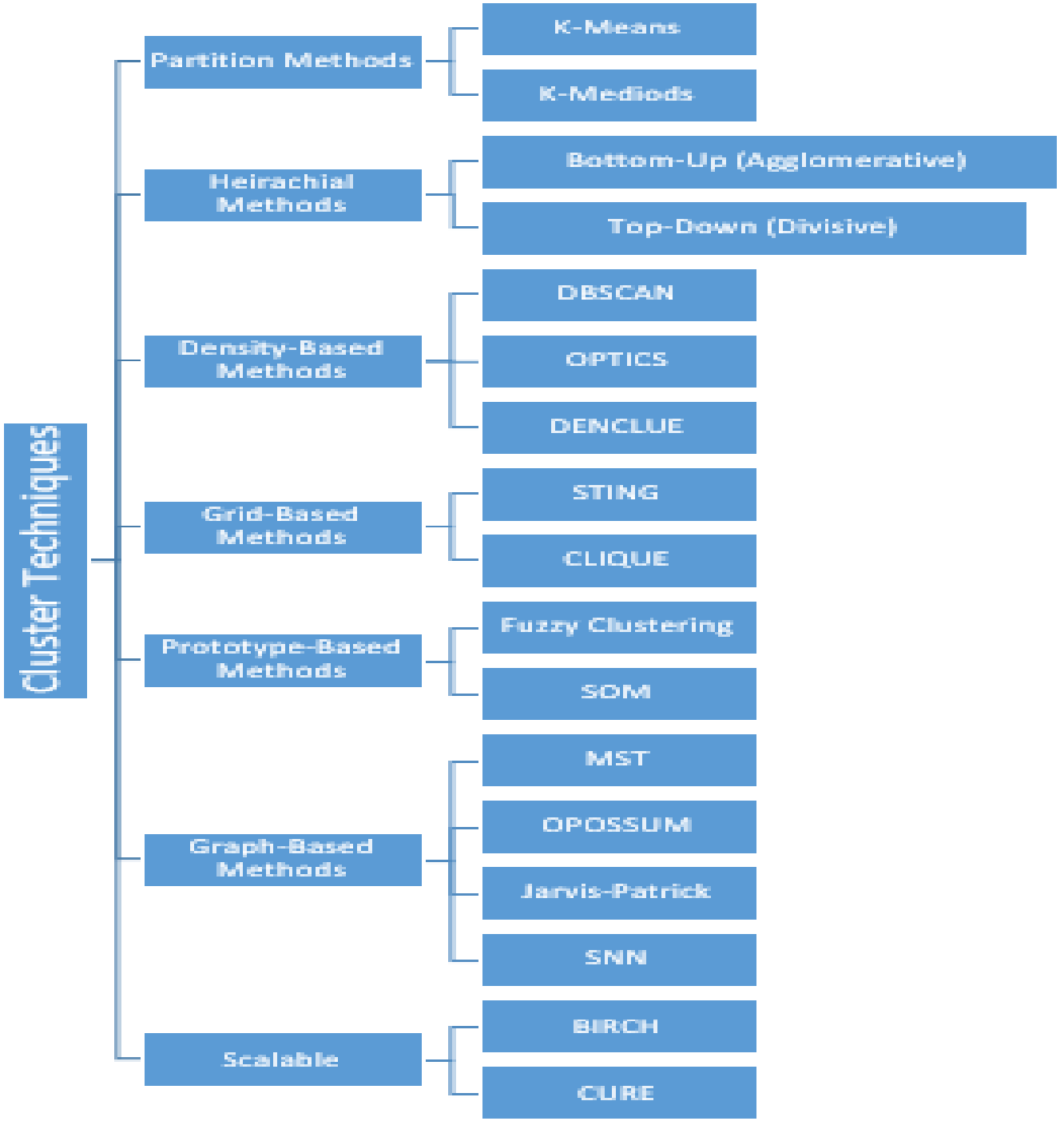





Benign (not cancer) tumor cells grow only locally and cannot spread by invasion or metastasis
Malignant (cancer) cells invade neighboring tissues, enter blood vessels, and metastasize to different sites
Stage I
Stage II
Stage III
Stage IV

Cluster for Analysis




Cluster Techniques
Partition Methods
K-Means
K-Mediods
Heirachial Methods
Bottom-Up (Agglomerative)
Top-Down (Divisive)
Density-Based Methods
DBSCAN
OPTICS
DENCLUE
Grid-Based Methods
STING
CLIQUE
Prototype-Based Methods
Fuzzy Clustering
SOM
Graph-Based Methods
MST
OPOSSUM
Jarvis-Patrick
SNN
Scalable
BIRCH
CURE

KM

K-Means: Within and Between Cluster

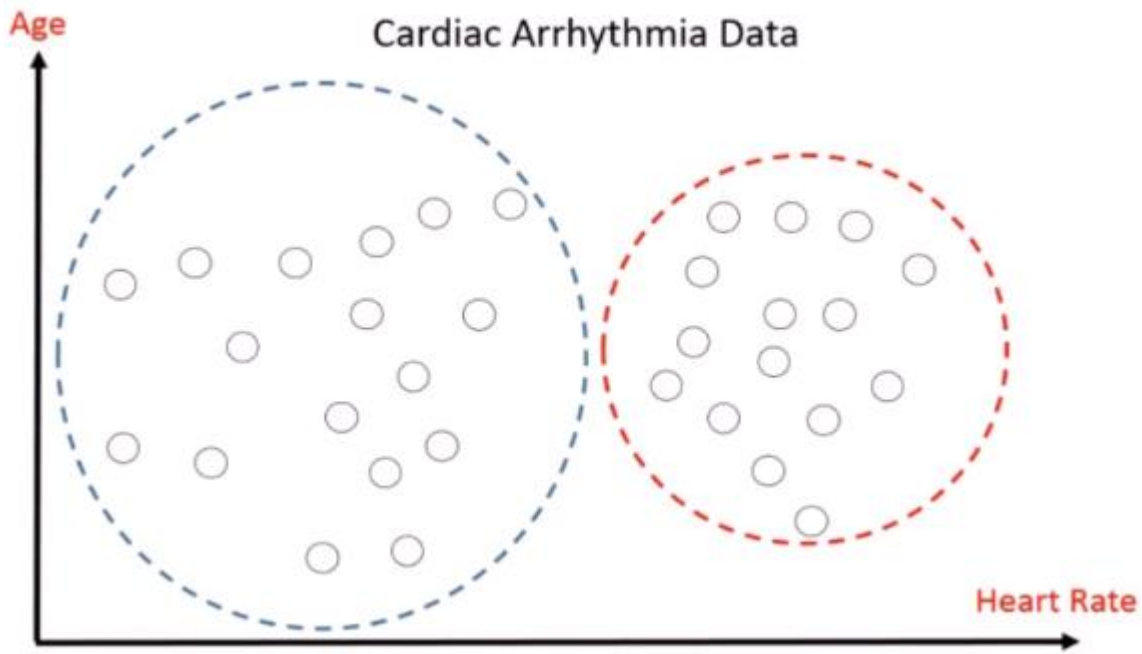
$$T = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d(x_i, x_j)$$

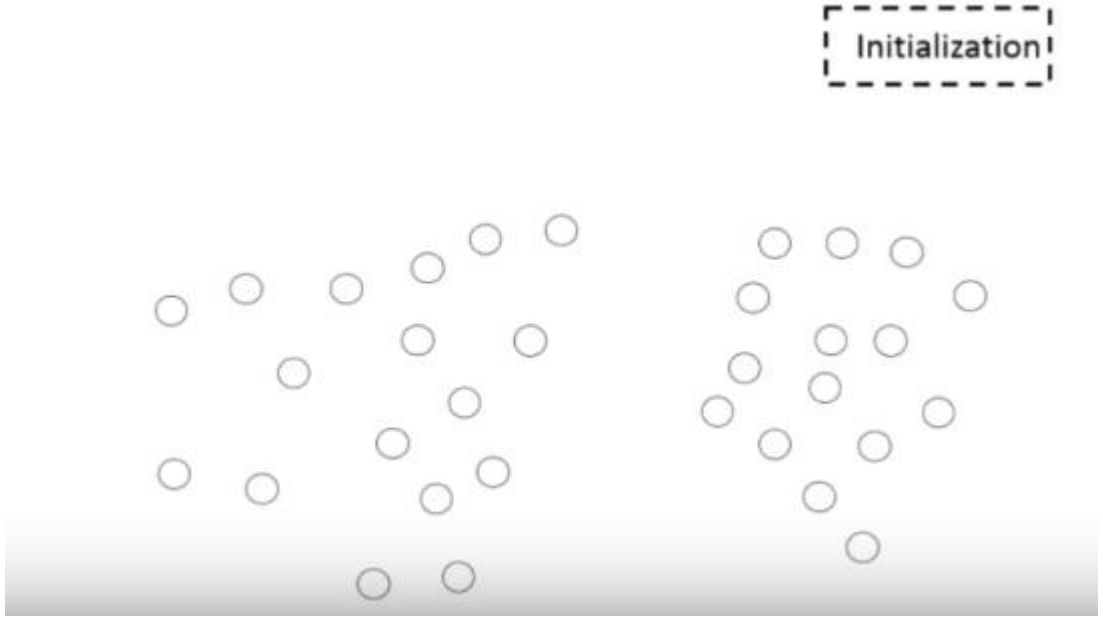
$$T = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\left(\sum_{C(j)=k} d(x_i, x_j) + \sum_{C(j)\neq k} d(x_i, x_j)\right)$$

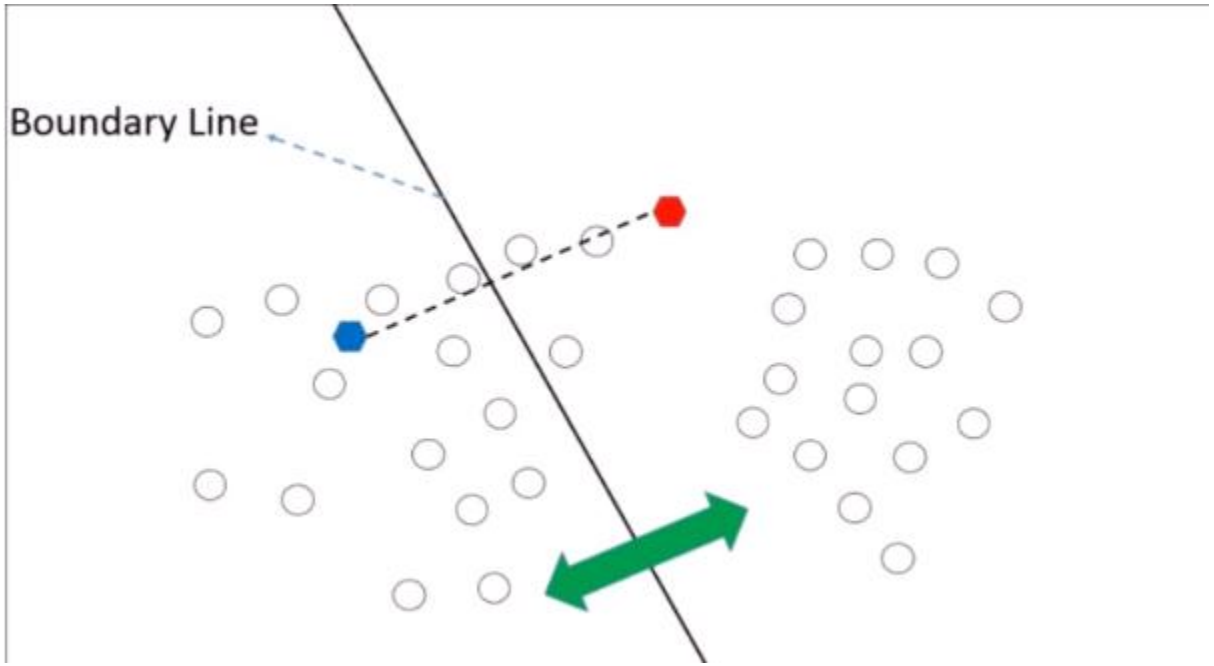
$$T = W(C) + B(C)$$

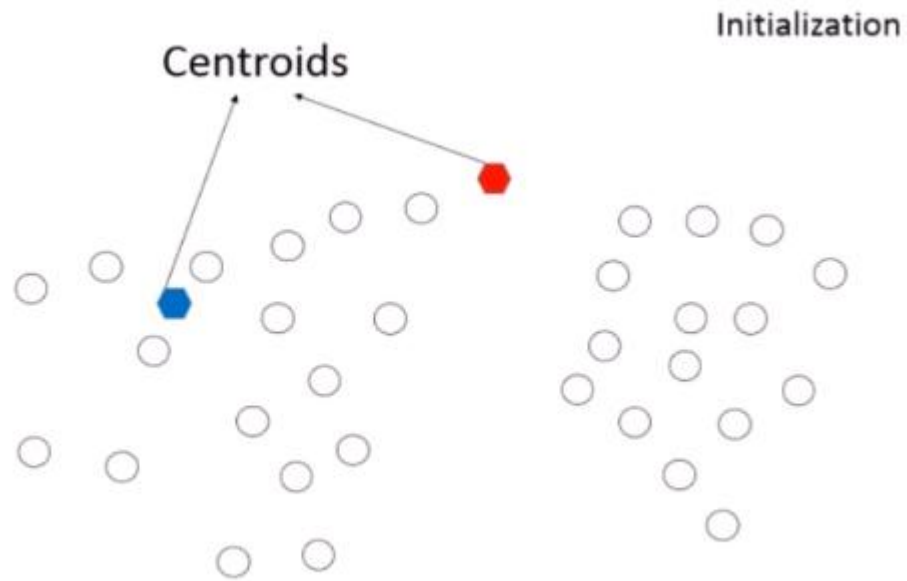
Within Cluster

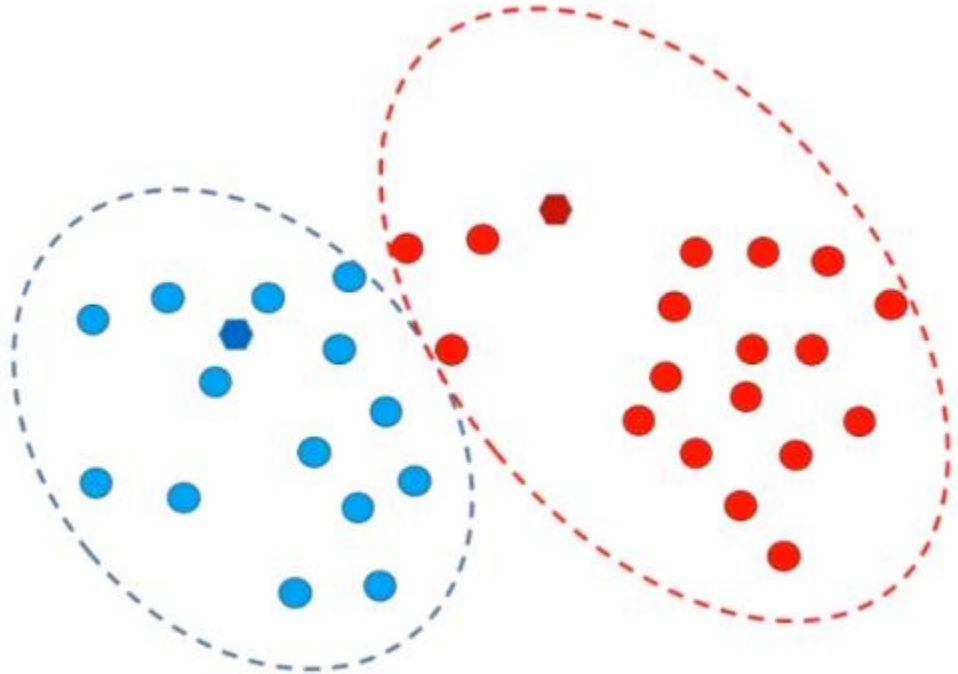
Between Clusters

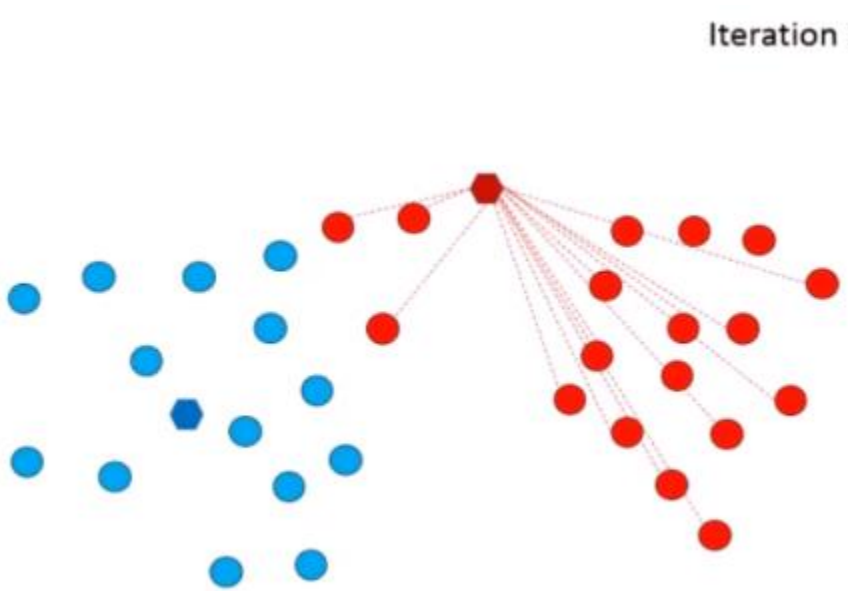

Age
Cardiac Arrhythmia Data
Heart Rate

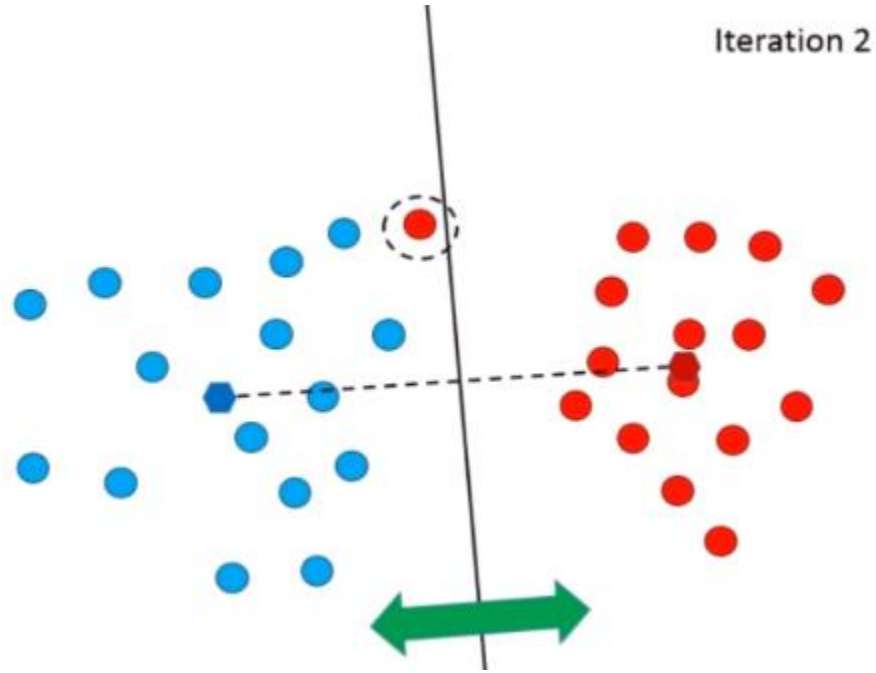

Initialization

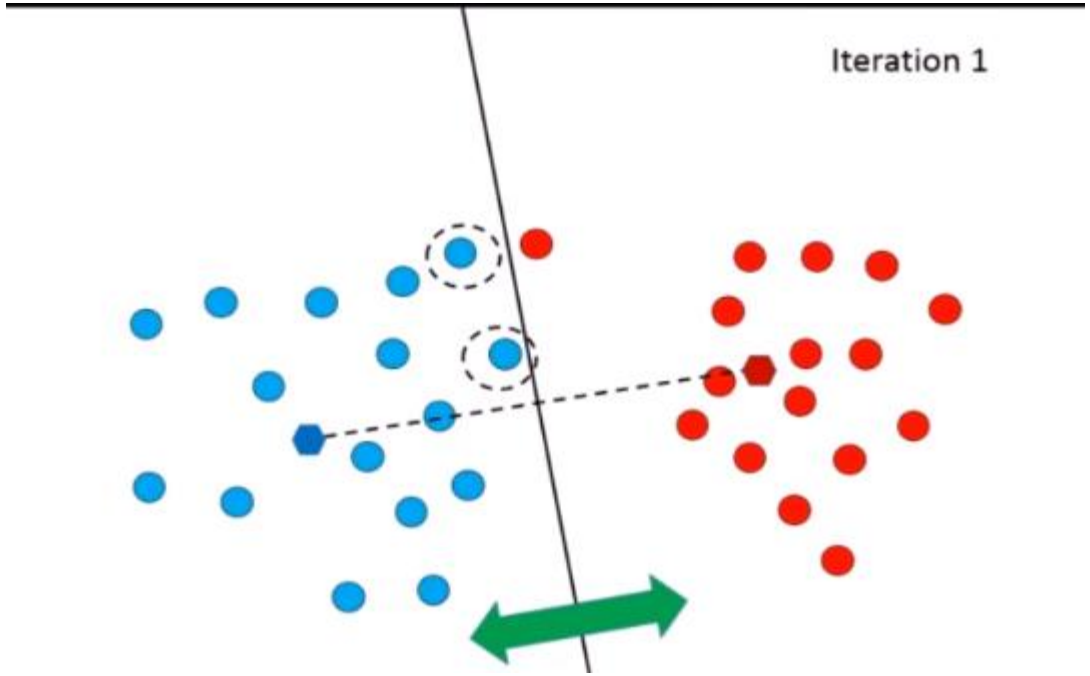

Boundary Line

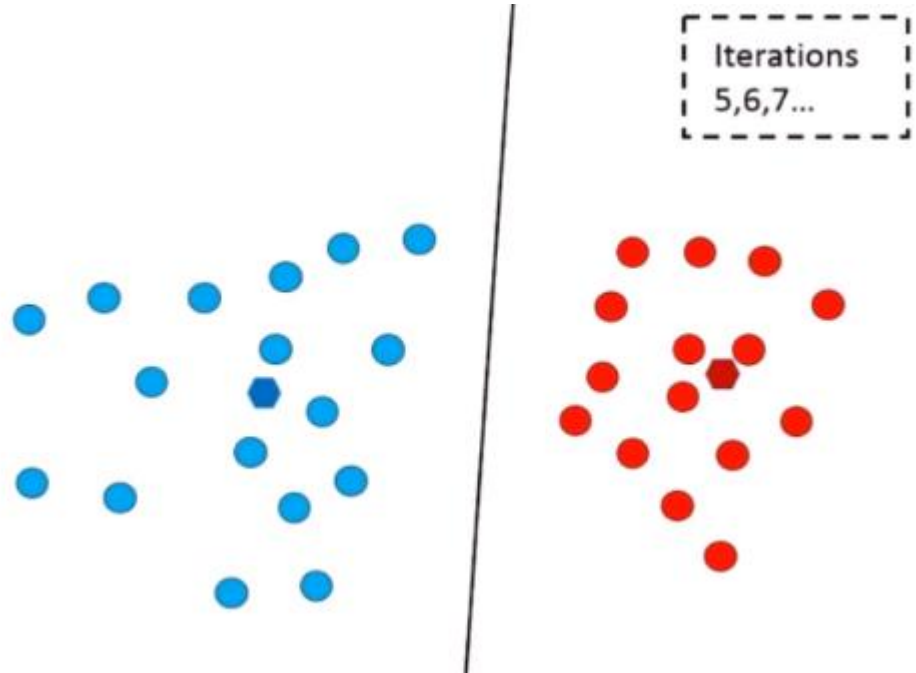

Initialization
Centroids



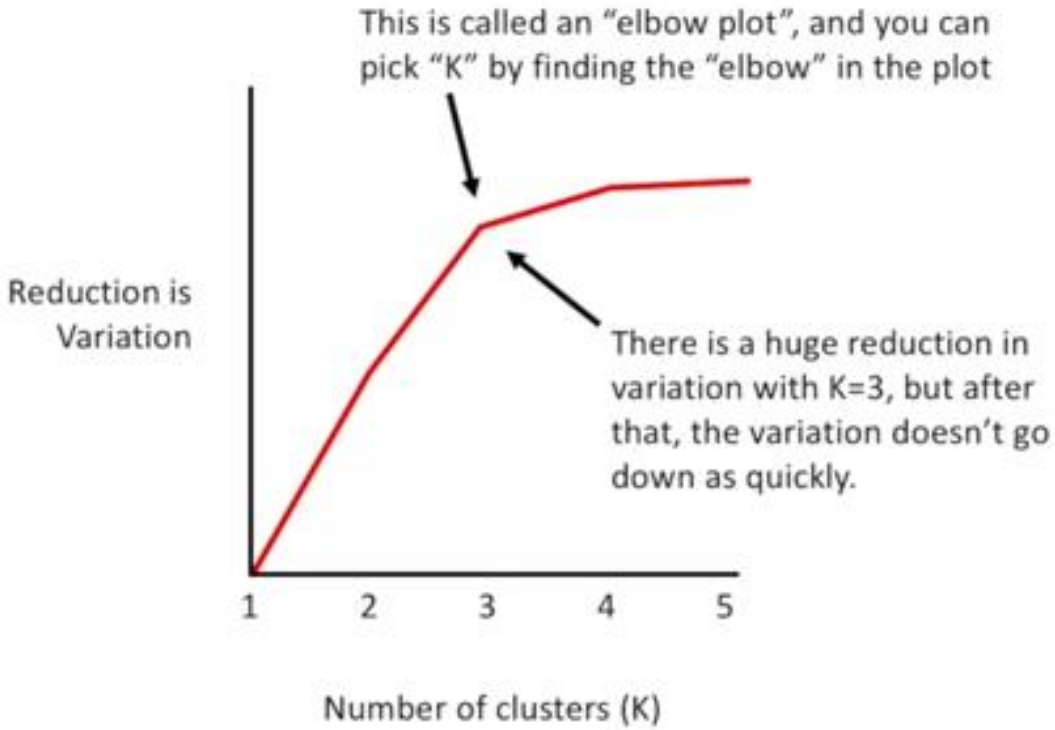

Iteration 1


Iteration 2


Iteration 1


Iterations
5,6,7...

How to Calculate K

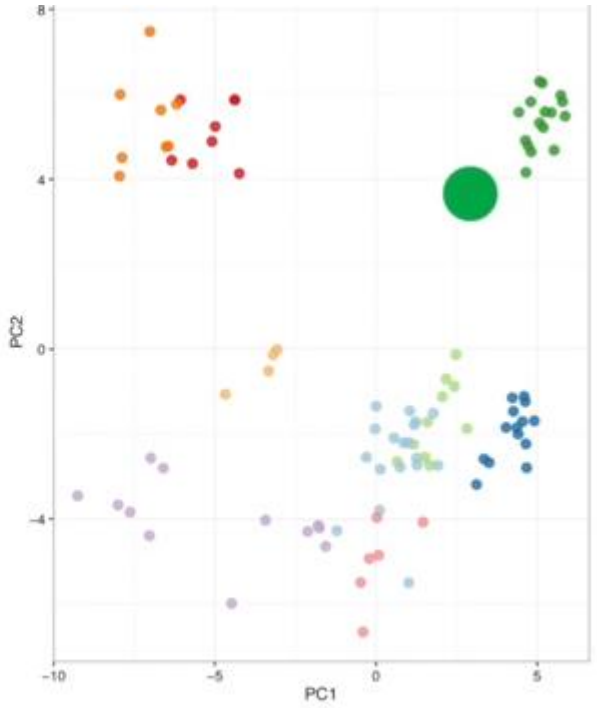


K-Nearest Neighbor (KNN)

In this case, the category is **GREEN**.

If K=11, we would use the 11 nearest neighbors.

In this case, the category is still **GREEN**.



If K=11 and the new cell is between two (or more) categories, we simply pick the category that "gets the most votes".

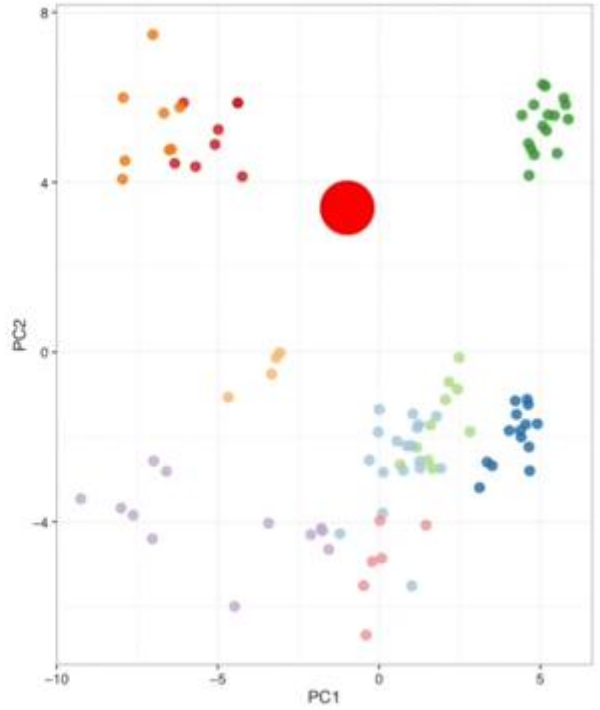
In this case....

7 nearest neighbors are **RED**.

3 nearest neighbors are **ORANGE**.

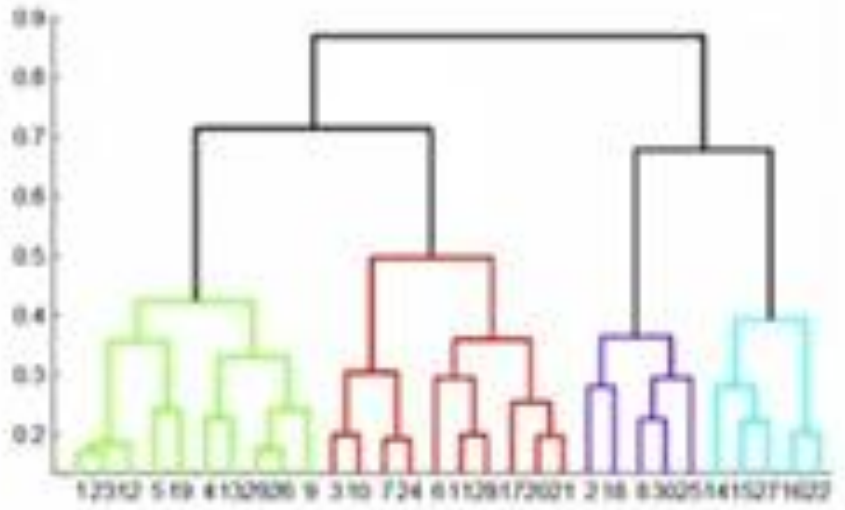
1 nearest neighbor is **GREEN**.

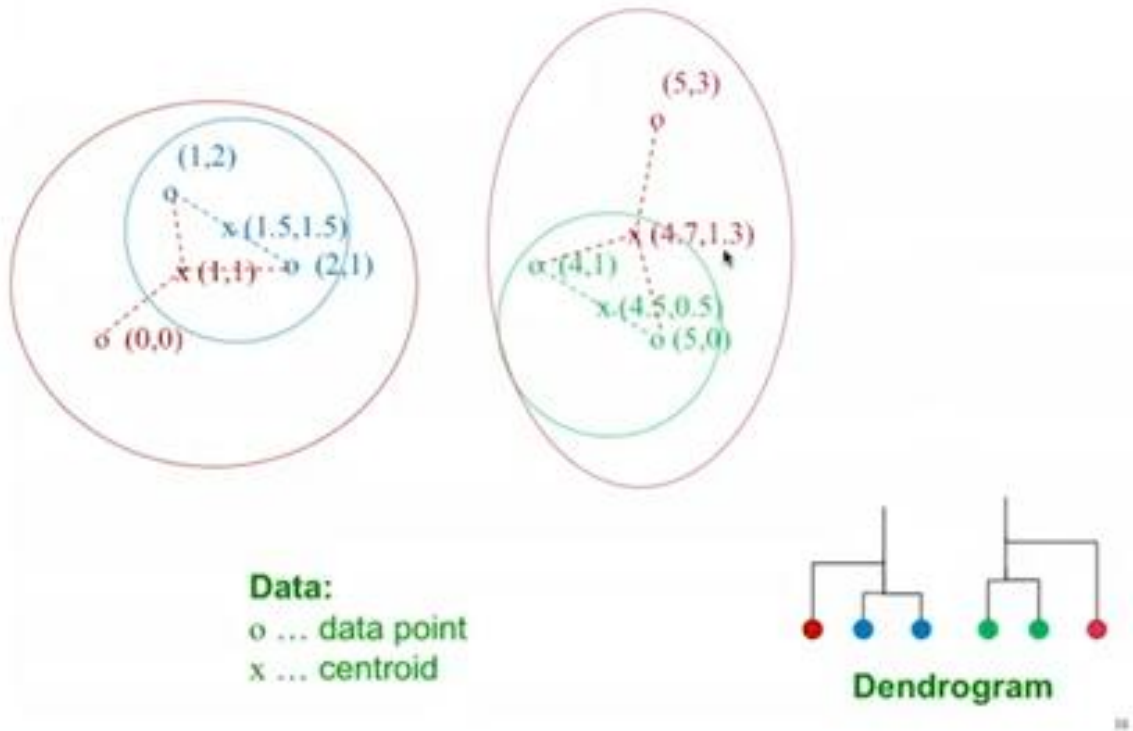
Since **RED** got the most votes, the final assignment is **RED**.



Hierarchical Clustering

- **Hierarchical:**
  - **Agglomerative** (bottom up):
    - Initially, each point is a cluster
    - Repeatedly combine the two "nearest" clusters into one
  - **Divisive** (top down):
    - Start with one cluster and recursively split it




(1,2)
x (1.5,1.5)
x (1,1)
(2,1)
(0,0)
(5,3)
x (4.7,1.3)
(4,1)
x (4.5,0.5)
(5,0)
Data:
o … data point
x … centroid
Dendrogram

Distance Functions

- Distance or similarity functions play a central role in all clustering algorithms.
- Numerous distance functions have been reported.
  - Euclidean Distance
  - Manhattan Distance
  - Minkowski Distance
  - Chebychev Distance
  - Cosine Distance