BIOINFORMATICS

How to cluster biological data?

Marco Beccuti

Università degli Studi di Torino

Dipartimento di Informatica

May 2019

Outline

1. Gene expression and clustering;
2. Clustering as optimization problem;
3. k-means clustering;
4. Hierarchical Clustering.

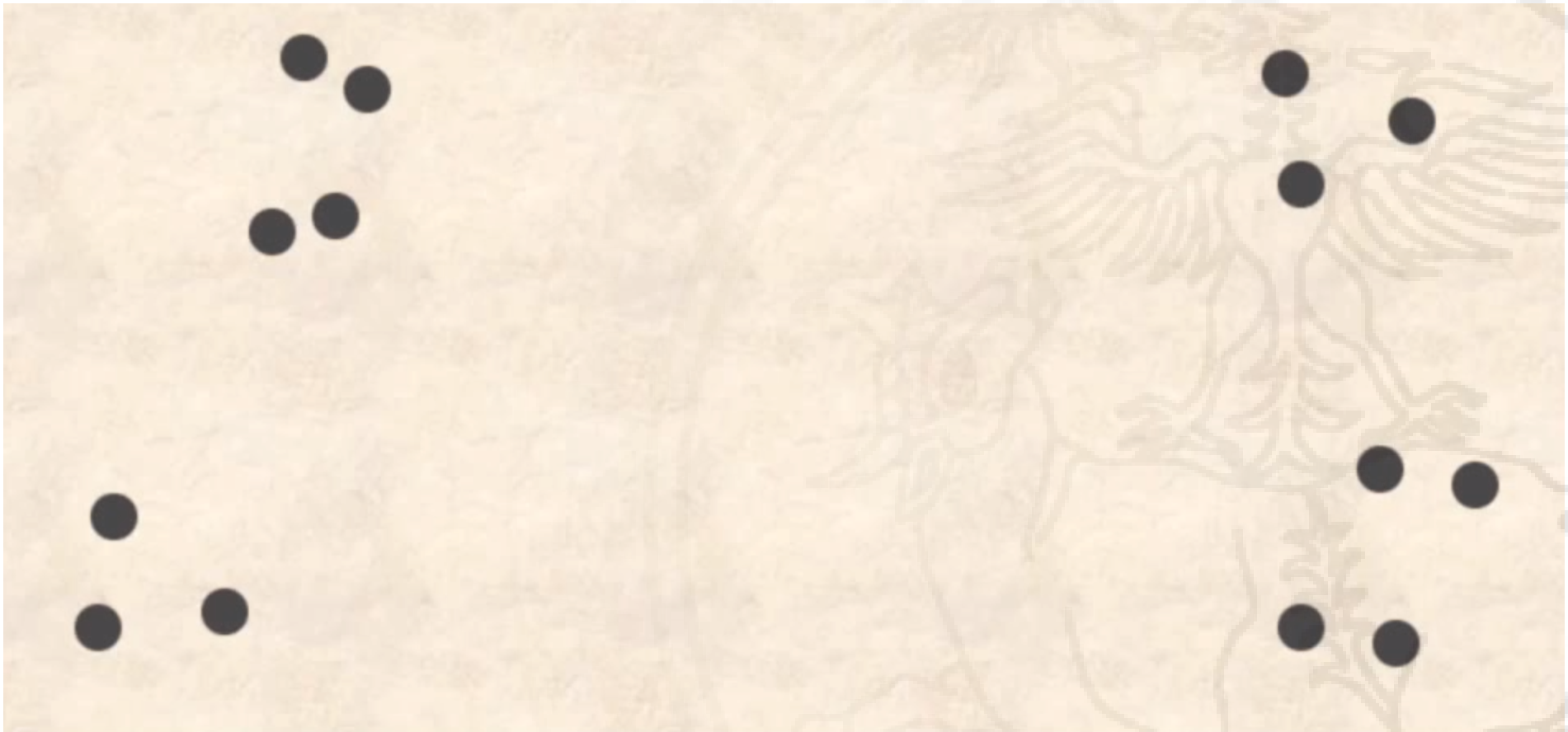
Chapter 8 in ***Bioinformatics Algorithms: An active Learning Approach (Vol.2).***

Part 1
Hierarchical Clustering

Hierarchical Clustering

Stratification of Clusters

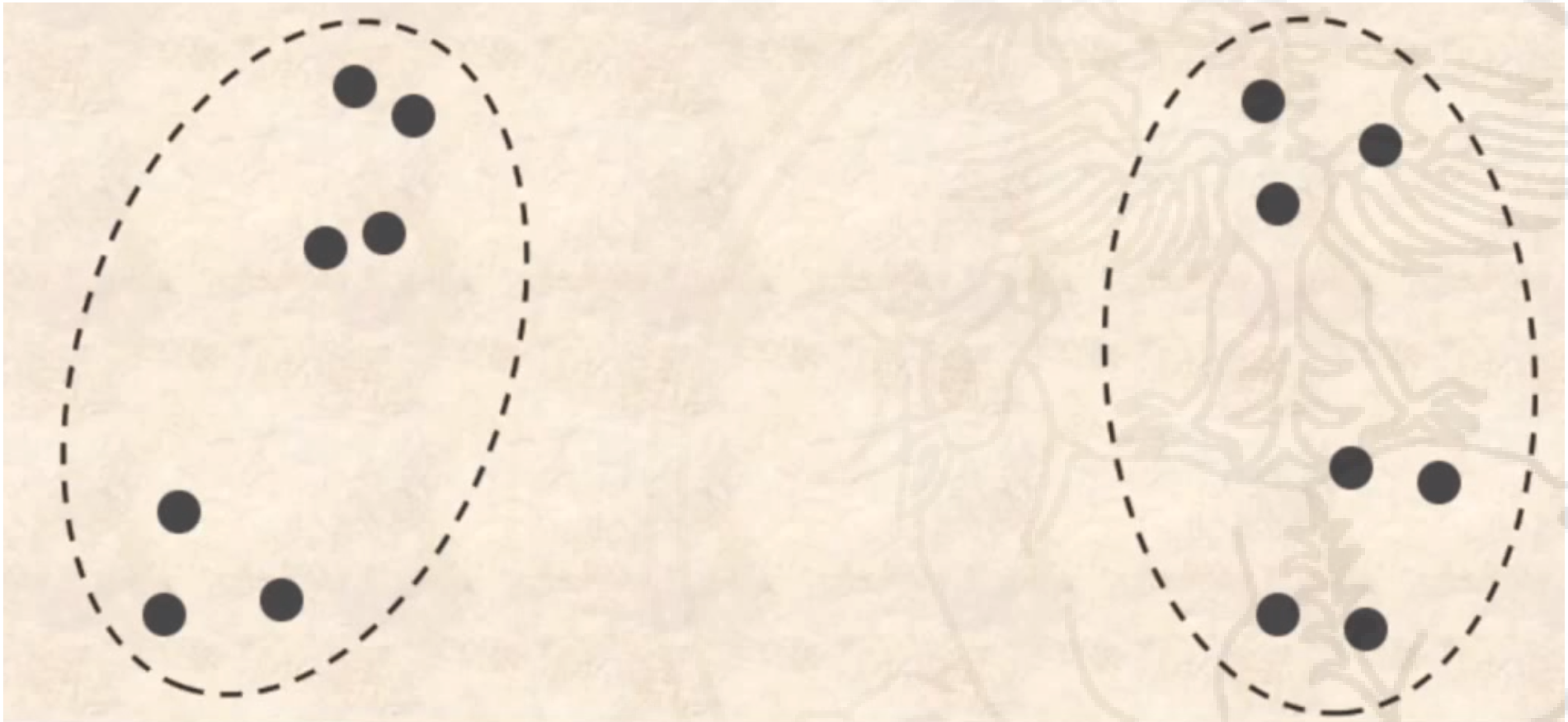
- Clusters often have **sub-cluster**, which have sub-clusters, and so on.



Hierarchical Clustering

Stratification of Clusters

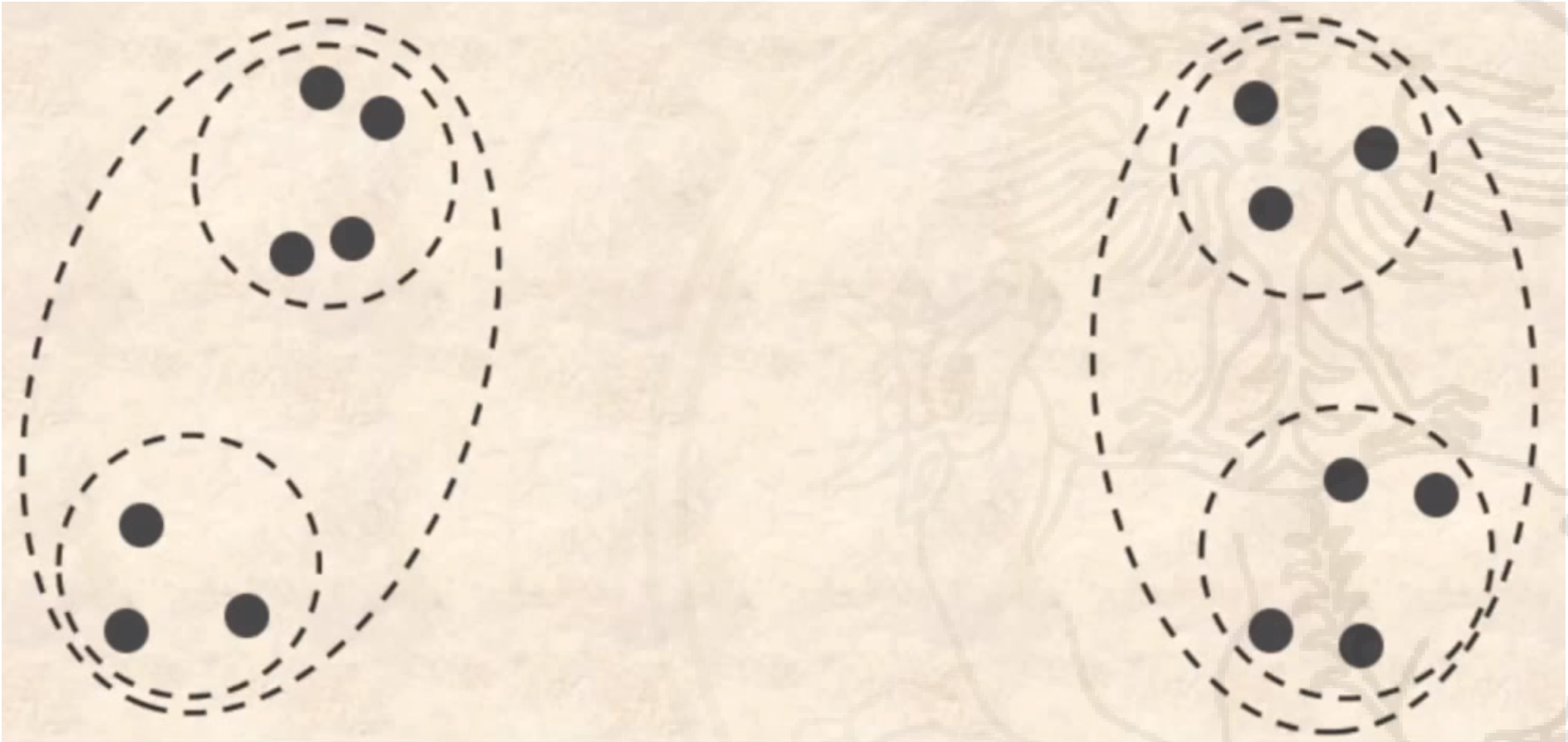
- Clusters often have **sub-cluster**, which have sub-clusters, and so on.



Hierarchical Clustering

Stratification of Clusters

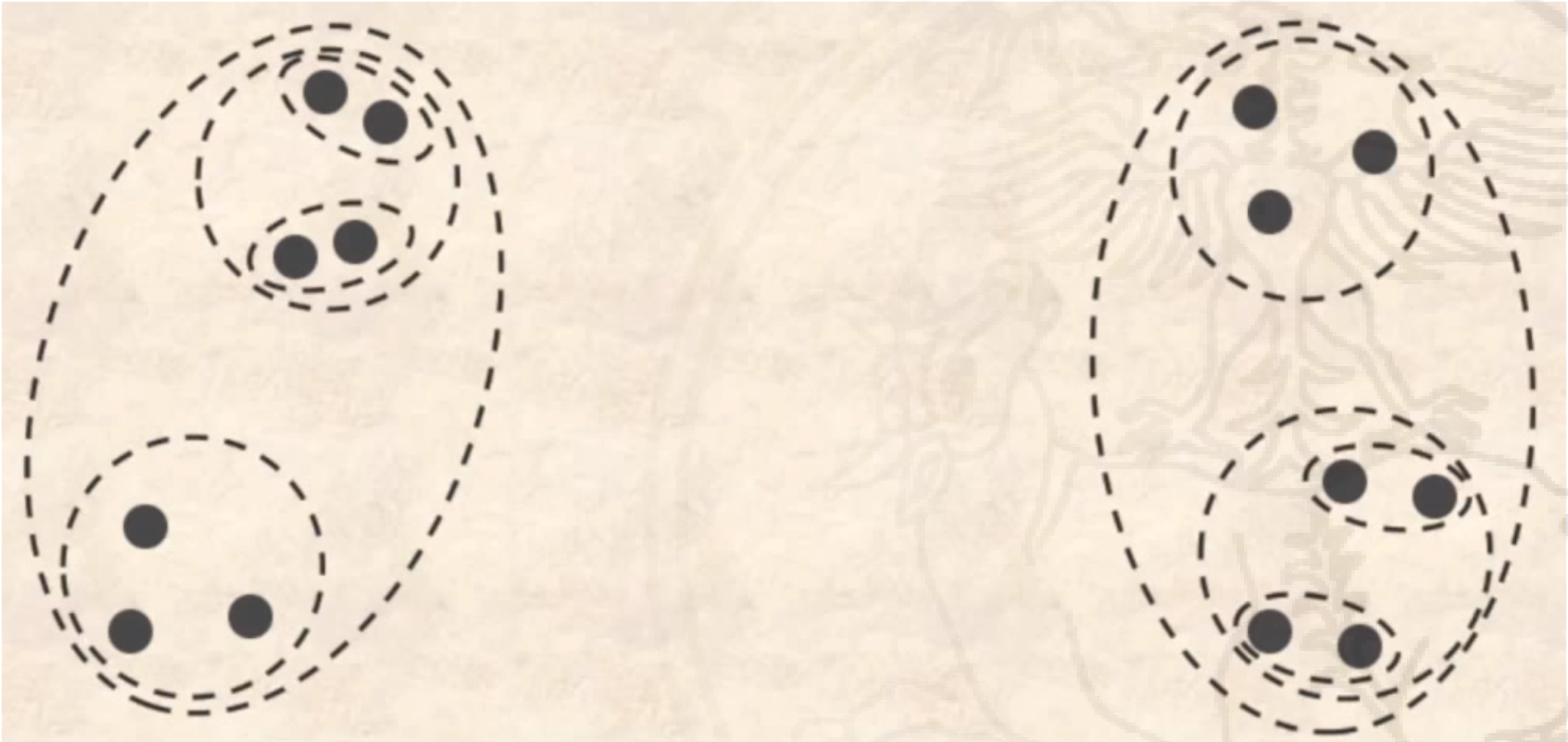
- Clusters often have **sub-cluster**, which have sub-clusters, and so on.



Hierarchical Clustering

Stratification of Clusters

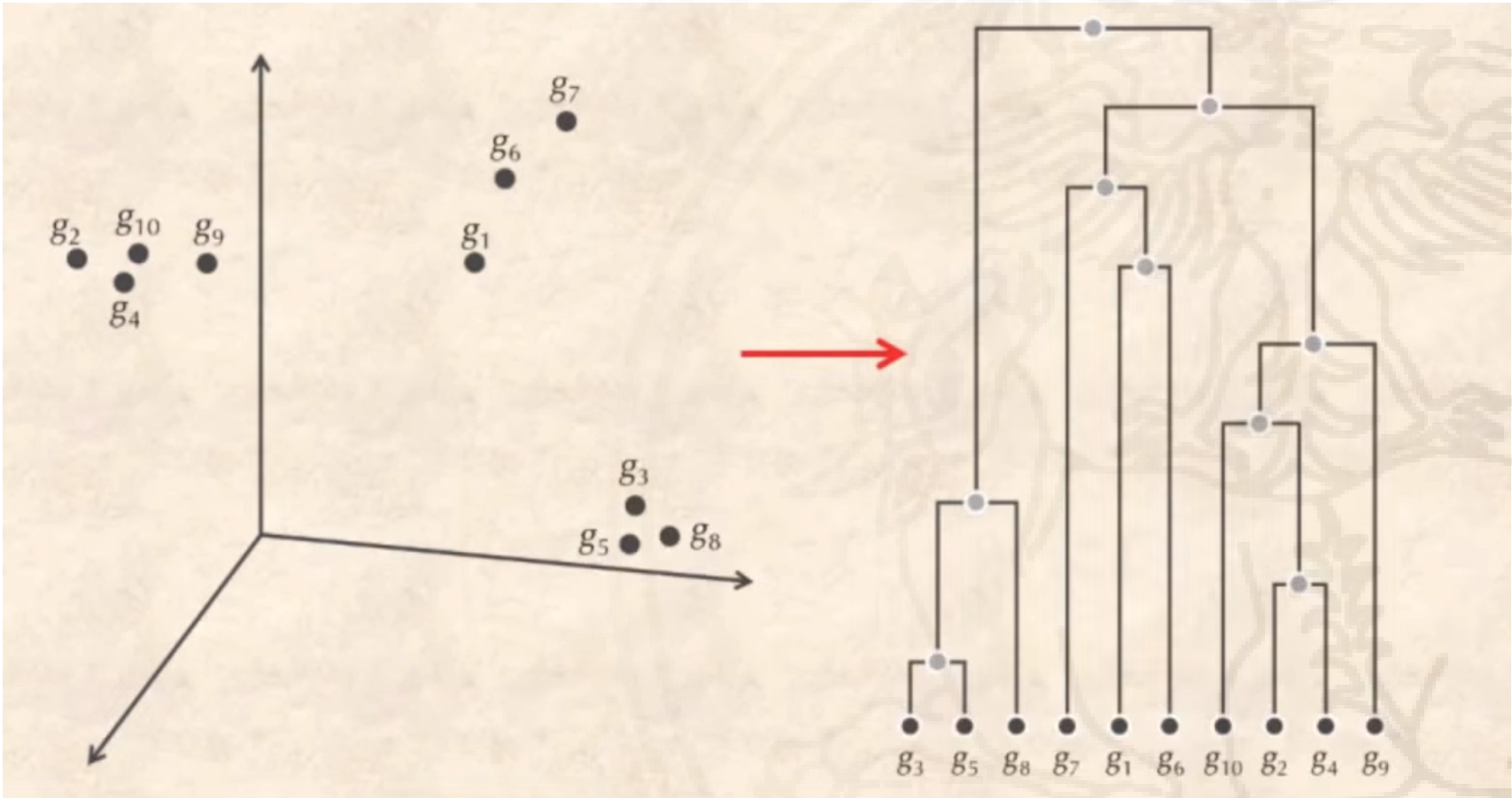
- Clusters often have **sub-cluster**, which have sub-clusters, and so on.



Hierarchical Clustering

From Data to a Tree

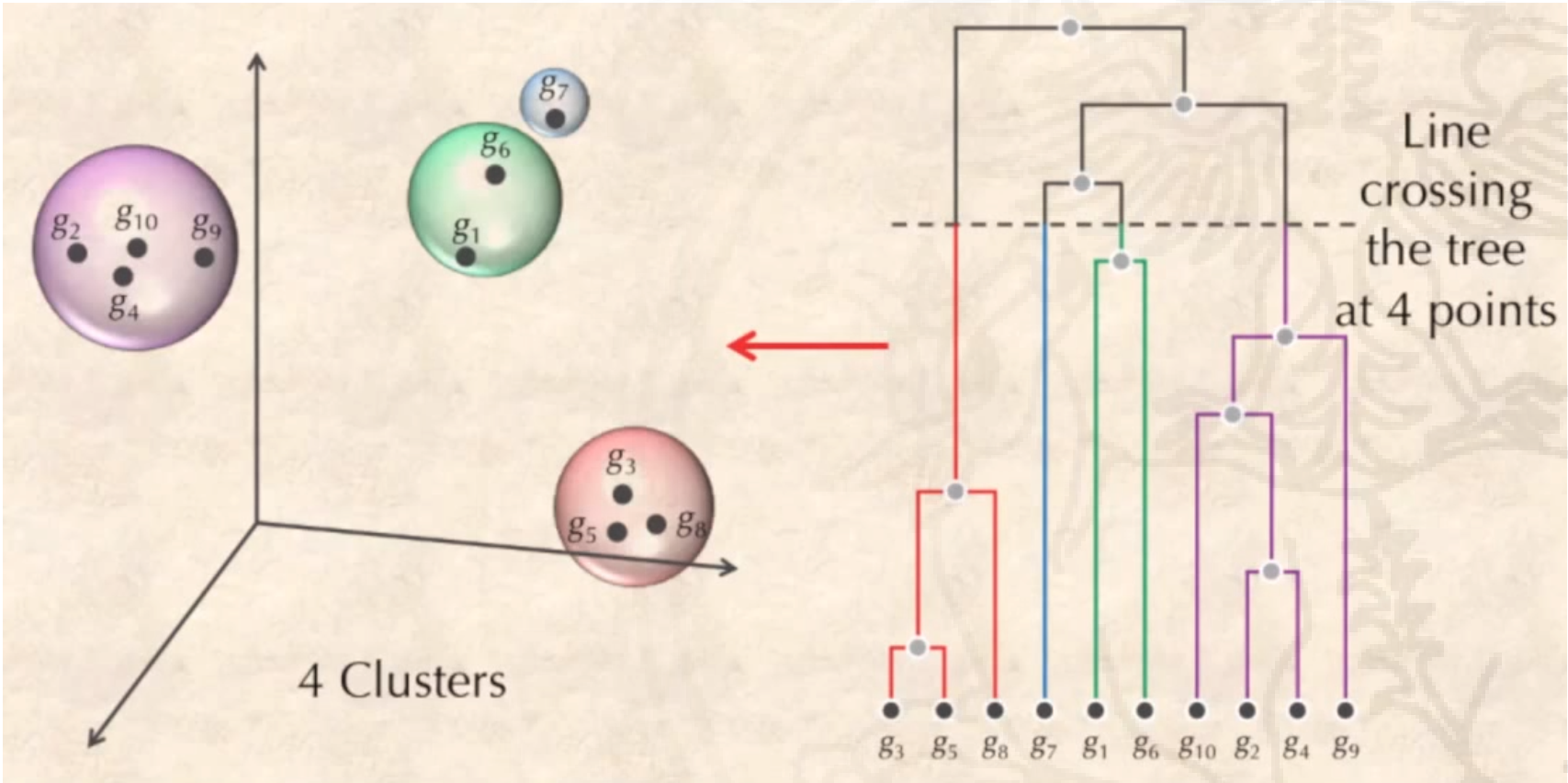
- To capture stratification, the **hierarchical clustering** algorithm organizes n data points into a tree (namely **Dendrogram**).



Hierarchical Clustering

From Data to a Tree

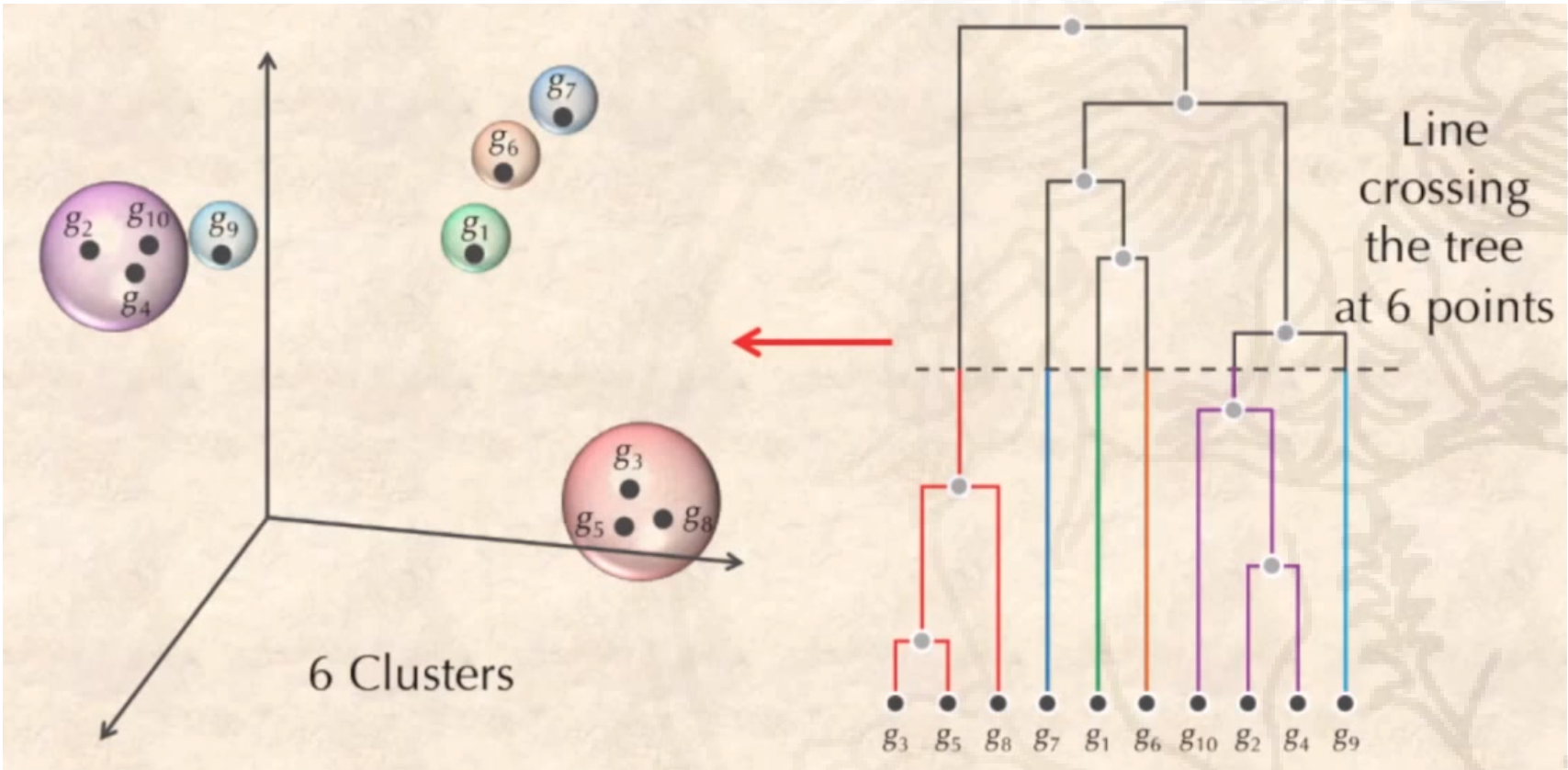
- To capture stratification, the **hierarchical clustering** algorithm organizes n data points int a tree.



Hierarchical Clustering

From Data to a Tree

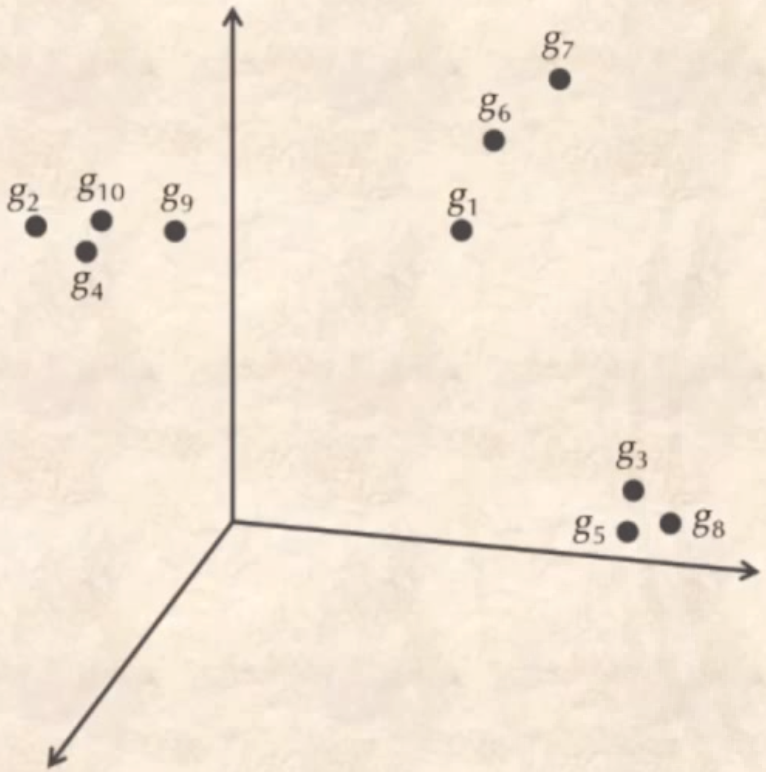
- To capture stratification, the **hierarchical clustering** algorithm organizes n data points int a tree.



Hierarchical Clustering

Constructing the Tree

- Hierarchical clustering starts from a trasformation of $n \times m$ expression matrix into $n \times n$ **similarity matrix** or **Distance matrix**;
- it can be obtained by simply computing Euclidean/Manhattan distance between genes.



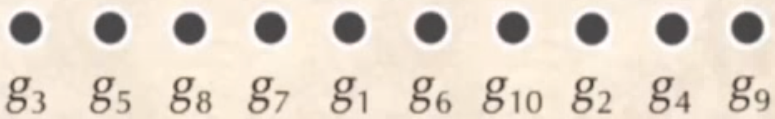
Distance Matrix

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

Hierarchical Clustering

Constructing the Tree

- Create a node (i.e. a single element cluster) for every gene.



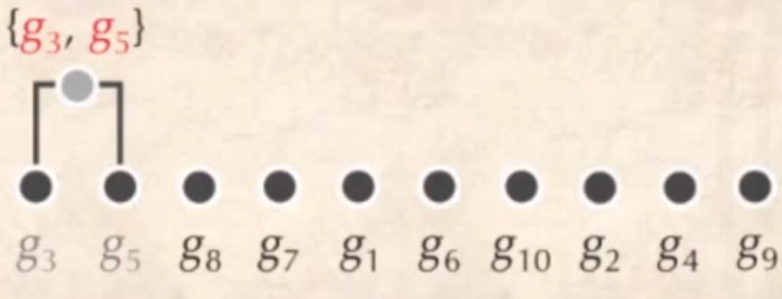
Distance Matrix

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

Hierarchical Clustering

Constructing the Tree

- Identify the two ***closest*** clusters and merge them.



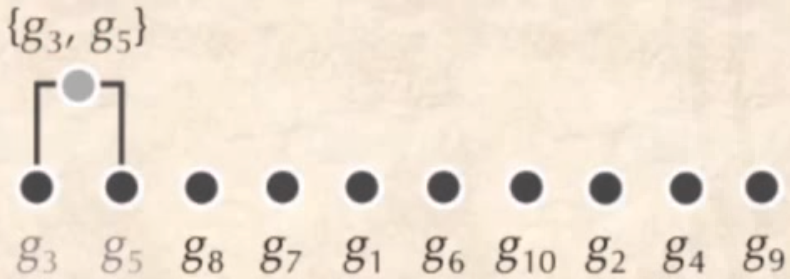
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

Hierarchical Clustering

Constructing the Tree

- Recompute the distance between two clusters as the minimal distance between the elements in the clusters:

$$D(C_1, C_2) = \min_{\forall i \in C_1, j \in C_2} D_{i,j}$$

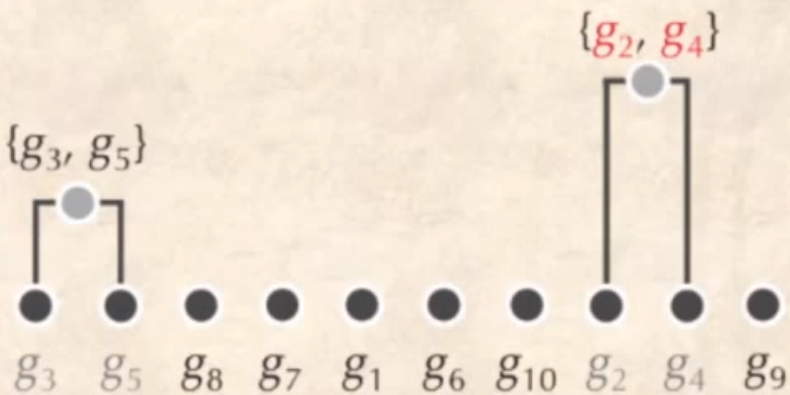


	g_1	g_2	g_3, g_5	g_4	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	9.5	10.1	12.8	2.0	1.0
g_3, g_5	9.2	12.0	0.0	11.2	11.1	8.1	1.0	10.5	11.5
g_4	7.7	0.9	11.2	0.0	9.2	9.5	12.0	1.6	1.1
g_6	2.3	9.5	11.1	9.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.0	12.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	8.5	9.3	12.4	1.1	0.0

Hierarchical Clustering

Constructing the Tree

- Identify the two ***closest*** clusters and merge them.



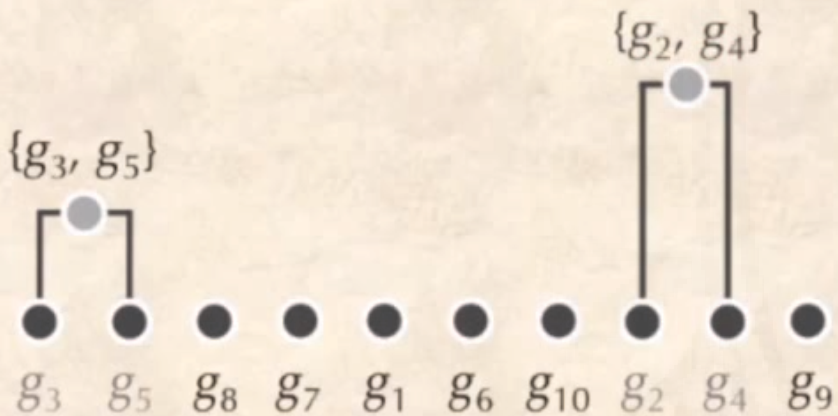
	g_1	g_2	g_3, g_5	g_4	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	9.5	10.1	12.8	2.0	1.0
g_3, g_5	9.2	12.0	0.0	11.2	11.1	8.1	1.0	10.5	11.5
g_4	7.7	0.9	11.2	0.0	9.2	9.5	12.0	1.6	1.1
g_6	2.3	9.5	11.1	9.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.0	12.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	8.5	9.3	12.4	1.1	0.0

Hierarchical Clustering

Constructing the Tree

- Recompute the distance between two clusters as the minimal distance between the elements in the clusters:

$$D(C_1, C_2) = \min_{\forall i \in C_1, j \in C_2} D_{i,j}$$

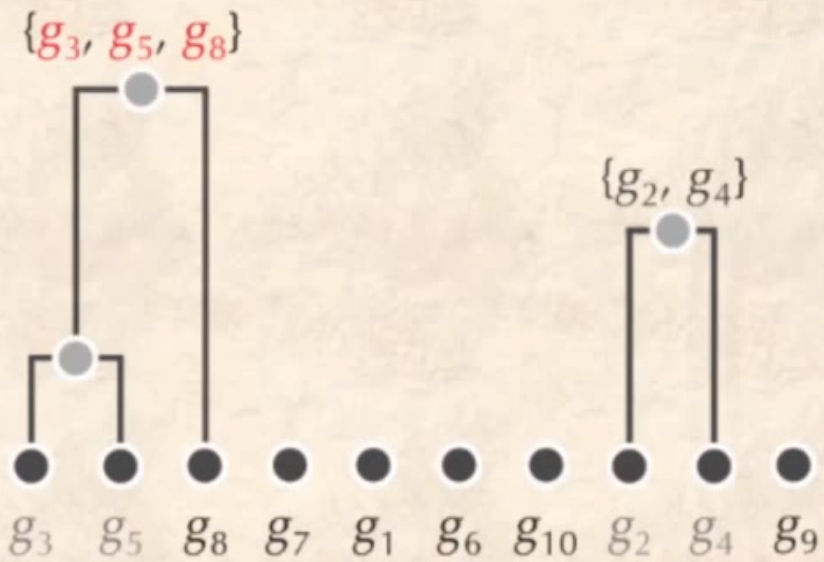


	g_1	g_2, g_4	g_3, g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	7.7	9.2	2.3	5.1	10.2	6.1	7.0
g_2, g_4	7.7	0.0	11.2	9.2	9.5	12.0	1.6	1.0
g_3, g_5	9.2	11.2	0.0	11.1	8.1	1.0	10.5	11.5
g_6	2.3	9.2	11.1	0.0	5.6	12.1	7.7	8.5
g_7	5.1	9.5	8.1	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	1.6	10.5	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	8.5	9.3	12.4	1.1	0.0

Hierarchical Clustering

Constructing the Tree

- Identify the two ***closest*** clusters and merge them.

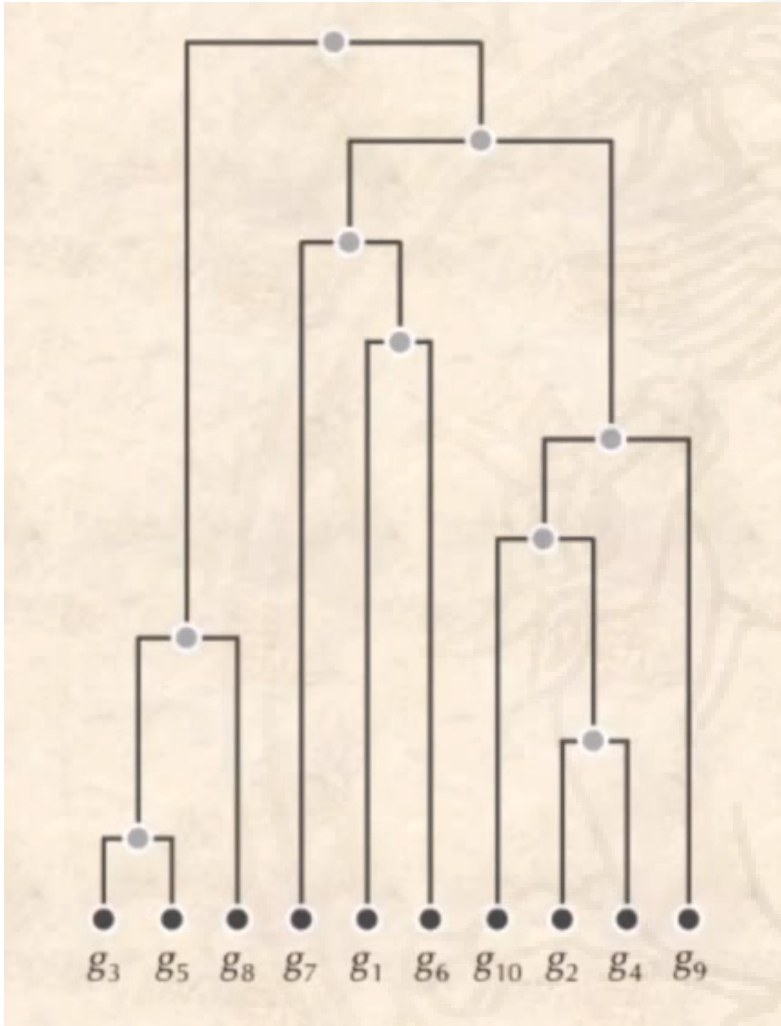


	g_1	g_2, g_4	g_3, g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	7.7	9.2	2.3	5.1	10.2	6.1	7.0
g_2, g_4	7.7	0.0	11.2	9.2	9.5	12.0	1.6	1.0
g_3, g_5	9.2	11.2	0.0	11.1	8.1	1.0	10.5	11.5
g_6	2.3	9.2	11.1	0.0	5.6	12.1	7.7	8.5
g_7	5.1	9.5	8.1	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	1.6	10.5	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	8.5	9.3	12.4	1.1	0.0

Hierarchical Clustering

Constructing the Tree

- Iterate until all elements form a single cluster(i.e. ***root***).



Hierarchical Clustering

Constructing the Tree

HierarchicalClustering (D, n)
  $Clusters \leftarrow n$ single-element clusters labeled 1 to n
  $T \leftarrow$ a graph with the n isolated nodes labeled 1 to n
  while there is more than one cluster
    find the two closest clusters C_i and C_j
    merge C_i and C_j into a new cluster C_{new} with $|C_i| + |C_j|$ elements
    add a new node labeled by cluster C_{new} to T
    connect node C_{new} to C_i and C_j by directed edges
    remove the rows and columns of D corresponding to C_i and C_j
    remove C_i and C_j from *Clusters*
    add a row and column to D for the cluster C_{new} by computing
      $D(C_{new}, C)$ for each cluster C in *Clusters*
    add C_{new} to *Clusters*
  assign root in T as a node with no incoming edges
  return T

Hierarchical Clustering

Different distance functions

Minimum distance between elements of two clusters:

$$D_{\text{min}}(C_1, C_2) = \min_{\text{all points } i \text{ and } j \text{ in clusters } C1 \text{ and } C2, \text{ respectively}} D_{i,j}$$

Average distance between elements of two clusters:

$$D_{\text{avg}}(C_1, C_2) = \left(\sum_{\text{all points } i \text{ and } j \text{ in clusters } C1 \text{ and } C2, \text{ respectively}} D_{i,j}\right) / \left(|C_1| * |C_2|\right)$$

Difference between K-means and Hierarchical clustering

- Hierarchical clustering can not handle big data as well as K-means clustering.
- In K-means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering;
- K-means clustering requires prior knowledge of K.

Exercises

- Try to apply K-means algorithm on the following inputs:
  - $k = 3$;
  - $P_1(1,1)$, $P_2(2,1)$, $P_3(3,4)$, $P_4(1,6)$ $P_5(2,3)$, $P_6(3,3)$, $P_7(4,5)$, $P_8(2,5)$.

- Try to apply hierarchical clustering algorithm on the same input points.

$$\mathcal{MD} = \begin{array}{c} \\ g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ g_7 \\ g_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} g_1 & g_2 & g_3 & g_4 & g_5 & g_6 & g_7 & g_8 \end{array} \\ \begin{pmatrix} 0.000 & 1.000 & 3.605 & 5.000 & 2.236 & 2.828 & 5.000 & 4.123 \\ 1.000 & 0.000 & 3.162 & 5.099 & 2.000 & 2.236 & 4.472 & 4.000 \\ 3.605 & 3.162 & 0.000 & 2.828 & 1.414 & 1.000 & 1.414 & 1.414 \\ 5.000 & 5.099 & 2.828 & 0.000 & 3.162 & 3.605 & 3.162 & 1.414 \\ 2.236 & 2.000 & 1.414 & 3.162 & 0.000 & 1.000 & 2.828 & 2.000 \\ 2.828 & 2.236 & 1.000 & 3.605 & 1.000 & 0.000 & 2.236 & 2.236 \\ 5.000 & 4.472 & 1.414 & 3.162 & 2.828 & 2.000 & 0.000 & 2.000 \\ 4.123 & 4.000 & 1.414 & 1.414 & 2.000 & 2.236 & 2.000 & 0.000 \end{pmatrix} \end{array}$$