UNSUPERVISED LEARNING

Sourav Karmakar

souravkarmakar29@gmail.com

UNSUPERVISED LEARNING

- The Data has no target attribute or class labels.
- We want to explore the data to find some intrinsic structures in them.
- Usually the objects / data are grouped into two or more groups based on the similarity or dissimilarity on a particular feature.
- Can produce completely different results based on the feature being used for grouping.

Grouping of objects into two or more groups based on the similarity / dissimilarities of objects such that each object fall into exactly one group is called **Clustering.**

SIMILARITY & DISSIMILARITY

How similar / dissimilar are following two objects?

- **Definition (Webster's Dictionary):**

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

- Similarity is hard to define but

"We know when we see it"

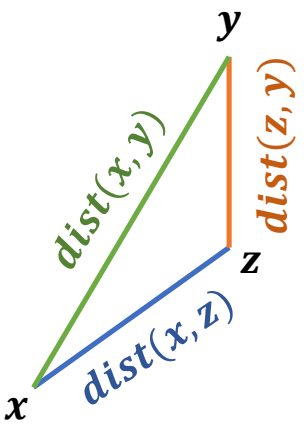
- The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

SIMILARITY & DISSIMILARITY

- **Similarity:**
  - Numerical measure of how similar two data /objects are.
  - Is higher when objects are more alike.
- **Dissimilarity:**
  - Numerical measure of how different two data /objects are.
  - Is lower when objects are more alike.
- So between two data objects if similarity increases then dissimilarity decreases and vice versa.
- Usually dissimilarity between two objects / data points can be defined in terms of the **distance** between those two objects / data points. For distant two objects are more is the dissimilarity and less is the similarity.
- There are different **distance** measures for both quantitative and categorical variables.
- The **definition** of *Distance function* or *Distance Metric* is following:

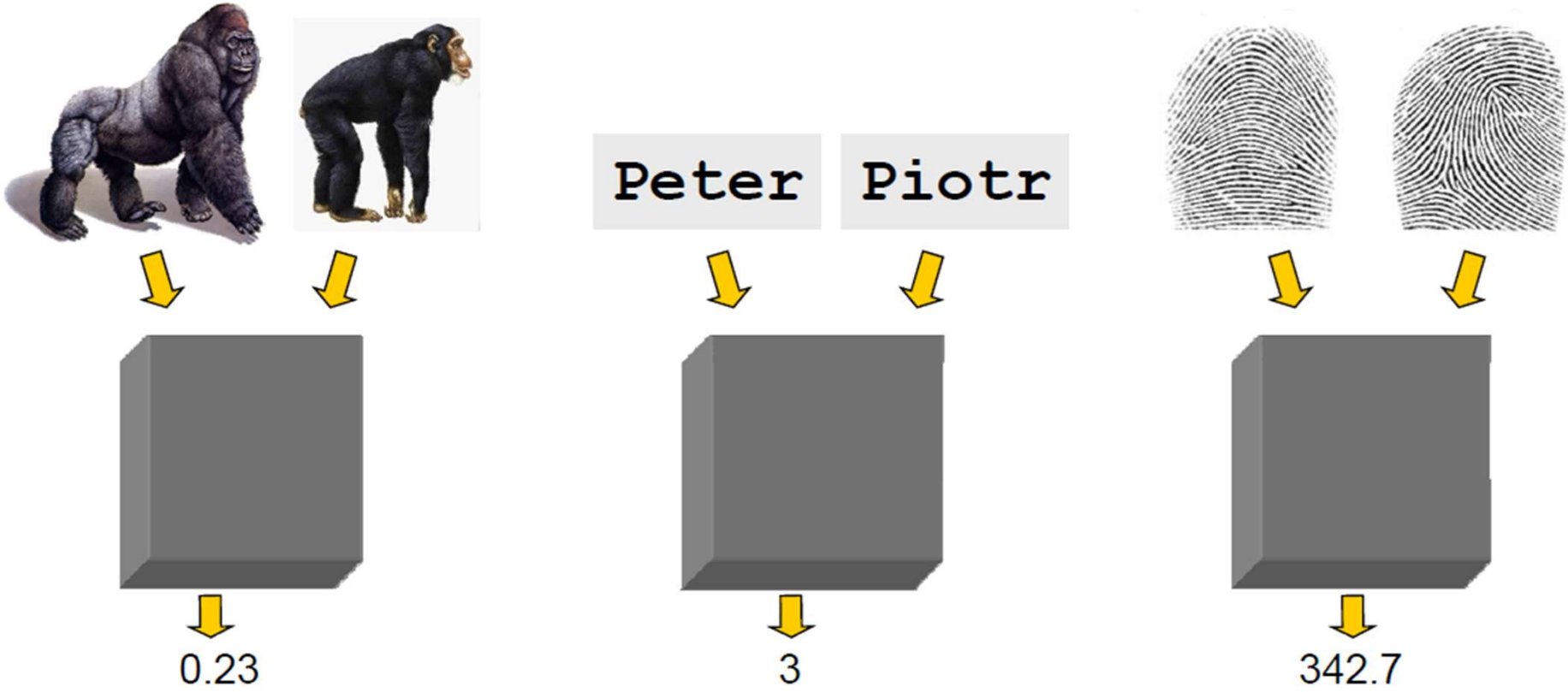
Let $\boldsymbol{x}, \boldsymbol{y}$ are vectors denoting two different objects, then $dist(\boldsymbol{x}, \boldsymbol{y})$ is a real number such that:

- $dist(\boldsymbol{x}, \boldsymbol{x}) = 0$
- $dist(\boldsymbol{x}, \boldsymbol{y}) = dist(\boldsymbol{y}, \boldsymbol{x})$ [Commutative property]
- For some object denoted by $\boldsymbol{z}$, $dist(\boldsymbol{x}, \boldsymbol{y}) \leq dist(\boldsymbol{x}, \boldsymbol{z}) + dist(\boldsymbol{z}, \boldsymbol{y})$ [Triangle inequality]



UNSUPERVISED LEARNING

Distance Measure for different objects



DISTANCE METRICS & SIMILARITY

- **Euclidean Distance:** For two data points denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$ the Euclidean distance is defined as:

$$dist_{euclidean}\,(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

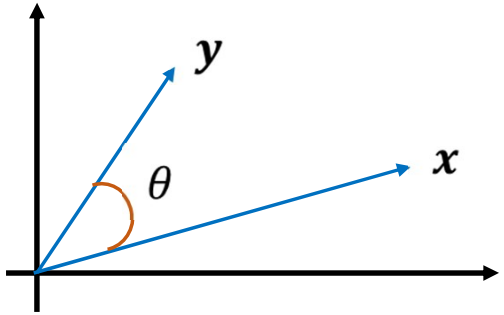
- **Manhattan Distance:** For two data points denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$ the Manhattan distance is defined as:

$$dist_{manhattan}\,(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n}|x_i - y_i|$$

- **Minkowski Distance:** For two data points denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$ the Minkowski distance is defined as:

$$dist_{minkowski}(x,y) = [\sum_{i=1}^{n}(x_i - y_i)^h]^{\left(\frac{1}{h}\right)}$$

- **Cosine Similarity:** For two data points denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$ the cosine similarity is defined as:



$$CosineSim(\boldsymbol{x},\boldsymbol{y}) = Cos(\angle\,\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

$$CosineDistance(\boldsymbol{x},\boldsymbol{y}) = 1 - CosineSim(\boldsymbol{x},\boldsymbol{y})$$

CLUSTERING ANALYSIS

- **What is Cluster Analysis?**

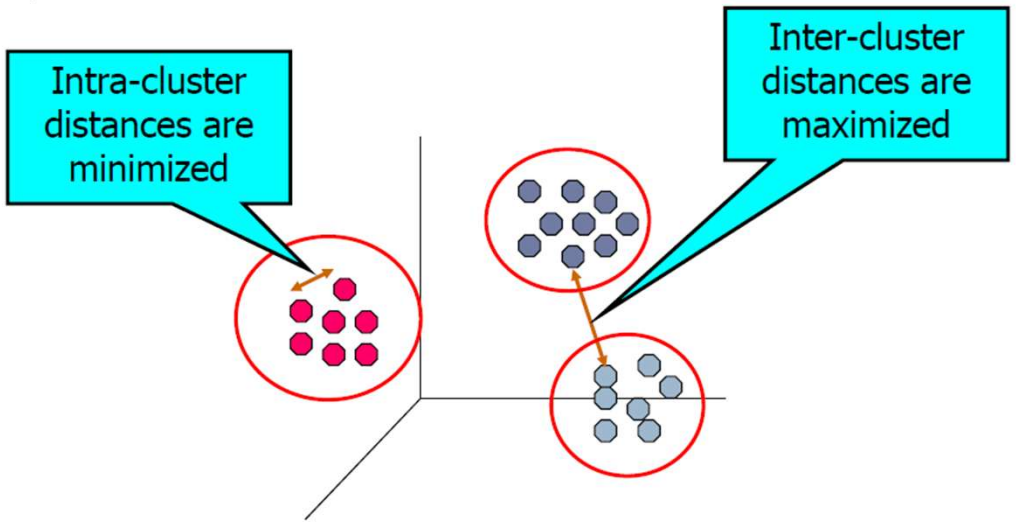
  Finding groups of objects in data such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups

- **Formal Definition**

  Let $X \subseteq \mathbb{R}^n$ be a dataset. A collection of subsets $\{C_1, C_2, C_3, \dots, C_k\}$; $C_i \subseteq X$ and $C_i \neq \emptyset \; \forall i$ is called a clustering of X if

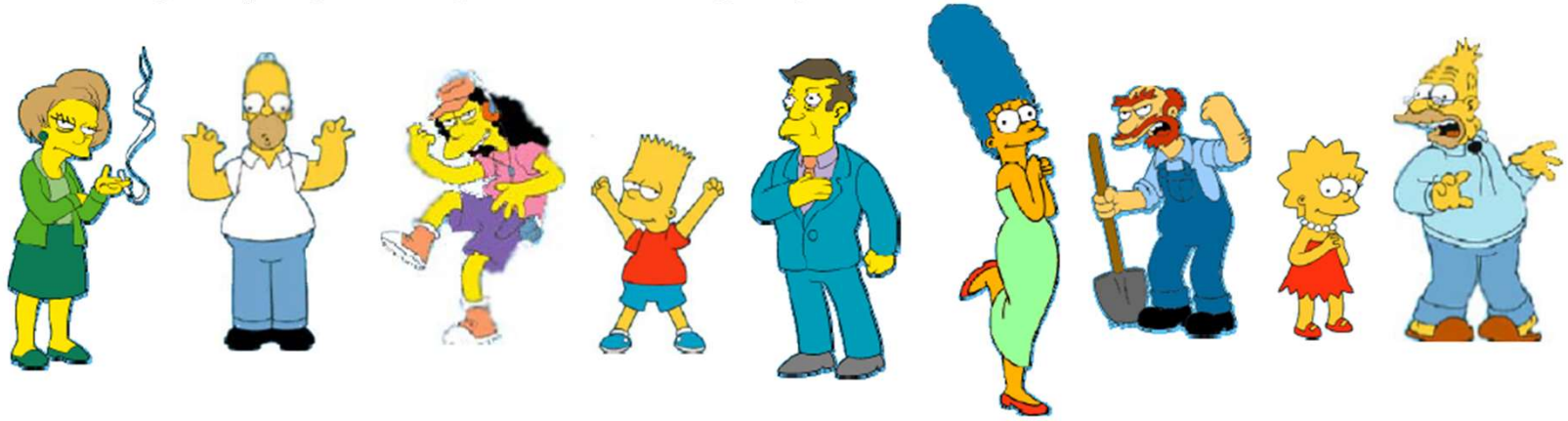
  - $C_i \cap C_j = \emptyset$ and
  - $\bigcup_{i=1}^{k} C_i = X$

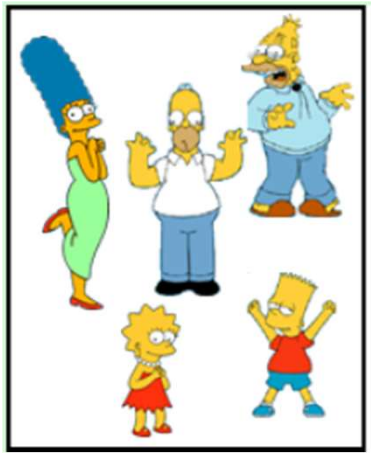
  Such that the **Inter-cluster distances** are maximized and **Intra-cluster distances** are minimized.



CLUSTERING – AN EXAMPLE

What is the natural grouping among the following objects



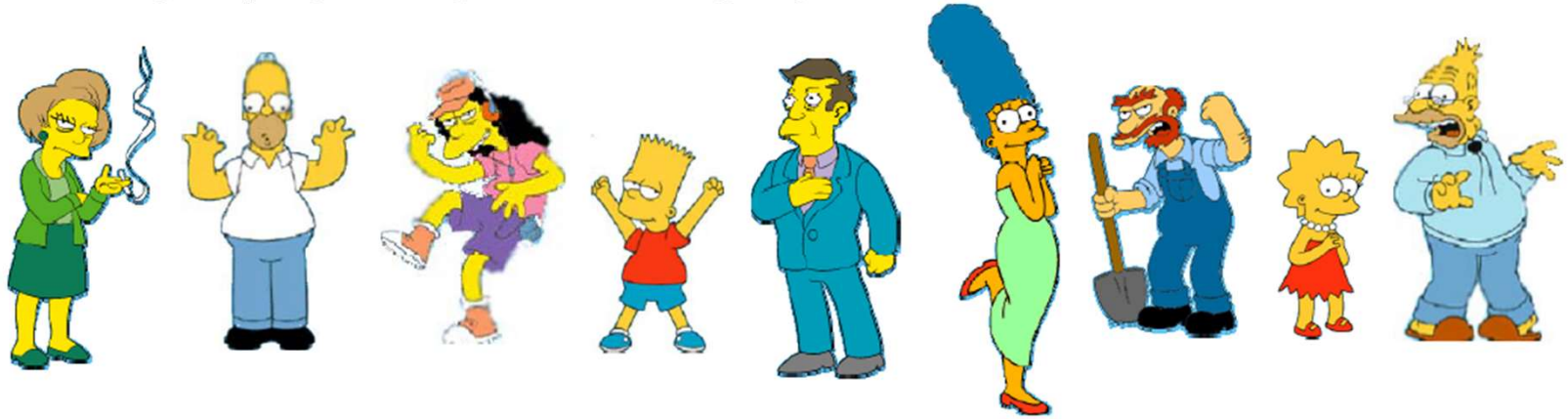


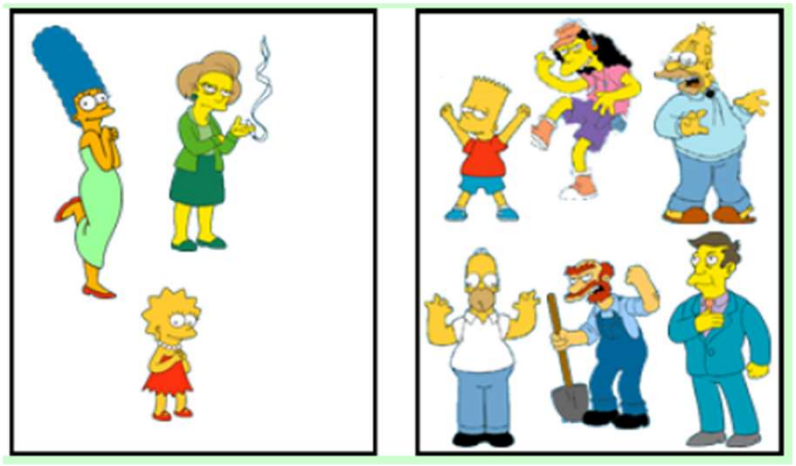
Simpson's Family

School Stuffs

CLUSTERING – AN EXAMPLE

What is the natural grouping among the following objects

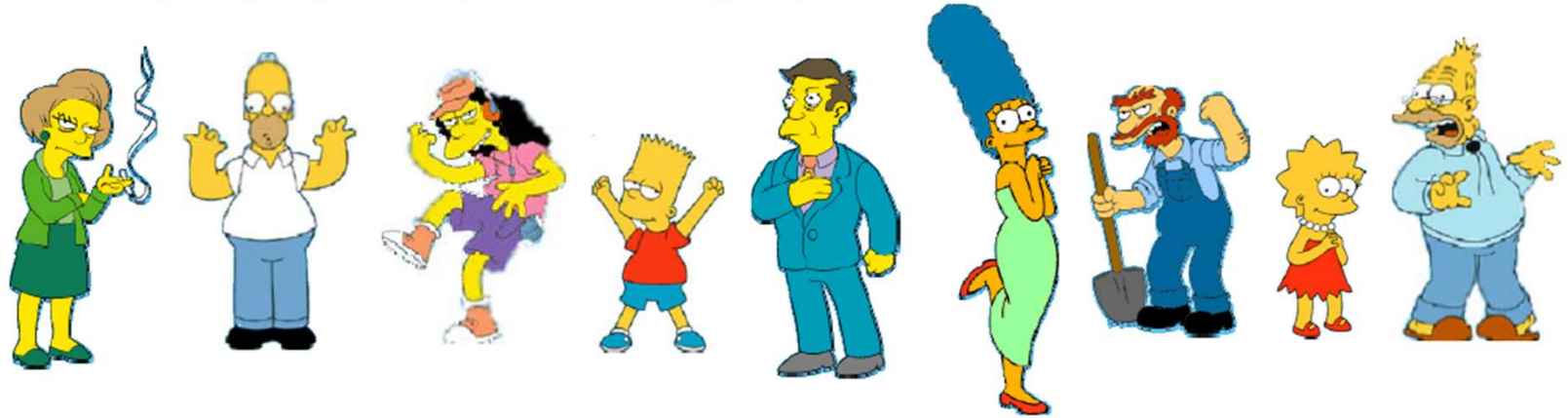


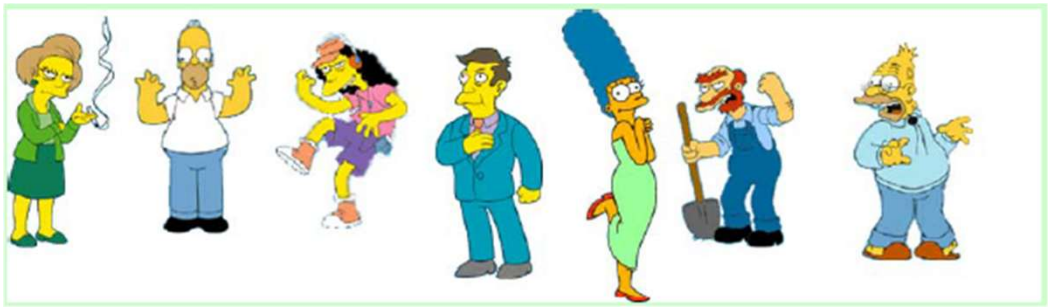


Females Males

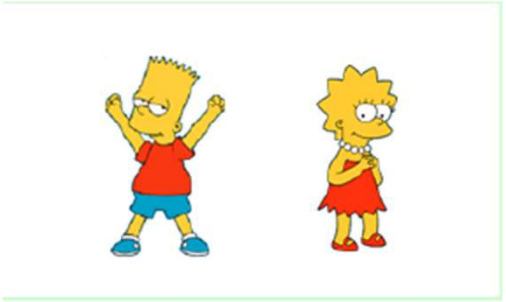
CLUSTERING – AN EXAMPLE

What is the natural grouping among the following objects





Adults



Children

Hence Clustering is Subjective to the choice of feature(s)

APPLICATIONS OF CLUSTERING

- **Customer Segmentation:** Grouping of customers based on their spending habits and other features.
- **Image Segmentation:** Break up the image into meaningful or perceptually similar regions.

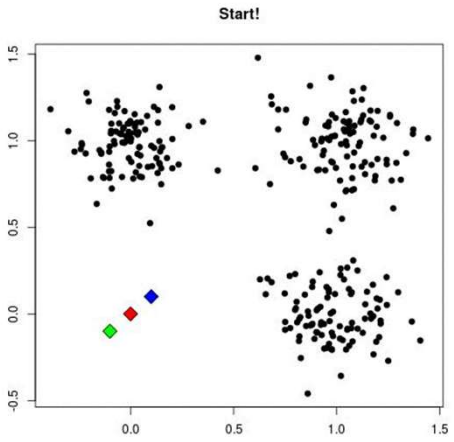
Original | Segmented Image | Original | Segmented

- **Social Network Analysis:** In the study of social networks, clustering may be used to recognize communities within large groups of people.
- **Medical Imaging:** On PET (Positron Emission Tomography) scans, cluster analysis can be used to differentiate between different types of tissue and blood in a three-dimensional image.
- **Anomaly Detection:** Anomalies often don't belong to any cluster of the data.

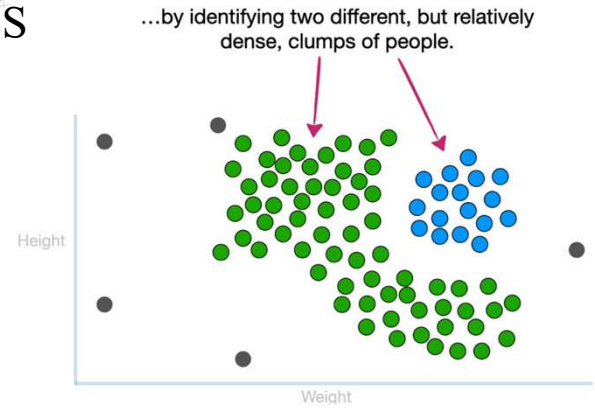
And Many More…

DIFFERENT TYPES OF CLUSTERING

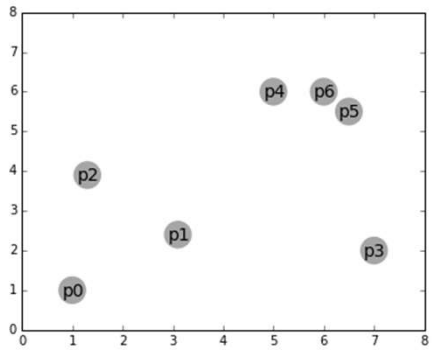
- **Partitional Clustering:**
  - K-Means
  - K-Medoids

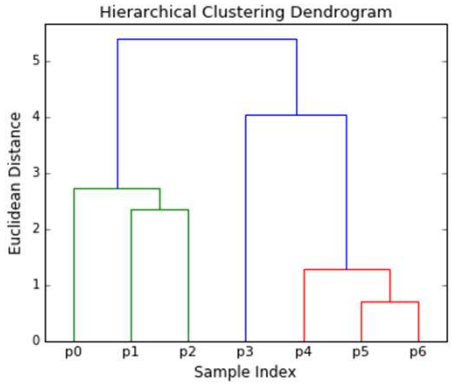


- **Density Based Clustering:**
  - DBSCAN
  - OPTICS

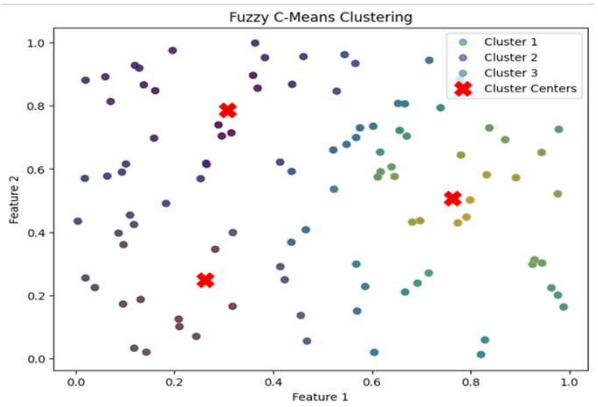


- **Hierarchical Clustering:**
  - Agglomerative
  - Divisive





- **Fuzzy Clustering:**
  - Fuzzy C-Means



METRICS OF CLUSTERING

- **Notations**
  - Dataset: $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$.
  - Clusters: $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$.
  - $c(i)$: cluster index containing x_i.
  - Dissimilarity (distance): $D(x_i, x_j) \geq 0$, often Euclidean.

- **Silhouette Score**

  The **Silhouette score** is a clustering validation measure that quantifies how well each data point lies within its cluster relative to other clusters. It ranges between -1 and $+1$ and can be applied with any dissimilarity measure (commonly Euclidean distance). Higher the value of Silhouette Score, better the clustering.

- **Other Clustering Validation Techniques**
  - **Dunn Index (DI):** Higher the DI, better the cluster (well separated and compact)
  - **Davis-Bouldin Index (DBI):** Lower the better.

POINT-WISE SILHOUETTE SCORE

For a point x_i:

1. **Cohesion Term:**

   The average distance to other points in the same cluster: $$a(i) = \frac{1}{|C_{c(i)}| - 1} \sum_{\substack{j \in C_{c(i)} \\ j \neq i}} D(x_i, x_j).$$

2. **Separation Term:**

   For another cluster $C_m \neq C_{c(i)}$: $$d(i, C_m) = \frac{1}{|C_m|} \sum_{j \in C_m} D(x_i, x_j).$$

   The nearest-cluster distance is: $$b(i) = \min_{m \neq c(i)} d(i, C_m).$$

3. **Silhouette Score of Point $'x_i'$:** $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad s(i) \in [-1, 1].$$

Interpretation of Silhouette Score:

- $s(i) \approx 1$: Well clustered, far from other cluster.
- $s(i) \approx 0$: on the border of two clusters.
- $s(i) < 0$: Likely mis-clustered.

CLUSTER AND OVERALL SILHOUETTE SCORE

Once point-wise Silhouette Score is calculated, one can calculate the cluster level and overall (dataset) silhouette scores.

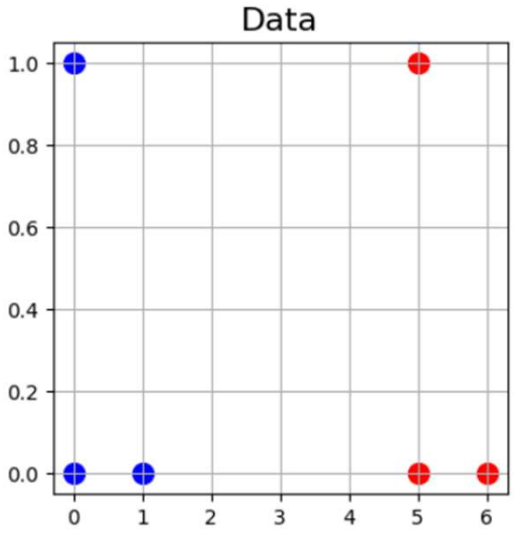
- **Cluster Level Silhouette Score:** $$S(C_m) = \frac{1}{|C_m|} \sum_{i \in C_m} s(i)$$

- **Overall Silhouette Score:** $$S = \frac{1}{n} \sum_{i=1}^{n} s(i).$$

Example Calculation of Silhouette Score:

Consider following two clusters in two-dimensions:

$$C_1 = \{(0,0), (0,1), (1,0)\}, \quad C_2 = \{(5,0), (5,1), (6,0)\}.$$



SILHOUETTE SCORE CALCULATION

Example Silhouette Score Calculation for $A_1 = (0, 0)$

$$a(A_1) = \frac{1+1}{2} = 1,$$

$$b(A_1) = \frac{5+\sqrt{26}+6}{3} \approx 5.366,$$

$$s(A_1) = \frac{5.366-1}{5.366} \approx 0.814.$$

Silhouette Score for all the points in the data:

Point	$a(i)$	$b(i)$	$s(i)$
A_1	1.000	5.366	0.814
A_2	1.207	5.394	0.776
A_3	1.207	4.374	0.724
B_1	1.000	4.700	0.787
B_2	1.207	4.741	0.745
B_3	1.207	5.694	0.788

Overall Silhouette Score:

$$S = \frac{1}{6}\sum_{i} s(i) \approx 0.772$$

Thank You