Unsupervised Learning

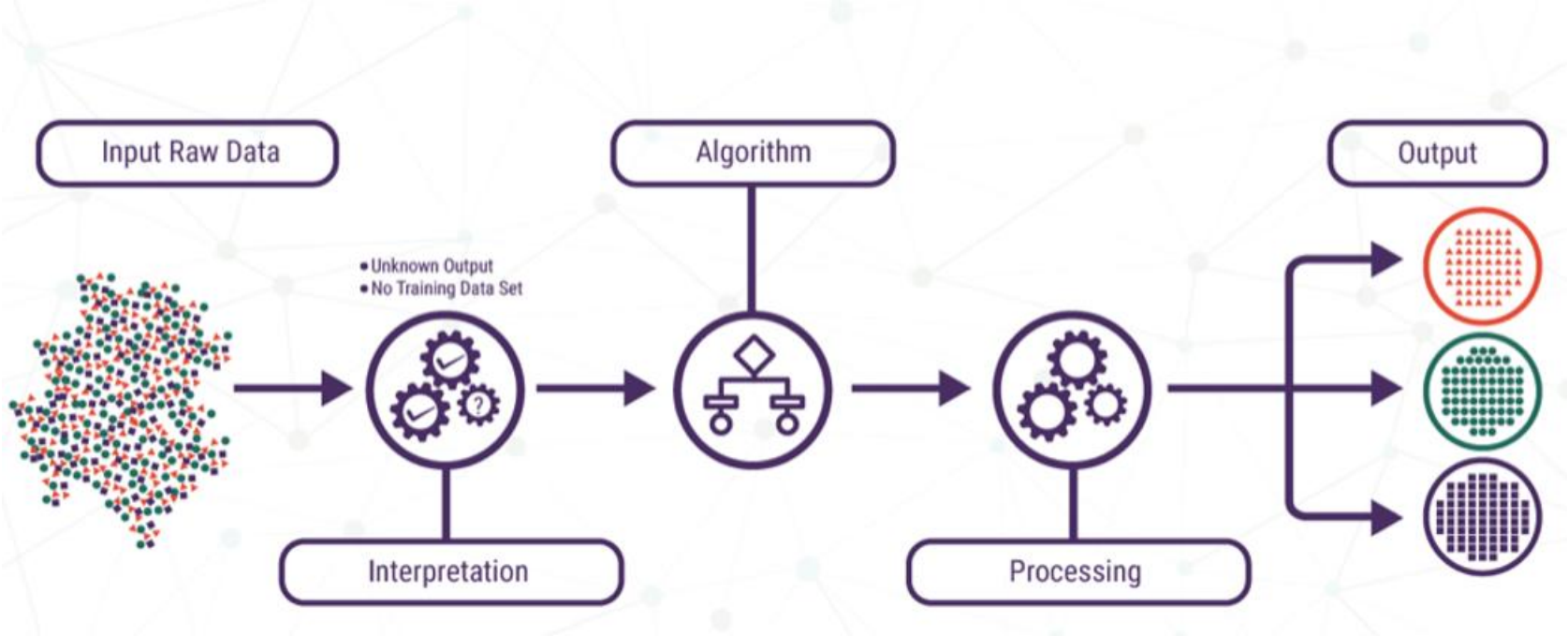
Dr. Tapan Kumar Jain,

HoD & Assistant Professor (ECE), IIIT Nagpur

Senior Member, IEEE

Unsupervised Learning

1. A machine learning technique, where you do not need to supervise the model.
2. Instead, allow the model to work on its own to discover information.
3. It mainly deals with the unlabeled data.



Unsupervised Algorithms

1. **Clustering**

   Automatically split the dataset into groups base on their similarities

2. **Anomaly Detection**

   Discover unusual data points in your dataset. It is useful for finding fraudulent transactions

3. **Association mining**

   Identifies sets of items which often occur together in your dataset

Clustering and its type

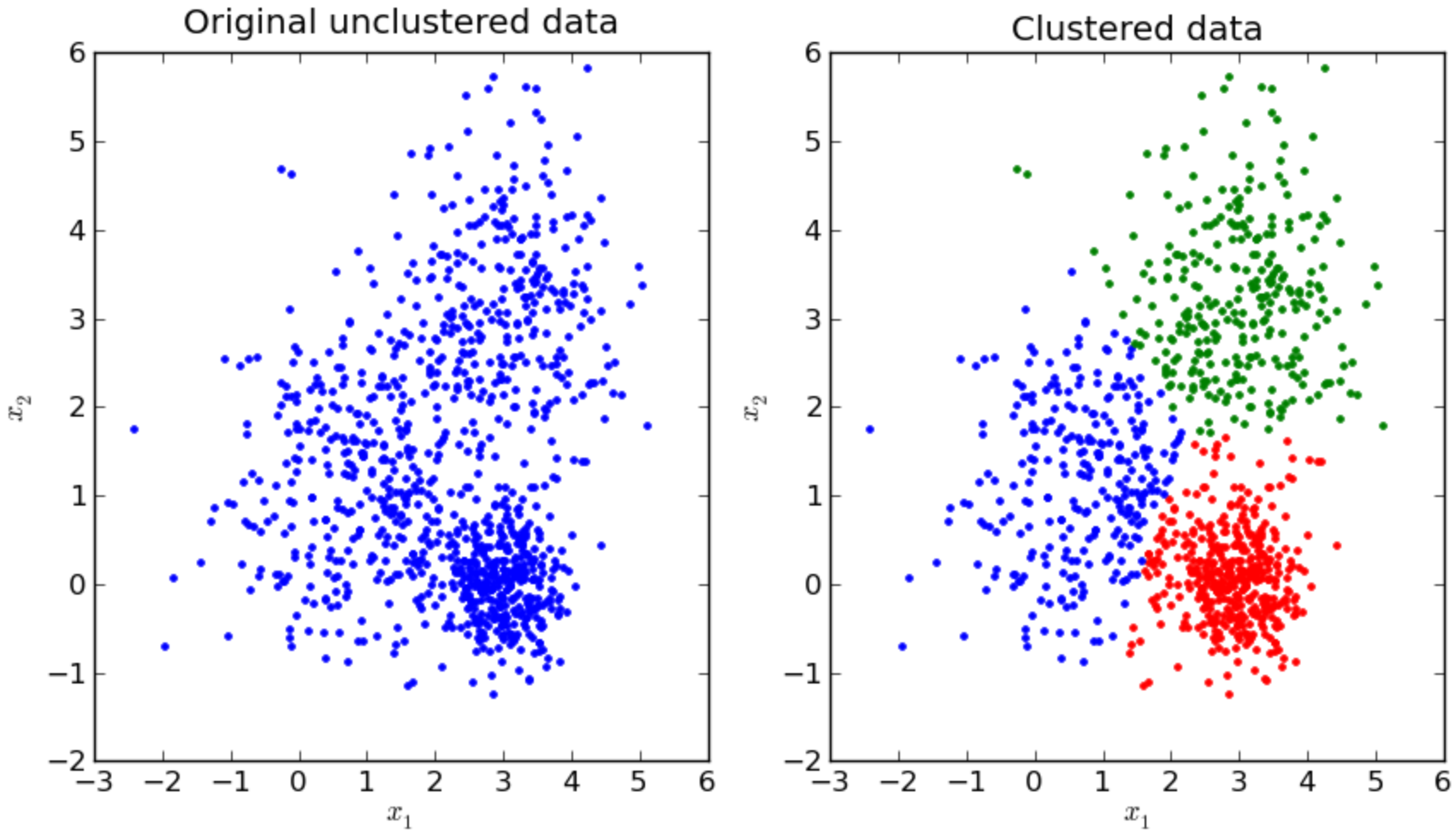
Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data.

Definition:

The task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different.

In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar according to similarity measure : Euclidean distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering Contd.



Clustering Algorithms

1. K-means clustering (we will focus on this)
2. Mean-shift clustering
3. Kernel K-means clustering
4. K-mediods clustering

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

K-means - I

- **K-means** : Iterative algorithm
- Algorithm partitions the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters)
- Each data point belongs to only one group.

K-means : II

- It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.
- It assigns data points to a cluster such that the Euclidean-distance between the data points and the cluster's centroid is minimum.
- The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

K-means : III

- The objective function is,

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \qquad (1)$$

  - K – clusters
  - μ_k centroid of cluster k
  - x^i data point – total m datapoints
  - w_{ik} = 1, if xi belongs to cluster k

    = 0, if xi belongs to cluster k

We want to minimize the objective function.

K-means : IV

Procedure

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.

K-means : V

ITERATE (until there is no change in Centroids)

E – Step

Compute the sum of the squared distance between data points and all centroids.

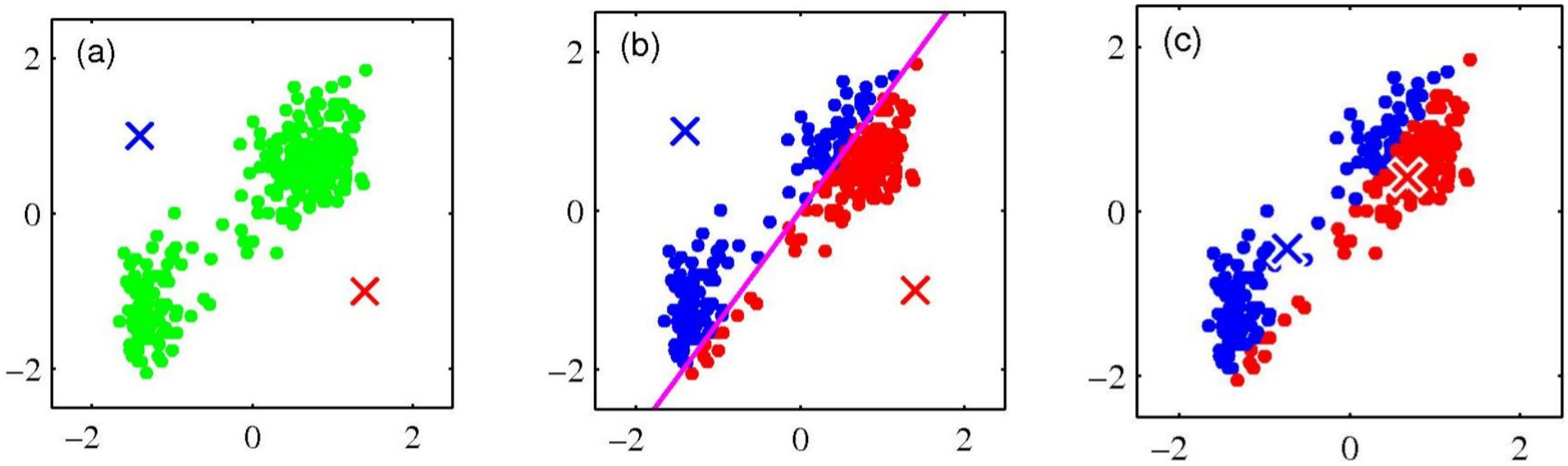
Assign each data point to the closest cluster (centroid).

M – Step

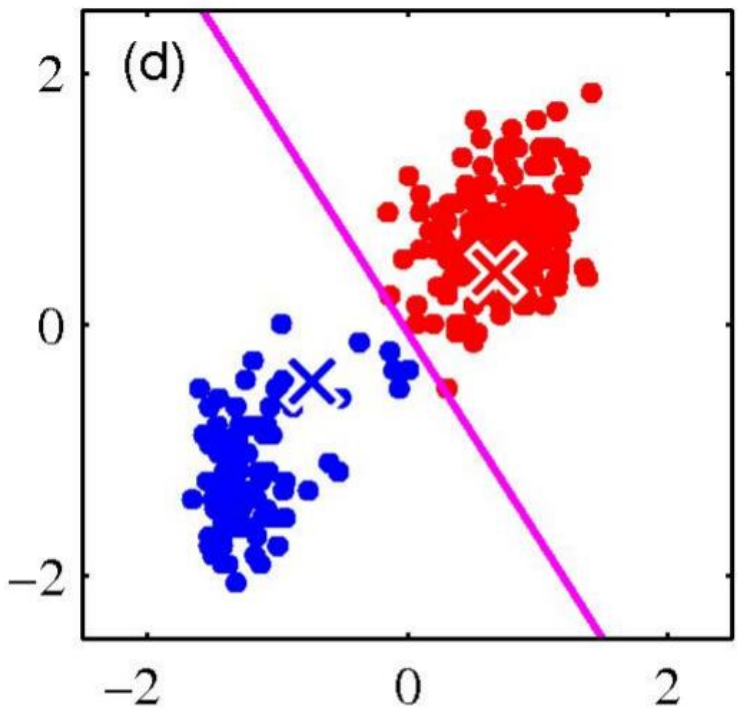
Compute the centroids for the clusters by taking the

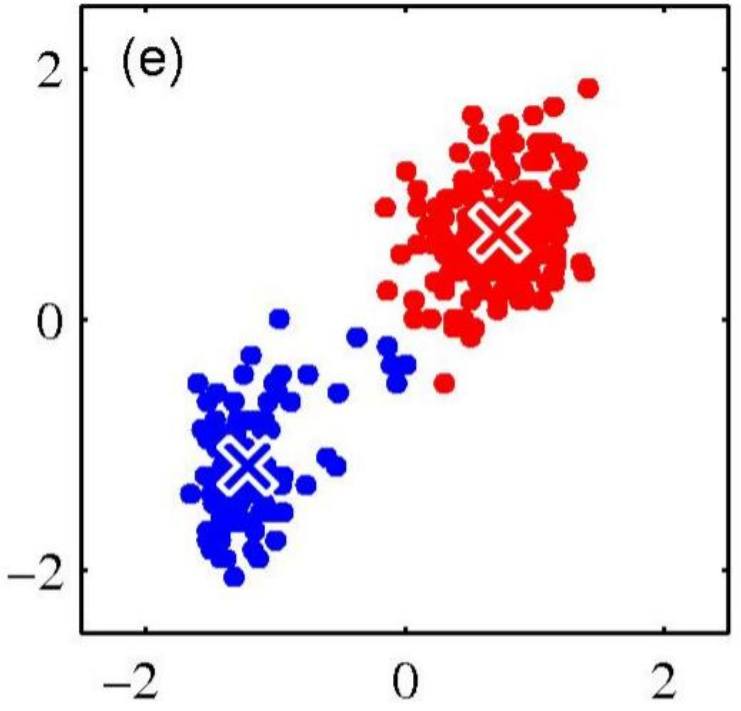
average of the all data points that belong to each cluster.

K-means : Example

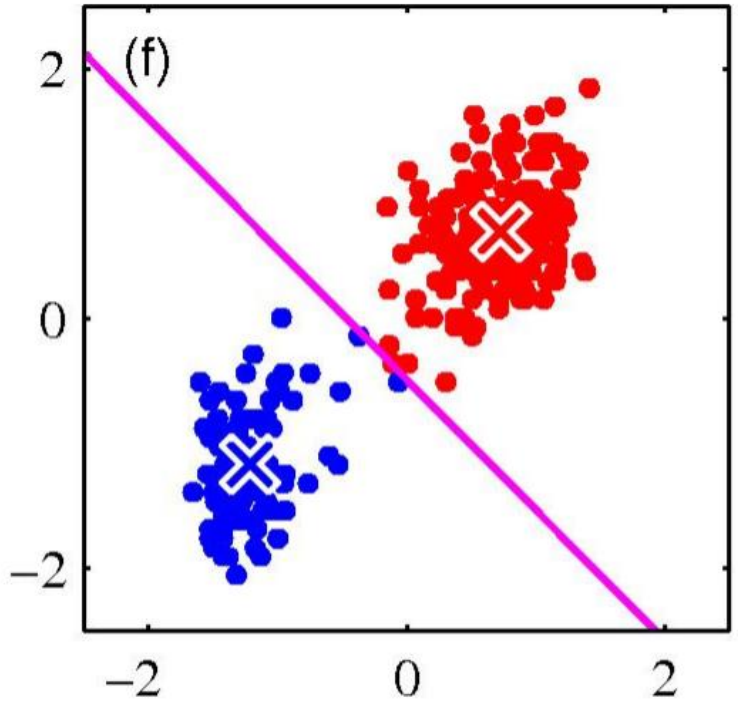


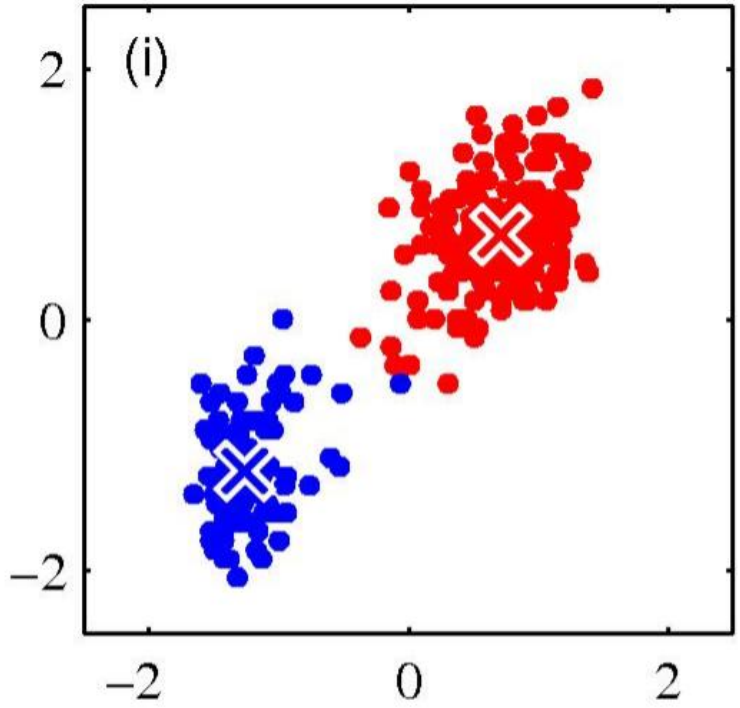
K-means : Example





K-means : Example





K-means : Example - II

Data Points	Weight (Kg)	Height (cm)
1	72	185
2	56	170
3	60	168
4	68	179
5	72	182
6	77	188
7	71	180
8	70	180
9	84	183
10	88	180
11	67	180
12	76	177

1. Initially we assume that two clusters are K_1 and K_2, based on initial two values. [Total DP is 12]
2. Take next data point and calculate the centroid value with the help of Euclidean distance.

$$D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

3. K_1 = (72,185) and K_2 = (56,170) now calculate the distance from data point 3 i.e. (60,168)
4. D_{13} = sqrt((72-60)^2 + (185-168)^2) = 20.8
5. D_{23} = sqrt((56-60)^2 + (170-168)^2) = 4.5

K-means : Example - II

Data Points	Weight (Kg)	Height (cm)
1	72	185
2, 3	58	169
4	68	179
5	72	182
6	77	188
7	71	180
8	70	180
9	84	183
10	88	180
11	67	180
12	76	177

	Weight (Kg)	Height (cm)
1	72	185
2	56	170
3	60	168

	Weight (Kg)	Height (cm)
1	72	185
2,3	58	169

6. D_{13} = 20.8, and D_{23} = 4.5, Minimum (D_{13}, D_{23}) so DP 2 and 3
7. Calculate new centroid values, K_1 remain same and modified value of K_2 is avg of DP 2 and 3 i.e. (56+60)/2 and (170+168)/2
8. Now take DP 4 and calculate the Euclidean distance from K_1 and K_2
9. D_{14} = 7.21, and D_{24} = 14.14,

Data Point = 11

K-means : Example - II

Data Points	Weight (Kg)	Height (cm)
1,4	70	182
2, 3	58	169
5	72	182
6	77	188
7	71	180
8	70	180
9	84	183
10	88	180
11	67	180
12	76	177

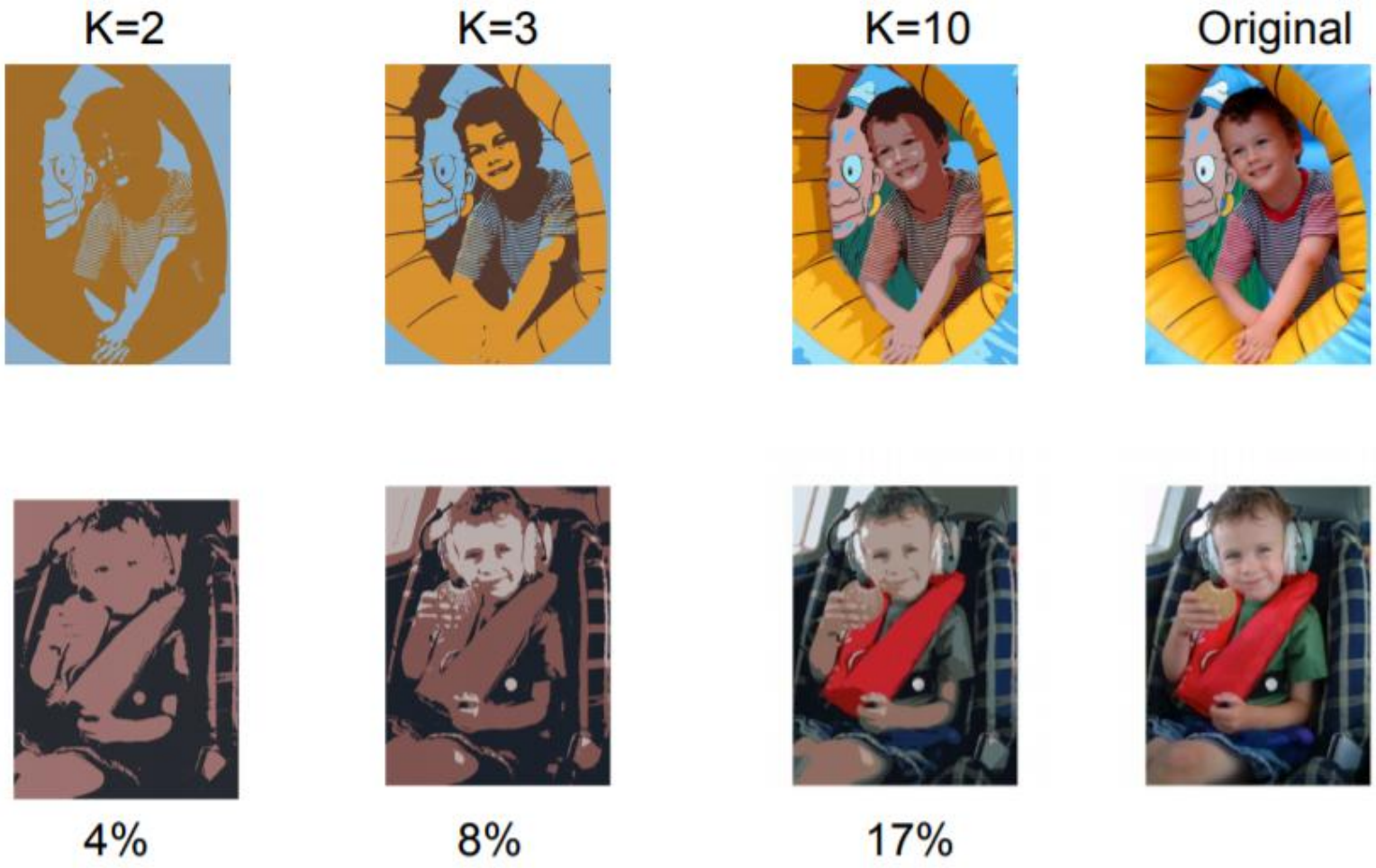
	Weight (Kg)	Height (cm)
1	72	185
2,3	58	169
4	68	179

	Weight (Kg)	Height (cm)
1,4	70	182
2,3	58	169

6. D_{14} = 7.21, and D_{24} = 14.14, Minimum(D_{14}, D_{24}) so DP 1and 4
7. Calculate new centroid values, K_2 remain same and modified value of K_1 is avg of DP 1 and 4 i.e. (72+68)/2 and (185+179)/2
8. Now take DP 5 and calculate the Euclidean distance from K_1 and K_2
9. Repeat

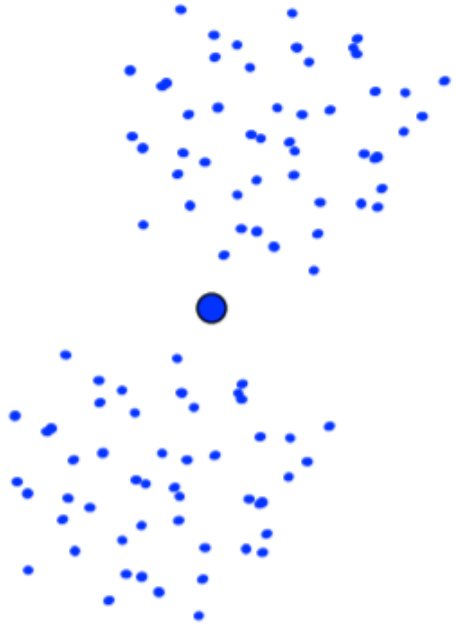
Data Point = 10

Image Segmentation



Limitation K-means

Would be better to have
one cluster here

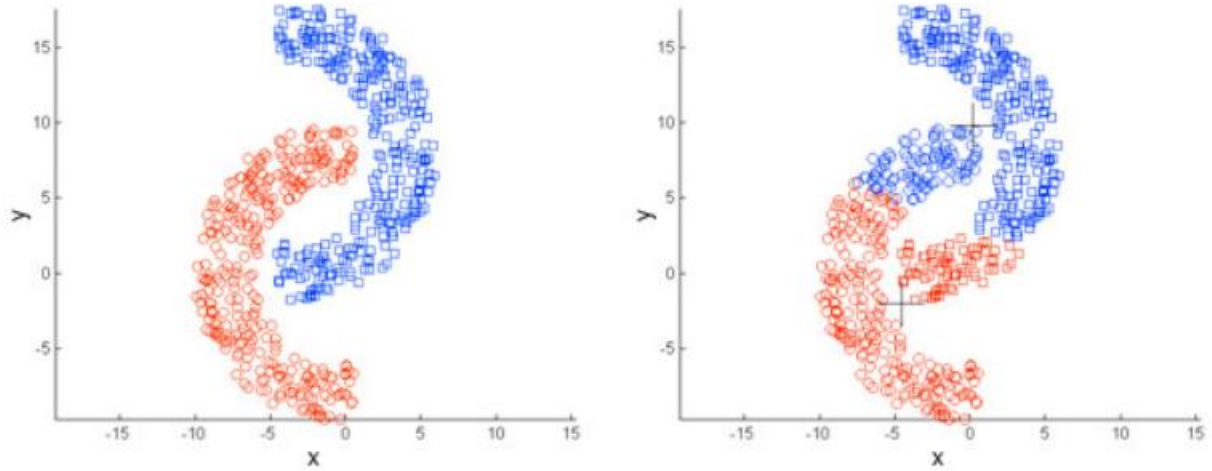


… and two clusters here

Limitation K-means

Similarity is subjective. Two Samples might be close in feature space but actually be very different.

Limitation K-means



K-means also works well only when the clusters are round-shaped and does badly if the clusters have non-convex shapes

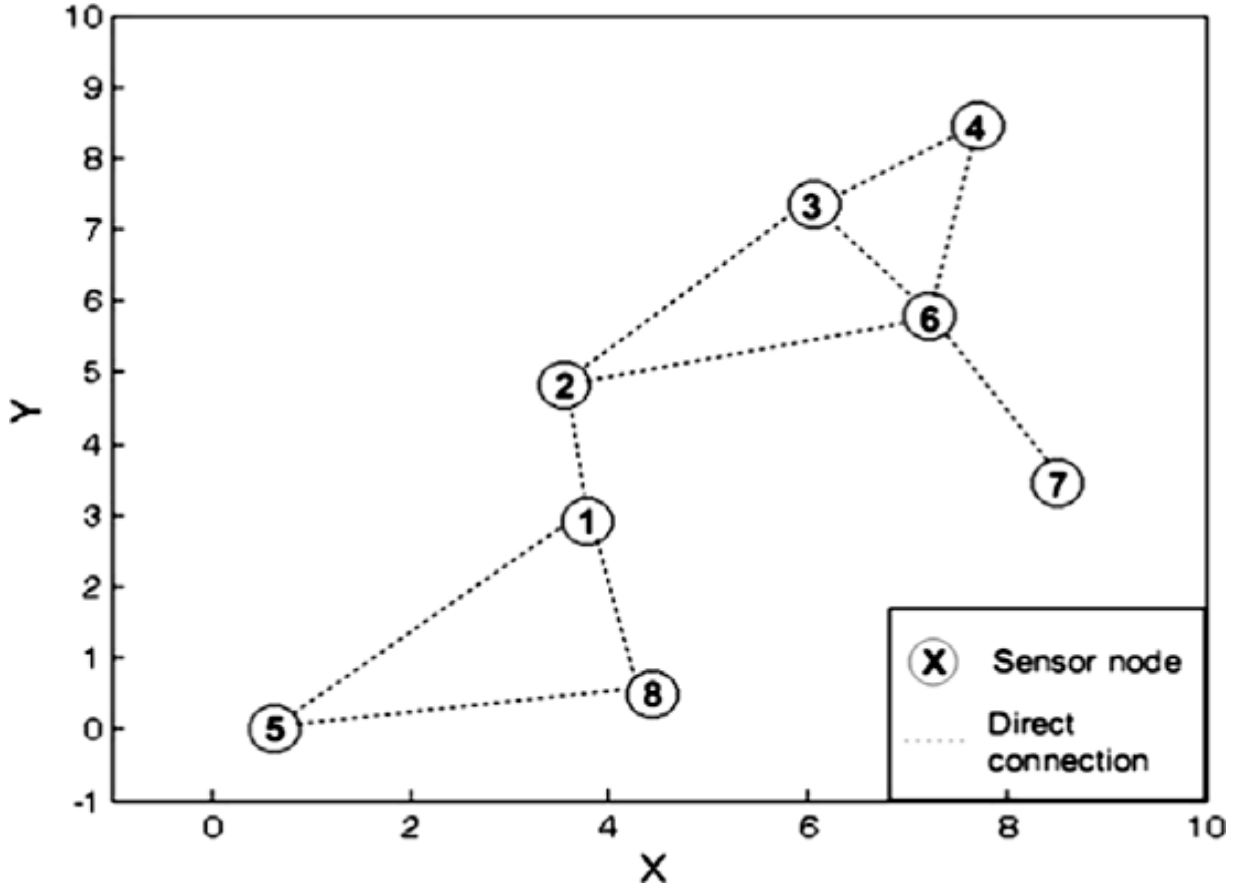
Hierarchical Clustering

- Divisive (Top to bottom fashion)
- Agglomerative (Bottom up fashion)

Hierarchical Agglomerative Clustering

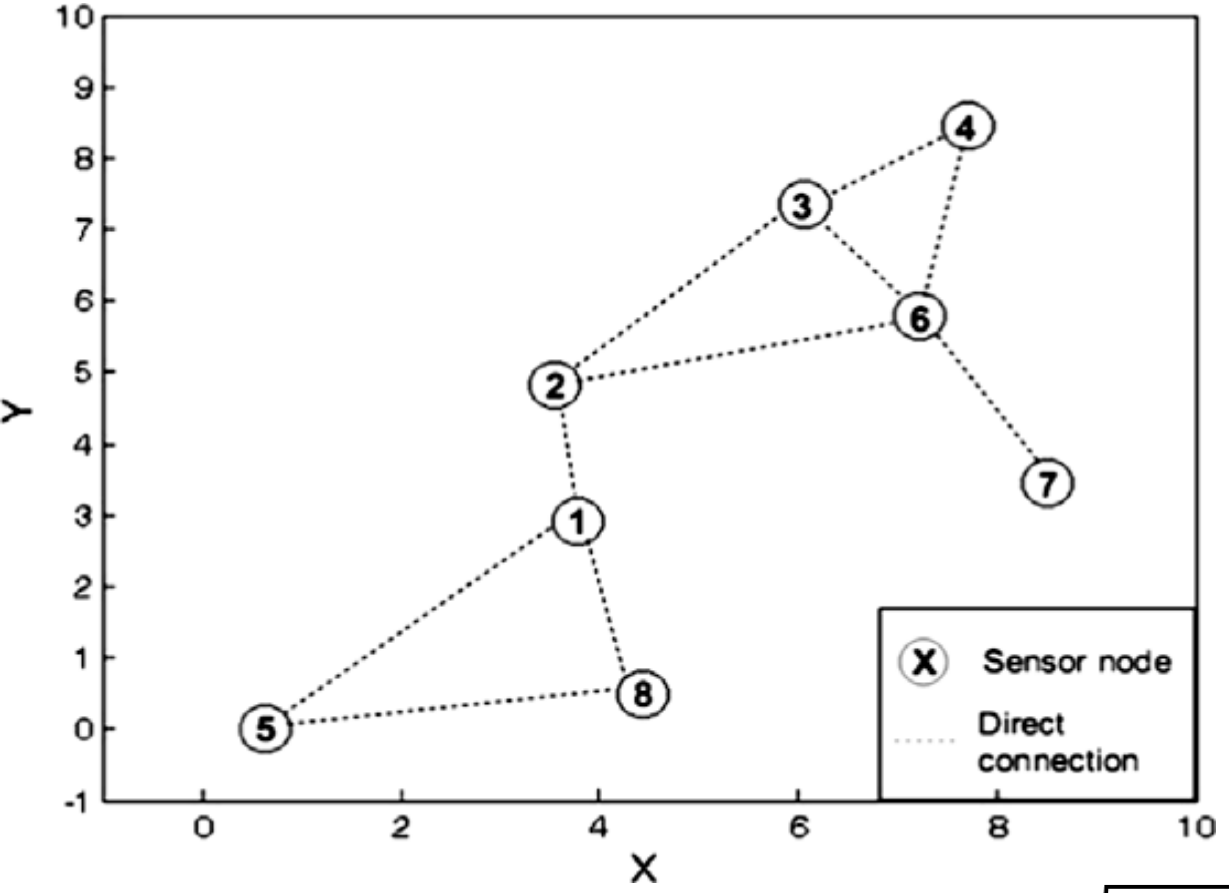
- Agglomerative (Bottom up fashion)
  - We want to group based on their similarity
    - Quantitative
    - Qualitative
  - Construct the Resemblance Coefficient Matrix
    - Euclidean distance
    - Sorenson
  - Merging the two clusters
    - Single, Complete and Average Link
    - and update the Resemblance Coefficient till [2x2]
  - Draw the tree (dendrogram)

A simple 8-node network



Component (node)	Attribute	
	x-Axis	*y*-Axis
(a) Quantitative data: node location data matrix		
{1}	3.78	2.9
{2}	3.56	4.83
{3}	6.06	7.34
{4}	7.71	8.46
{5}	0.63	0.01
{6}	7.23	5.78
{7}	8.52	3.46
{8}	4.43	0.48

Resemblance matrix: Euclidean distance

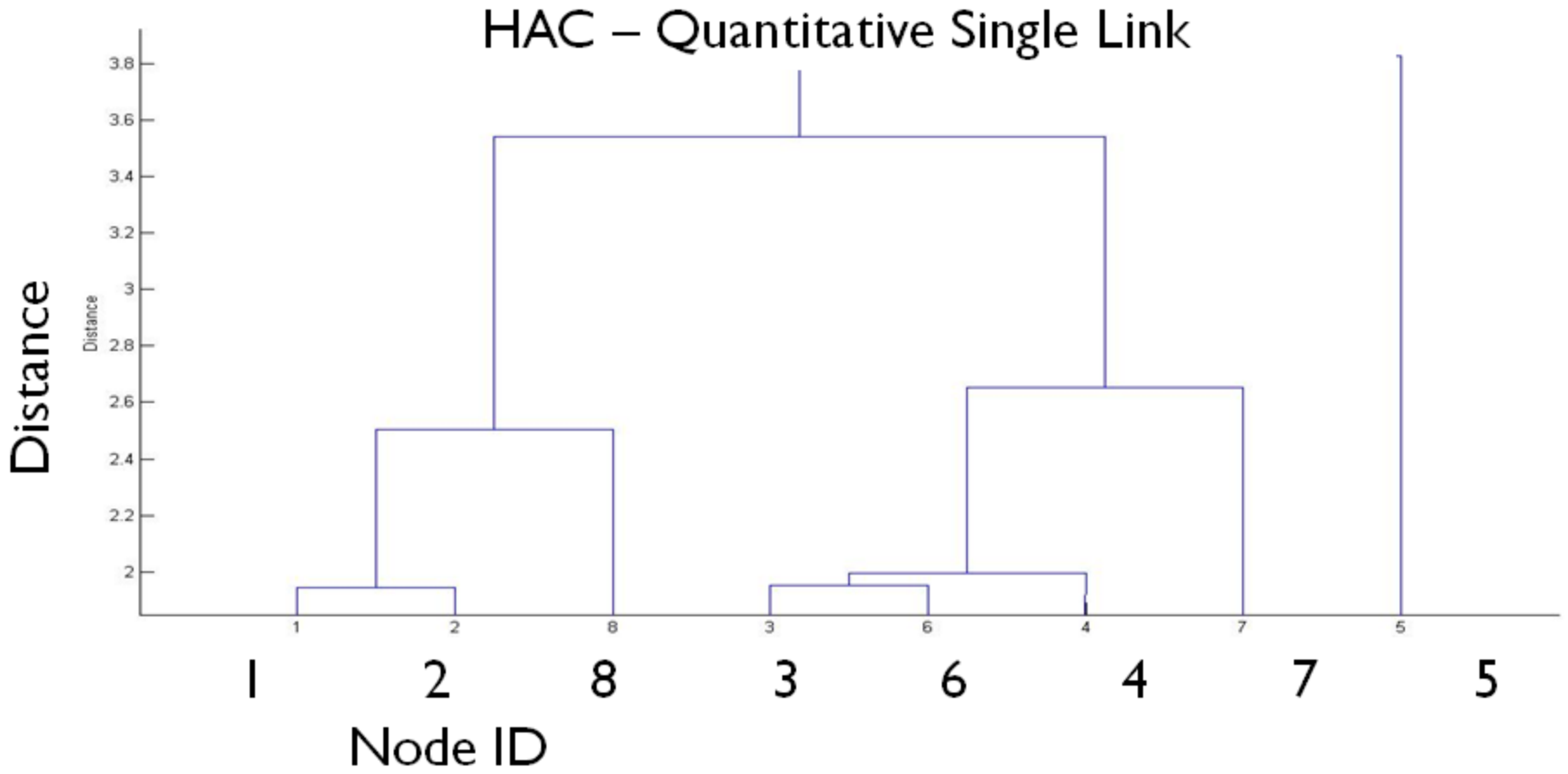


	{2}	{3}	{4}	{5}	{6}	{7}	{8}
{1},{2}	1.99	3.54	4.99	5.51	6.81	4.27	4.77	2.51
{2}	-	3.54	5.51	5.64	3.79	5.15	4.44
{3}	-	-	1.99	9.12	1.95	4.59	7.05
{4}	-	-	-	11	2.72	5.07	8.63
{5}	-	-	-	-	8.77	8.61	3.83
{6}	-	-	-	-	-	2.65	5.99
{7}	-	-	-	-	-	-	5.06

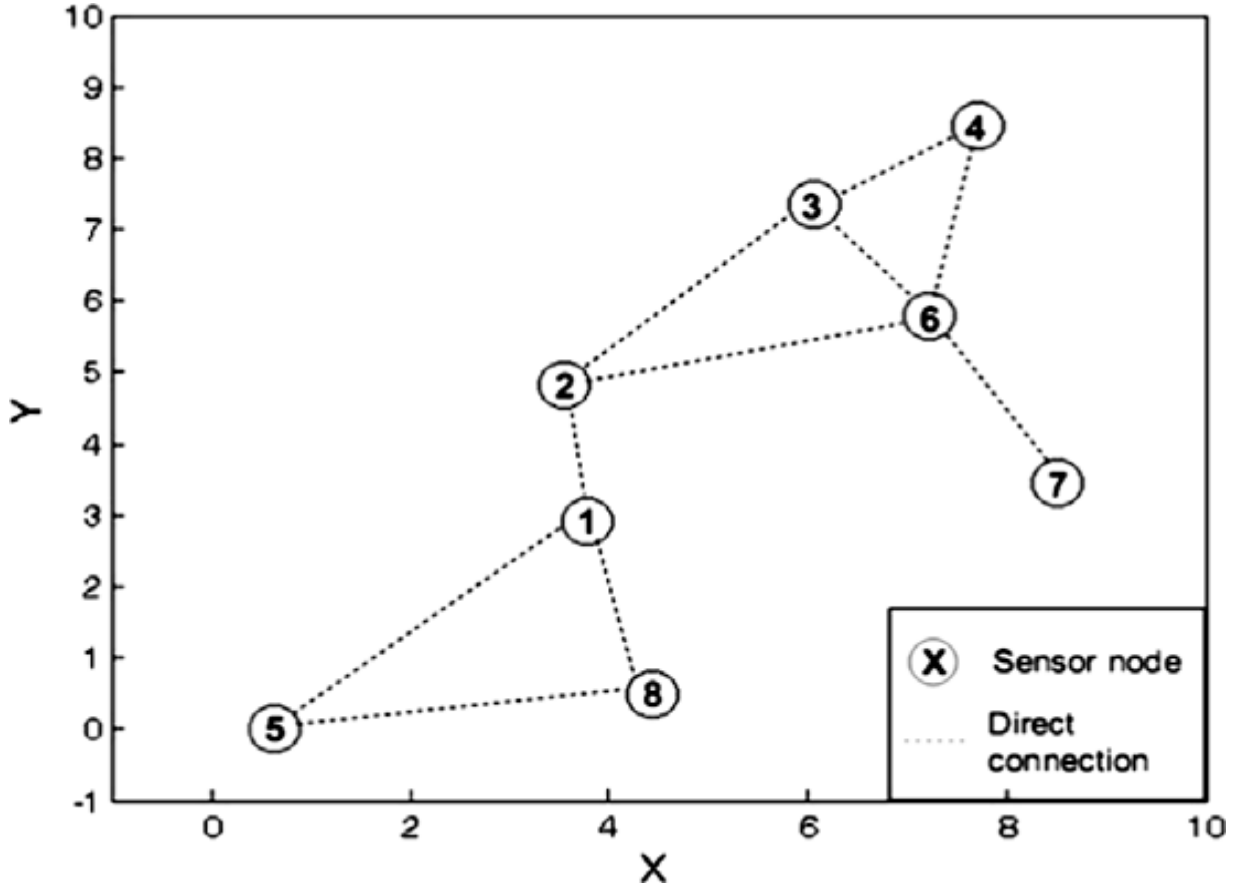
$$D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

[4]

Dendrogram: HAC Quantitative using single link



A simple 8-node network



	{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
(b) Qualitative data: one-hop network connectivity data matrix								
{1}	1	1	0	0	1	0	0	1
{2}	1	1	1	0	0	1	0	0
{3}	0	1	1	1	0	1	0	0
{4}	0	0	1	1	0	1	0	0
{5}	1	0	0	0	1	0	0	1
{6}	0	1	1	1	0	1	1	0
{7}	0	0	0	0	0	1	1	0
{8}	1	0	0	0	1	0	0	1

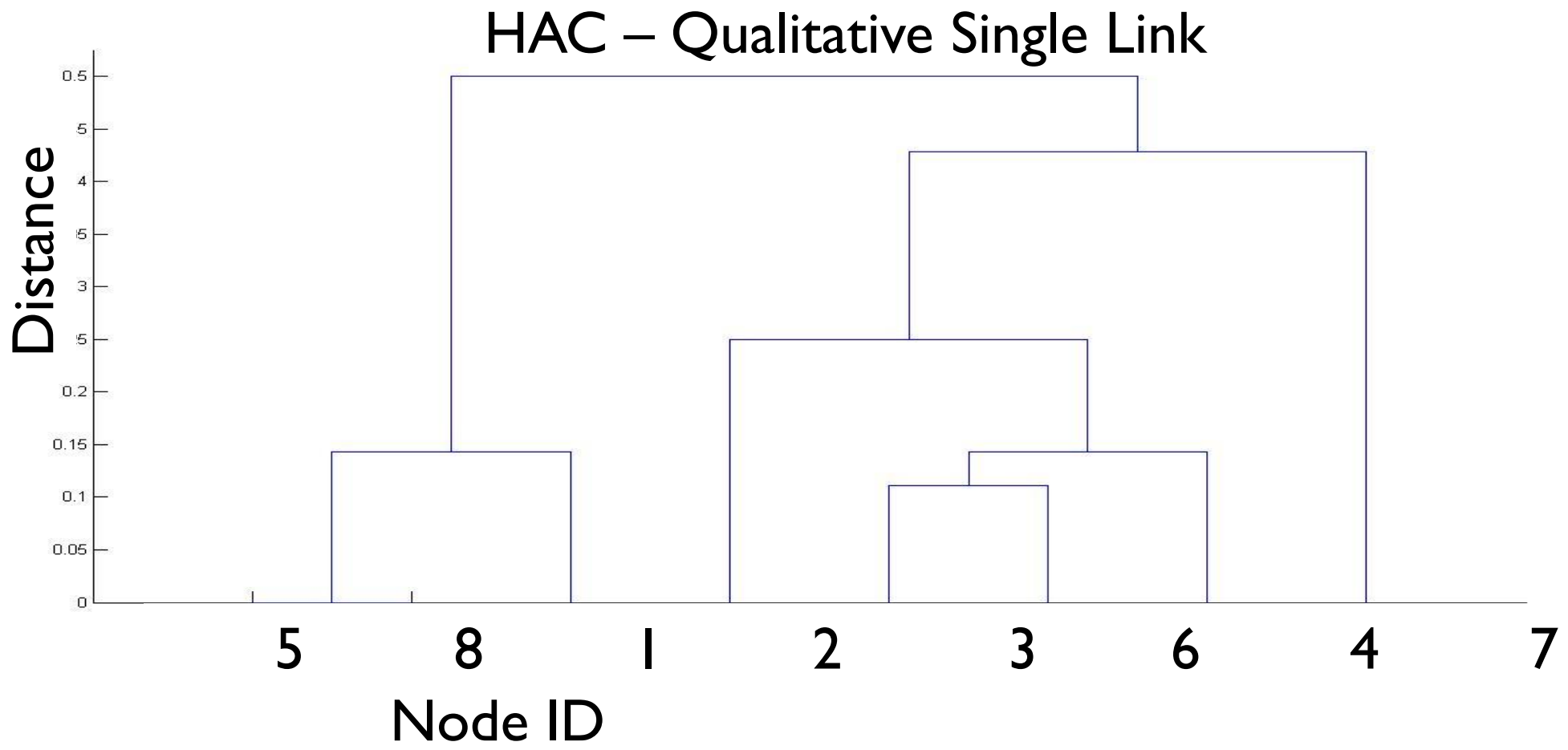
Resemblance matrix with qualitative data using SORENSON dissimilarity coefficients.

	{2}	{3}	{4}	{5}	{6}	{7}	{8}
{1}	0.5	0.75	1	0.143	0.778	1	0.143
{2}	–	0.25	0.429	0.714	0.333	0.667	0.714
{3}	–	–	0.143	1	0.111	0.667	1
{4}	–	–	–	1	0.25	0.6	1
{5}	–	–	–	–	1	1	0
{6}	–	–	–	–	–	0.429	1
{7}	–	–	–	–	–	–	1

$$\boldsymbol{Sorenson}: \boldsymbol{SIM}(a,b) = \frac{2\,M_{1-1}}{2\,M_{1-1} + M_{1-0} + M_{0-1}}$$

$$\boldsymbol{Dissimmilarity}: \boldsymbol{DSIM}(a,b) = 1 - \boldsymbol{SIM}(a,b)$$

Dendrogram: HAC Qualitative using single link



Beyond K – means

Reference: Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Thank You

tapankumarjain@gmail.com