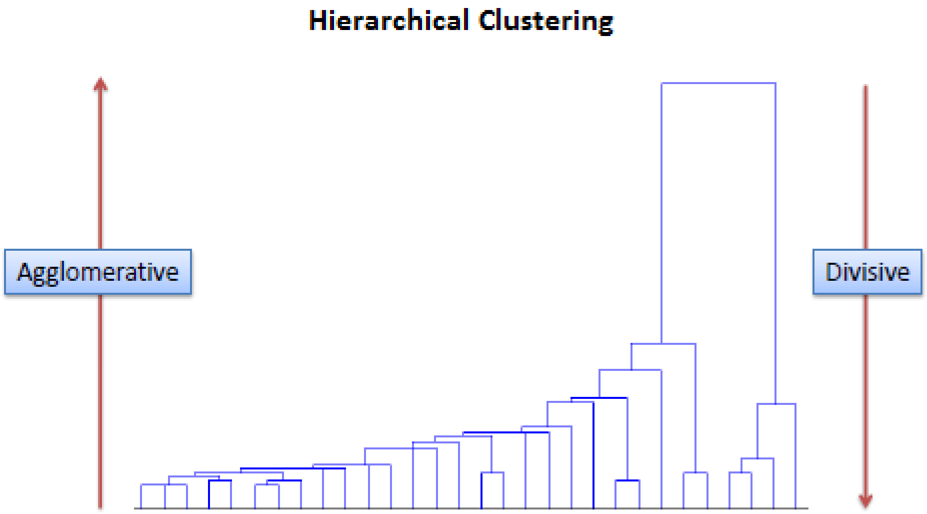
Hierarchical Clustering

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, *Divisive* and *Agglomerative*.

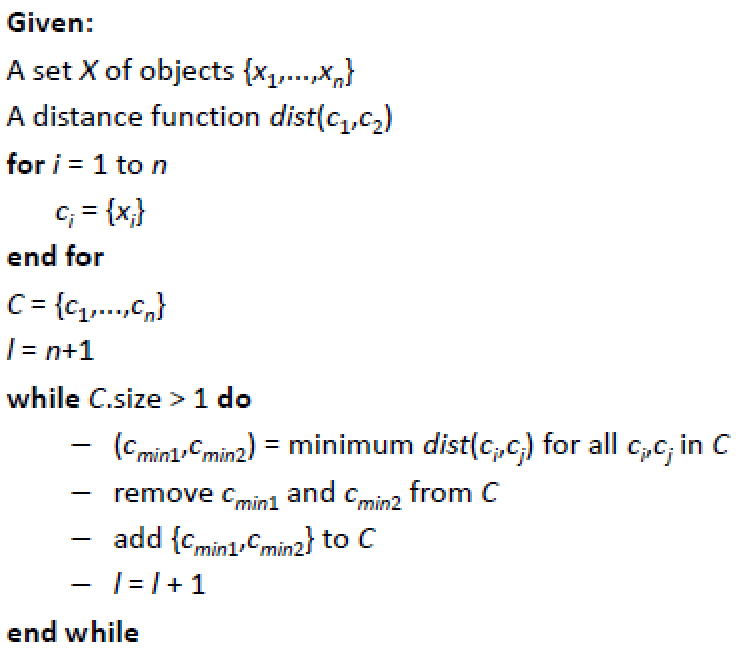


Divisive method

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

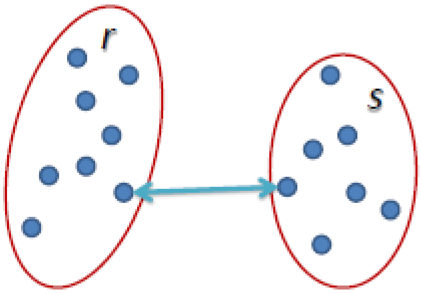


```
Given:
A set $X$ of objects $\{x_1,\ldots,x_n\}$
A distance function $dist(c_1,c_2)$
for $i$ = 1 to $n$
    $c_i = \{x_i\}$
end for
$C = \{c_1,\ldots,c_n\}$
$l = n+1$
while $C$.size > 1 do
    – $(c_{min1},c_{min2})$ = minimum $dist(c_i,c_j)$ for all $c_i,c_j$ in $C$
    – remove $c_{min1}$ and $c_{min2}$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

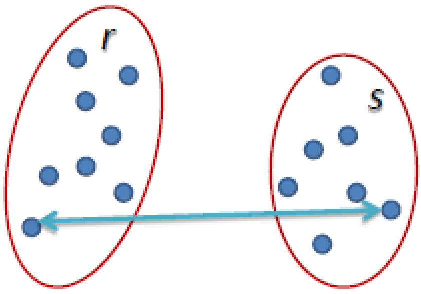
In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.



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

Complete Linkage

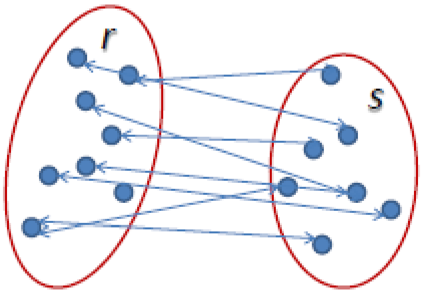
In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.



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

Average Linkage

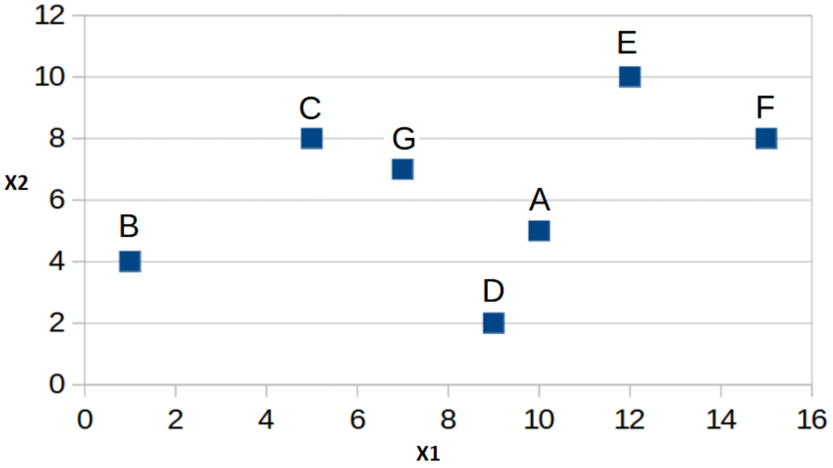
In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters “r” and “s” to the left is equal to the average length each arrow between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



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

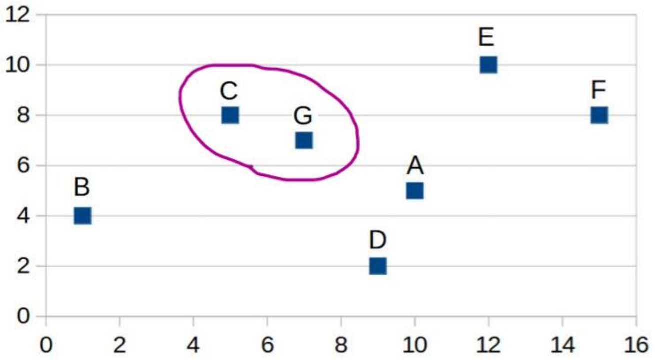
Example: Clustering the following 7 data points.

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



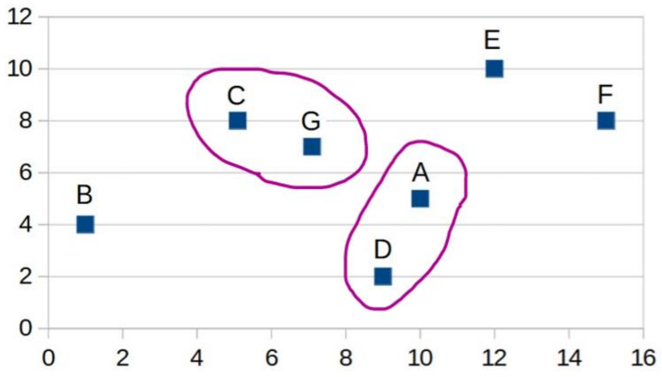
Step 1: Calculate distances between all data points using Euclidean distance function. The shortest distance is between data points C and G.

	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



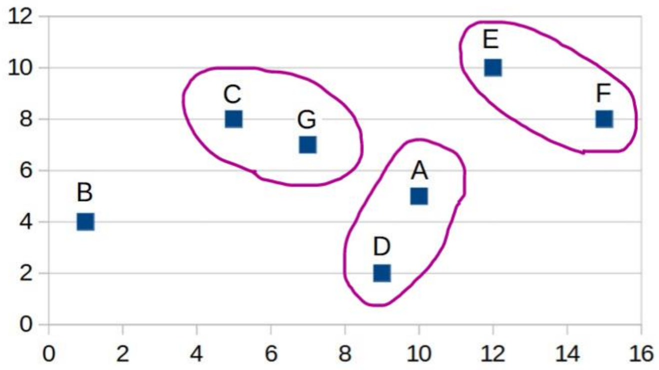
Step 2: We use "Average Linkage" to measure the distance between the "C,G" cluster and other data points.

	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



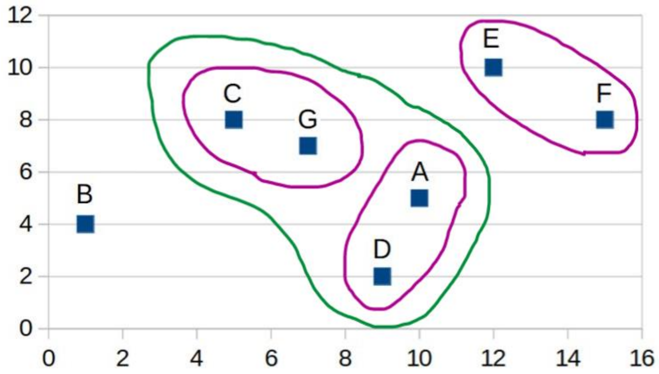
Step 3:

	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**



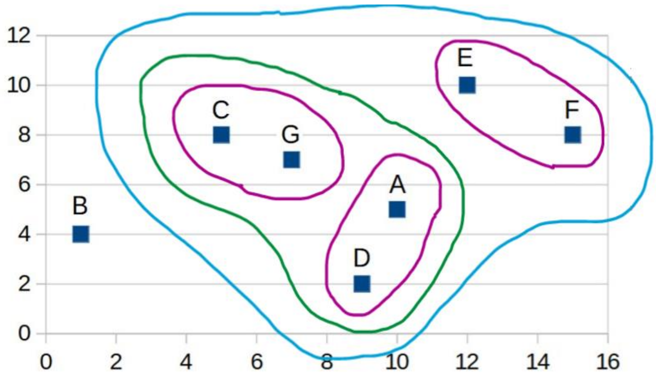
Step 4:

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



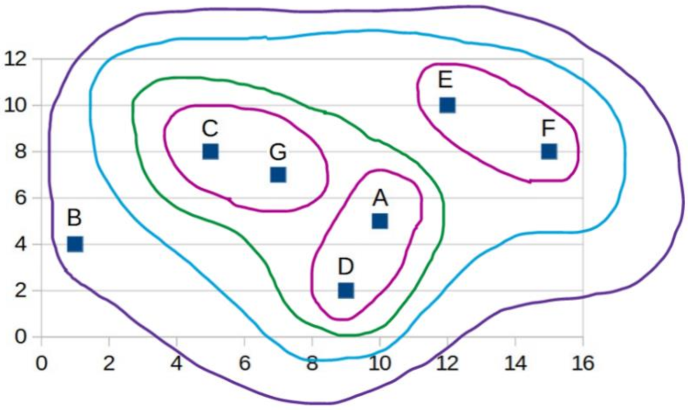
Step 5:

	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
```