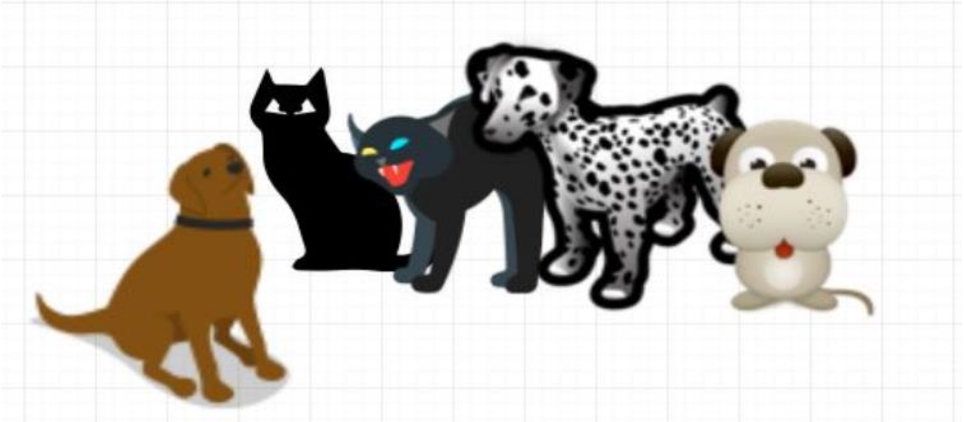
Clustering

- Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.
- Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.
- Unlike supervised learning, no teacher is provided that means no training will be given to the machine.
- Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.
- For instance, suppose it is given an image having both dogs and cats which it has never seen.



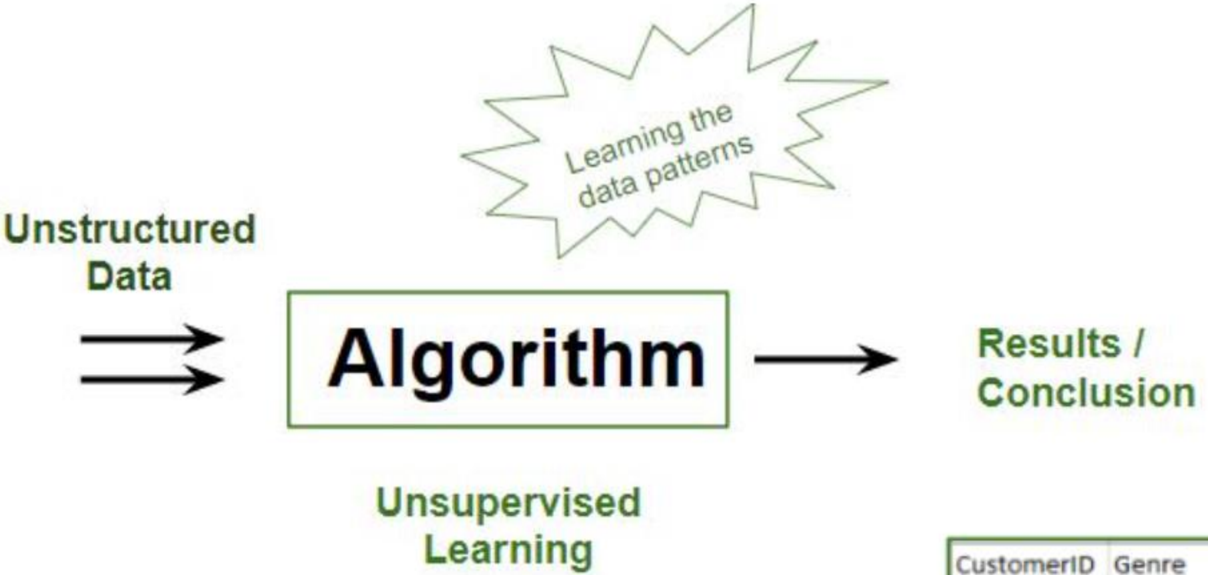
- Thus the machine has no idea about the features of dogs and cats so we can’t categorize it as ‘dogs and cats ‘.
- But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts.
- The first may contain all pics having dogs in them and the second part may contain all pics having cats in them.
- Here you didn’t learn anything before, which means no training data or examples.

- It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.

Unsupervised learning is classified into two categories of algorithms:

- ➢Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- ➢Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

- Unsupervised machine learning analyzes and clusters unlabeled datasets using machine learning algorithms.
- These algorithms find hidden patterns and data without any human intervention, i.e., we don’t give output to our model.
- The training model has only input parameter values and discovers the groups or patterns on its own.
- Data-set in Figure A is Mall data that contains information about its clients that subscribe to them. Once subscribed they are provided a membership card and the mall has complete information about the customer and his/her every purchase.
- Now using this data and unsupervised learning techniques, the mall can easily group clients based on the parameters we are feeding in.

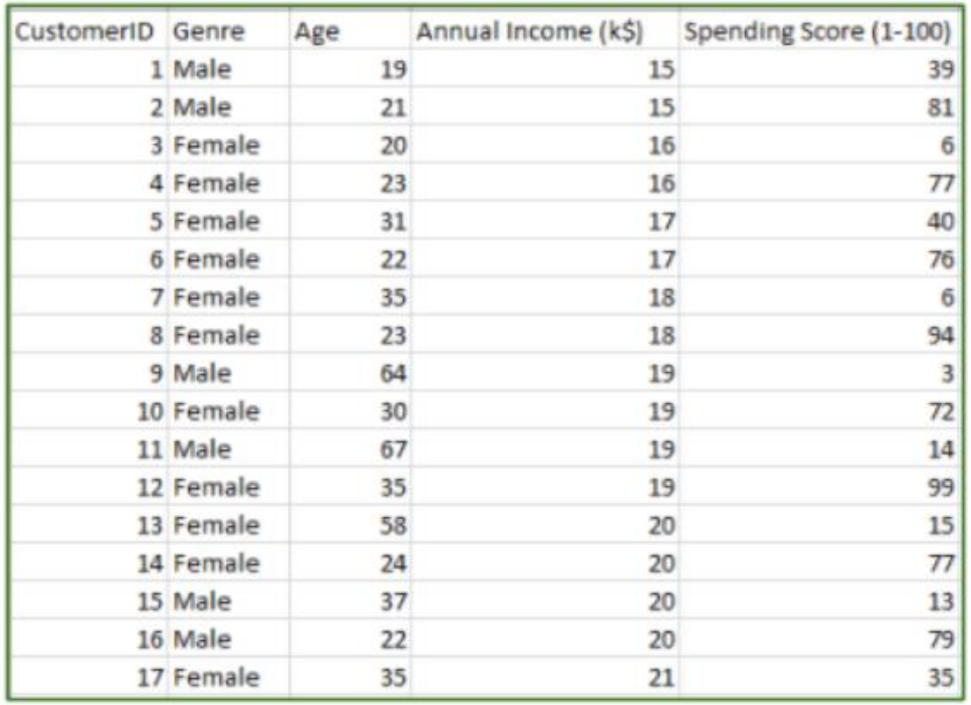

Unstructured Data

Algorithm

Learning the data patterns

Results / Conclusion

Unsupervised Learning

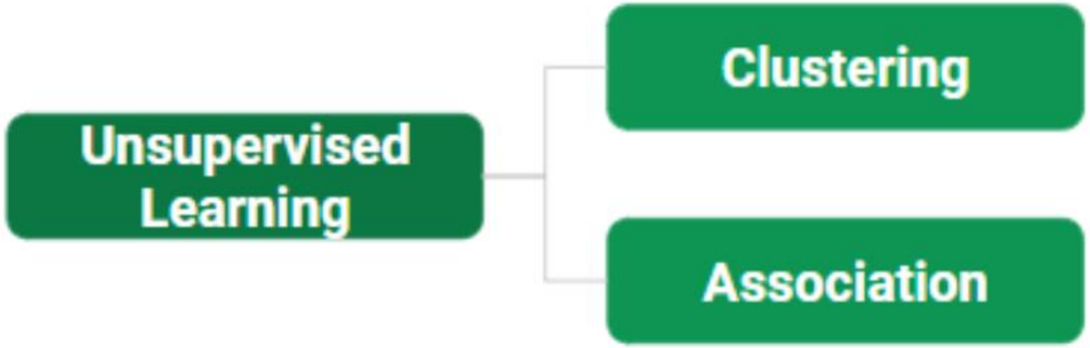


CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35

Figure A

The input to the unsupervised learning models is as follows:

- **Unstructured data**: May contain noisy(meaningless) data, missing values, or unknown data.
- **Unlabeled data**: Data only contains a value for input parameters, there is no targeted value(output). It is easy to collect as compared to the labeled one in the Supervised approach.



Types of Unsupervised Learning:-

Clustering

1.Exclusive (partitioning)
2.Agglomerative
3.Overlapping
4.Probabilistic

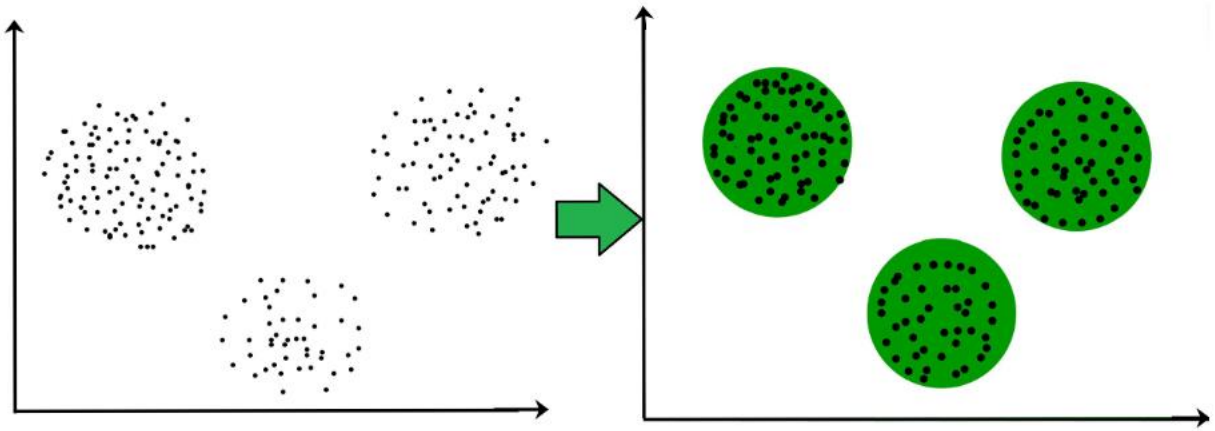
Clustering Types:-

1.Hierarchical clustering
2.K-means clustering
3.Principal Component Analysis
4.Singular Value Decomposition
5.Independent Component Analysis

Introduction to Clustering

- An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses.
- Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.
- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
- It is basically a collection of objects on the basis of similarity and dissimilarity between them.

- For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



Clustering Methods :

- **Density-Based Methods**: These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space.
- These methods have good accuracy and the ability to merge two clusters.
- Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), etc.
- **Hierarchical Based Methods**: The clusters formed in this method form a tree-type structure based on the hierarchy.
- New clusters are formed using the previously formed one. It is divided into two category
  - Agglomerative (bottom-up approach)
  - Divisive (top-down approach)

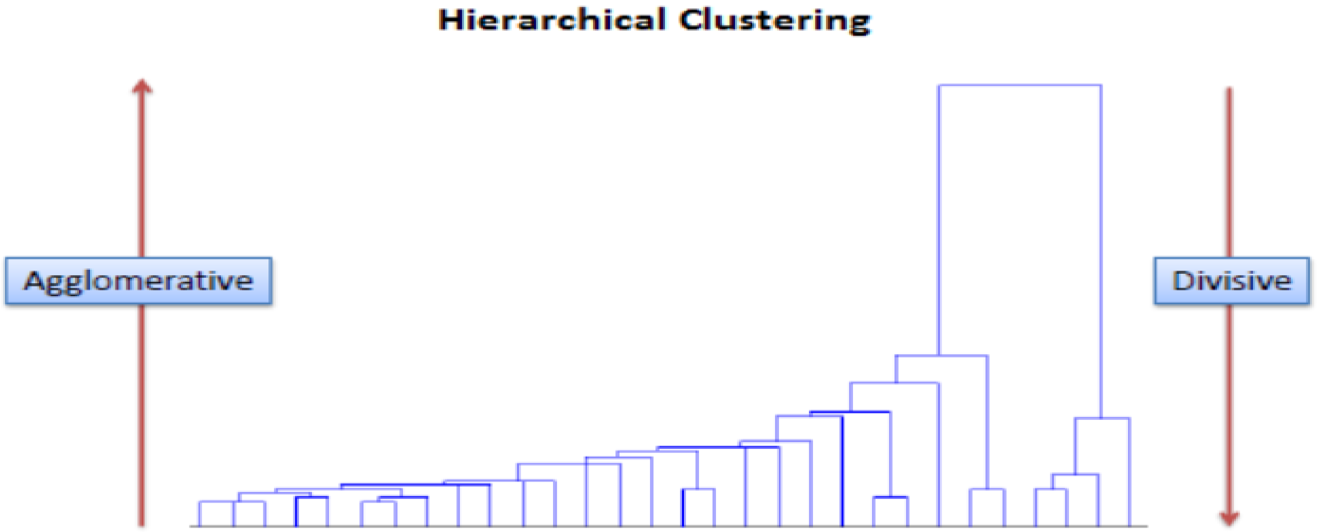
- **Partitioning Methods**: These methods partition the objects into k clusters and each partition forms one cluster.
- This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon Randomized Search), etc.
- **Grid-based Methods**: In this method, the data space is formulated into a finite number of cells that form a grid-like structure.
- All the clustering operations done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest), etc.

- **Hierarchical clustering**, also known as hierarchical cluster analysis or HCA, is an unsupervised machine learning approach for grouping unlabeled datasets into clusters.
- The hierarchy of clusters is developed in the form of a tree in this technique, and this tree-shaped structure is known as the dendrogram.
- Separating data into groups based on some measure of similarity, finding a technique to quantify how they're alike and different, and limiting down the data is what hierarchical clustering is all about.

- Hierarchical clustering method functions in two approaches-

➢ Agglomerative

➢ Divisive

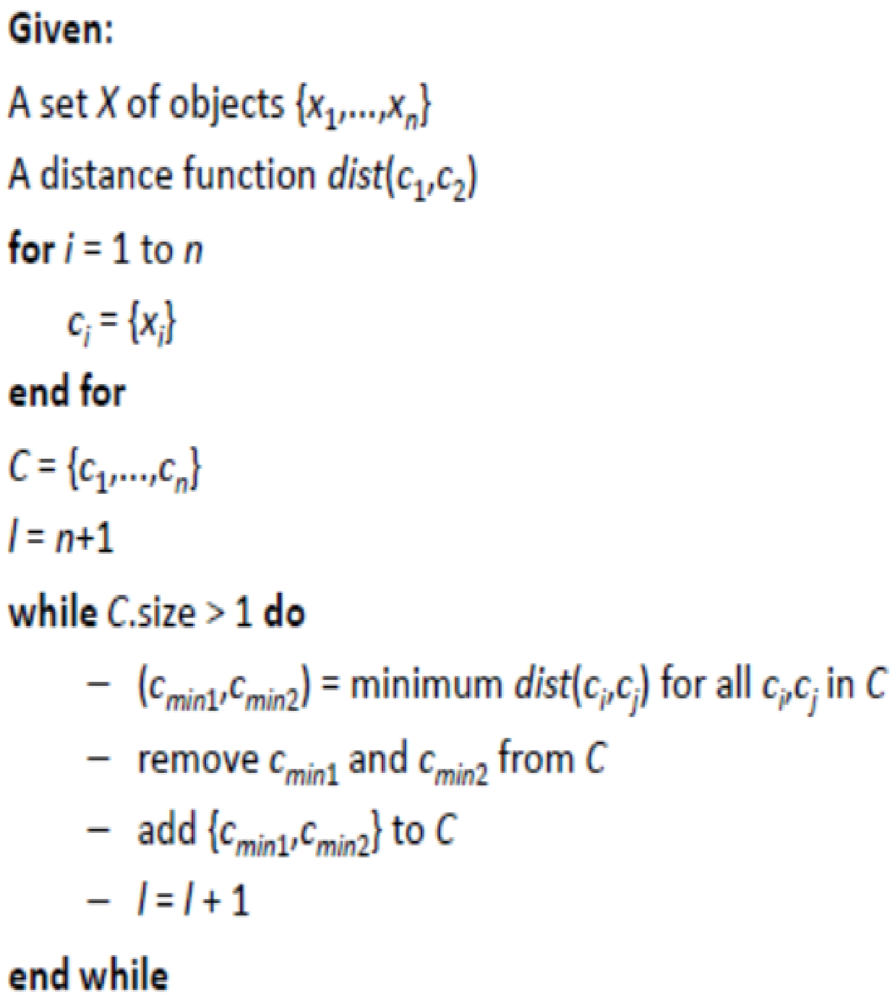


Divisive method

In *divisive* or *top-down clustering* method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters using a flat clustering method (e.g., K-Means). Finally, we proceed recursively on each cluster until there is one cluster for each observation. There is evidence that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in some circumstances but is conceptually more complex.

Agglomerative method

In *agglomerative* or *bottom-up clustering* method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single cluster left. The related algorithm is shown below.

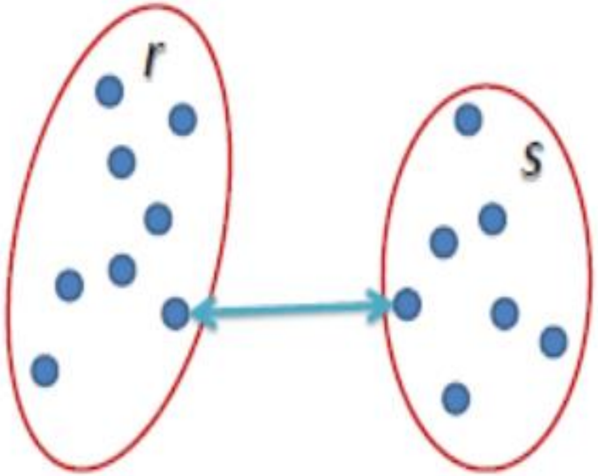


```
Given:
A set $X$ of objects $\{x_1,\ldots,x_n\}$
A distance function $dist(c_1,c_2)$
for $i = 1$ to $n$
    $c_i = \{x_i\}$
end for
$C = \{c_1,\ldots,c_n\}$
$l = n+1$
while $C$.size > 1 do
    – $(c_{min1},c_{min2})$ = minimum $dist(c_i,c_j)$ for all $c_i,c_j$ in $C$
    – remove $c_{min1}$ and $c_{min2}$ from $C$
    – add $\{c_{min1},c_{min2}\}$ to $C$
    – $l = l + 1$
end while
```

- Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function.
- Then, the matrix is updated to display the distance between each cluster.
- The following three methods differ in how the distance between each cluster is measured.

Single Linkage

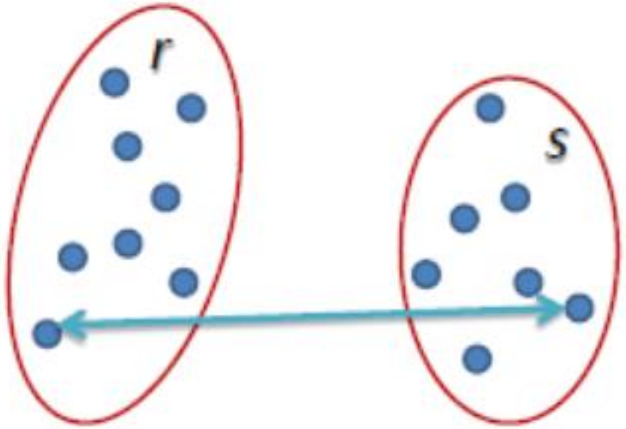
In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete Linkage

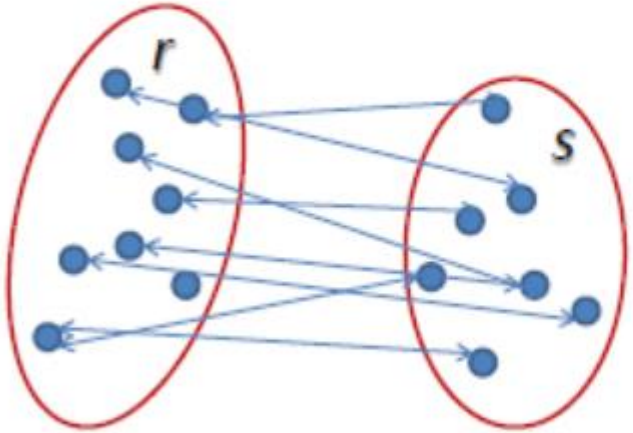
In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average Linkage

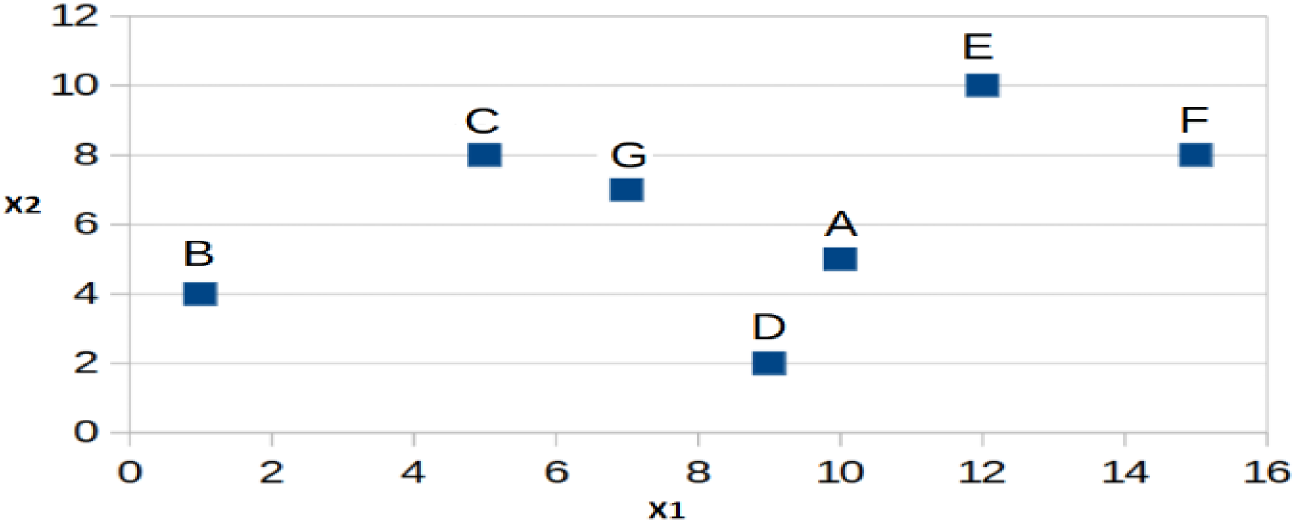
In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters “r” and “s” to the left is equal to the average length each arrow between connecting the points of one cluster to the other.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

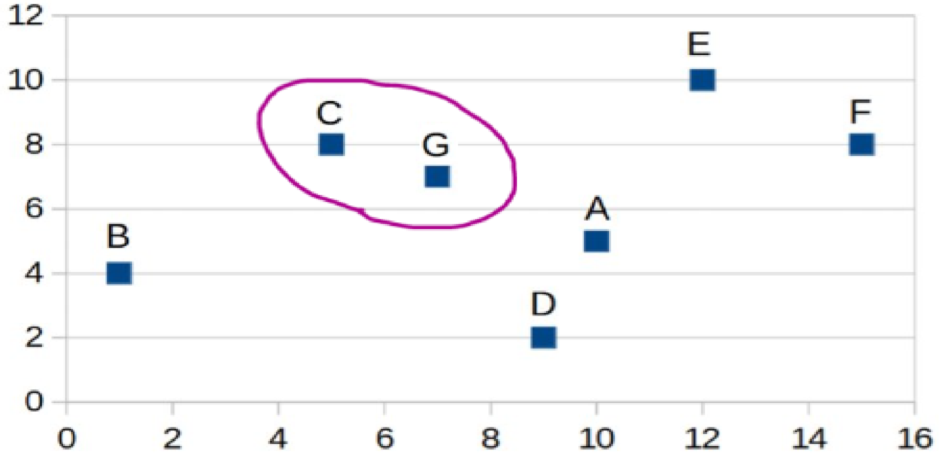
Example: Clustering the following 7 data points.

	X1	X2
A	10	5
B	1	4
C	5	8
D	9	2
E	12	10
F	15	8
G	7	7



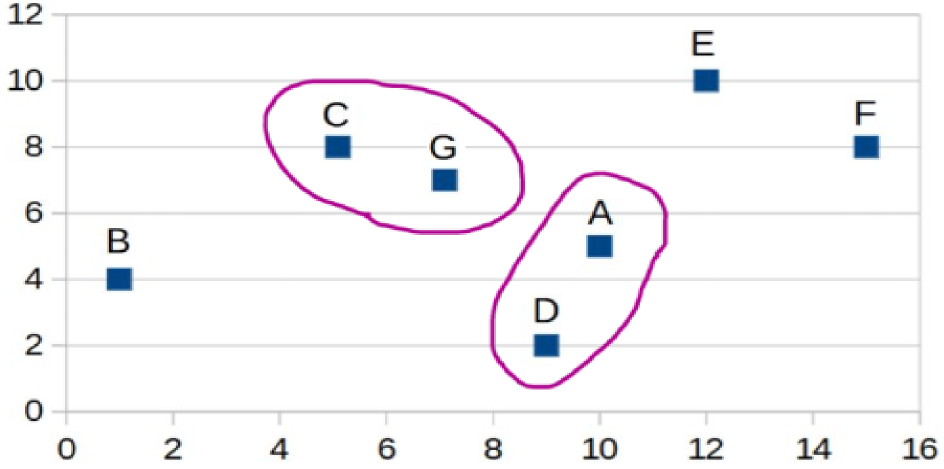
Step 1: Calculate distances between all data points using Euclidean distance function. The shortest distance is between data points C and G.

	A	**B**	**C**	**D**	**E**	**F**
B	9.06					
C	5.83	5.66				
D	3.16	8.25	7.21			
E	5.39	12.53	7.28	14.42		
F	5.83	14.56	10.00	16.16	3.61	
G	3.61	6.71	**2.24**	8.60	5.83	8.06



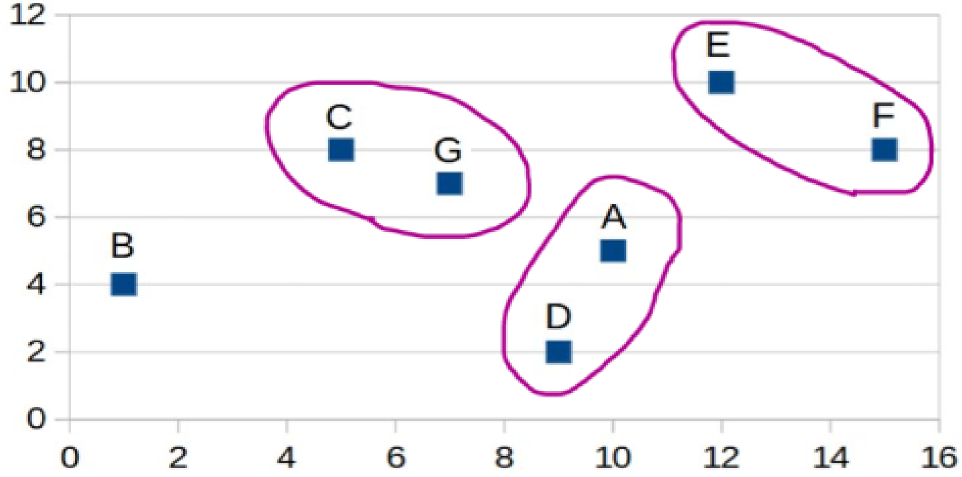
Step 2: We use "Average Linkage" to measure the distance between the "C,G" cluster and other data points.

	A	**B**	**C,G**	**D**	**E**
B	9.06				
C,G	4.72	6.10			
D	**3.16**	8.25	6.26		
E	5.39	12.53	6.50	14.42	
F	5.83	14.56	9.01	16.16	3.61



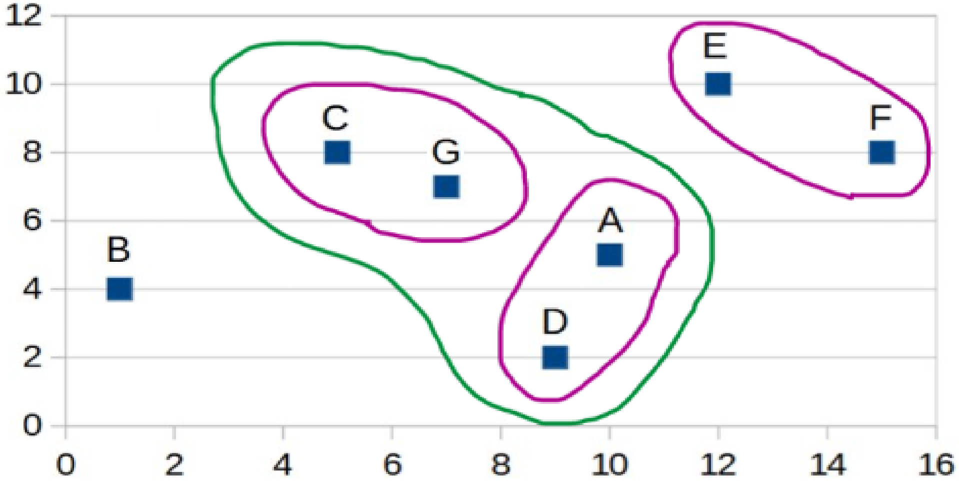
Step 3:

	A,D	B	C,G	E
B	8.51			
C,G	5.32	6.10		
E	6.96	12.53	6.50	
F	7.11	14.56	9.01	**3.61**



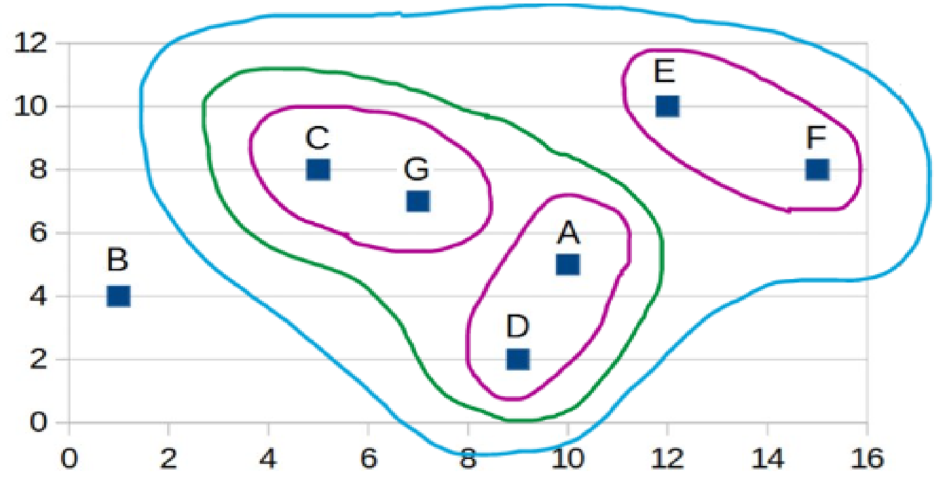
Step 4:

	A,D	B	C,G
B	8.51		
C,G	**5.32**	6.10	
E,F	6.80	13.46	7.65



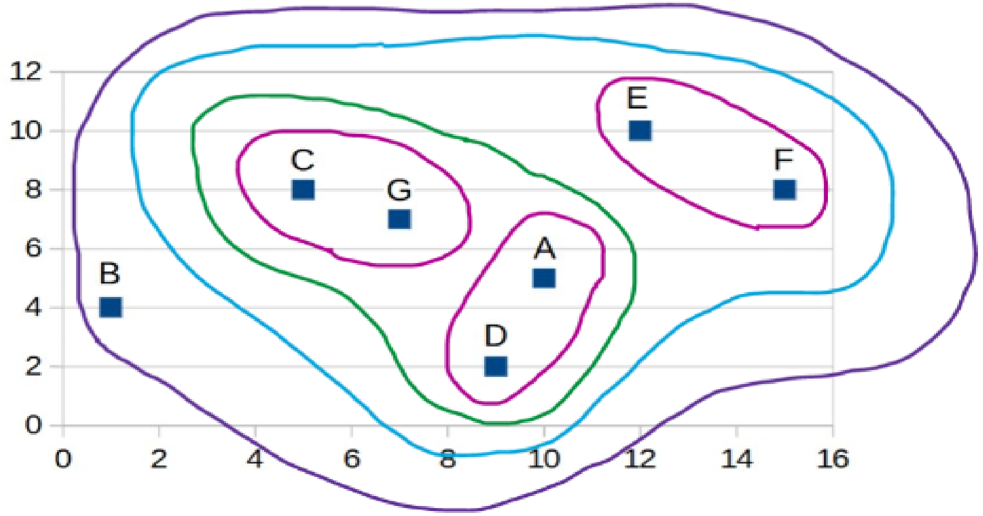
Step 5:

	A,D,C,G	**B**
B	6.91	
E,F	**6.73**	13.46

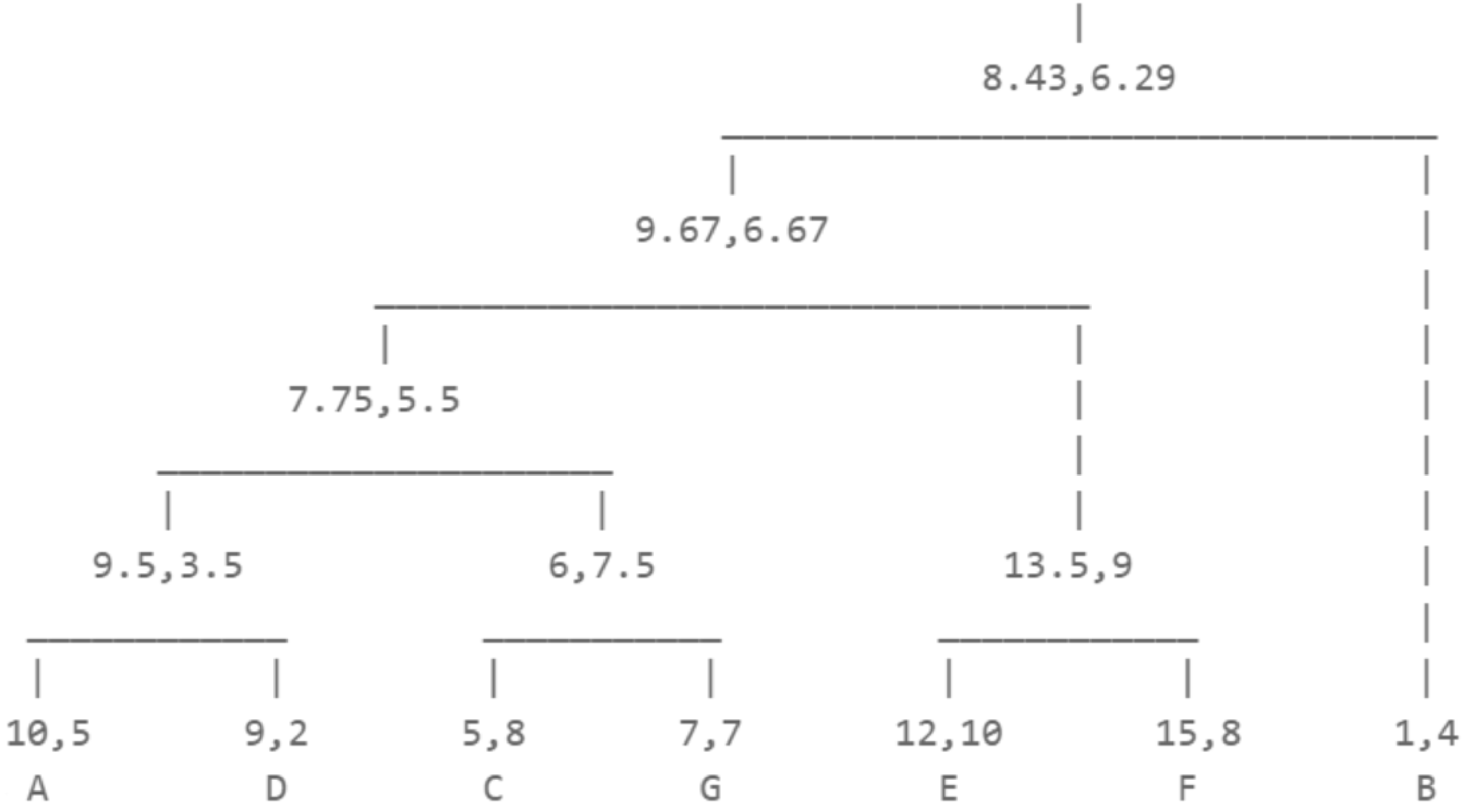


Step 6:

	A,D,C,G,E,F
B	**9.07**



Final dendrogram:



```
                                                  |
                                              8.43,6.29
                                  _________________________________
                                  |                               |
                              9.67,6.67                           |
                  _________________________________               |
                  |                               |               |
              7.75,5.5                            |               |
        _____________________                     |               |
        |                   |                     |               |
    9.5,3.5               6,7.5                13.5,9             |
  ____________         ___________          ____________          |
  |          |         |         |          |          |          |
10,5        9,2       5,8       7,7       12,10       15,8       1,4
  A          D         C         G          E          F          B
```