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.

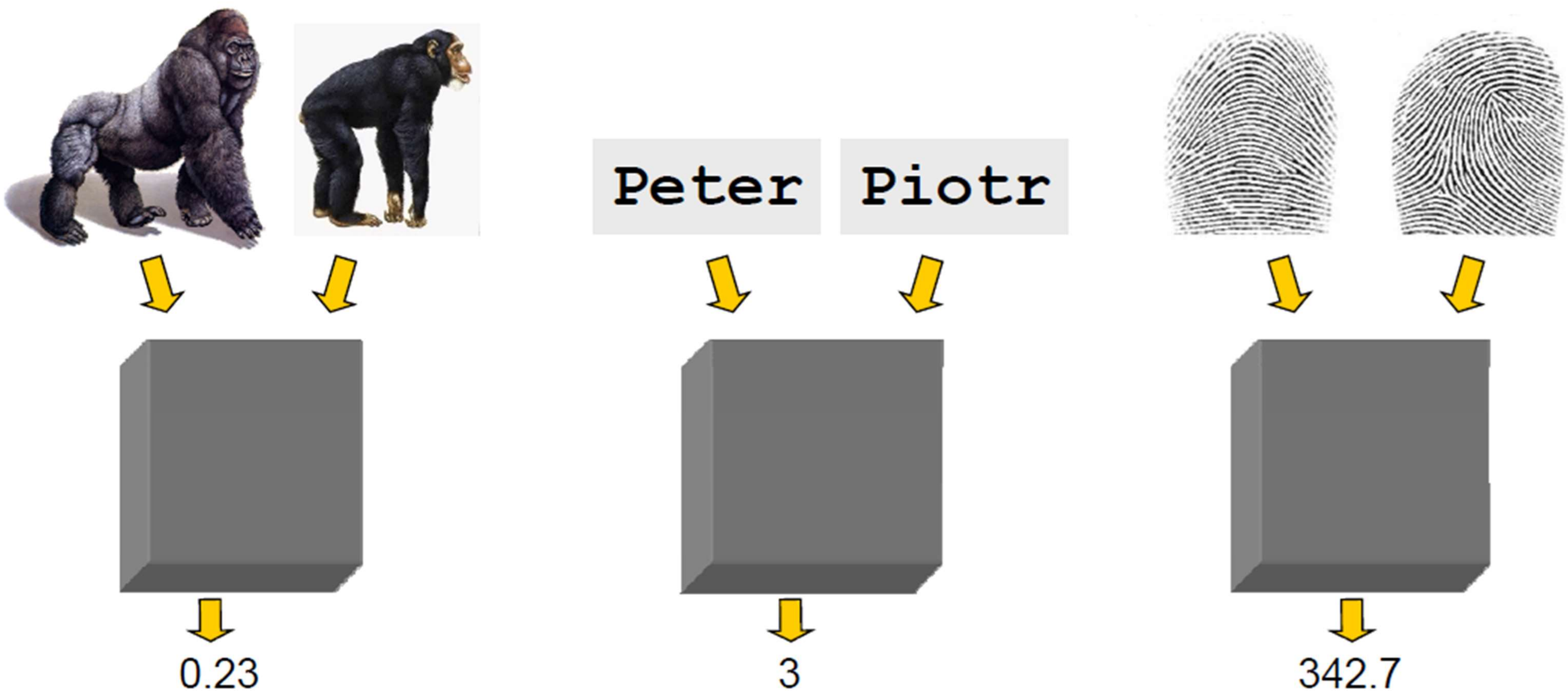
SIMILARITY & DISSIMILARITY

- **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 METRICS & SIMILARITY

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

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

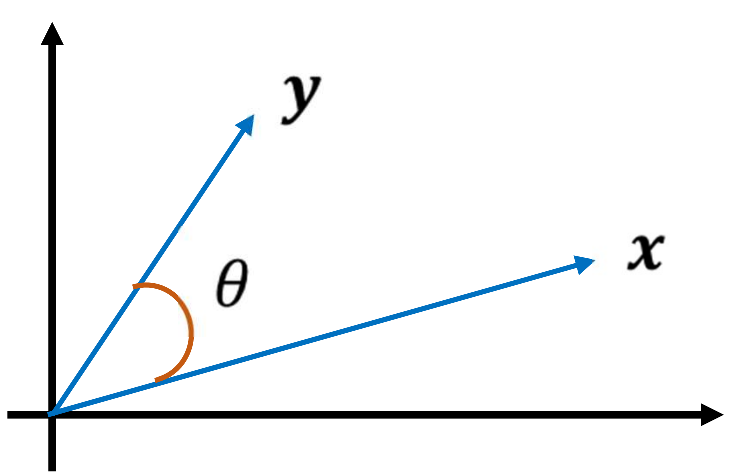
- **Manhattan Distance:** For two data points denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$ the Manhattan distance is defined as:

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

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

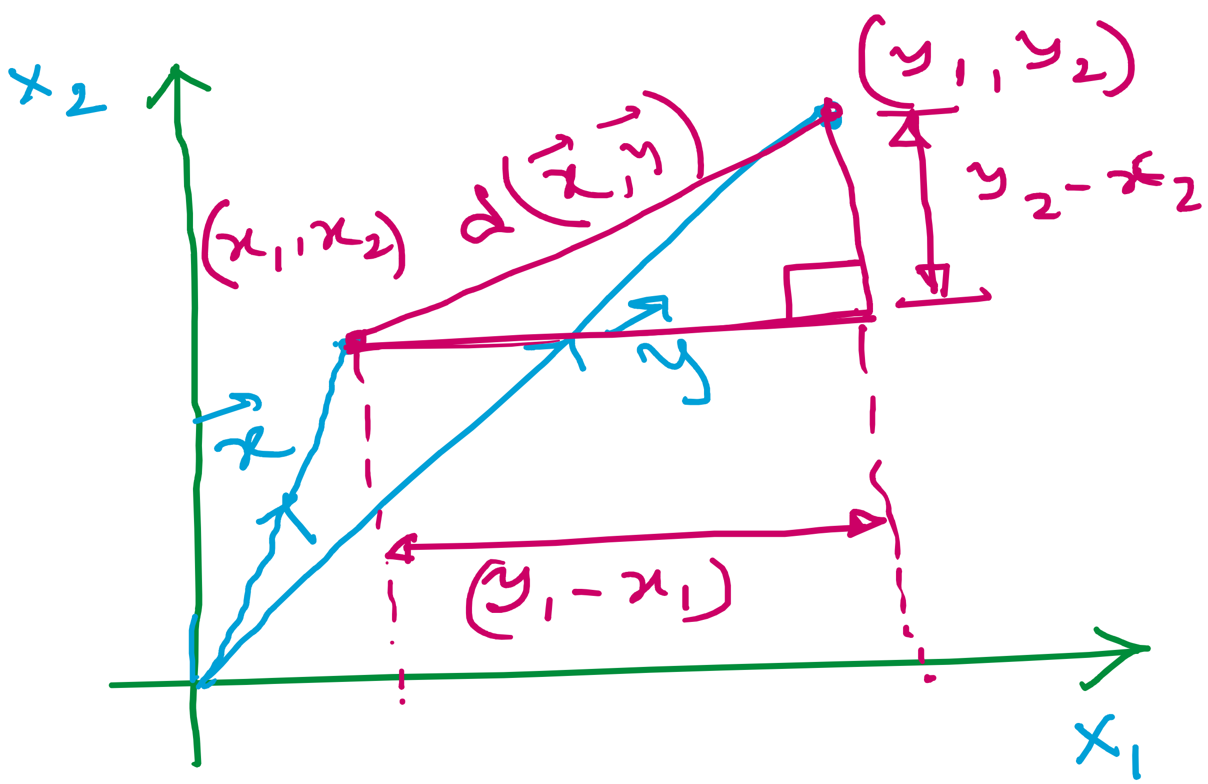
$$dist_{minkowski}\,(\boldsymbol{x},\boldsymbol{y}) = [\sum_{i=1}^{n}(x_i - y_i)^h]^{\left(\frac{1}{h}\right)}$$

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



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

Euclidean Distance :-



$$d_{euclidean}(\vec{x}, \vec{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

In two-dimension

Suppose I have n-dimensional case.

$\vec{x} \longrightarrow (x_1, x_2, x_3, \ldots, x_n)$

$\vec{y} \longrightarrow (y_1, y_2, y_3, \ldots, y_n)$

$$d_{euclidean}(\vec{x}, \vec{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

$\vec{x}$	$\vec{y}$	$\vec{x}-\vec{y}$
x_1	y_1	$x_1 - y_1$
x_2	y_2	$x_2 - y_2$
x_3	y_3	$x_3 - y_3$
x_4	y_4	$x_4 - y_4$
$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$x_n - y_n$

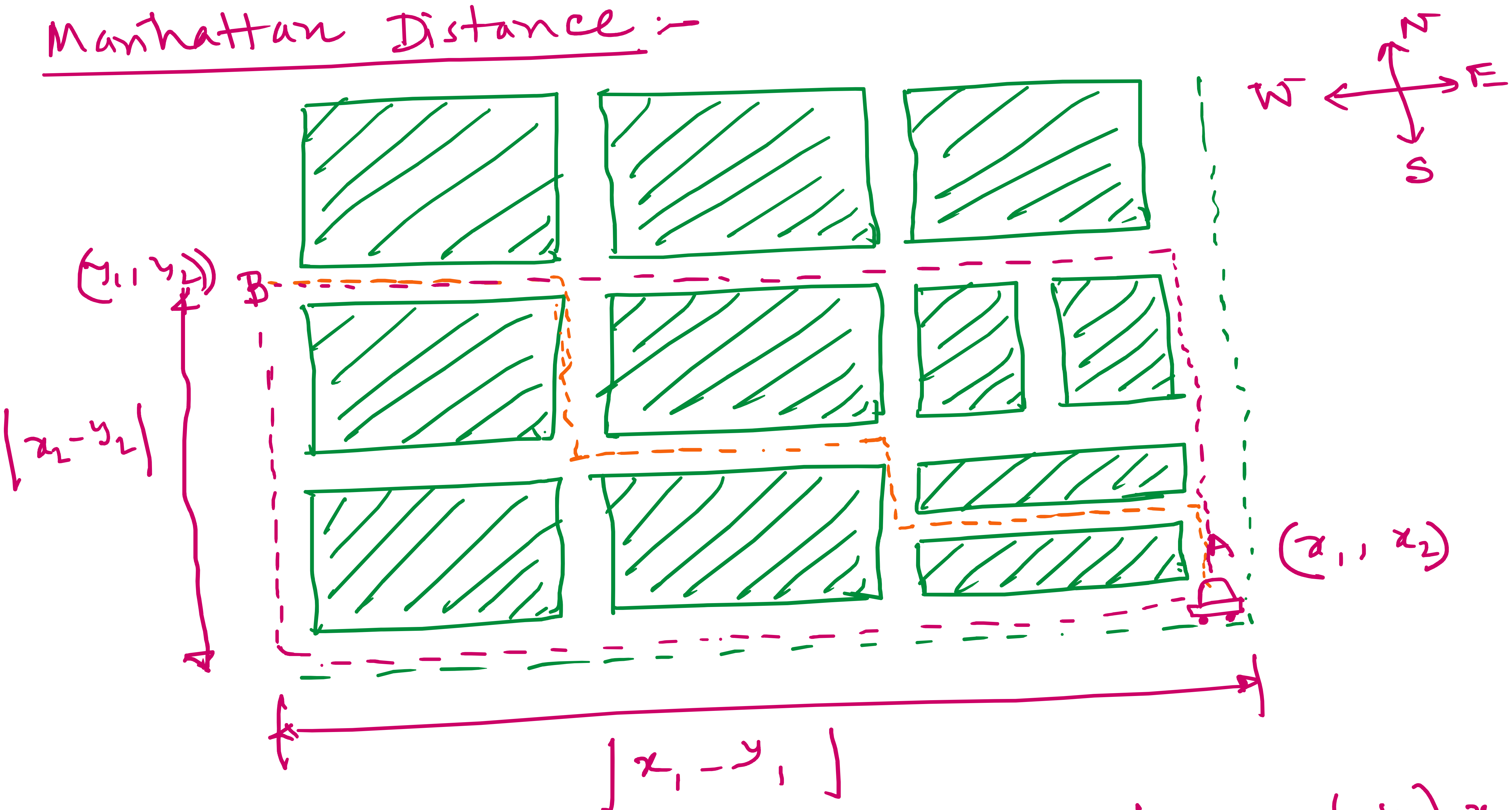
$$\underbrace{(\vec{x}-\vec{y})}_{\vec{z}}^T \underbrace{(\vec{x}-\vec{y})}_{\vec{z}} \qquad \vec{z}^T\vec{z} \qquad \vec{z}\cdot\vec{z}$$

$$(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_n-y_n)^2$$

$$d_{euclidean} = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_n-y_n)^2}$$

$$\therefore\ d_{euclidean}(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T(\vec{x}-\vec{y})}$$

Manhattan Distance :-



$$|x_1 - y_1| + |x_2 - y_2|$$

$$\sum_{i=1}^{n} |x_i - y_i| = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

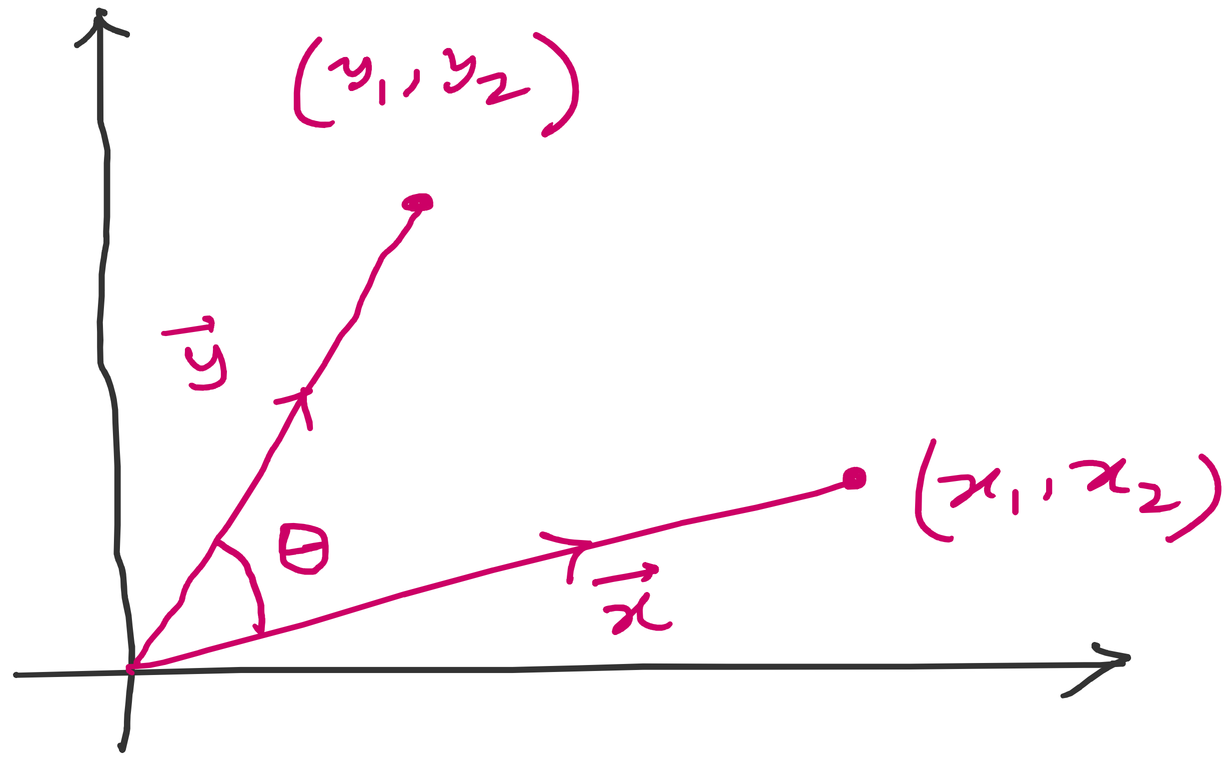
Cosine Similarity

$$\vec{x}^T \vec{y} \text{ or } \vec{x} \cdot \vec{y}$$

$$= (x_1 y_1 + x_2 y_2)$$

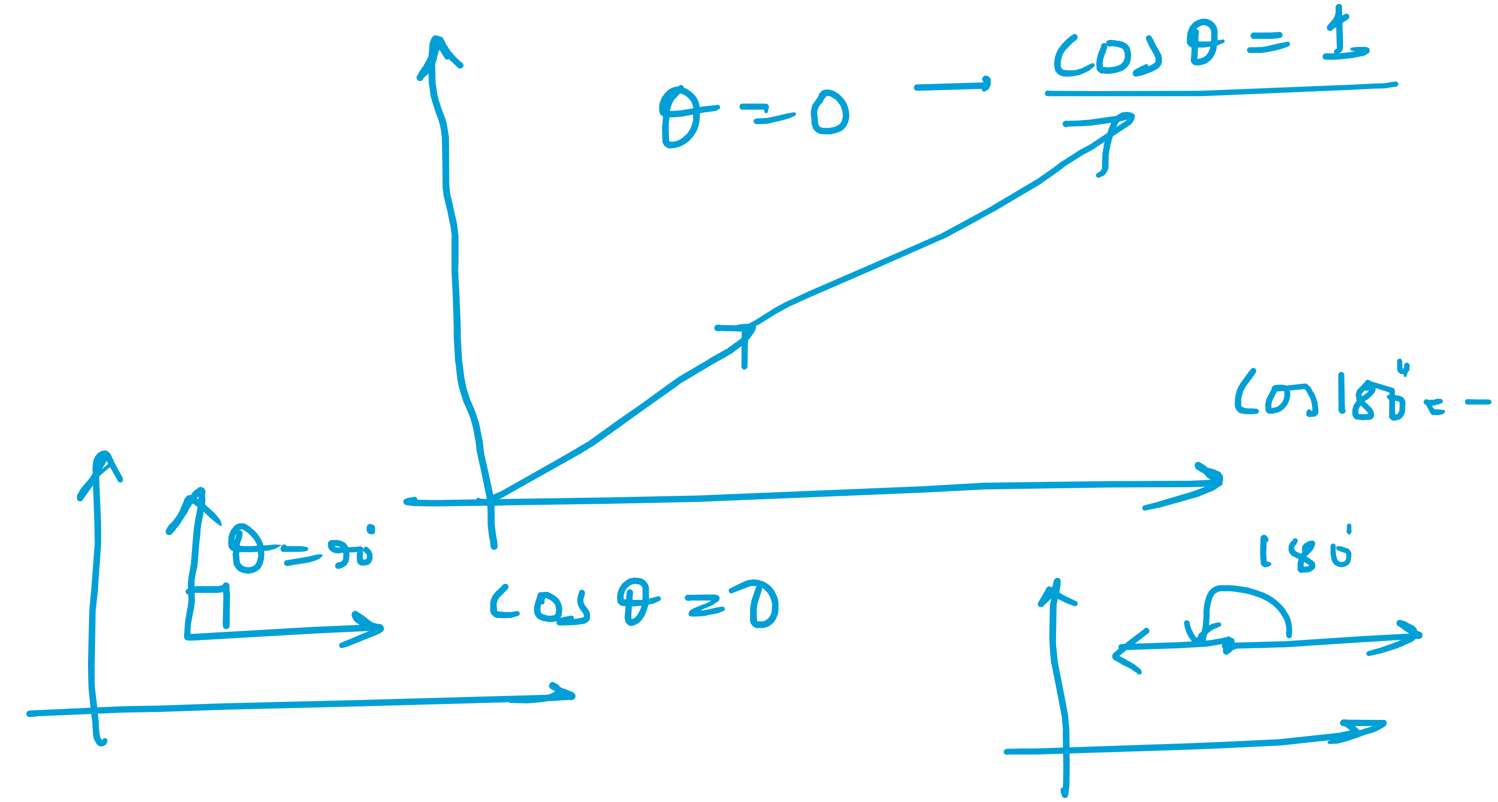
$$\|\vec{x}\| = \sqrt{x_1^2 + x_2^2}$$

$$\|\vec{y}\| = \sqrt{y_1^2 + y_2^2}$$



$$\vec{x} \cdot \vec{y} = \|\vec{x}\| \cdot \|\vec{y}\| \cdot \underline{\cos\theta}$$

$$\Rightarrow \quad \cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \, \|\vec{y}\|}$$



CLUSTERING ANALYSIS

- **What is Cluster Analysis?**

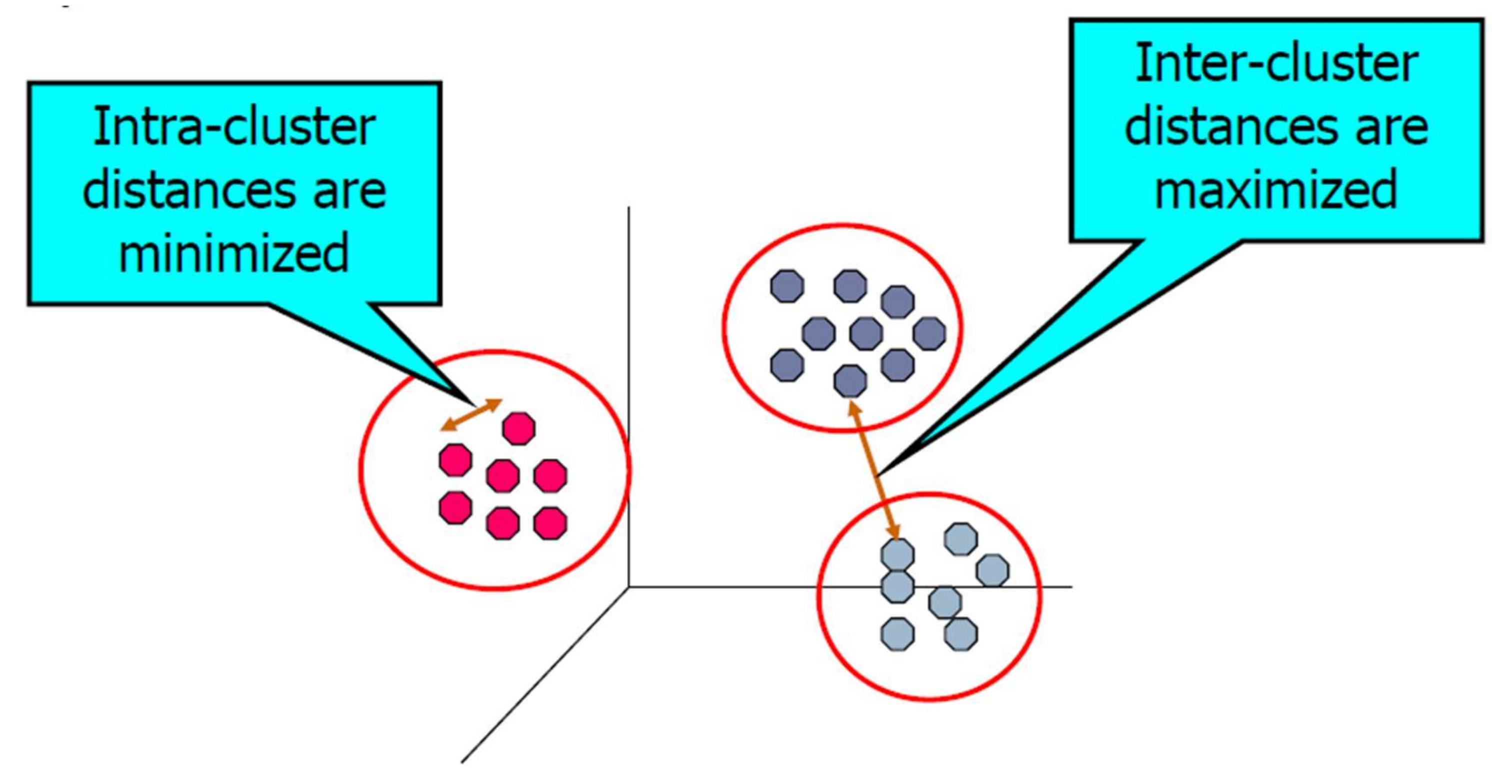
  Finding groups of objects in data such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups

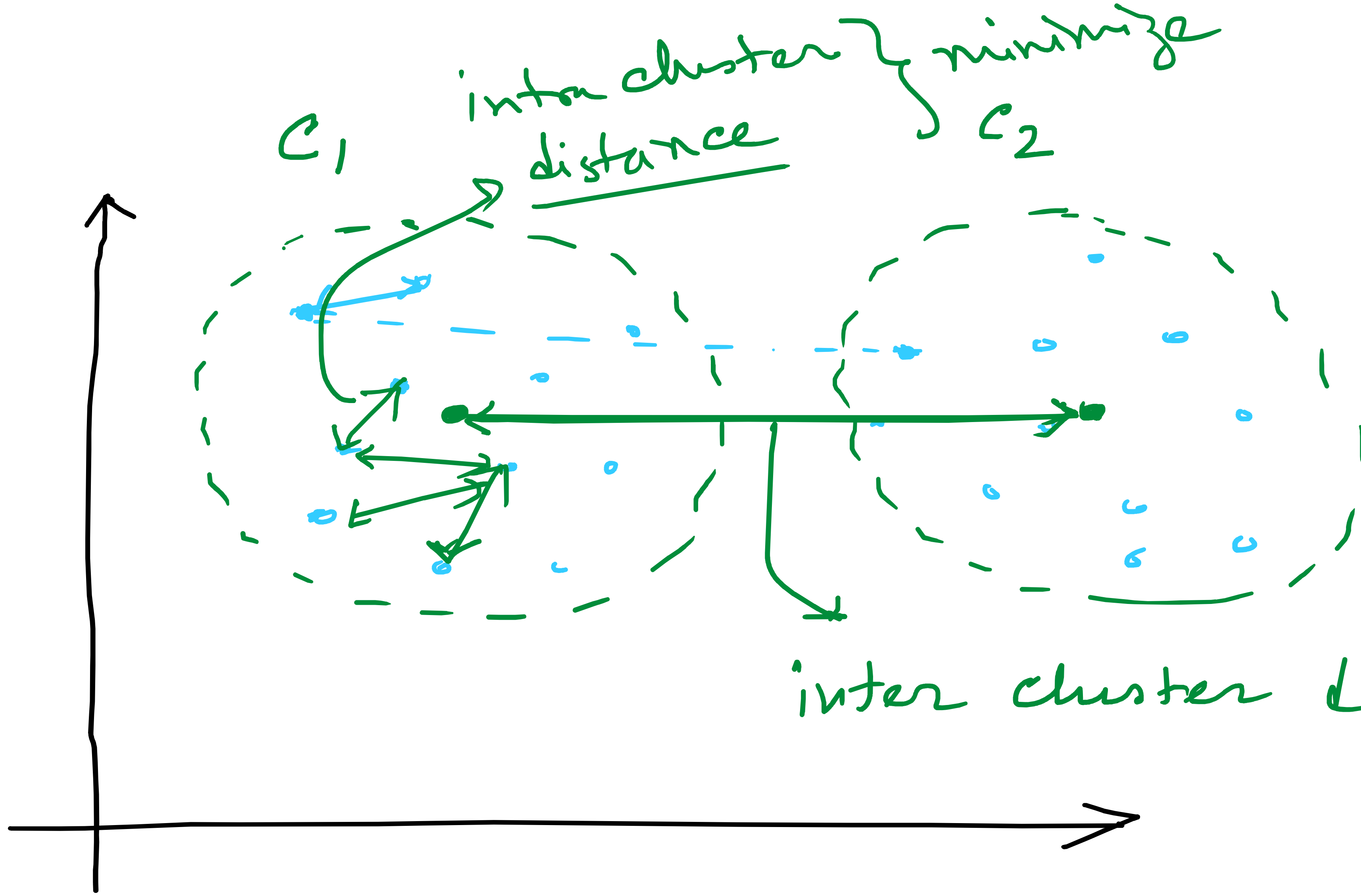
- **Formal Definition**

  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.





$C_1 \cup C_2 = X$

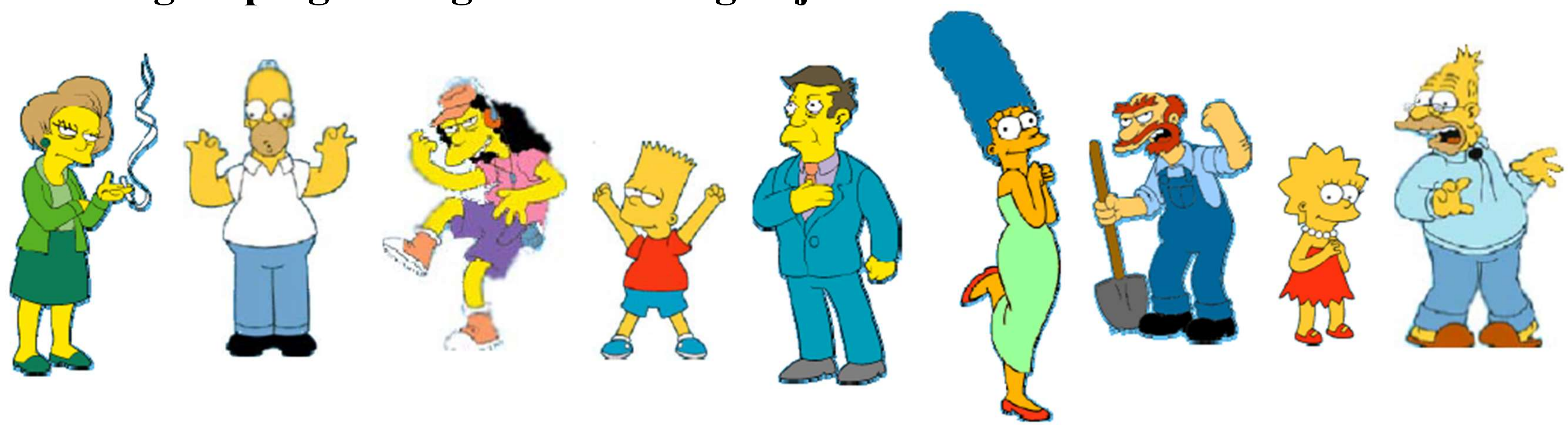
$C_1 \cap C_2 = \phi$

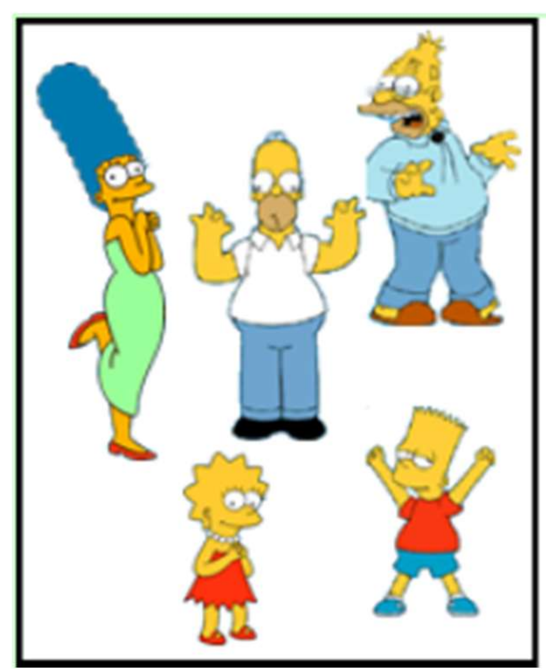
intra cluster distance $\downarrow$ $\longrightarrow$ intra cluster similarity $\uparrow$

inter cluster distance $\uparrow$ $\longrightarrow$ inter cluster similarity $\downarrow$

CLUSTERING – AN EXAMPLE

What is the natural grouping among the following objects



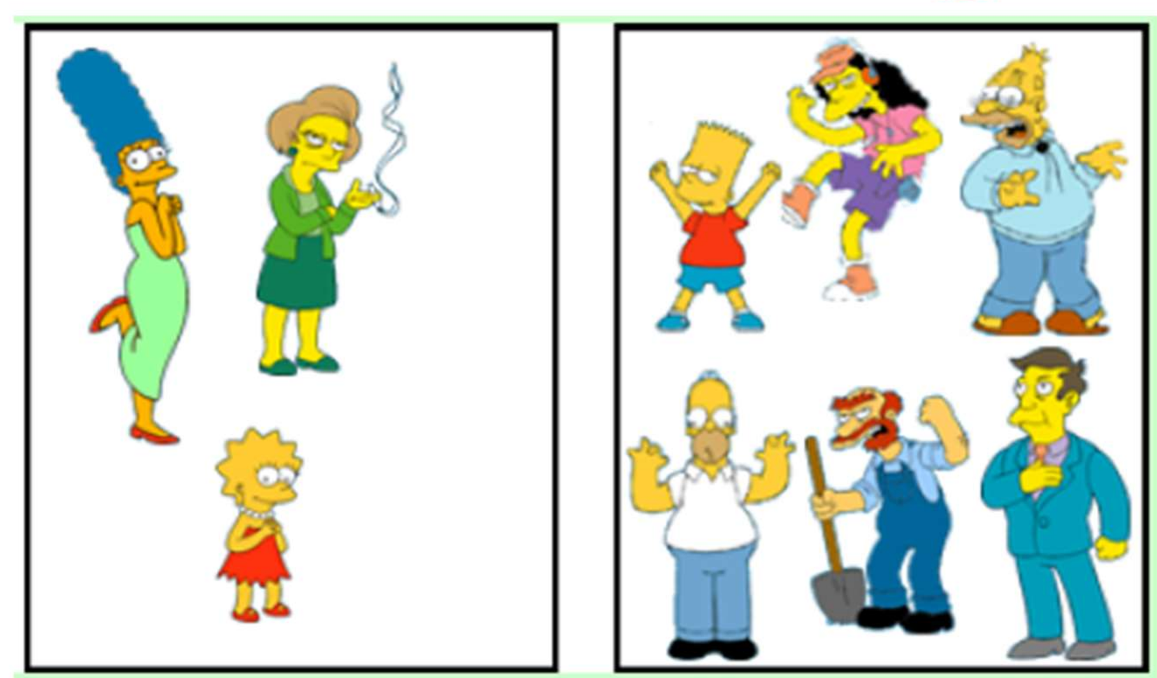


Simpson's Family

School Stuffs

CLUSTERING – AN EXAMPLE

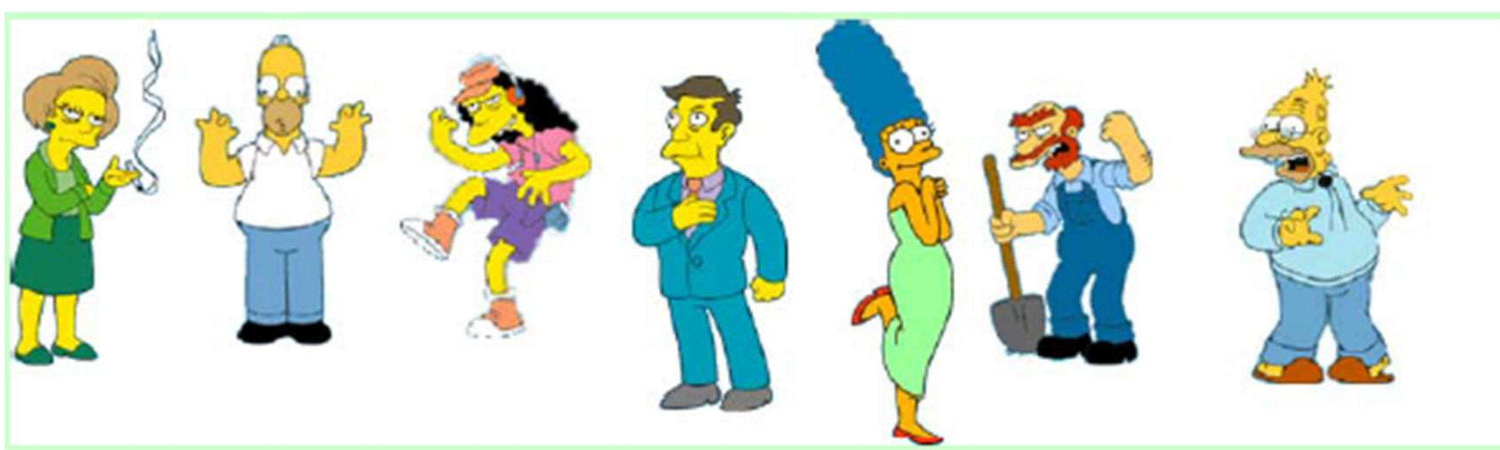
What is the natural grouping among the following objects



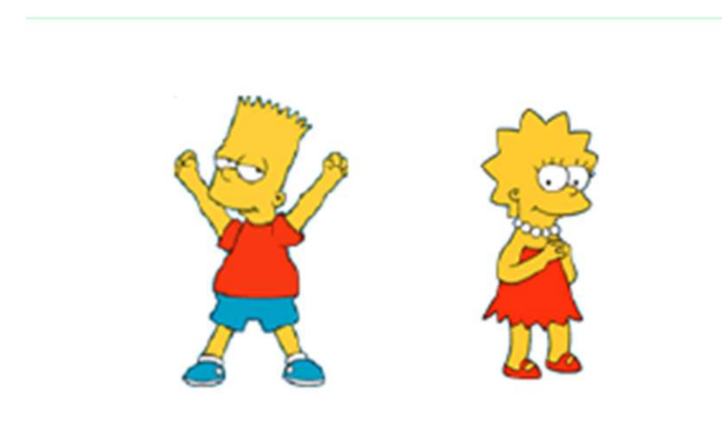
Females Males

CLUSTERING – AN EXAMPLE

What is the natural grouping among the following objects



Adults



Children

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

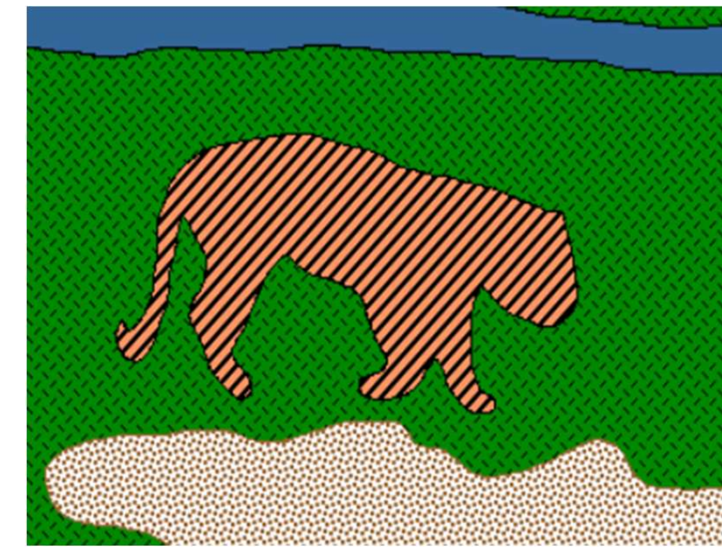
APPLICATIONS OF CLUSTERING

- **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.

And Many More…

DIFFERENT TYPES OF CLUSTERING

- There are several kinds of Clustering algorithms
- **Partitional Clustering:**
  - K-Means
  - K-Medoids
- **Hierarchical Clustering:**
  - Agglomerative
  - Divisive
- **Density Based Clustering:**
  - DBSCAN
  - OPTICS
- **Fuzzy Clustering:**
  - Fuzzy C-Means

Thank You