

SNS COLLEGE OF TECHNOLOGY

Coimbatore

An Autonomous Institution



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CST203 - DATA ANALYTICS

II YEAR /IV SEMESTER

Unit 3- CLUSTERING

SUPERVISED LEARNING



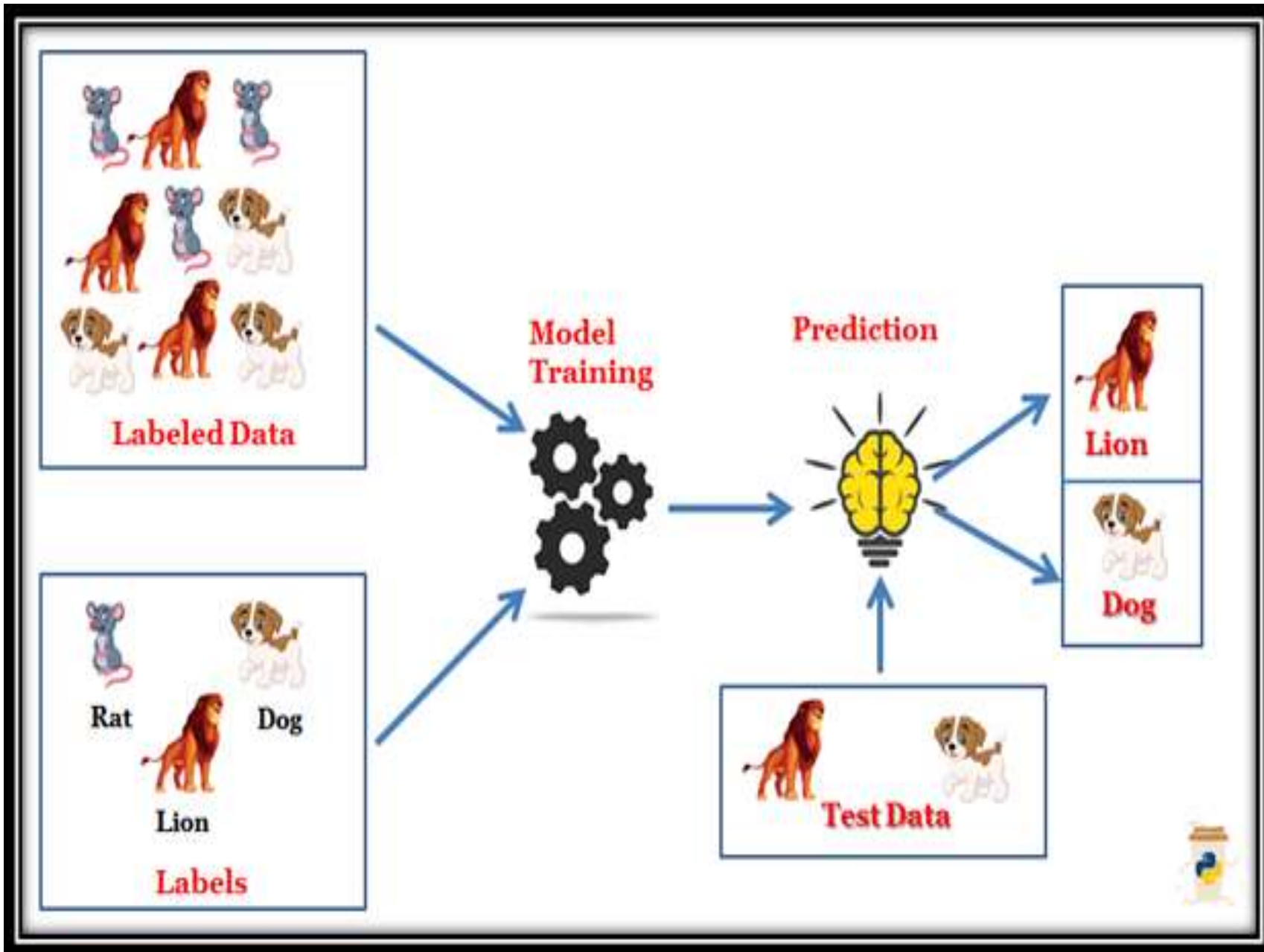
UNSUPERVISED LEARNING



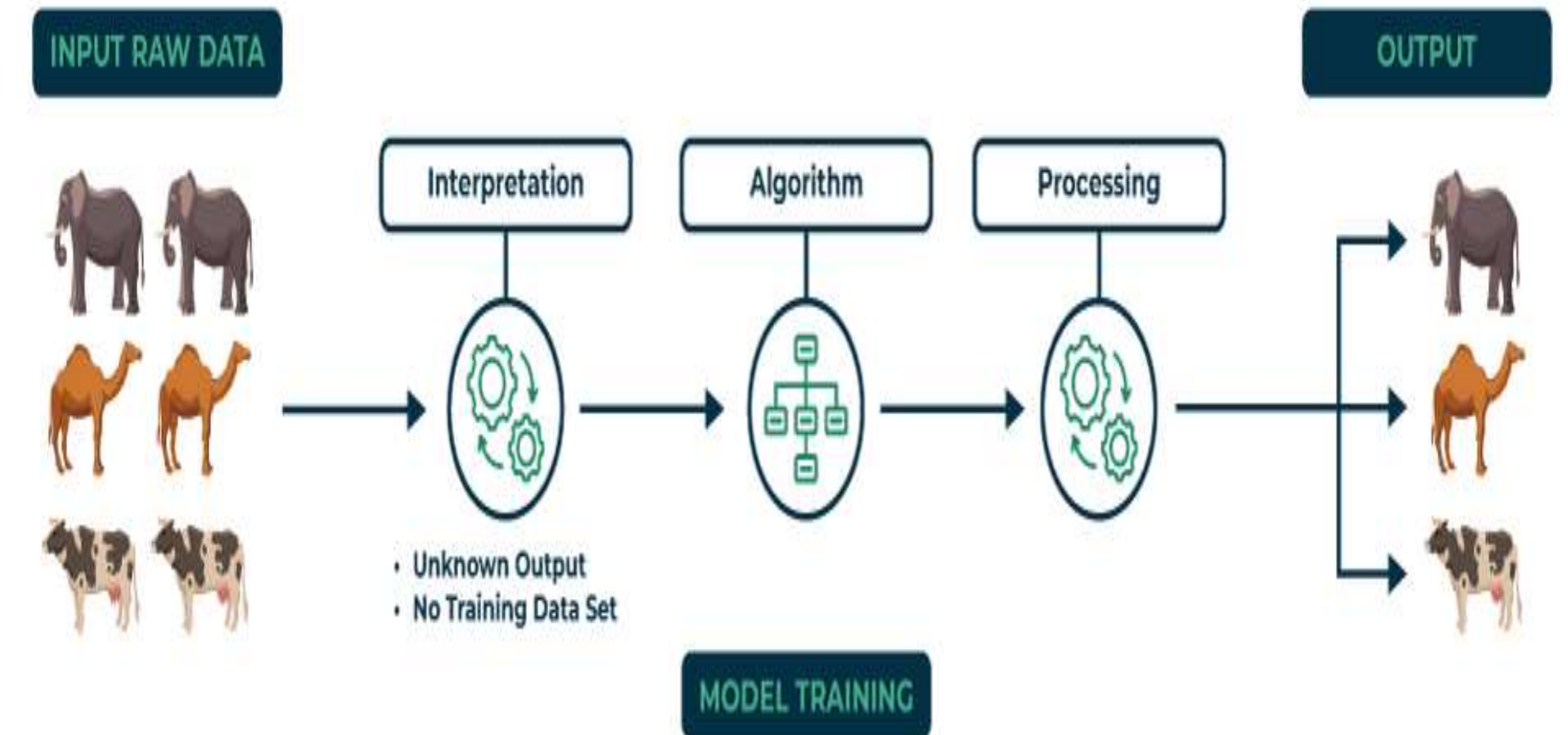
REINFORCEMENT LEARNING



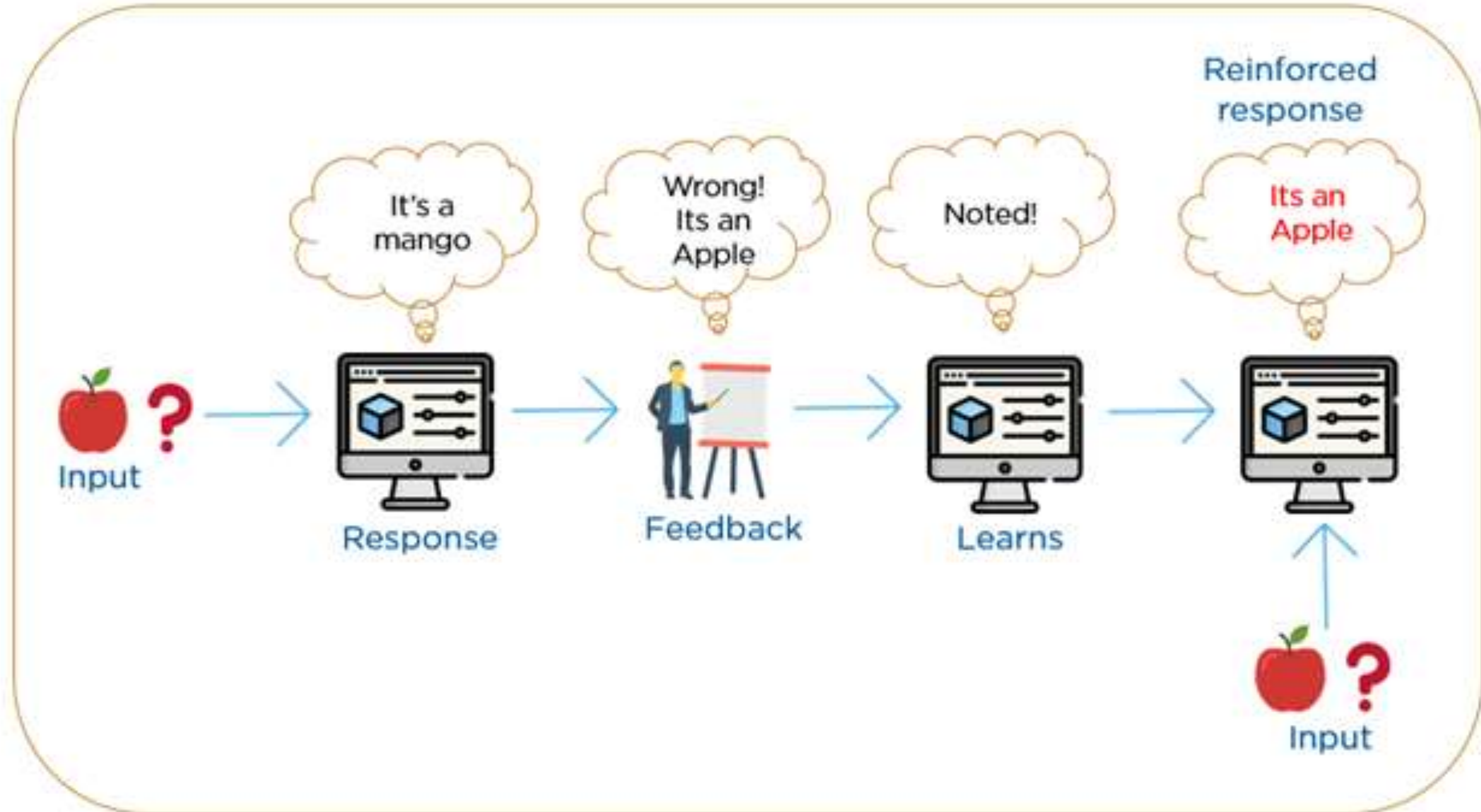
SUPERVISED LEARNING

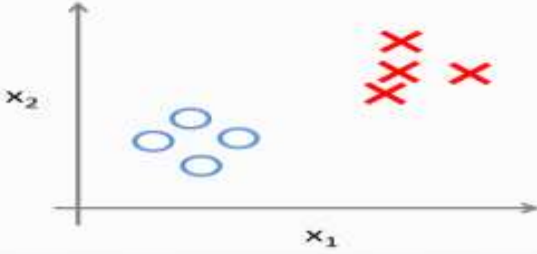
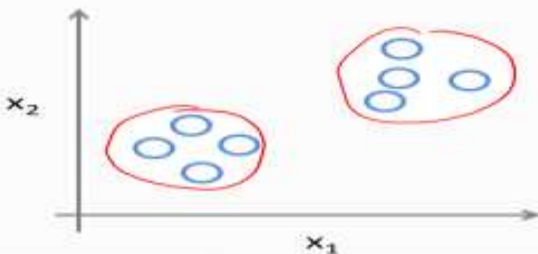


Unsupervised Learning



REINFORCEMENT LEARNING



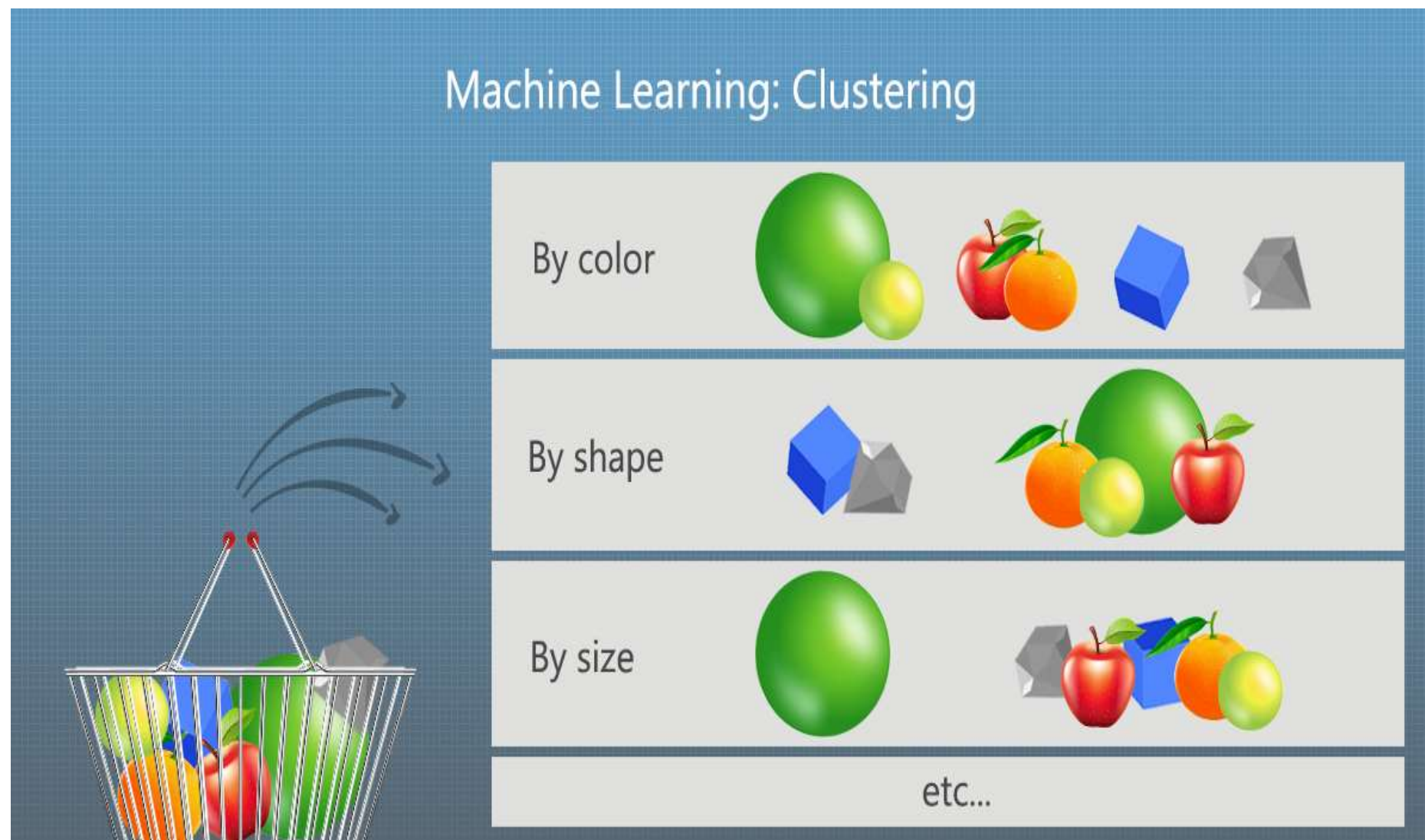
Supervised	Unsupervised
Input Data is labelled	Input Data is Unlabelled
Uses training Dataset	Uses just input dataset
Data is classified based on training dataset	Uses properties of given data to classify it.
Used for prediction	Used for Analysis
Divided into two types Regression & Classification	Divided into two types Clustering & Association
Known number of classes	Unknown number of classes
	
Use off-line analysis of data	Use Real-Time analysis of data

Brain Storming

1. What is Clustering?

What is a Clustering?

- ❖ **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).



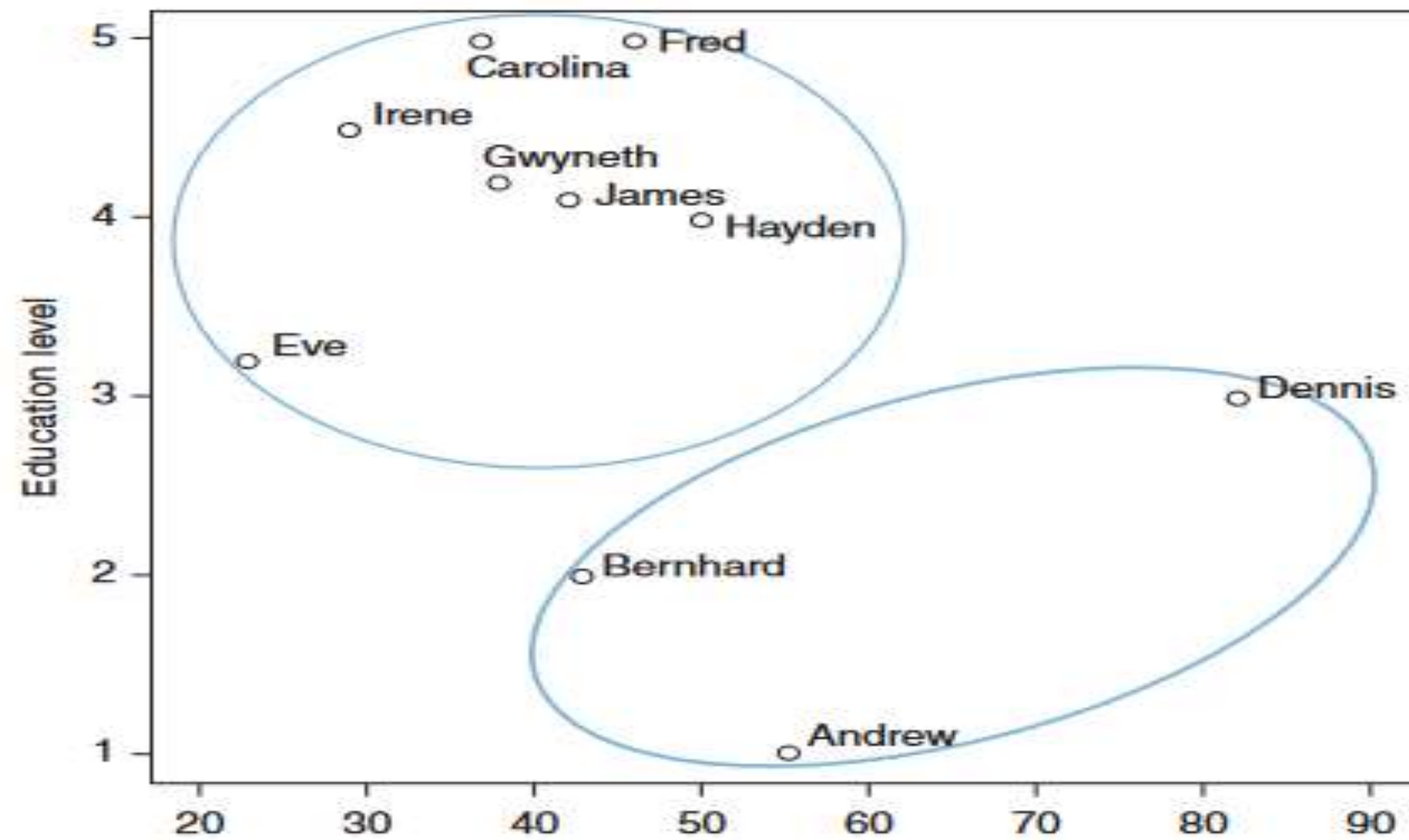
Clustering

- ❖ **Hard clustering** - grouping the data items such that each item is only assigned to one cluster
- ❖ **Soft clustering** - clustering in which each data point can belong to more than one cluster.

Clustering

Table 5.1 Simple social network data set.

Name	Age	Educational level
Andrew (A)	55	1
Bernhard (B)	43	2
Carolina (C)	37	5
Dennis (D)	82	3
Eve (E)	23	3.2
Fred (F)	46	5
Gwyneth (G)	38	4.2
Hayden (H)	50	4
Irene (I)	29	4.5
James (J)	42	4.1



Distance Measures

- ❖ what is similar
- ❖ what is not similar

The most similar objects have the smallest distances between them, and the most dissimilar have the largest distances

Differences between Values of Common Attribute Type

The difference between two values for the same attribute, here named **a** and **b**, will be denoted as **$d(a, b)$** .

1) For quantitative attributes, one can calculate the absolute difference:

$$d(a, b) = |a - b|$$

For example, the difference in age between Andrew ($a = 55$) and Carolina ($b = 37$) is **$|55 - 37| = 18$** . Note, that even if we change the order of the values ($a = 37$ and $b = 55$) the result is the same.

2) If the attribute type is qualitative, we use distance measures suitable for the given type. If the qualitative attribute has ordinal values, we can measure the difference in their positions as:

$$d(a, b) = (|\text{pos } a - \text{pos } b|) / (n - 1)$$

where n is the number of different values, and $\text{pos } a$ and $\text{pos } b$ are the positions of the values a and b , respectively, in a ranking of possible values

EX: In our data set, education level can be considered an ordinal attribute, with larger values meaning a higher level of education. Thus the distance between the education levels of Andrew and Caroline is

$$|\text{pos1} - \text{pos5}|/4 = |1 - 5|/4 = 1.$$

If a qualitative attribute has nominal values, in order to compute the distance between two values we simply determine if they are equal (in which case the difference, or dissimilarity, will be zero) or not (in which case the difference will be one).

$$\mathbf{d(a, b) = \begin{cases} 1, & \text{if } a \neq b \\ 0, & \text{if } a = b \end{cases}}$$

Clustering

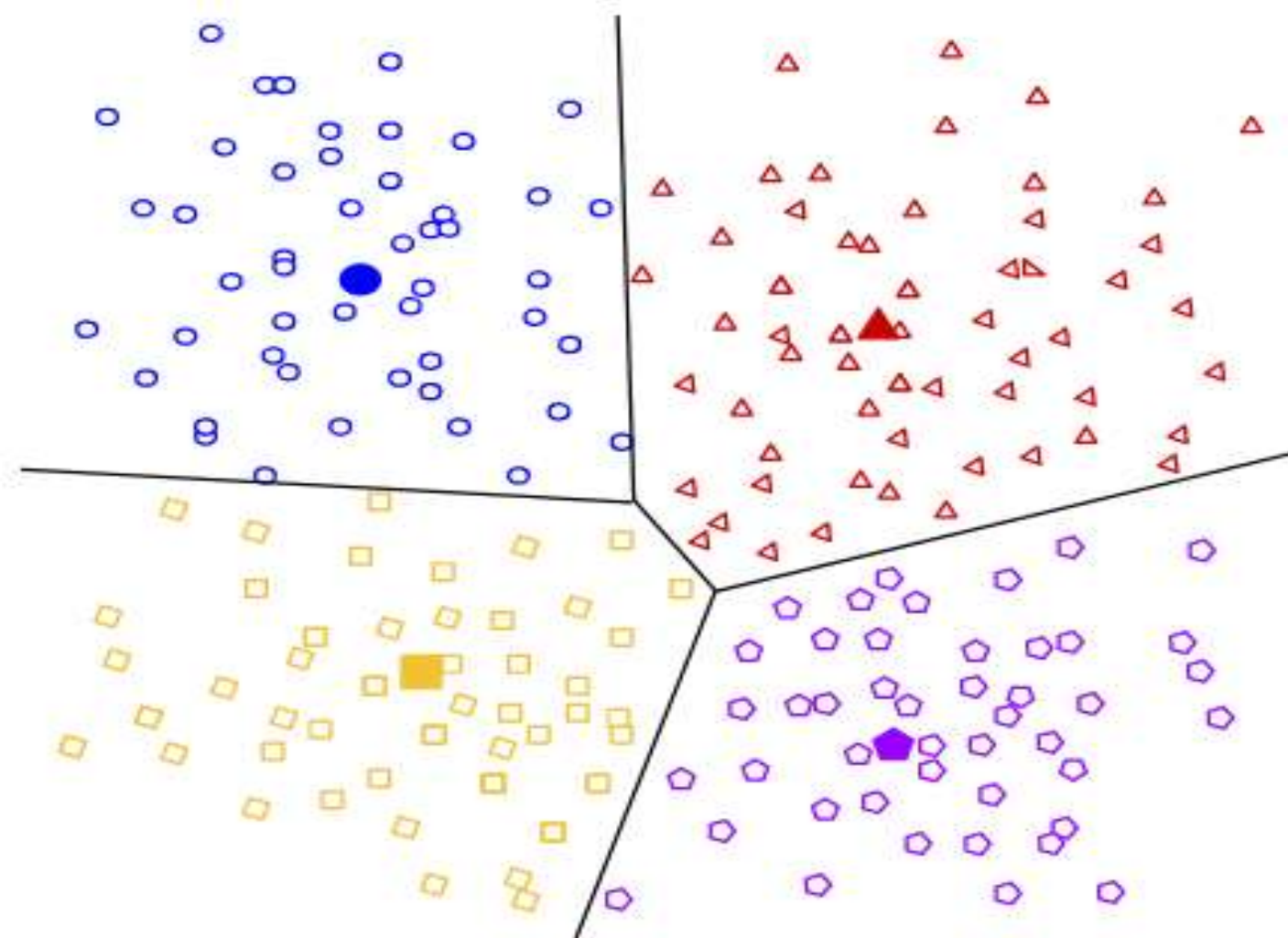
Types of Clustering :

- ❖ Centroid-based Clustering.
- ❖ Density-based Clustering.
- ❖ Distribution-based Clustering.
- ❖ Hierarchical Clustering.

Clustering

❖ Centroid-based Clustering

- K means clustering most widely used
- each cluster is represented by a central vector
- find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.



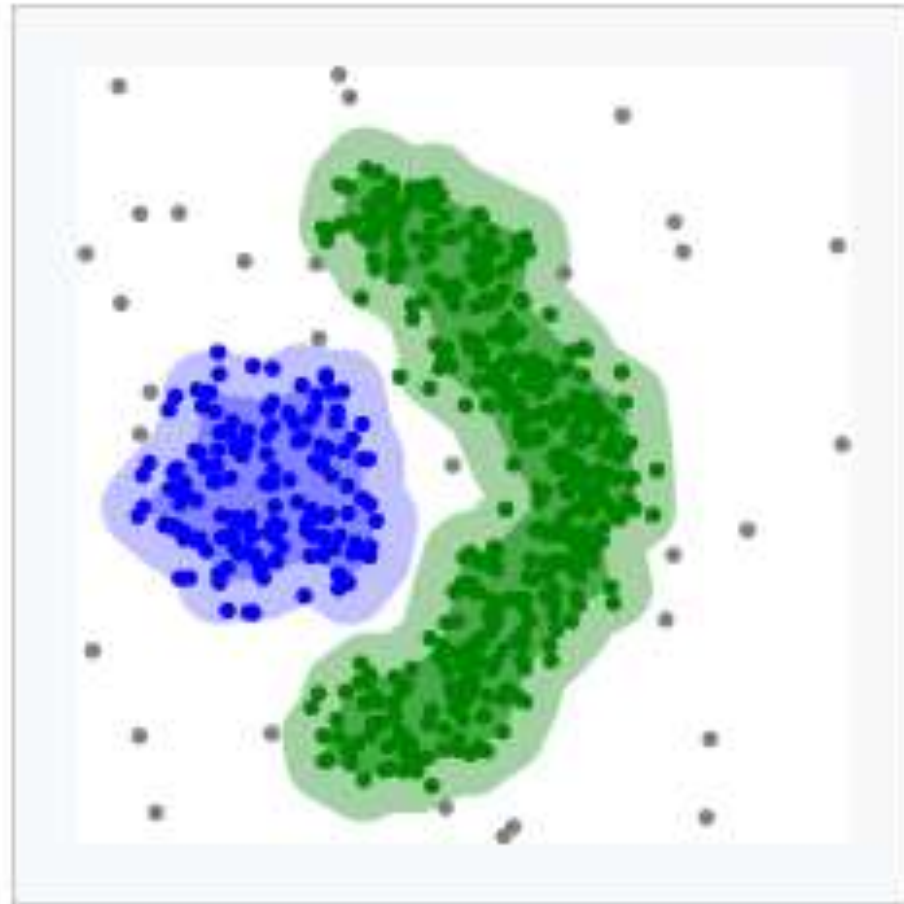
Clustering

❖ **Density-based Clustering :**

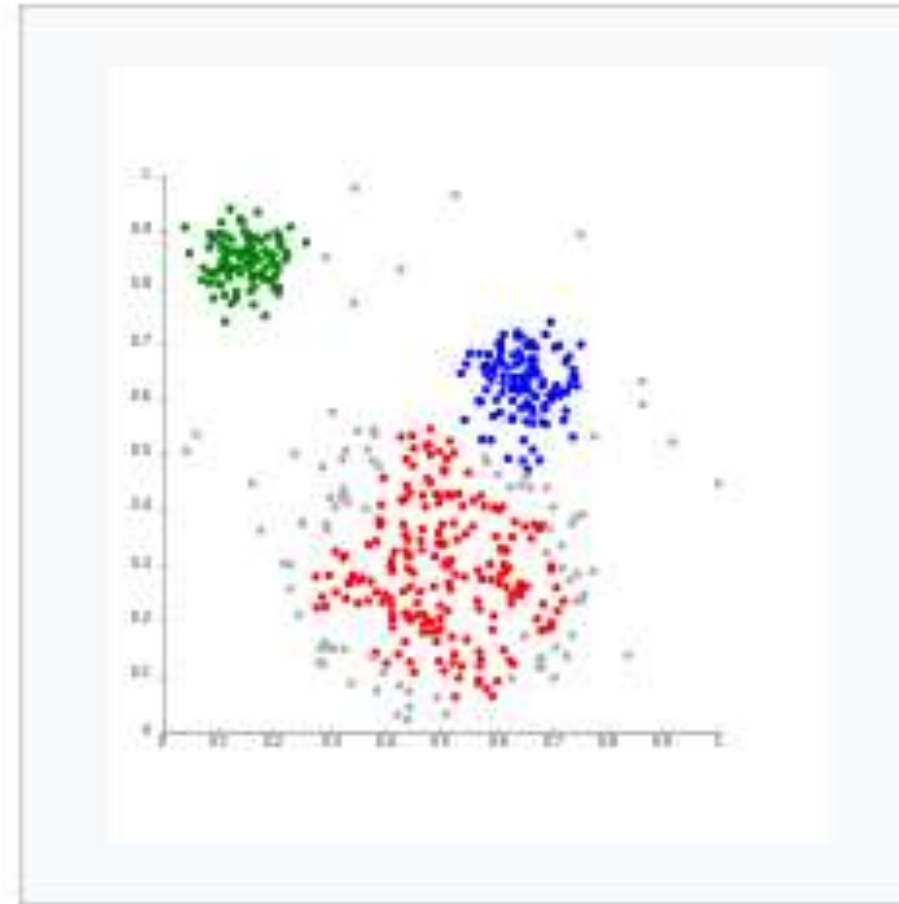
- clusters are defined as areas of higher density
- Objects in sparse areas – that are required to separate clusters – are usually considered to be noise and border points.
- density based clustering method is **DBSCAN** (Density-based spatial clustering of applications with noise)

Clustering

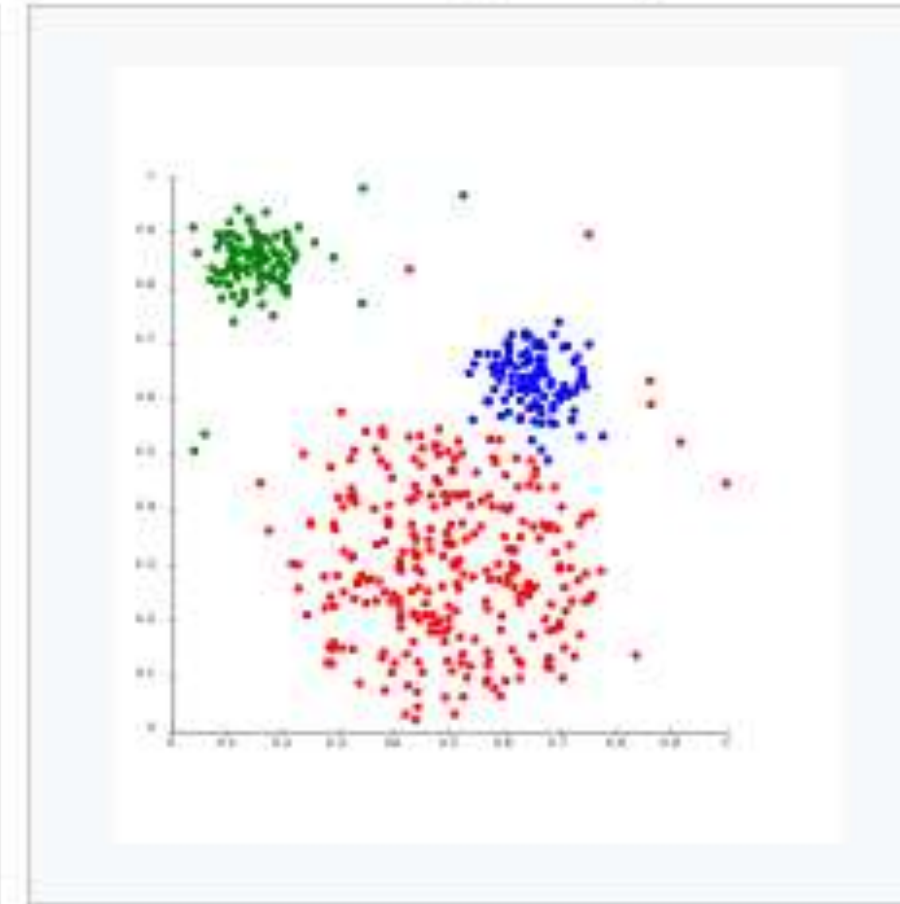
❖ Density-based Clustering :



Density-based clustering with DBSCAN.



DBSCAN assumes clusters of similar density, and may have problems separating nearby clusters



OPTICS is a DBSCAN variant, improving handling of different densities clusters

Clustering

❖ **Distribution-based Clustering :**

- This clustering approach assumes data is composed of distributions, such as Gaussian distributions.
- As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases.
- The bands show that decrease in probability.

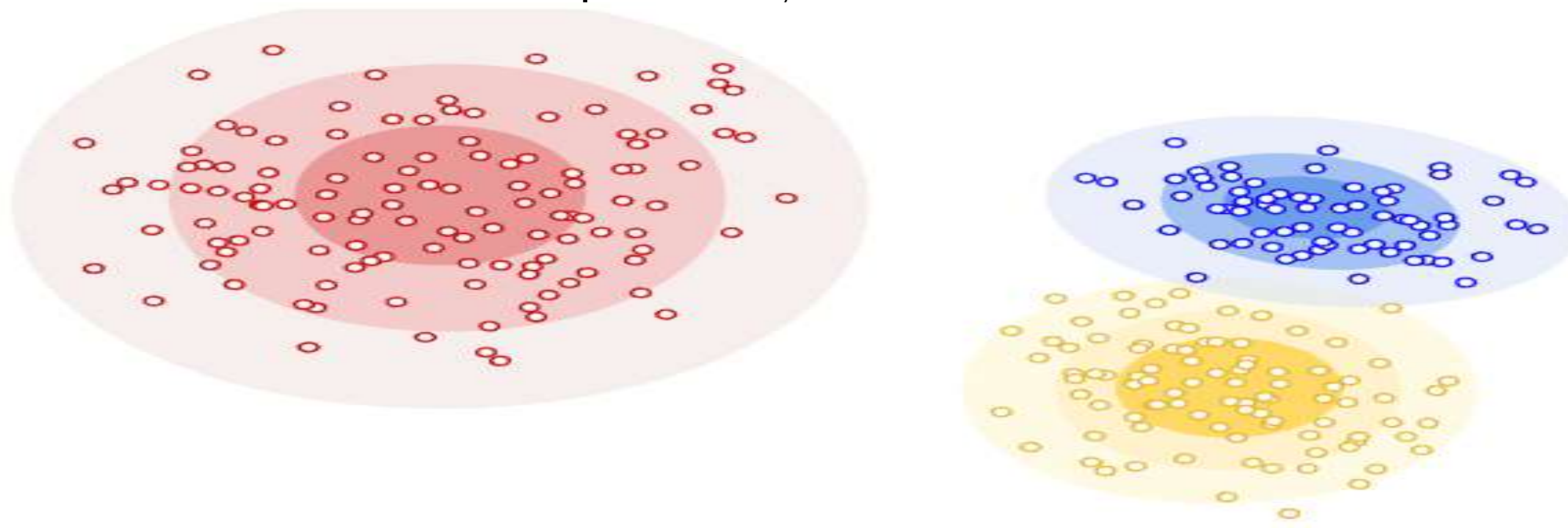
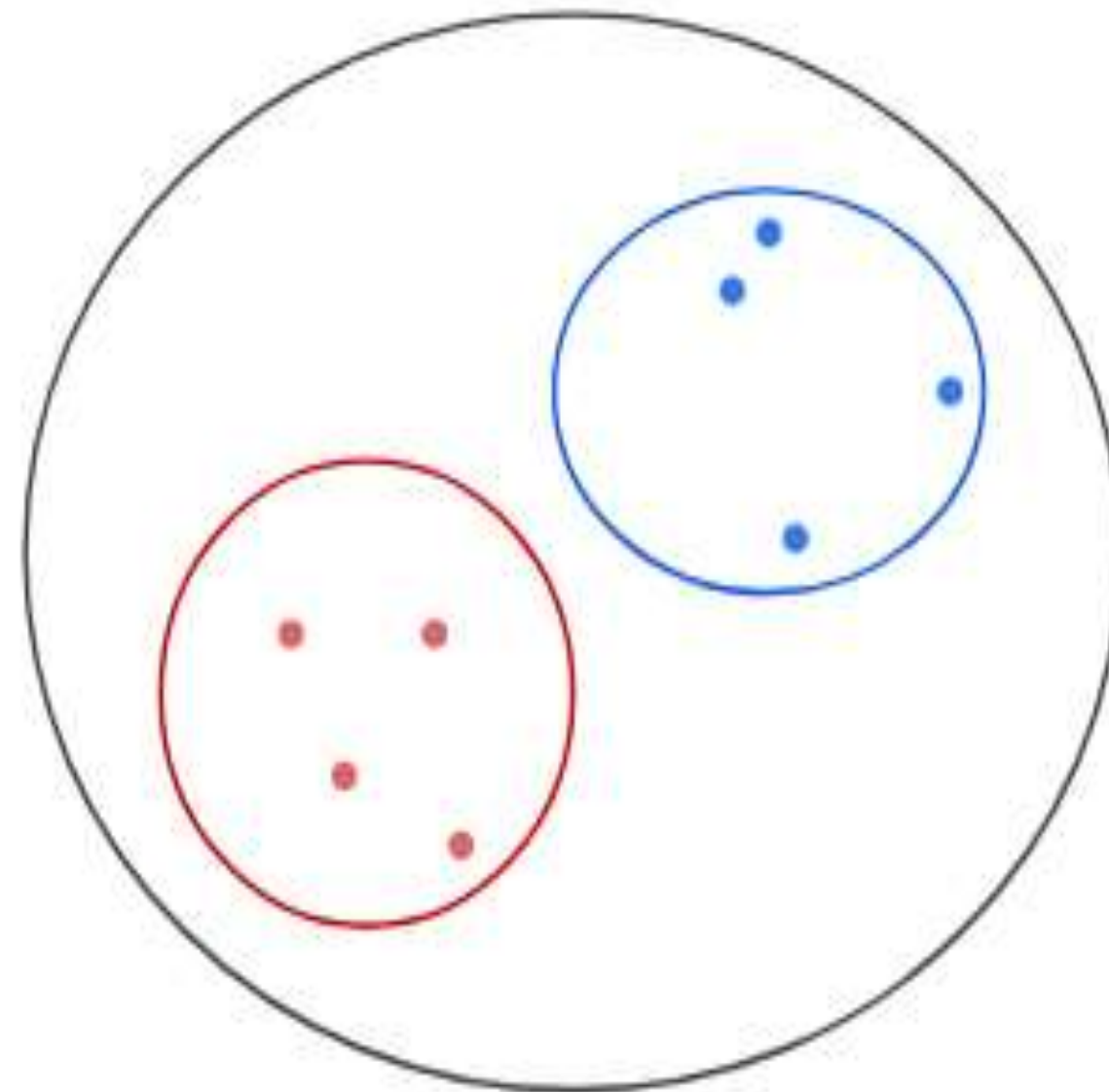


Figure 3: Example of distribution-based clustering.

Clustering

❖ Hierarchical Clustering :

- The core idea of objects being more related to nearby objects than to objects farther away.
- These algorithms connect "objects" to form "clusters" based on their distance.



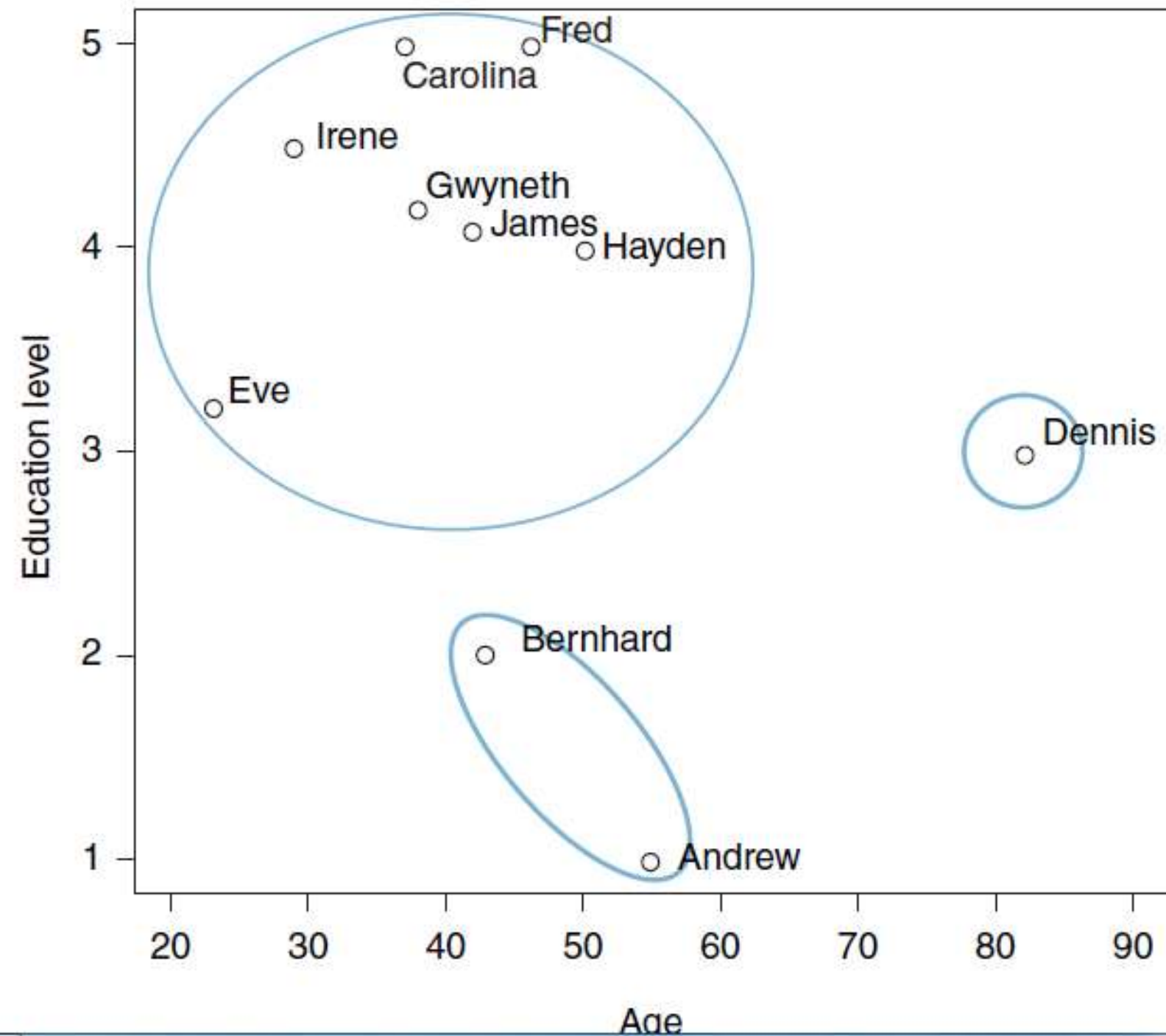
Clustering

1. Distance Measures
2. Clustering Validation
3. Clustering Techniques

Clustering

1. Distance Measures :

- ❖ what is similar and what is not similar (dissimilar).



Clustering

1. Distance Measures :

- ❖ what is similar and what is not similar (dissimilar).
- ❖ represent the similarity between two objects by a single number.
- ❖ A common approach to associate a number with the similarity (and dissimilarity) between two objects is to use distance measures.
- ❖ The most similar objects have the smallest distances between them, and the most dissimilar have the largest distances.
- ❖ scale type of its attributes: whether they are quantitative or qualitative.

Clustering

1.1 Differences between values of common Attribute Types :

- ❖ The difference between two values for the same attribute, here named *a* and *b*, will be denoted as $d(a, b)$.

$$d(a, b) = |a - b|$$

Clustering

1.1 Differences between values of common Attribute Types :

Example 5.1 For example, the difference in age between Andrew ($a = 55$) and Carolina ($b = 37$) is $|55 - 37| = 18$. Note, that even if we change the order of the values ($a = 37$ and $b = 55$) the result is the same.

If the attribute type is qualitative, we use distance measures suitable for the given type. If the qualitative attribute has ordinal values, we can measure the difference in their positions as:

$$d(a, b) = (|pos_a - pos_b|) / (n - 1)$$

where n is the number of different values, and pos_a and pos_b are the positions of the values a and b , respectively, in a ranking of possible values.

Clustering

1.1 Differences between values of common Attribute Types :

Example 5.2 In our data set, education level can be considered an ordinal attribute, with larger values meaning a higher level of education. Thus the distance between the education levels of Andrew and Caroline is $|pos1 - pos5|/4 = |1 - 5|/4 = 1$.

Note that ordinal attributes need not be expressed only by numbers. For example, the education level can have values such as “primary”, “high school”, “undergraduate”, “graduate” and “postgraduate”. However, these can be readily transformed into numbers (see Section 2.1).

If a qualitative attribute has nominal values, in order to compute the distance between two values we simply determine if they are equal (in which case the difference, or dissimilarity, will be zero) or not (in which case the difference will be one).

$$d(a, b) = \begin{cases} 1, & \text{if } a \neq b \\ 0, & \text{if } a = b \end{cases}$$

Clustering

1.2 Distance Measures for Objects with Quantitative Attributes :

- ❖ Several distance measures are particular cases of the Minkowski distance.

Minkowski distance for two m -dimensional objects p and q with quantitative attributes is given by:

$$d(p, q) = \sqrt[r]{\sum_{k=1}^m |p_k - q_k|^r}$$

where m is the number of attributes, while p_k and q_k are the values of the k th attribute for objects p and q , respectively. Variants are obtained using different values for r . For example, for the Manhattan distance, $r = 1$, and for the Euclidean distance, $r = 2$.

Clustering

1.2 Distance Measures for Objects with Quantitative Attributes :

- ❖ Several distance measures are particular cases of the Minkowski distance.

Minkowski distance for two m -dimensional objects p and q with quantitative attributes is given by:

$$d(p, q) = \sqrt[r]{\sum_{k=1}^m |p_k - q_k|^r}$$

where m is the number of attributes, while p_k and q_k are the values of the k th attribute for objects p and q , respectively. Variants are obtained using different values for r . For example, for the Manhattan distance, $r = 1$, and for the Euclidean distance, $r = 2$.

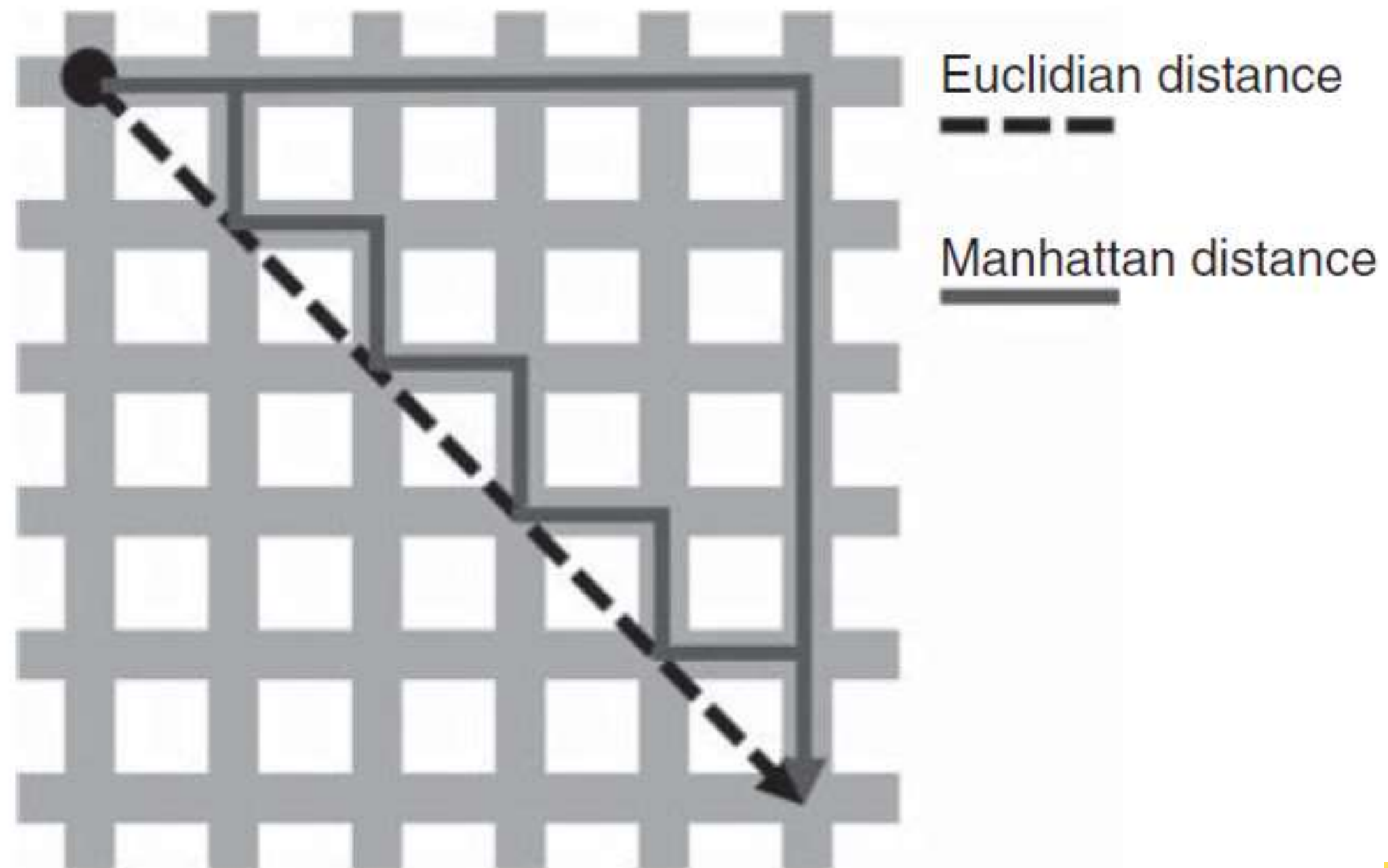
Clustering

1.2 Distance Measures for Objects with Quantitative Attributes :

- ❖ Manhattan distance - also known as the city block or taxicab distance
- ❖ Attribute types that are neither quantitative nor qualitative, termed “non-conventional”, include:
 - biological sequences
 - time series
 - images
 - sound
 - video.

Clustering

1.2 Distance Measures for Objects with Quantitative Attributes :



Clustering

1.3 Distance Measures for Non-conventional Attributes:

- ❖ Hamming distance - used for sequences of values and these values are usually characters or binary values.
- ❖ A binary value is either 1 or 0, meaning true or false.
- ❖ Hamming distance - number of positions at which the corresponding
- ❖ characters or symbols in the two strings are different.
- ❖ **Ex** - “James” and “Jimmy” is 3
“Tom” and “Tim” is 1

Clustering

1.3 Distance Measures for Non-conventional Attributes:

Example For example, for the two texts:

- A = “I will go to the party. But first, I will have to work.”
- B = “They have to go to the work by bus.”

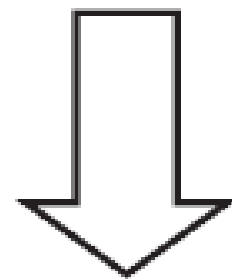
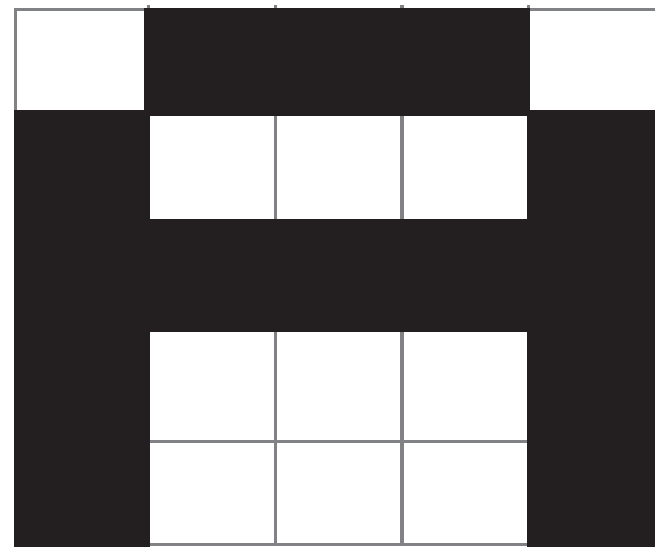
the bag of words will produce the vectors shown in Table

Example of bag of words vectors.

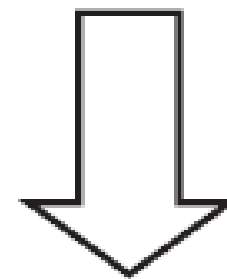
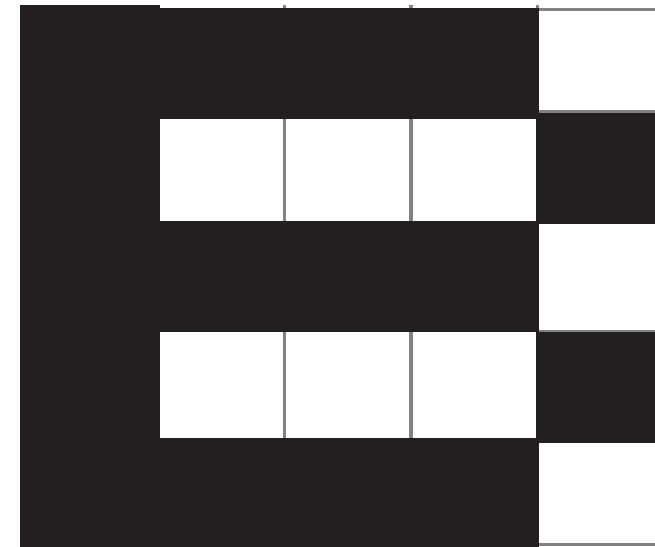
	I	will	go	to	the	party	but	first	have	work	they	by	bus
A	2	2	1	2	1	1	1	1	1	1	0	0	0
B	0	0	1	2	1	0	0	0	1	1	1	1	1

Clustering

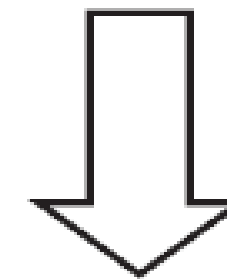
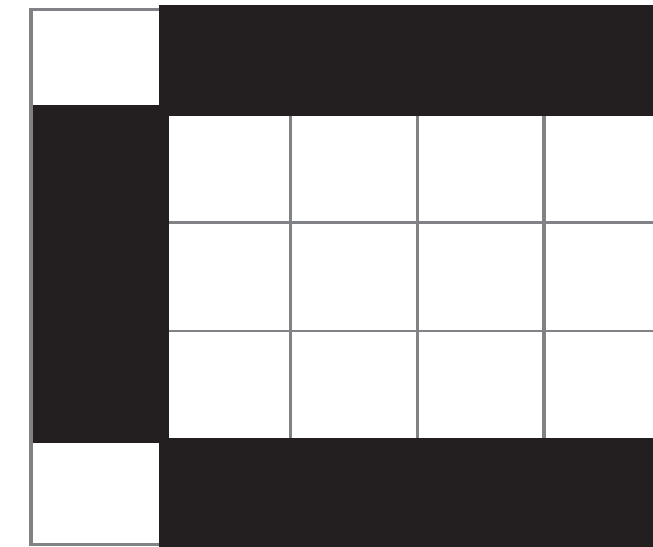
1.3 Distance Measures for Non-conventional Attributes:



0	1	1	1	0
1	0	0	0	1
1	1	1	1	1
1	0	0	0	1
1	0	0	0	1



1	1	1	1	0
1	0	0	0	1
1	1	1	1	0
1	0	0	0	1
1	1	1	1	0



0	1	1	1	1
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	1	1	1	1

Clustering

1.3 Distance Measures for Non-conventional Attributes:

	1 st row					2 nd row					3 rd row					4 th row					5 th row				
A	0	1	1	1	0	1	0	0	0	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1
B	1	1	1	1	0	1	0	0	0	1	1	1	1	1	0	1	0	0	0	1	1	1	1	1	0
C	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1

Transformation of images of size 5×5 pixels into matrices and vectors.

Clustering

2. Clustering Validation:

- ❖ To find good clustering partitions for a data set and the quality of the partitions must be evaluated.
- ❖ Automatic validation measures for clustering partition evaluation can be roughly divided into three categories:
 - External indices - uses external information, such as class label, if available, to define the quality of the clusters
 - Internal indices - looks for compactness inside each cluster
 - Relative indices - compares partitions found by two or more clustering techniques

Clustering

3. Clustering Techniques:

- ❖ **Separation-based:** objects in the cluster is closer to every other object in the cluster
- ❖ **Prototype-based:** each object in the cluster is closer to a prototype representing the cluster than to a prototype representing any other cluster
- ❖ **Graph-based:** represents the data set by a graph structure associating
- ❖ **Density-based:** a cluster is a region where the objects have a high number of close neighbors (i.e. a dense region), surrounded by a region of low density.
- ❖ **Shared-property:** a cluster is a group of objects that share a property.

Clustering

3. Clustering Techniques:

❖ Methods :

1. K-means
2. DBSCAN
3. Agglomerative hierarchical clustering

Clustering

1. K-means :

- Centroids are a key concept

1.1 Centroids and Distance Measures :

- A centroid can also be seen as a prototype or profile of all the objects in a cluster for example the average of all the objects.

Clustering

1. K-means :

1.2 How K-means Works :

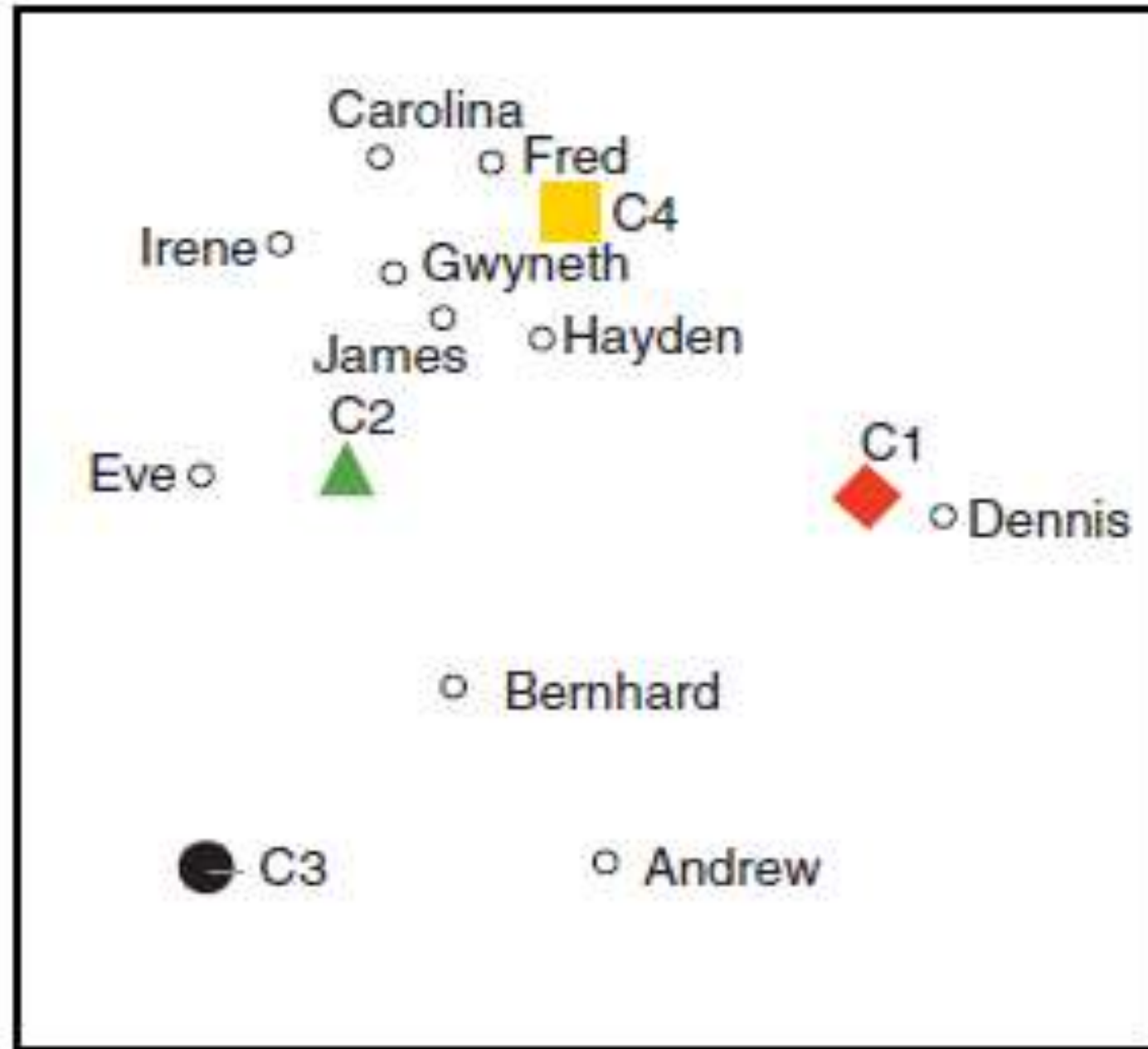
Algorithm K-means

- 1: INPUT D the data set
 - 2: INPUT d the distance measure
 - 3: INPUT K the number of clusters
 - 4: Define the initial K centroids (they are usually randomly defined, but can be defined explicitly in some software packages)
 - 5: **repeat**
 - 6: Associate each instance in D with the closest centroid according to the chosen distance measure d
 - 7: Recalculate each centroid using all instances from D associated with it.
 - 8: **until** No instances from D change of associated centroid.
-

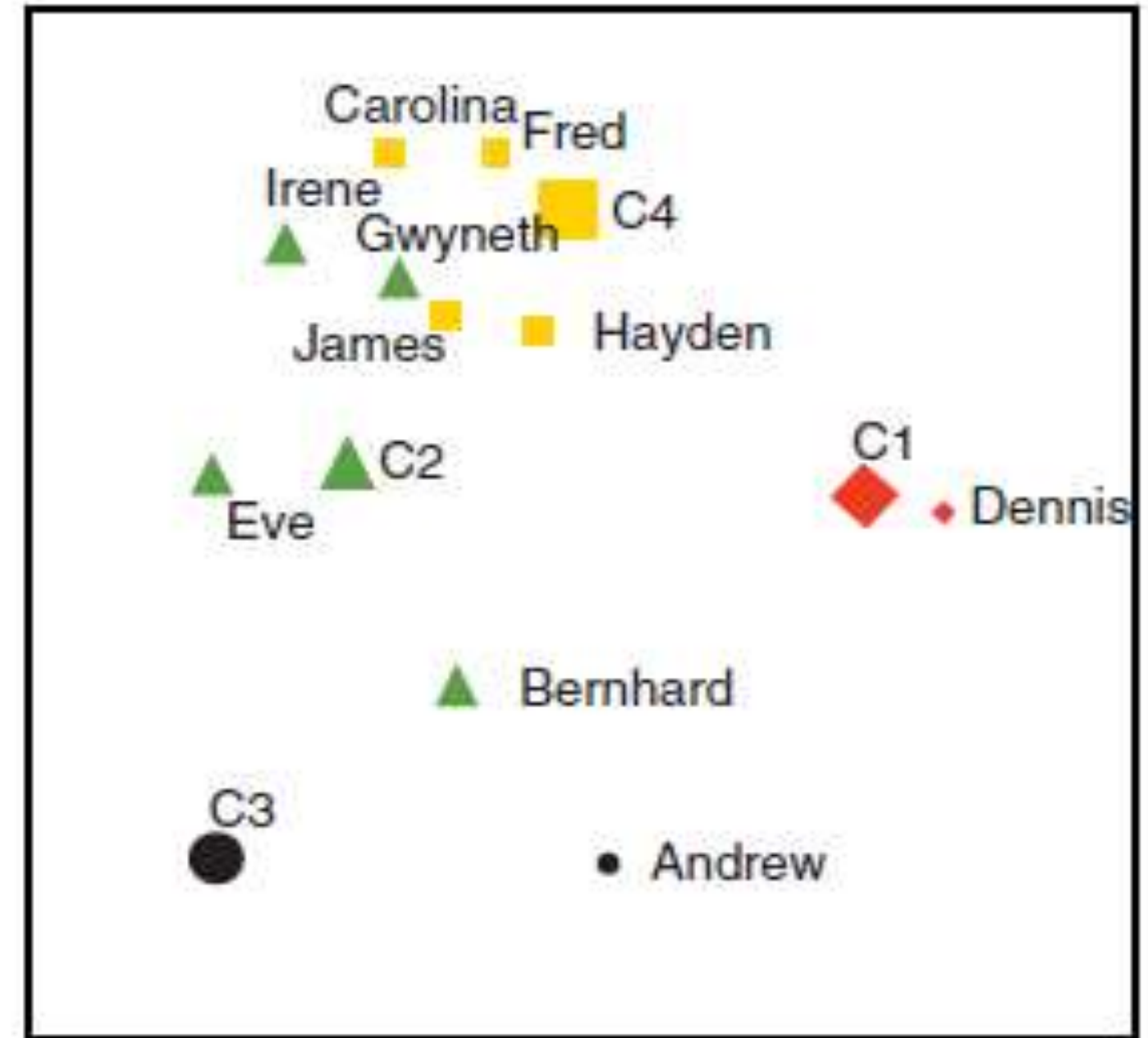
Clustering

1. K-means :

(a)

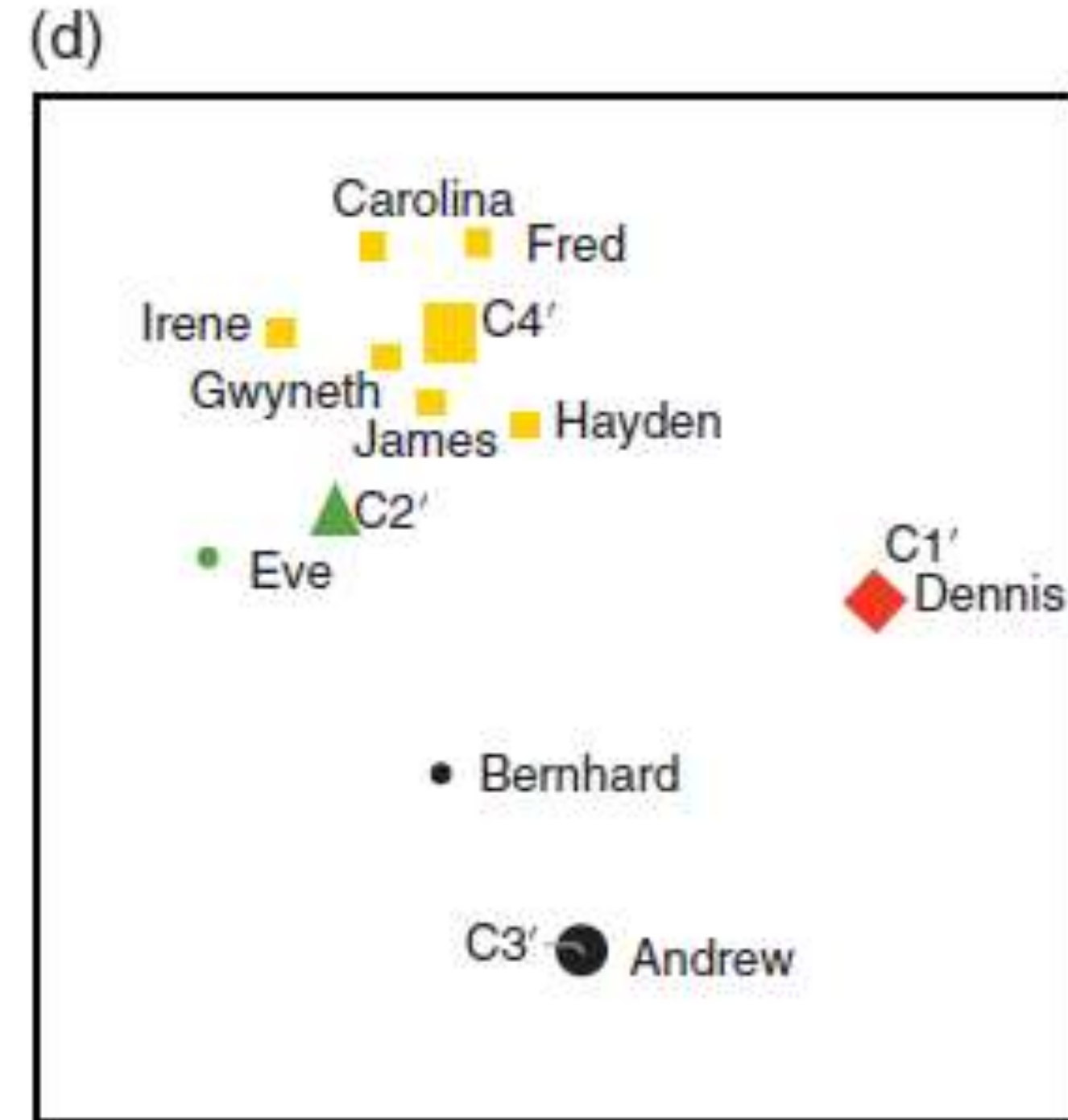
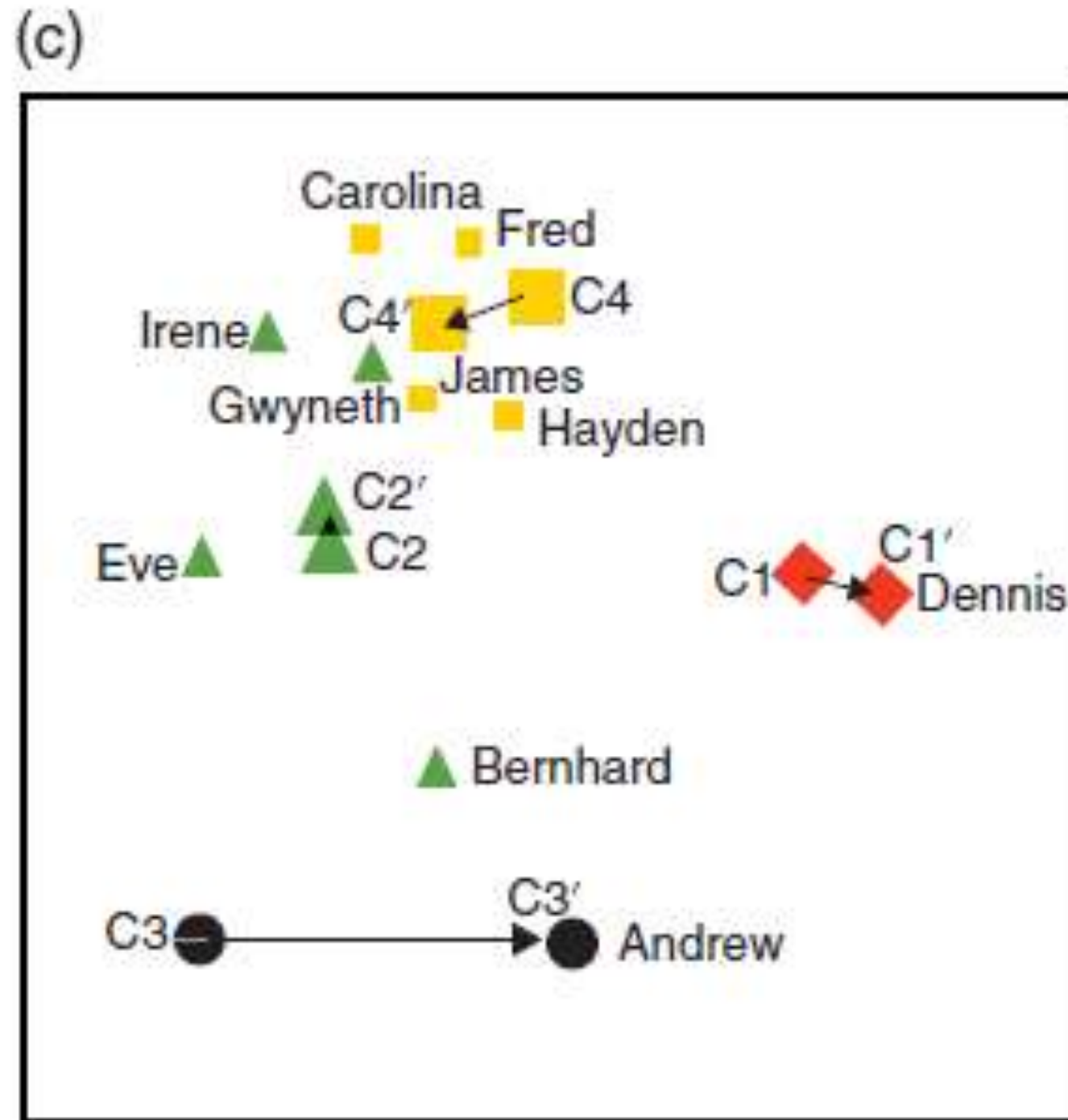


(b)



Clustering

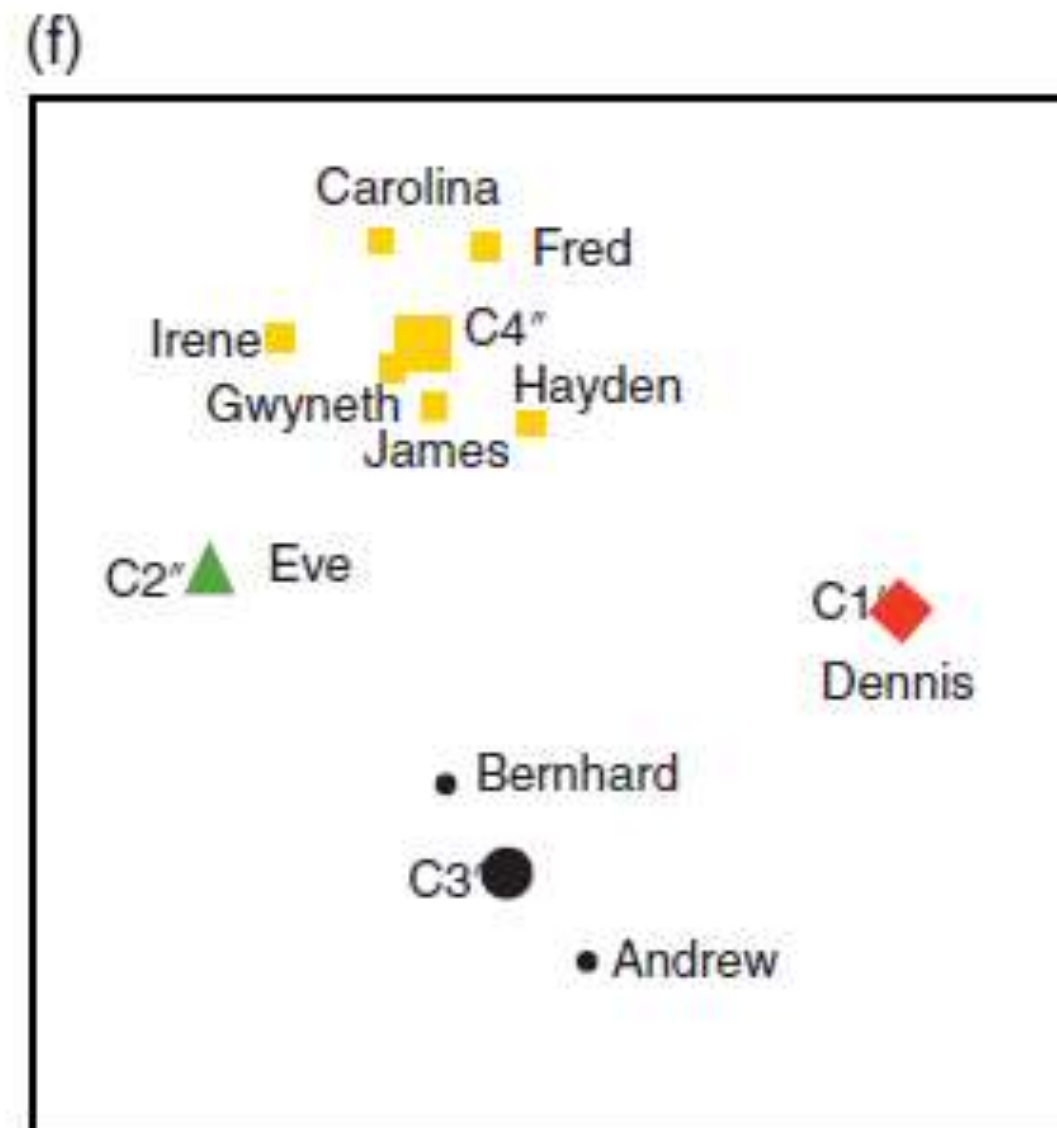
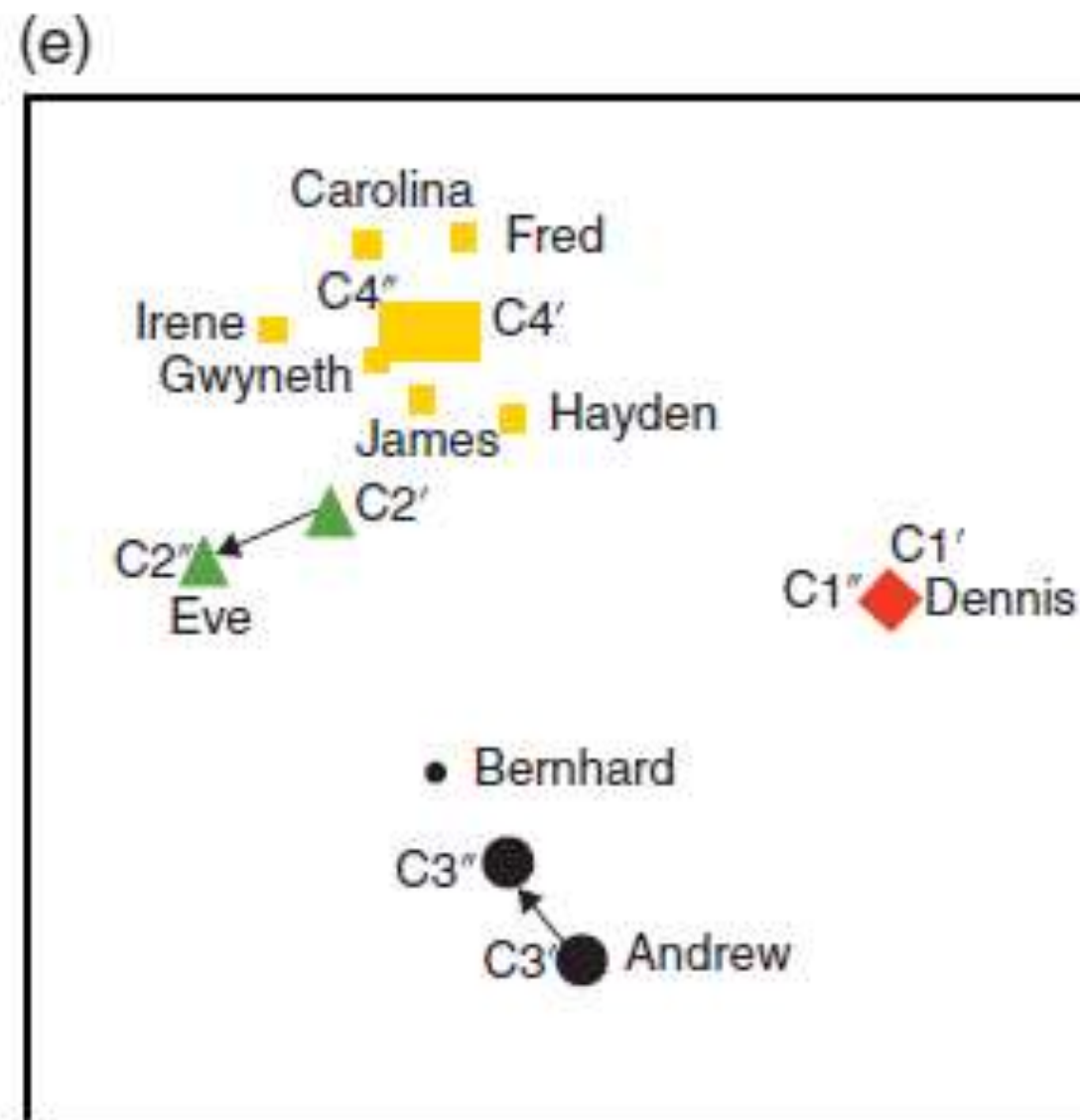
1. K-means :



Clustering

1. K-means :

K-means with $k = 4$. The big symbols represent the centroids. The example is step-by-step according to the Algorithm K-means: (a) step 4; (b) 1st iteration of step 6; (c) 1st iteration of step 7; (d) 2nd iteration of step 6; (e) 2nd iteration of step 7; and (f) 3rd iteration of step (6). The algorithm stops in the 3rd iteration (f) because there is no instance changing of symbol between the 2nd (d) and the 3rd (f) iterations of step (6).



Clustering

1. K-means :

Advantages and disadvantages of k-means.

Advantages

- Computationally efficient
- Good results obtained quite often: global optima

Disadvantages

- Typically, each time we run k-means the results are different due to the random initialization of the centroids text
 - Need to define number of clusters in advance
 - Does not deal with noisy data and outliers
 - k-means can only find partitions whose clusters have convex shapes
-

Clustering

2. DBSCAN :

Assessment 1

1. Write about clustering algorithms?

Ans : _____

