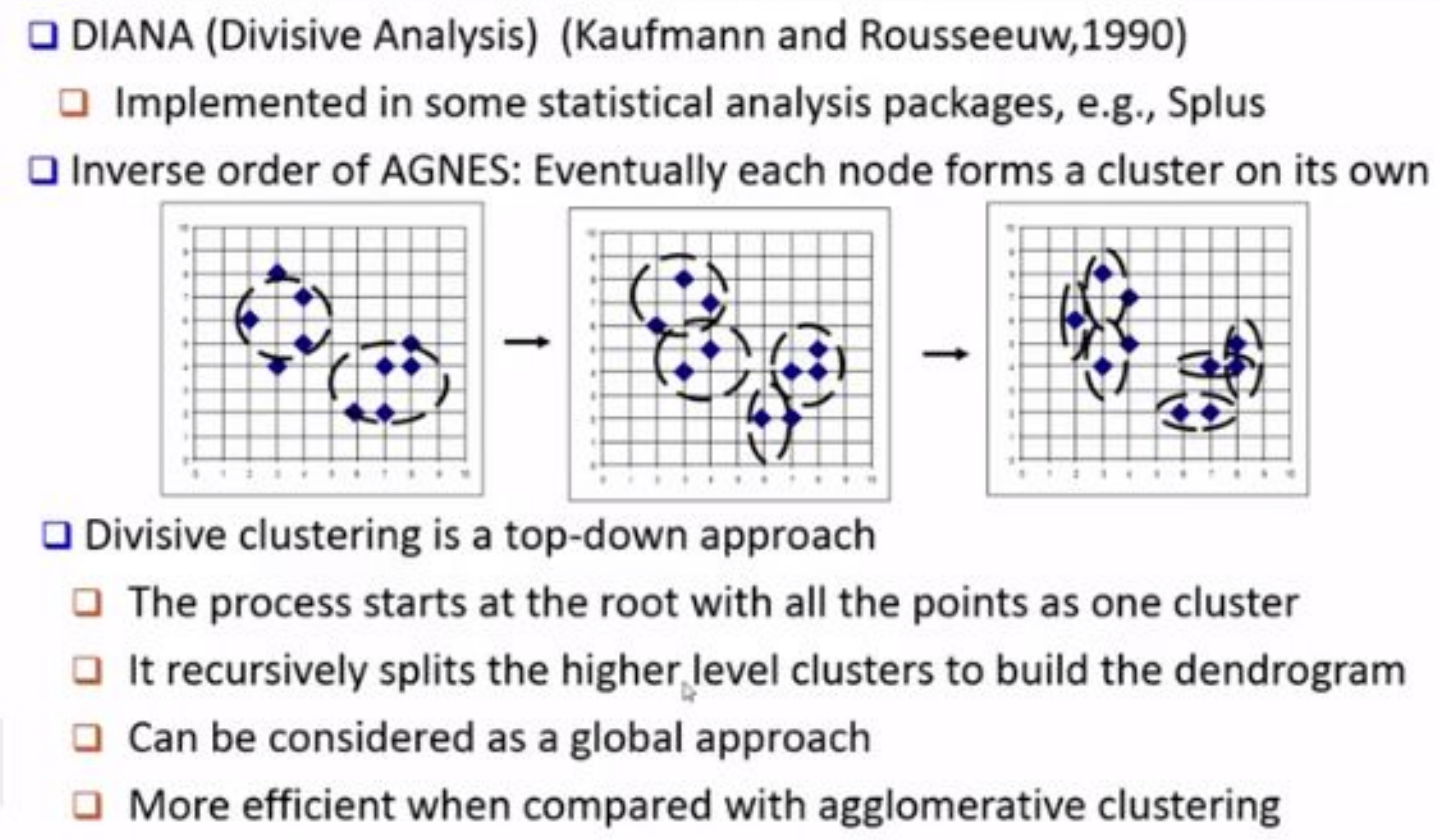
DIANA
Divisive Analysis

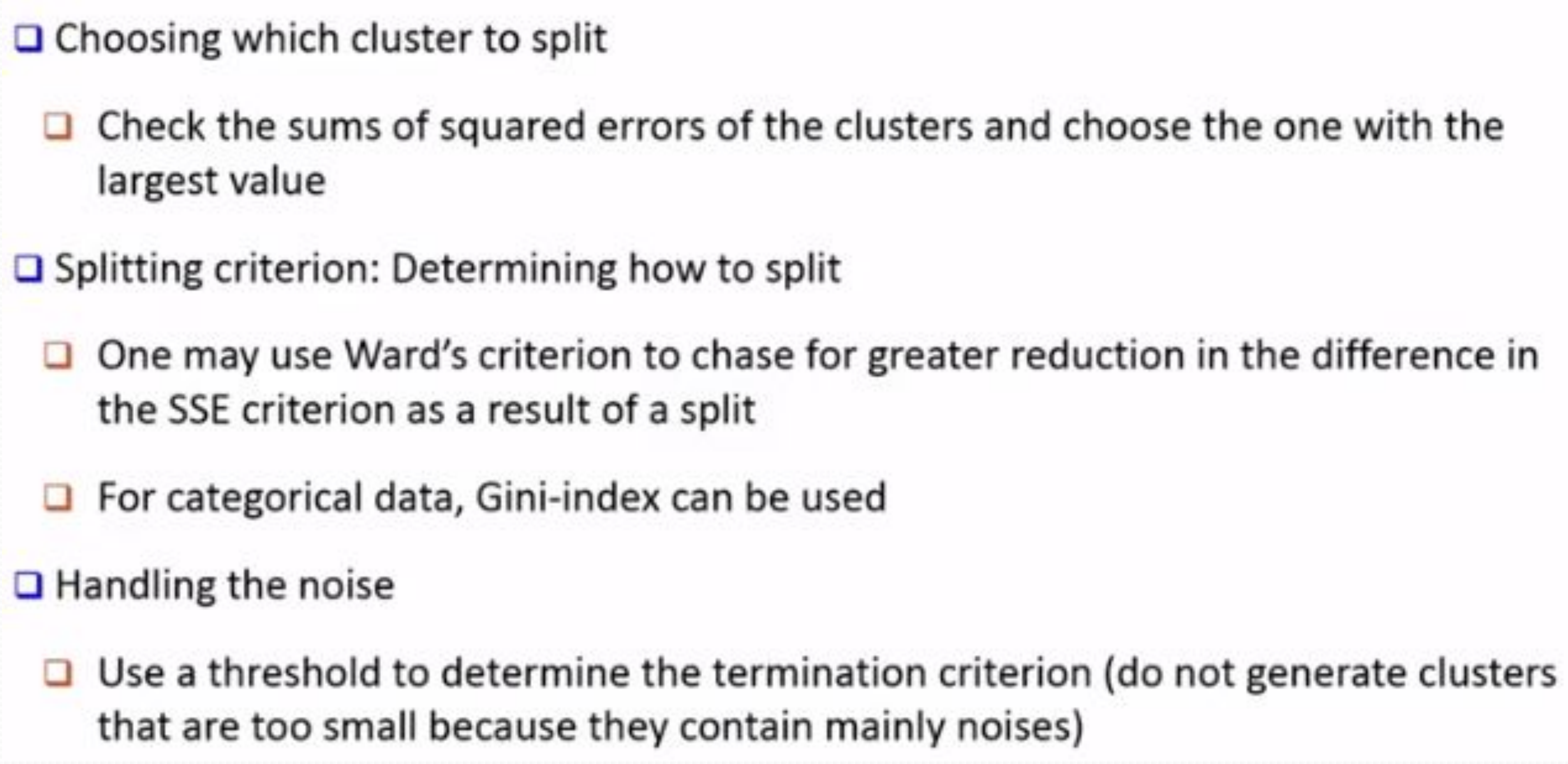
Dr.Mydhili K Nair
ISE Dept.
M S Ramaiah Institute of Technology

Introduction



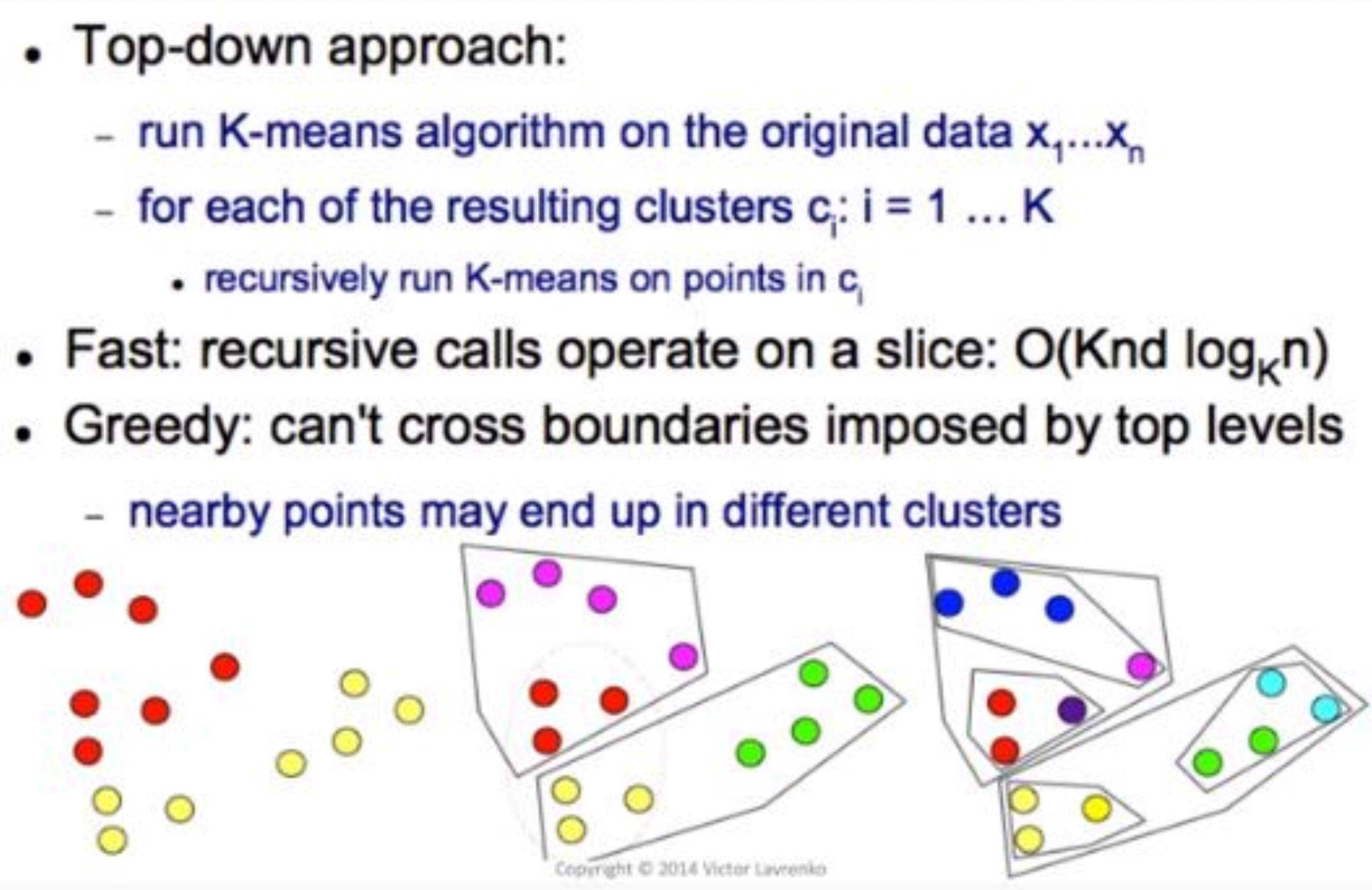
- DIANA (Divisive Analysis) (Kaufmann and Rousseeuw,1990)
  - Implemented in some statistical analysis packages, e.g., Splus
- Inverse order of AGNES: Eventually each node forms a cluster on its own
- Divisive clustering is a top-down approach
  - The process starts at the root with all the points as one cluster
  - It recursively splits the higher level clusters to build the dendrogram
  - Can be considered as a global approach
  - More efficient when compared with agglomerative clustering

Split criteria



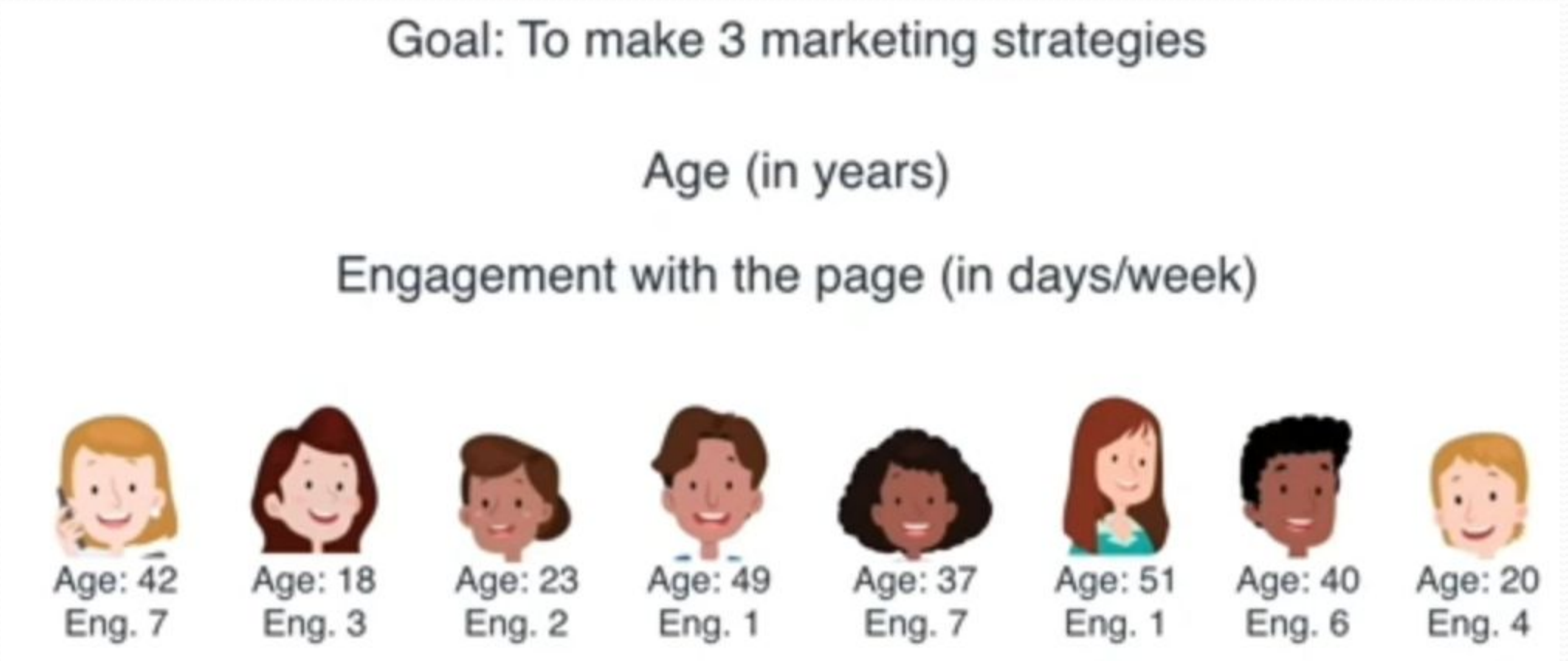
- ❑ Choosing which cluster to split
  - ❑ Check the sums of squared errors of the clusters and choose the one with the largest value
- ❑ Splitting criterion: Determining how to split
  - ❑ One may use Ward's criterion to chase for greater reduction in the difference in the SSE criterion as a result of a split
  - ❑ For categorical data, Gini-index can be used
- ❑ Handling the noise
  - ❑ Use a threshold to determine the termination criterion (do not generate clusters that are too small because they contain mainly noises)

Simplest Algorithm to split



- Top-down approach:
  - run K-means algorithm on the original data $x_1 \ldots x_n$
  - for each of the resulting clusters c_i: $i = 1 \ldots K$
    - recursively run K-means on points in c_i
- Fast: recursive calls operate on a slice: $O(Knd \log_K n)$
- Greedy: can't cross boundaries imposed by top levels
  - nearby points may end up in different clusters

Customer clustering



Goal: To make 3 marketing strategies

Age (in years)

Engagement with the page (in days/week)

Age: 42
Eng. 7

Age: 18
Eng. 3

Age: 23
Eng. 2

Age: 49
Eng. 1

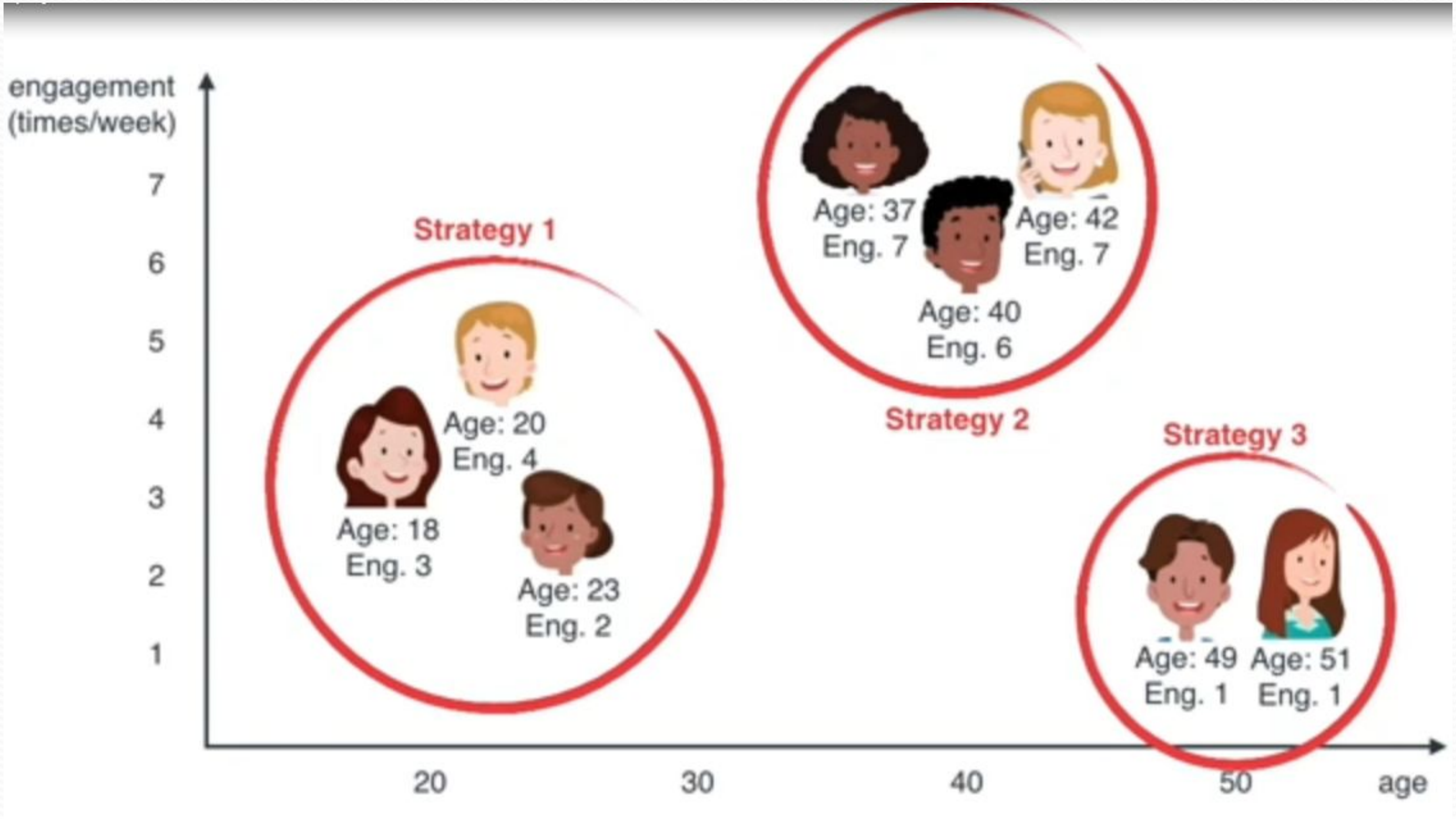
Age: 37
Eng. 7

Age: 51
Eng. 1

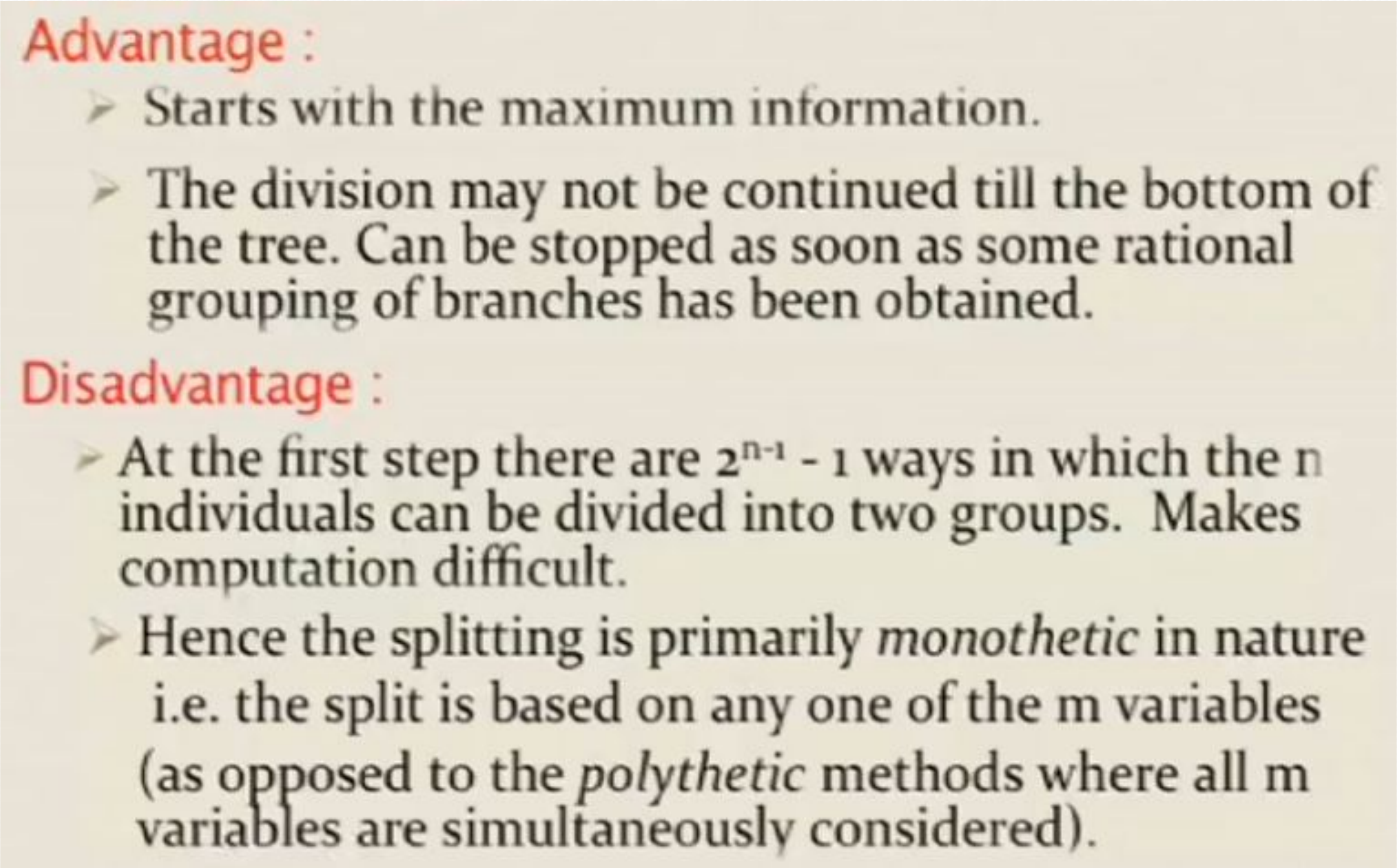
Age: 40
Eng. 6

Age: 20
Eng. 4

Data plot


engagement
(times/week)
7
6
5
4
3
2
1
Strategy 1
Age: 20
Eng. 4
Age: 18
Eng. 3
Age: 23
Eng. 2
Age: 37
Eng. 7
Age: 42
Eng. 7
Age: 40
Eng. 6
Strategy 2
Strategy 3
Age: 49
Eng. 1
Age: 51
Eng. 1
20
30
40
50
age

Pros and cons



Advantage :

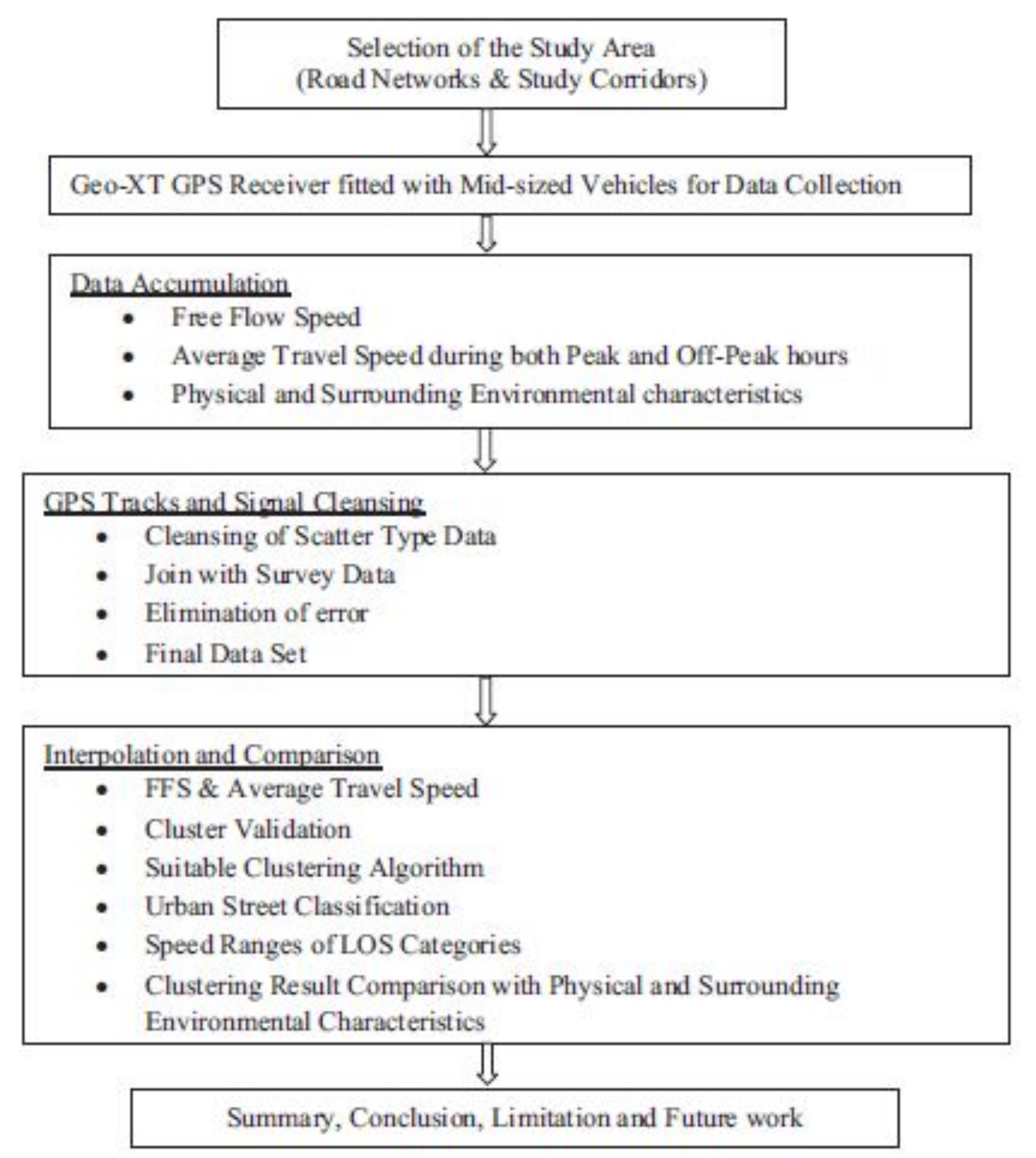
- Starts with the maximum information.
- The division may not be continued till the bottom of the tree. Can be stopped as soon as some rational grouping of branches has been obtained.

Disadvantage :

- At the first step there are $2^{n-1} - 1$ ways in which the n individuals can be divided into two groups. Makes computation difficult.
- Hence the splitting is primarily *monothetic* in nature i.e. the split is based on any one of the m variables (as opposed to the *polythetic* methods where all m variables are simultaneously considered).

An example

Level of Service (LOS) for heterogeneous traffic flow on urban streets



Selection of the Study Area
(Road Networks & Study Corridors)

Geo-XT GPS Receiver fitted with Mid-sized Vehicles for Data Collection

Data Accumulation

- Free Flow Speed
- Average Travel Speed during both Peak and Off-Peak hours
- Physical and Surrounding Environmental characteristics

GPS Tracks and Signal Cleansing

- Cleansing of Scatter Type Data
- Join with Survey Data
- Elimination of error
- Final Data Set

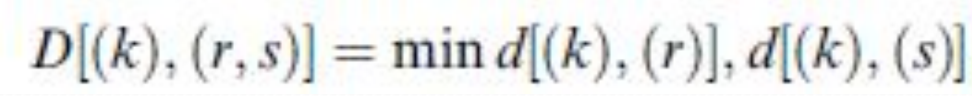
Interpolation and Comparison

- FFS & Average Travel Speed
- Cluster Validation
- Suitable Clustering Algorithm
- Urban Street Classification
- Speed Ranges of LOS Categories
- Clustering Result Comparison with Physical and Surrounding Environmental Characteristics

Summary, Conclusion, Limitation and Future work

Algorithm steps

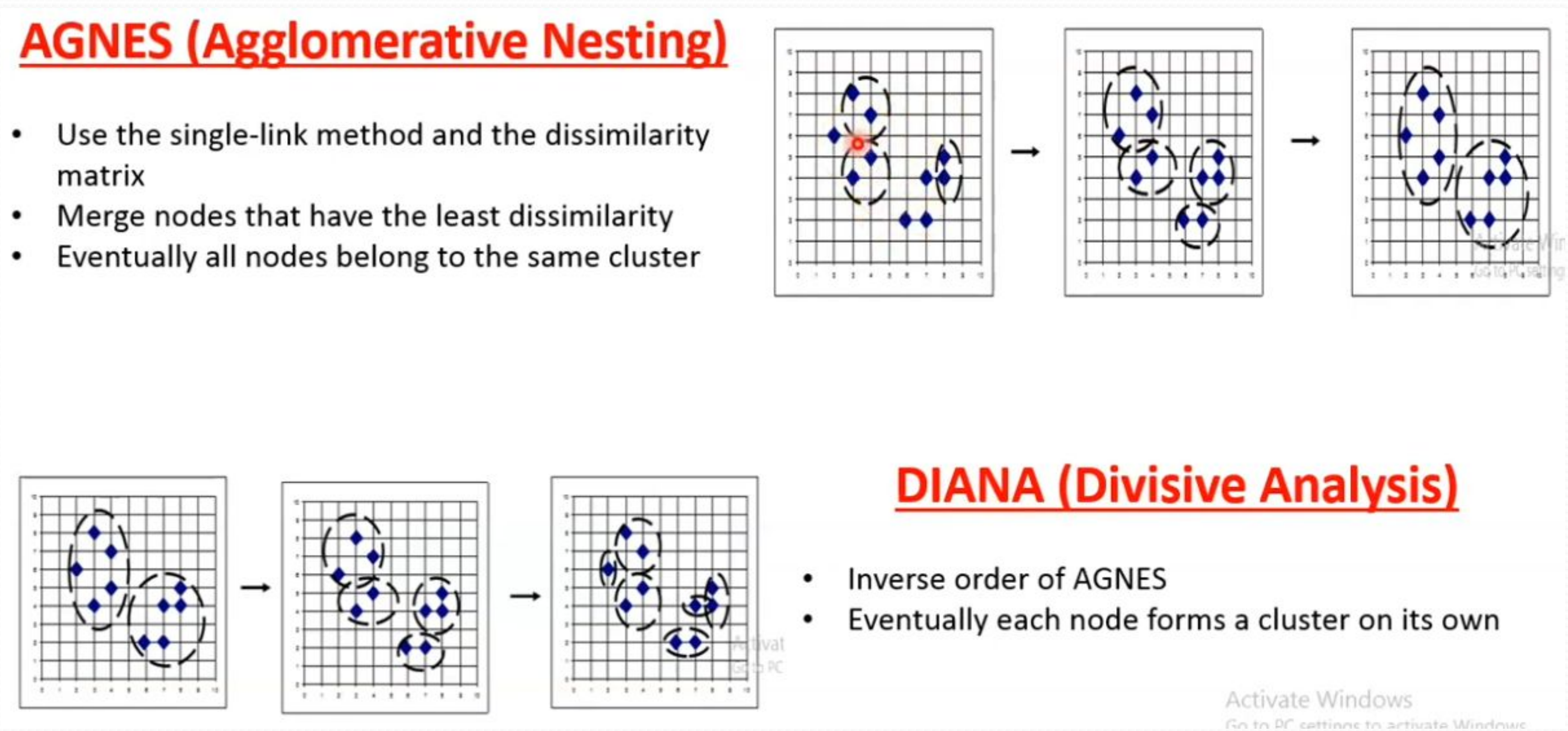
- Step-1: The DIANA clustering is followed by Agglomerative Hierarchical Clustering up to the cluster contains all the objects. Then the Divisive Analysis Clustering (DIANA) follows the top-down approach assuming it single cluster having level L (0) =n and sequence number m= 0.
- Step-2: The most dissimilar pair of clusters in the current cluster is found out; that is (r), (s) in which d [(r), (s)] =min d [(i), (j)], where min is the complete pairs of cluster in the current cluster.
- Step-3: The sequence number is incremented in the manner m= m+ 1. The cluster is broken into clusters (r) and (s) to form next cluster to make the level of clustering: L (m1) =d [(r)] and L (m2)= d [(s)].
- Step-4: The distance matrix (D) is updated by adding the rows and columns corresponding to clusters (r) and (s). The similarity between the new cluster, denoted by (r, s) and old cluster (k) is defined in this way:



$$D[(k),(r,s)] = \min d[(k),(r)], d[(k),(s)]$$

- If all objects are distinct clusters, then stop; otherwise proceed to step-2.

AGNES vs. DIANA

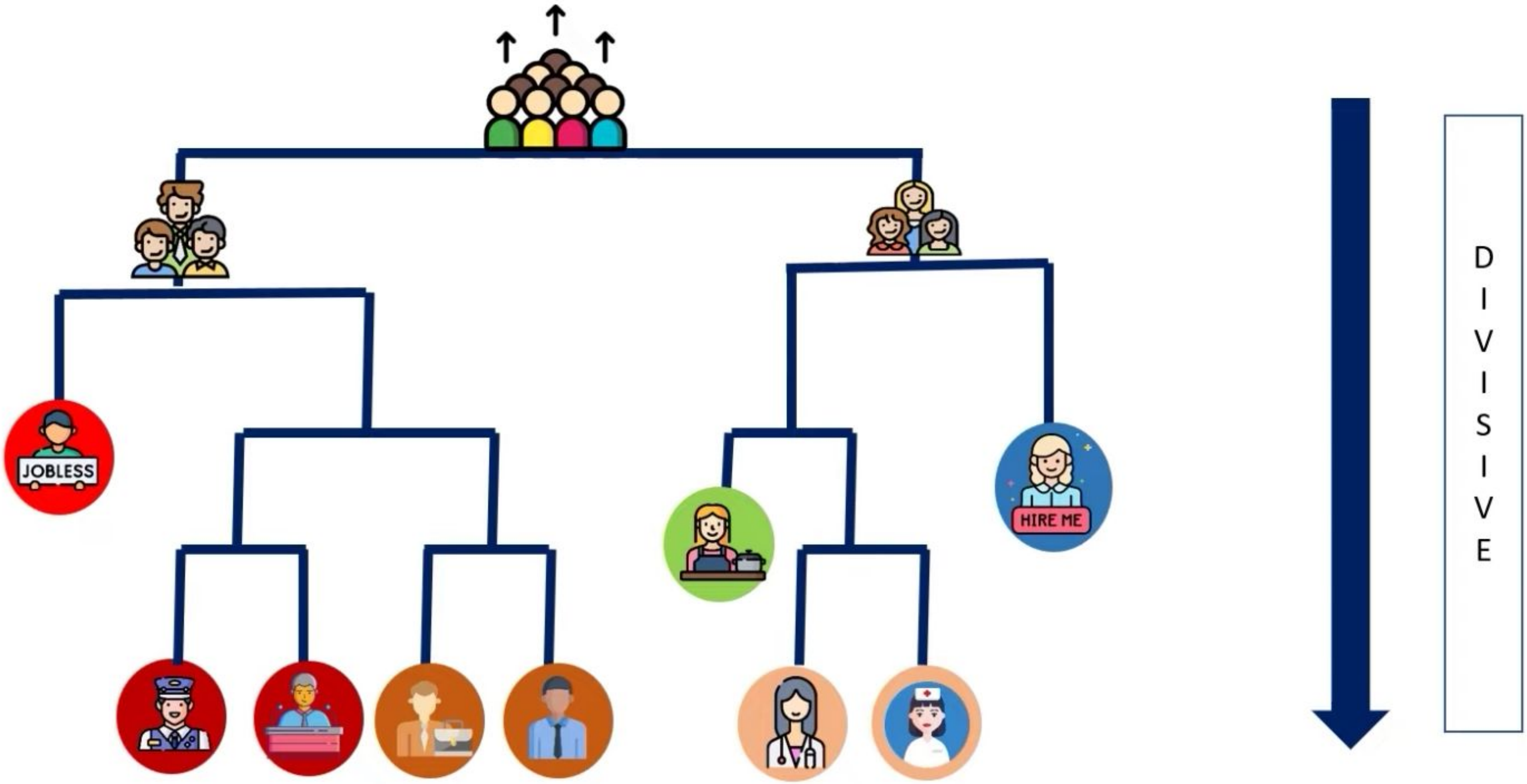


AGNES (Agglomerative Nesting)

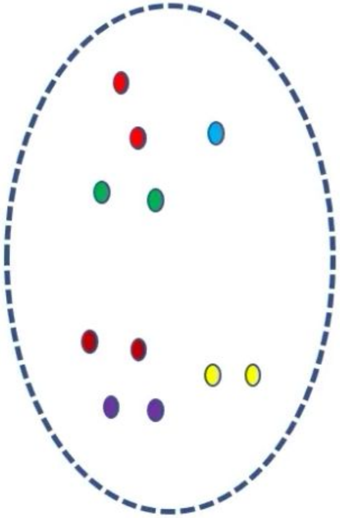
- Use the single-link method and the dissimilarity matrix
- Merge nodes that have the least dissimilarity
- Eventually all nodes belong to the same cluster

DIANA (Divisive Analysis)

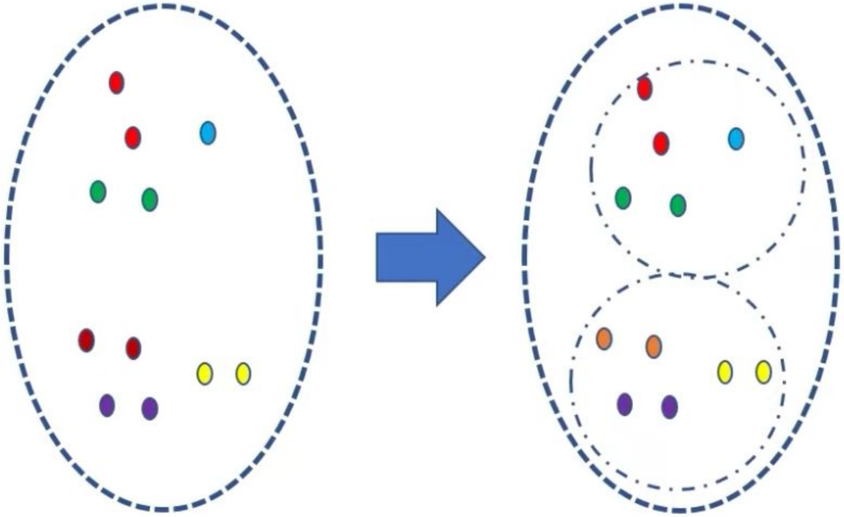
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



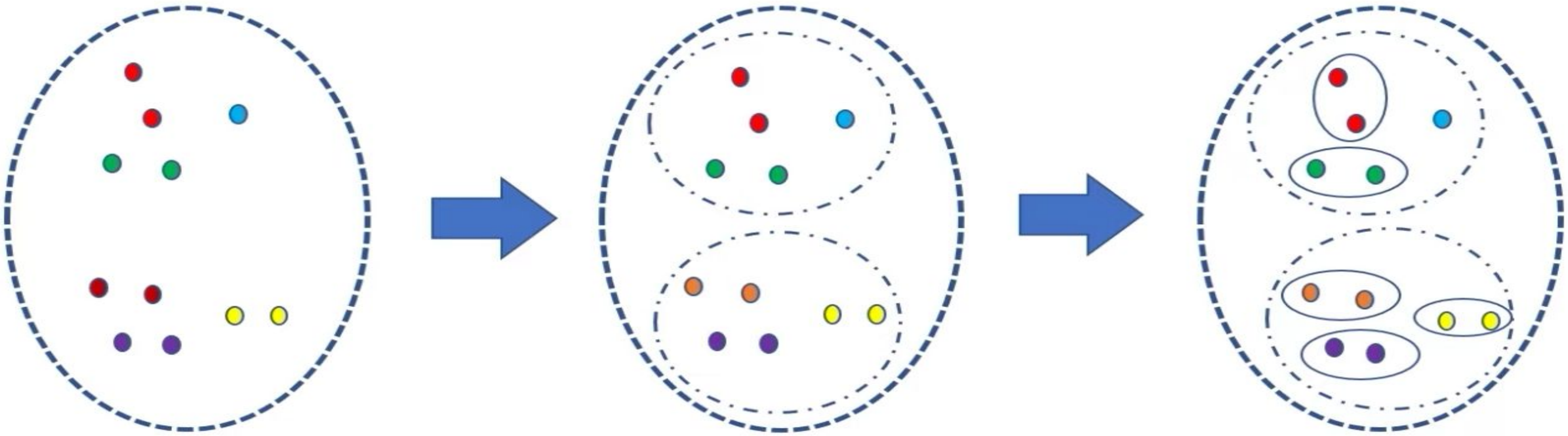
DIVISIVE

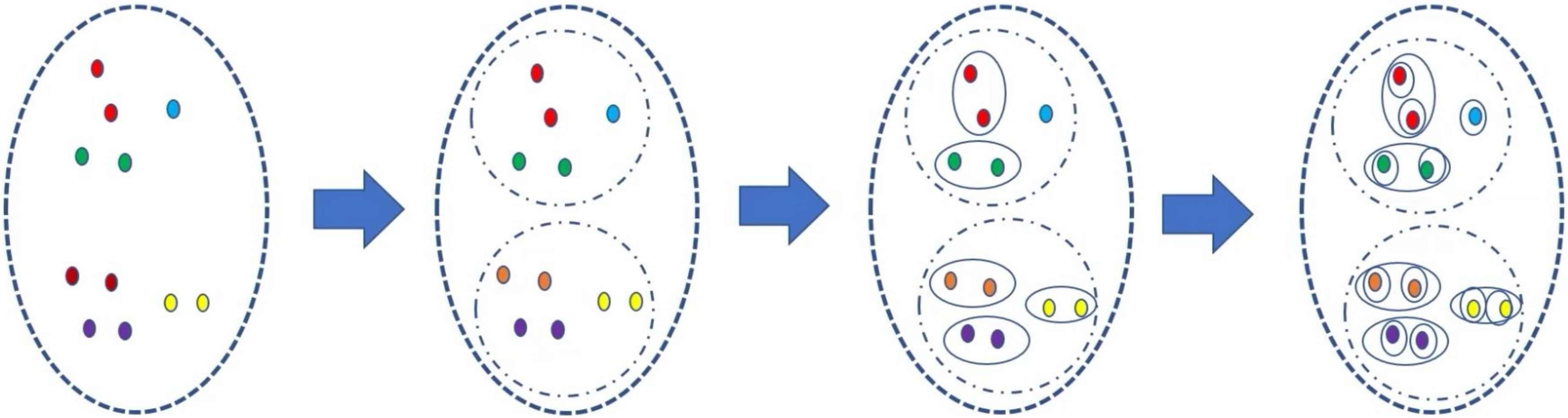


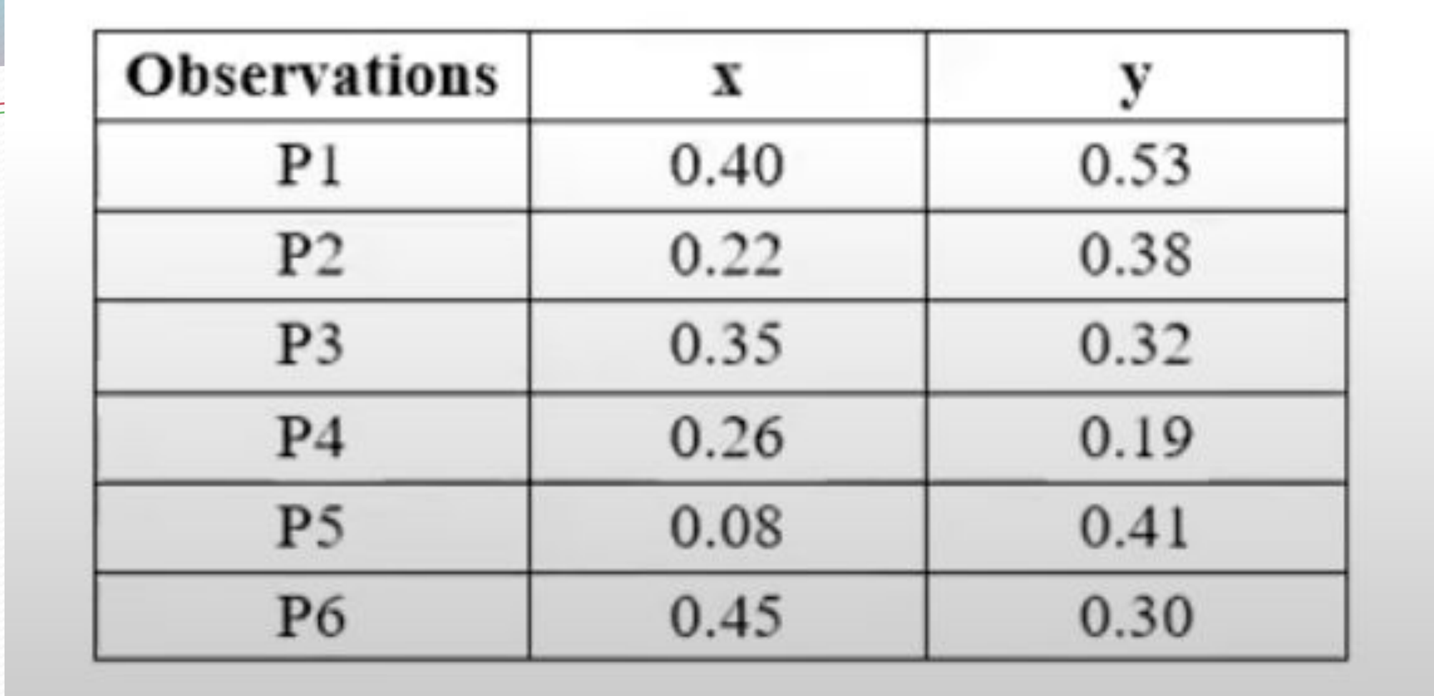
All data points are treated as a single cluster.



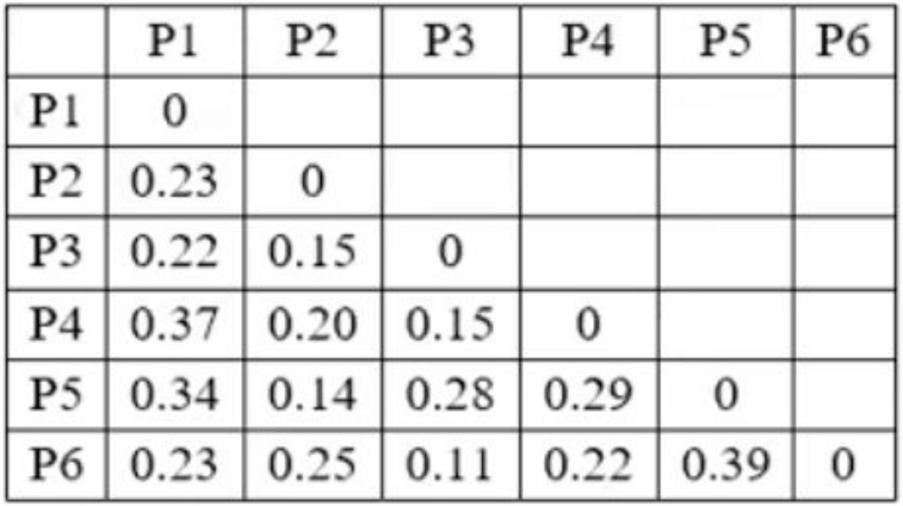
Cluster is partitioned into two least similar cluster.







Observations	x	y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30



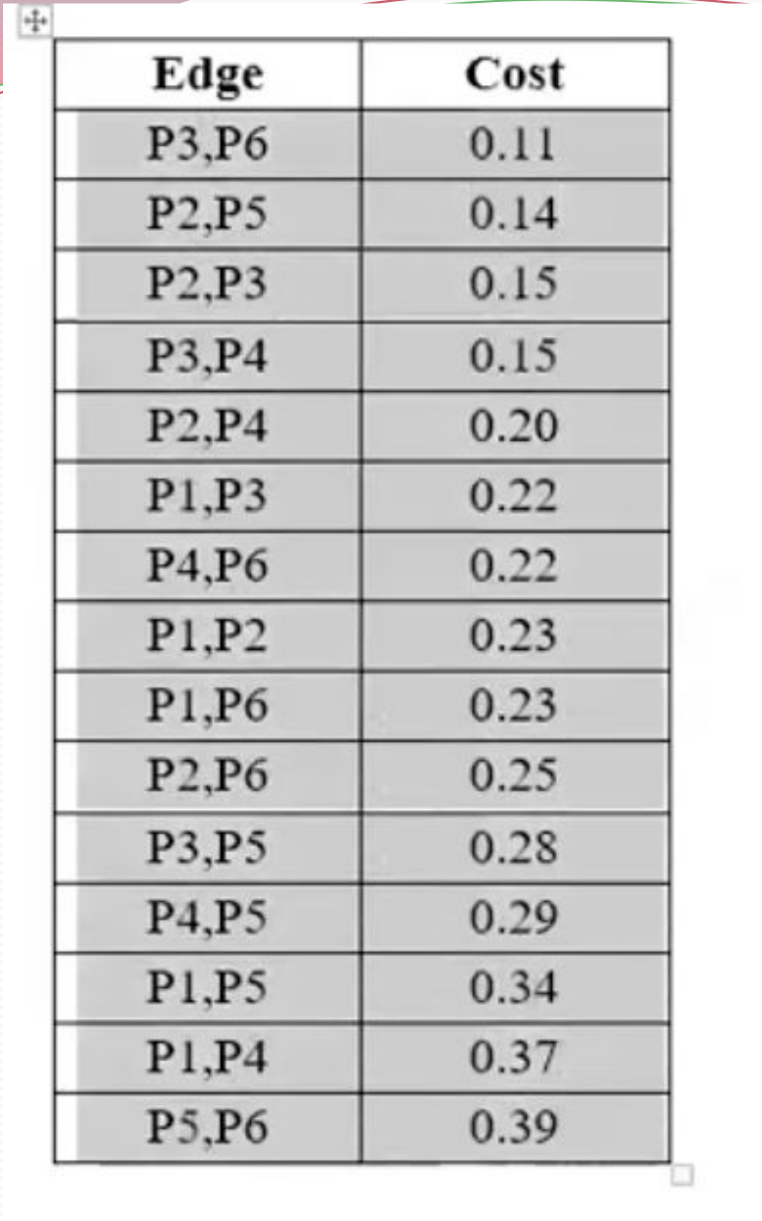
	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

Source: https://www.youtube.com/watch?v=p-1NFmePDp4

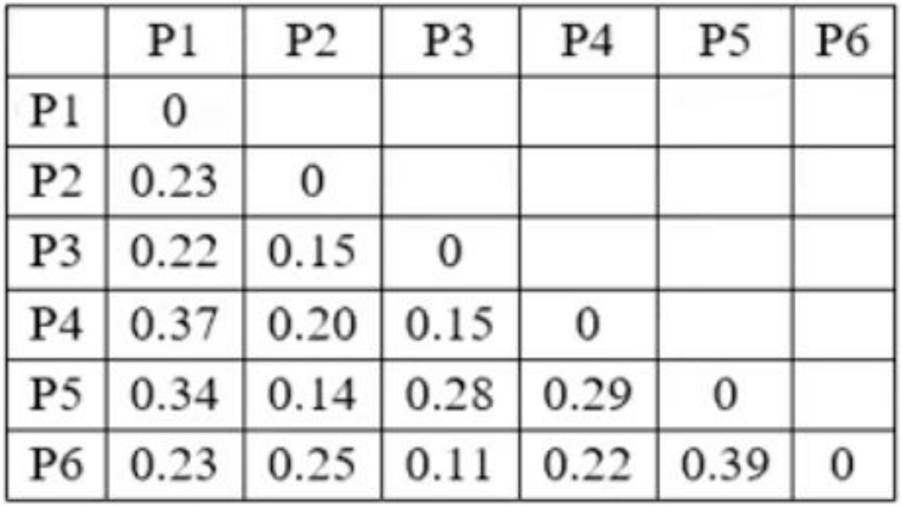
Minimum Spanning Tree (MST)

- **Prims**
- **Kruskal**

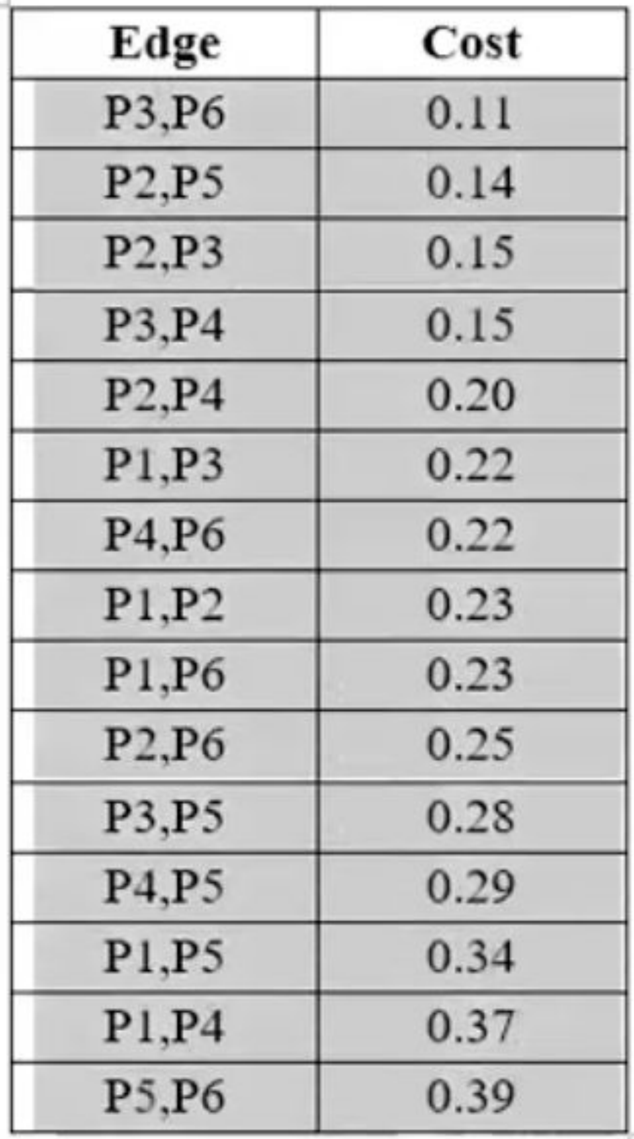
Greedy Algorithm



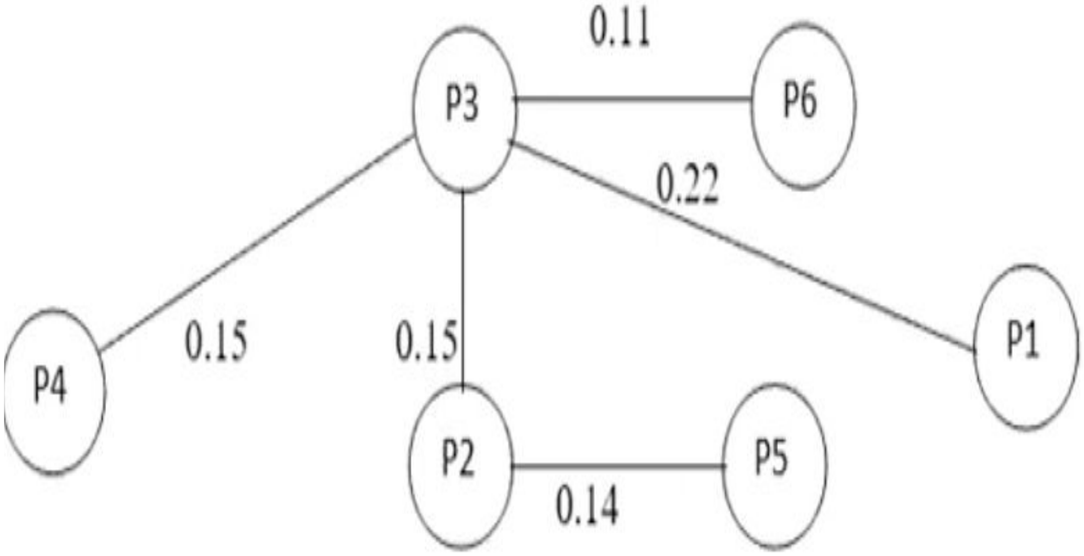
Edge	Cost
P3,P6	0.11
P2,P5	0.14
P2,P3	0.15
P3,P4	0.15
P2,P4	0.20
P1,P3	0.22
P4,P6	0.22
P1,P2	0.23
P1,P6	0.23
P2,P6	0.25
P3,P5	0.28
P4,P5	0.29
P1,P5	0.34
P1,P4	0.37
P5,P6	0.39



	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0



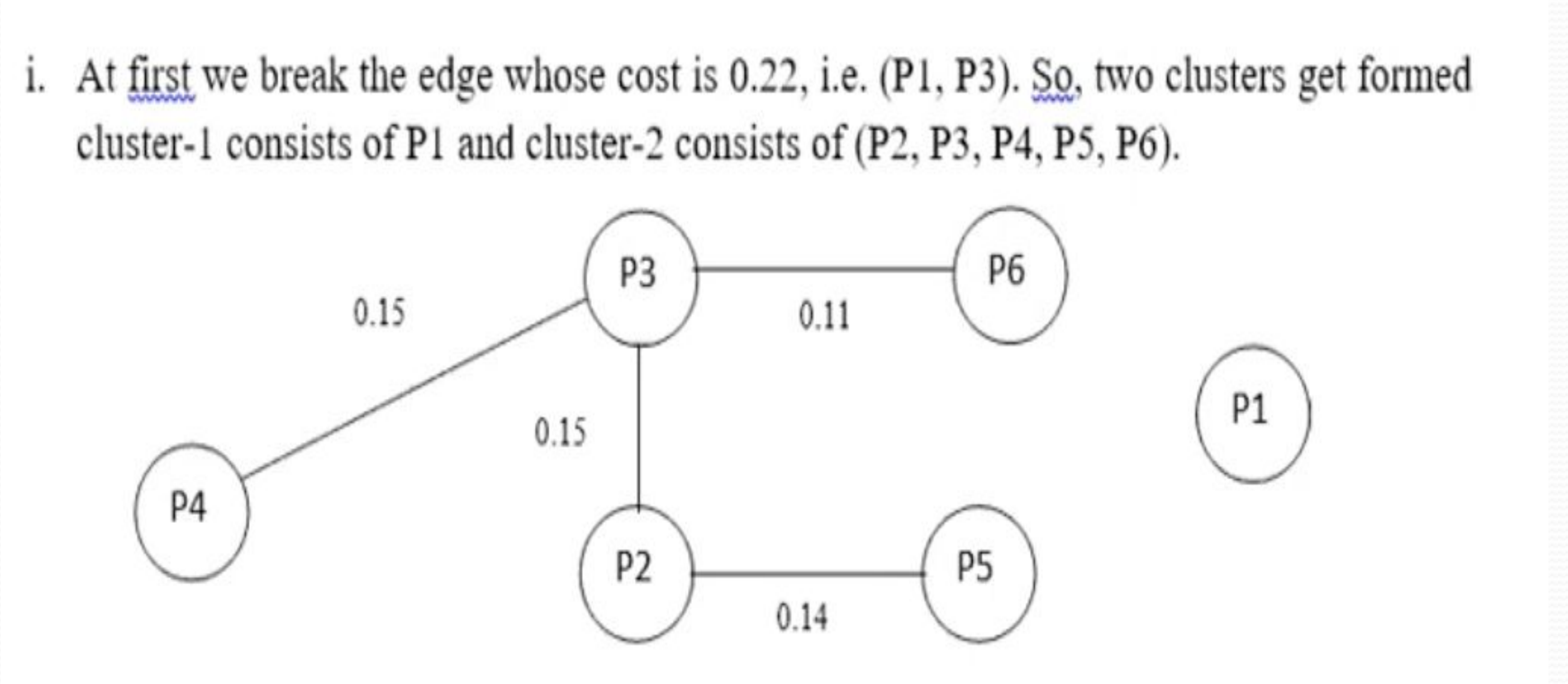
Edge	Cost
P3,P6	0.11
P2,P5	0.14
P2,P3	0.15
P3,P4	0.15
P2,P4	0.20
P1,P3	0.22
P4,P6	0.22
P1,P2	0.23
P1,P6	0.23
P2,P6	0.25
P3,P5	0.28
P4,P5	0.29
P1,P5	0.34
P1,P4	0.37
P5,P6	0.39



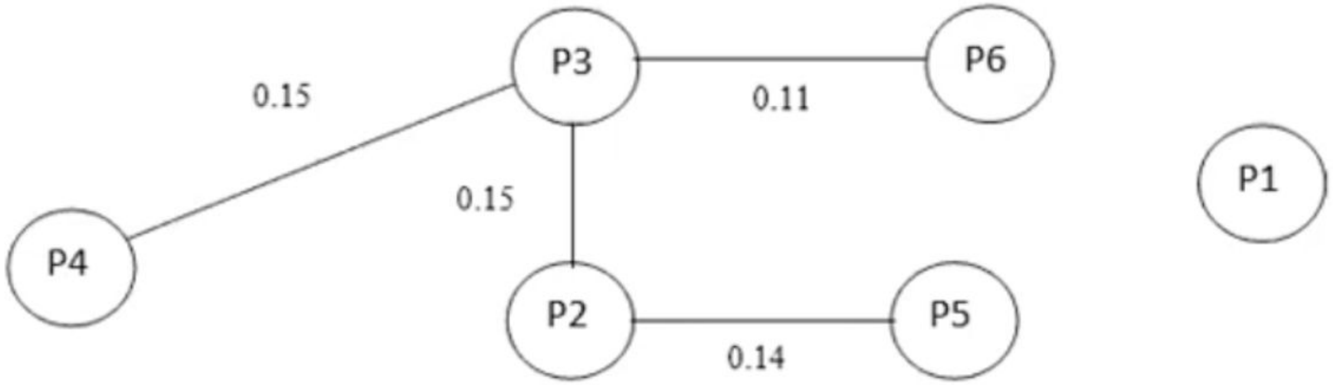
MST: It is a ***Greedy Approach*** because it always starts with the minimum value of distance/cost.

All edges between the points are connected,until all the points are included / made to participate,***BUT*** *there must be no closed loop or circuit.*

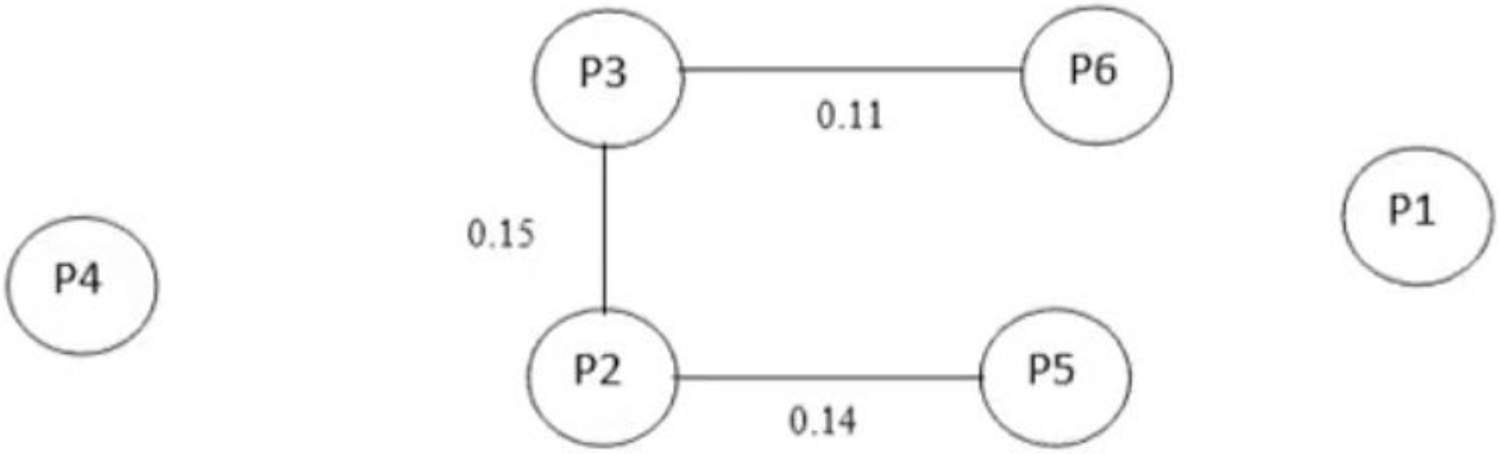
Complete Linkage : Max Value Split

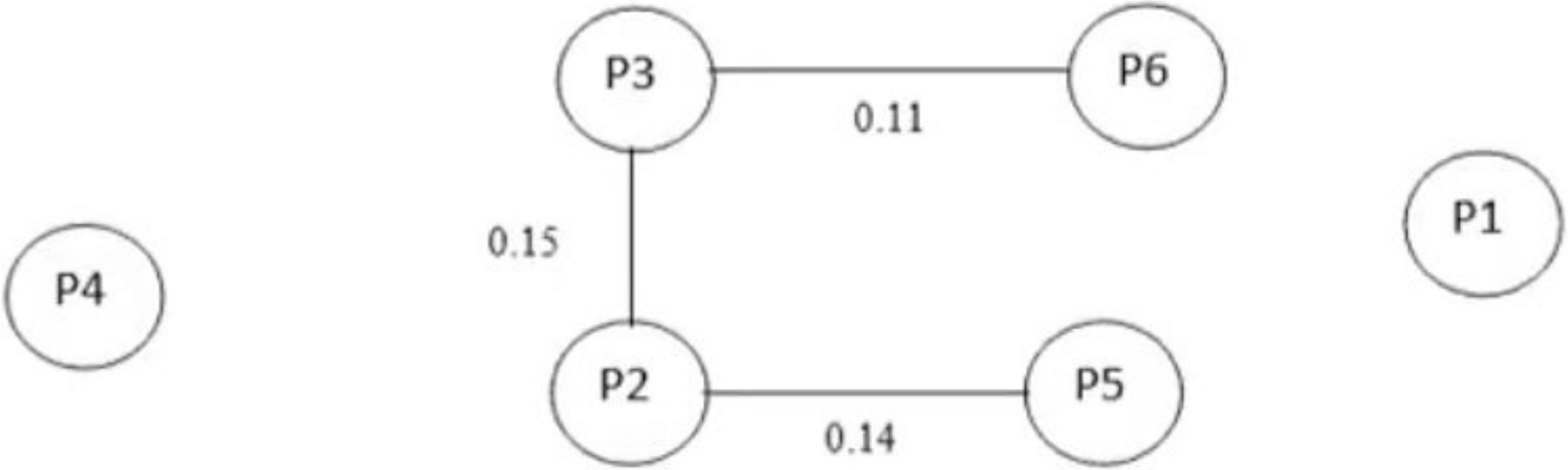


i. At first we break the edge whose cost is 0.22, i.e. (P1, P3). So, two clusters get formed cluster-1 consists of P1 and cluster-2 consists of (P2, P3, P4, P5, P6).

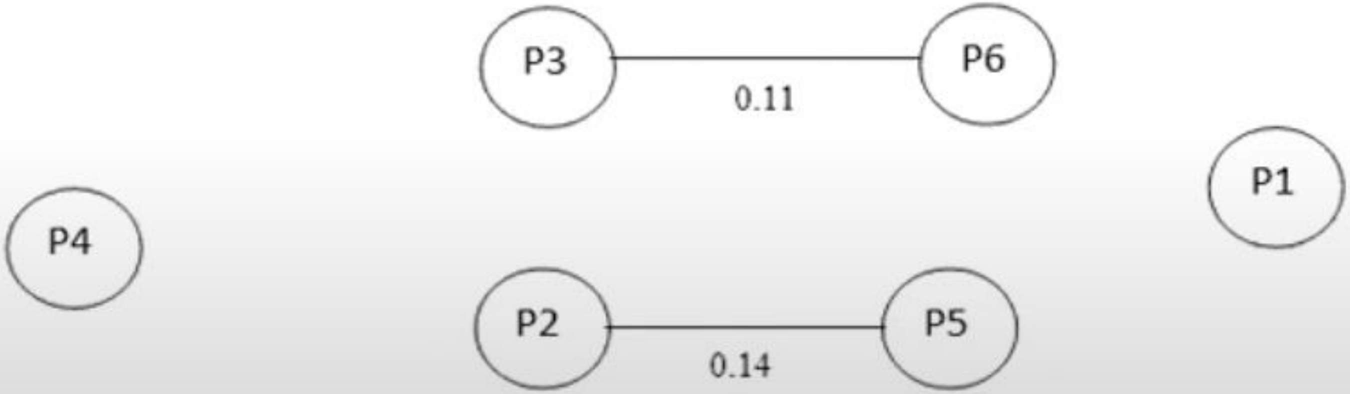


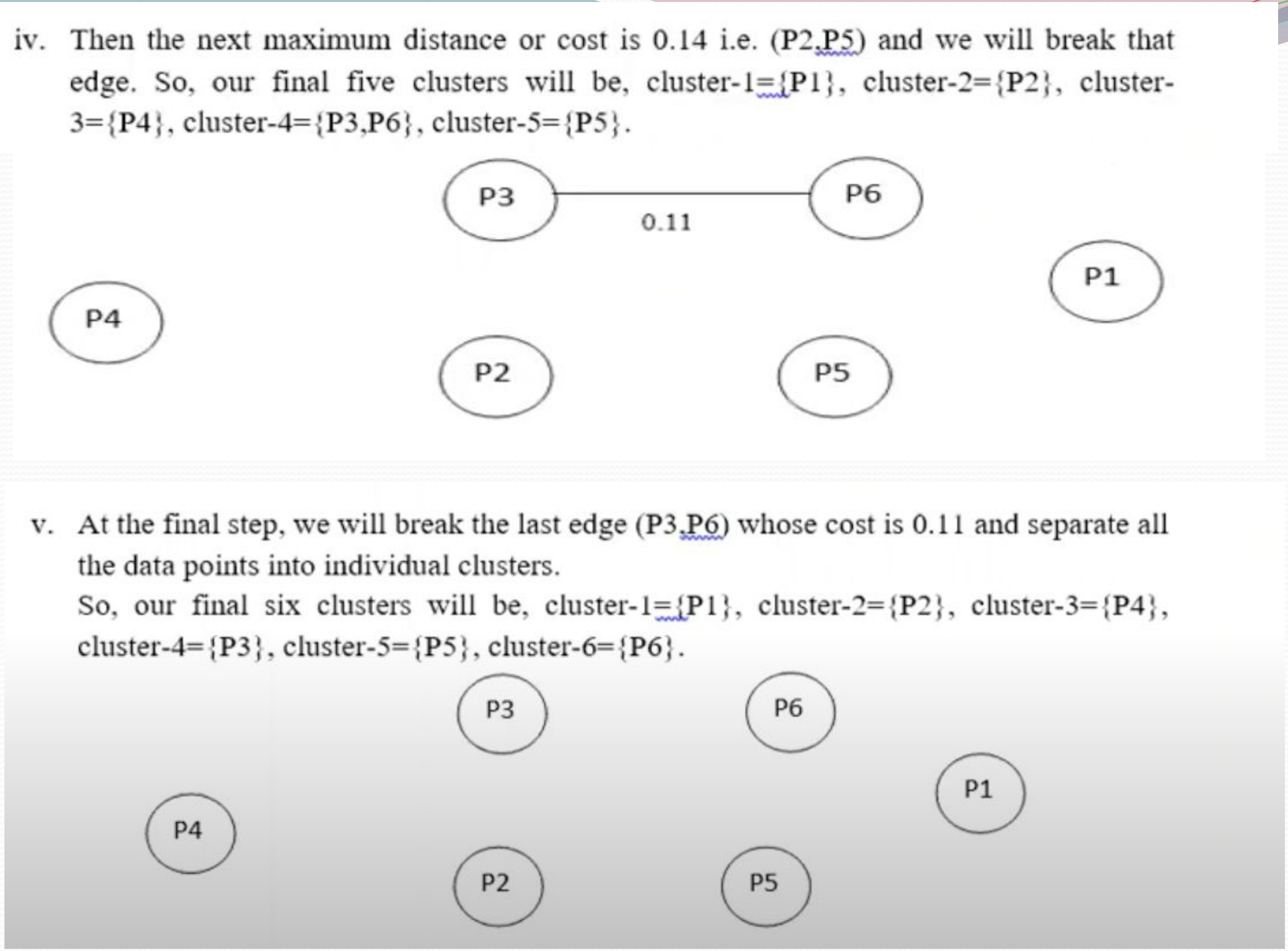
ii. Now, we will seek for the next maximum distance greedily and we will choose (P3,P4) as the next highest cost i.e. 0.15 and break that edge to create three separate clusters, cluster-1={P1}, cluster-2={P2,P3,P5,P6}, cluster-3={P4}. We will continue this splitting iteration until each new cluster contains only a single object.





iii. Now, the next maximum cost is (P2,P3) i.e. 0.15 and we will break that edge. So, our final four clusters will be, cluster-1={P1}, cluster-2={P2,P5}, cluster-3={P4}, cluster-4={P3,P6}.

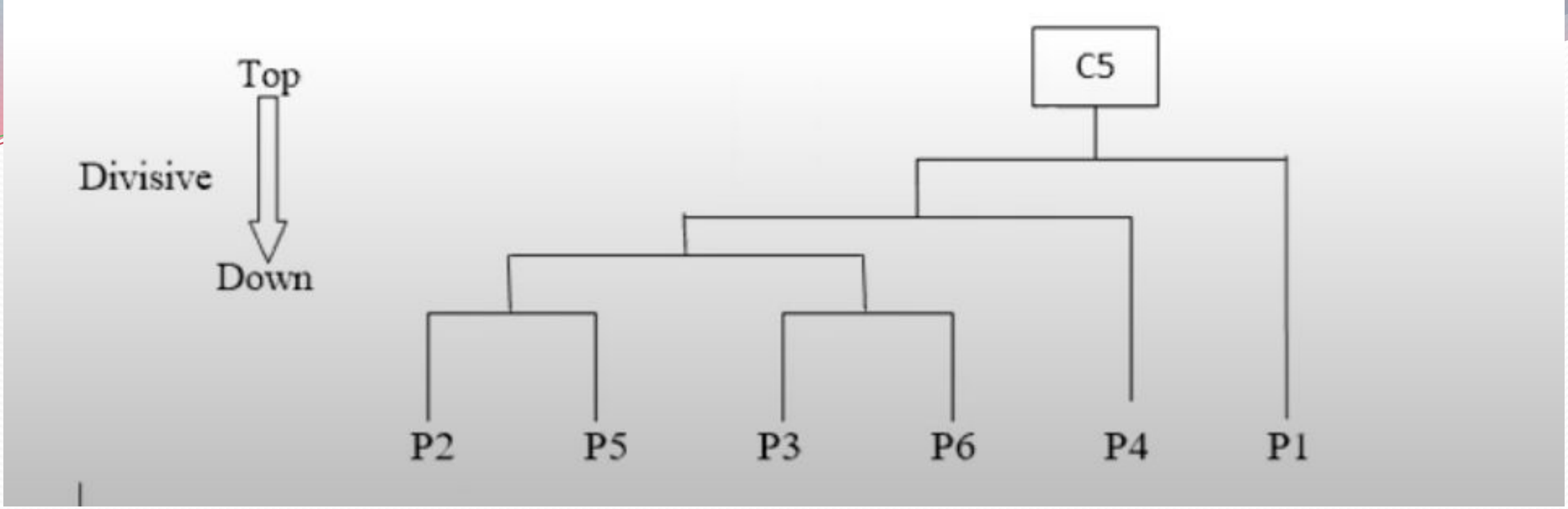




iv. Then the next maximum distance or cost is 0.14 i.e. (P2,P5) and we will break that edge. So, our final five clusters will be, cluster-1={P1}, cluster-2={P2}, cluster-3={P4}, cluster-4={P3,P6}, cluster-5={P5}.

v. At the final step, we will break the last edge (P3,P6) whose cost is 0.11 and separate all the data points into individual clusters.

So, our final six clusters will be, cluster-1={P1}, cluster-2={P2}, cluster-3={P4}, cluster-4={P3}, cluster-5={P5}, cluster-6={P6}.


Top
Divisive
Down
C5
P2
P5
P3
P6
P4
P1

Divisive Approach by Minimum Spanning Tree (MST) Concept

1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0
	1	**2**	**3**	**4**	**5**

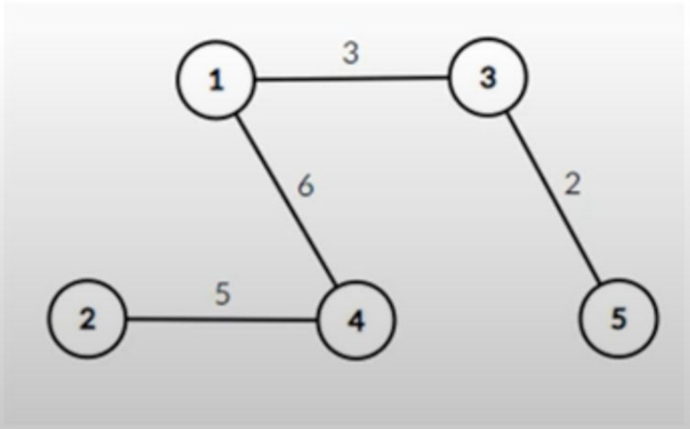
Draw MST by either Kruskal's or Prim's Algorithm

Kruslak's Method: Arrange edge in ascending order of their cost.

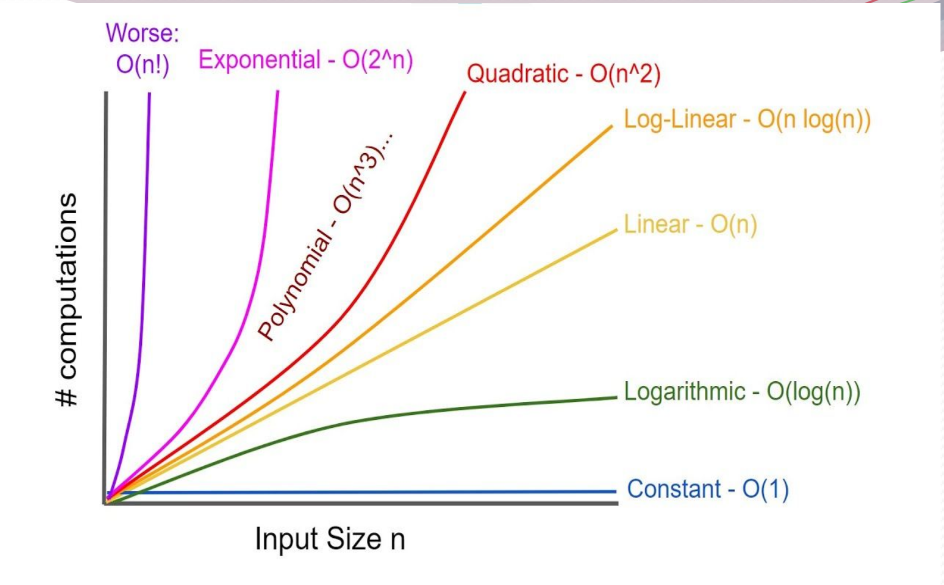
Edge	Cost
3-5	2
1-3	3
2-4	5
1-4	6
2-3	7
4-5	8
1-2	9
3-4	9
2-5	10
4-5	11

Minimum Spanning Tree (MST)

Edge	Cost
3-5	2
1-3	3
2-4	5
1-4	6

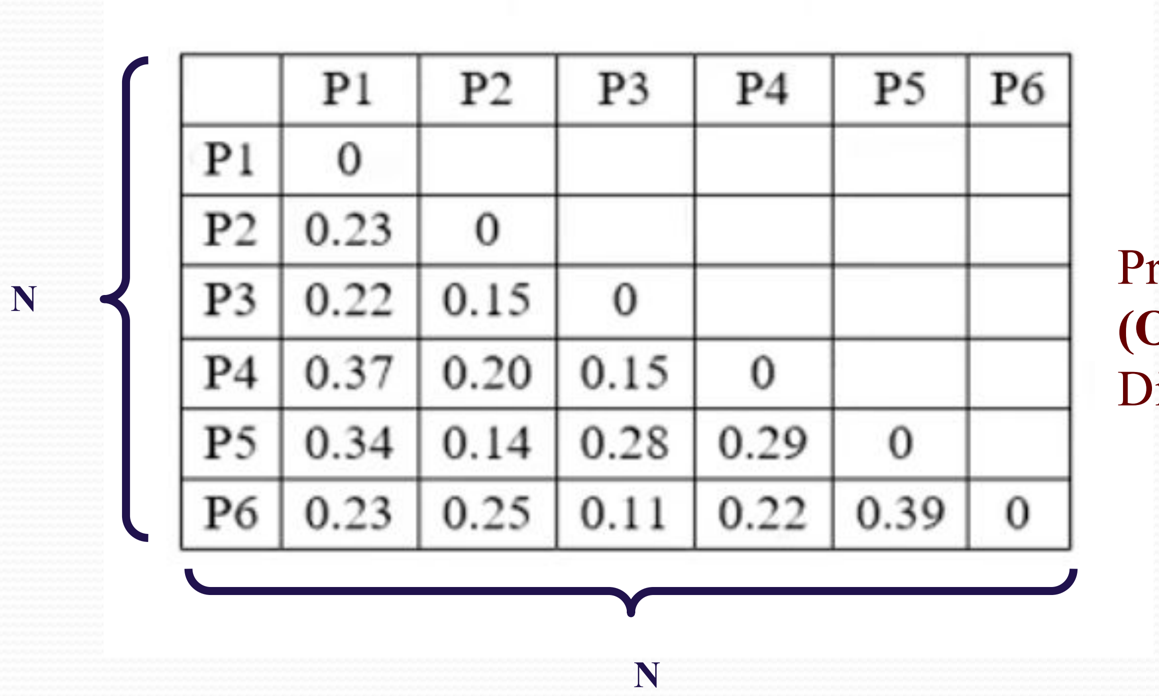


TIME COMPLEXITY ANALYSIS



Given: Time Complexity of Hierarchial Clustering is $O(n^3)$

How?



	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

Proximity Matrix
(OR)
Distance Matrix

We exhaustively scan the N x N matrix dist_mat for

- lowest distance(in SLINK-default in AGNES)
- highest distance(in CLINK-default in DIANA)

in each of N-1 iterations.

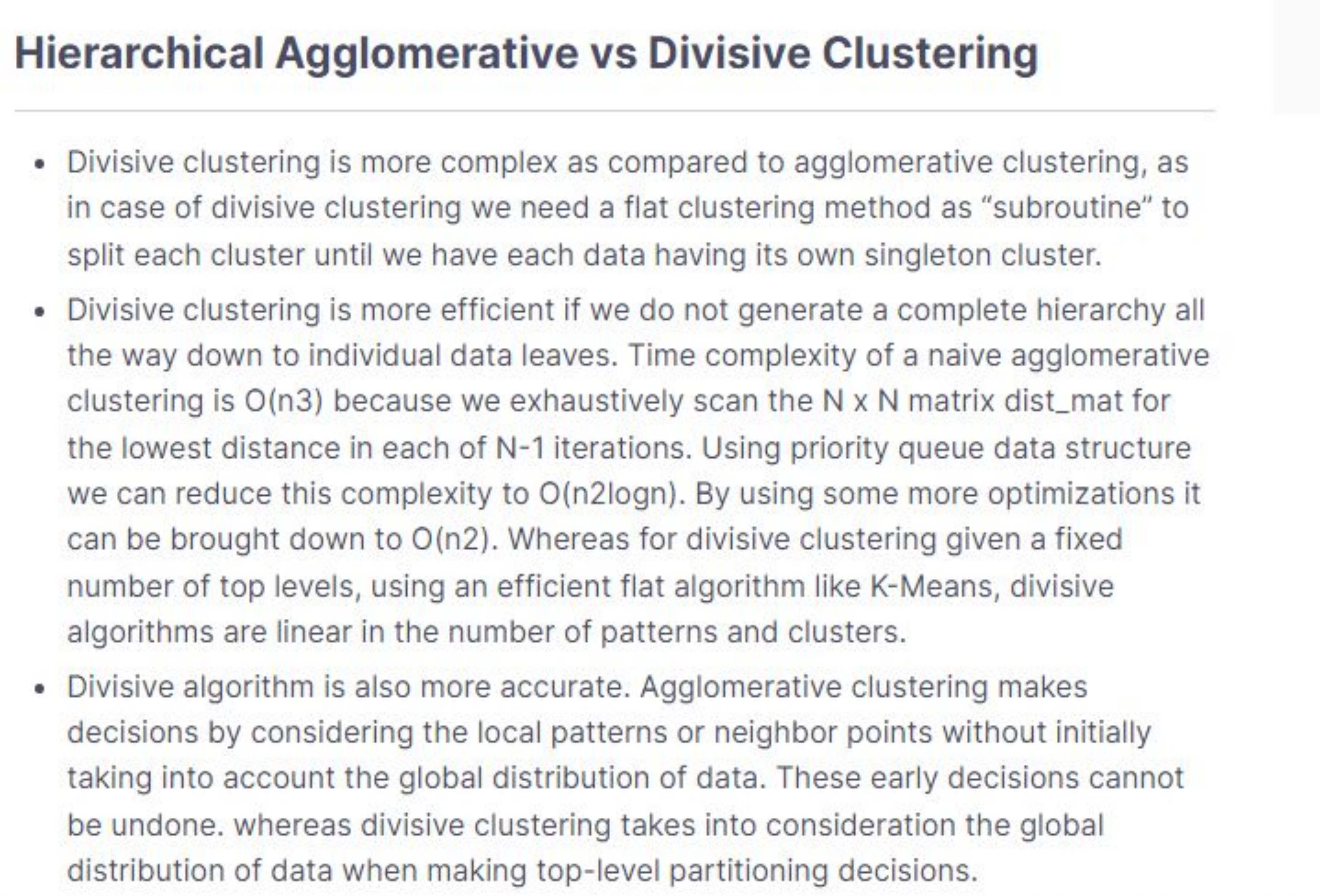
The time complexity of Prim’s algorithm is O(V^2).

The time complexity of Kruskal’s algorithm is O(E log V).

Points to Ponder:

For ***Agglomaterive Clustering*** using priority queue data structure we can reduce this complexity to **$O(n^2 logn)$**. By using some more optimizations it can be brought down to **$O(n^2)$**.

For ***Divisive Clustering*** given a fixed number of top levels, using an efficient flat algorithm like K-Means, divisive algorithms can be made linear in the number of patterns and clusters



Hierarchical Agglomerative vs Divisive Clustering

- Divisive clustering is more complex as compared to agglomerative clustering, as in case of divisive clustering we need a flat clustering method as “subroutine” to split each cluster until we have each data having its own singleton cluster.
- Divisive clustering is more efficient if we do not generate a complete hierarchy all the way down to individual data leaves. Time complexity of a naive agglomerative clustering is O(n3) because we exhaustively scan the N x N matrix dist_mat for the lowest distance in each of N-1 iterations. Using priority queue data structure we can reduce this complexity to O(n2logn). By using some more optimizations it can be brought down to O(n2). Whereas for divisive clustering given a fixed number of top levels, using an efficient flat algorithm like K-Means, divisive algorithms are linear in the number of patterns and clusters.
- Divisive algorithm is also more accurate. Agglomerative clustering makes decisions by considering the local patterns or neighbor points without initially taking into account the global distribution of data. These early decisions cannot be undone. whereas divisive clustering takes into consideration the global distribution of data when making top-level partitioning decisions.

Reference for DIANA

- https://books.google.co.in/books?id=ZXLSVPN1X1sC&pg=PA146&lpg=PA146&dq=chapter+6+of+Kaufman+and+Rousseeuw+(1990)&source=bl&ots=lqgaN_96WY&sig=ACfU3U1kDiQYh9COEsODdQvaFOSPFiTGEQ&hl=en&sa=X&ved=2ahUKEwiSiveb6tLpAhX3wjgGHS7iDMwQ6AEwBnoECAgQAQ#v=onepage&q=chapter%206%20of%20Kaufman%20and%20Rousseeuw%20(1990)&f=false
- Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets
  - Ashish Kumar Patnaik a,*, Prasanta Kumar Bhuyan a,1, K.V. Krishna Rao

diana is fully described in chapter 6 of Kaufman and Rousseeuw (1990)