QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 2019

Consensus clustering[1]

Assume, we are given a dataset for the purpose of clustering analysis

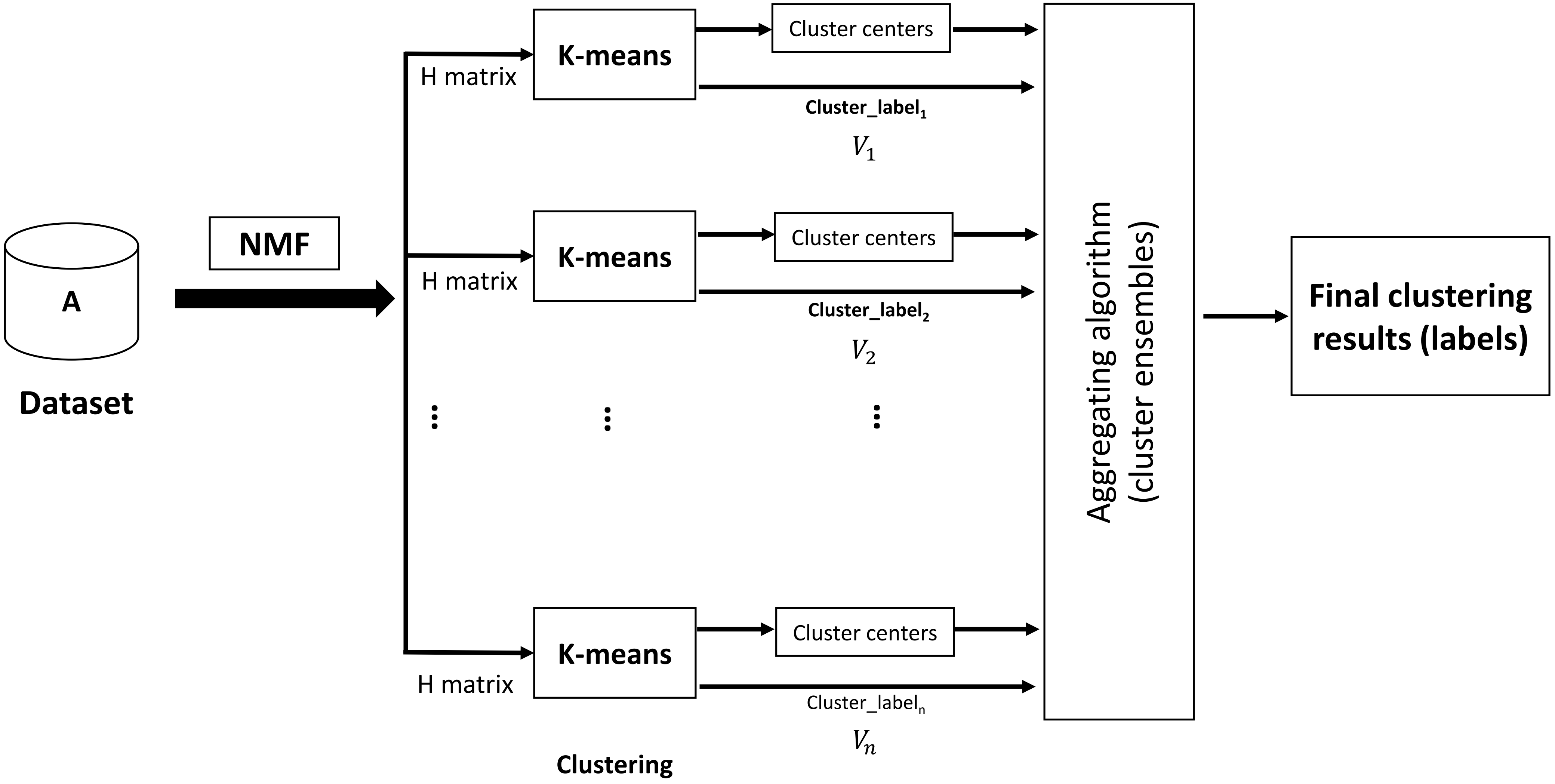
1) Multiple runs of a clustering algorithm
   a) Determine the number of clusters and assess the stability of the discovered clusters
   b) In k-means clustering: with using random restart

2) Aggregating the cluster (label) results of different clustering algorithms

[1] Ensemble clustering

Summary: consensus approach

1) Multiple runs of a clustering algorithm



A comprehensive Ensemble approach for unsupervised clustering using NMF projection and k-means clustering

Main clustering approaches

- **Partitioning algorithms**: Make different partitions heuristically and then evaluate them by some criteria (k-means & PAM)
- **Hierarchy algorithms**: Create a hierarchical decomposition of data using some criteria
- **Density-based algorithms**: based on connectivity and density functions
- **Model-based algorithms**: A model is assumed for each cluster and the idea is to find the best fit of a model

Hierarchical clustering

A **hierarchical clustering** is a set of nested clusters that are organized as a tree

Two types of algorithms

Agglomerative ("Bottom-up")

Start with the points as individual clusters. Then at each step, **merge** the closest pair of clusters.

Divisive ("Top-down")

Start with one, all-inclusive cluster. Then at each step, **split** a cluster until only singleton clusters of individual points remain.

Proximity between clusters

The definition of cluster proximity differentiates the various agglomerative hierarchical techniques.

MIN (single link)
MAX (complete link)
Group average proximity (group-based or average)
Ward's method (Prototype-based or centroid-based)

Proximity between clusters

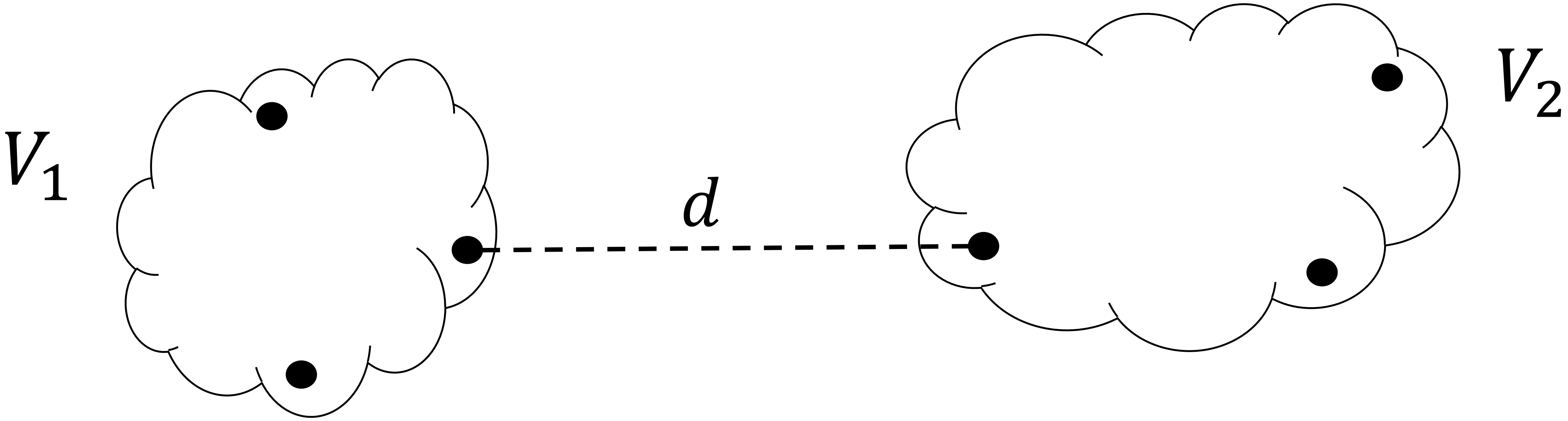
The definition of cluster proximity differentiates the various agglomerative hierarchical techniques.

MIN (single link)
MAX (complete link)
Group average proximity (group-based or average)
Ward's method (Prototype-based or centroid-based)

Min (single link) proximity

Cluster proximity is defined as the shortest distance between two points, x and y, that are in different clusters, V_1 and V_2:

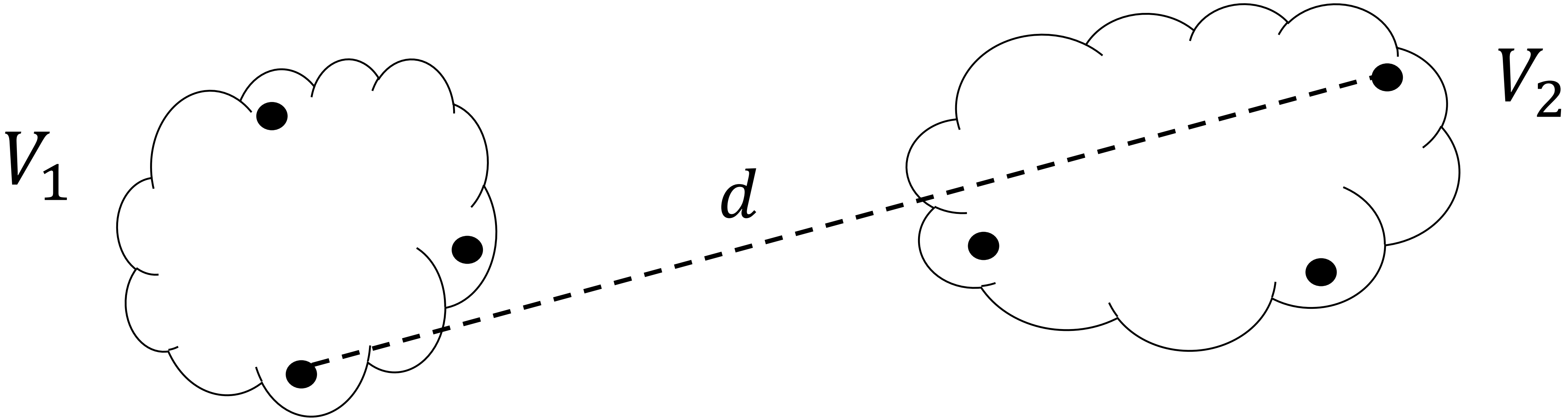
$$d(V_1, V_2) = \min_{x \in V_1, y \in V_2} d(x - y)$$



Max (complete link) proximity

Cluster proximity is defined as the furthest distance between two points, x and y, that are in different clusters, V_1 and V_2:

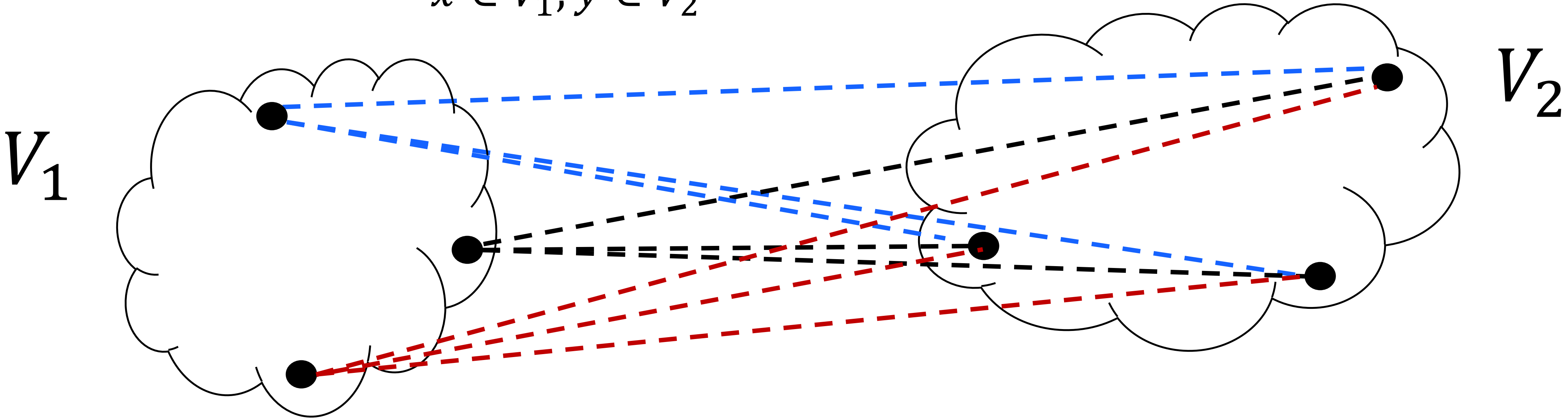
$$d(V_1, V_2) = \max_{x \in V_1, y \in V_2} d(x - y)$$



Group average proximity

Cluster proximity is defined as the average distance between two points, x and y, that are in different clusters, V_1 and V_2:

$$d(V_1, V_2) = \sum_{x \in V_1, y \in V_2} d(x - y) / [n(V_1) \times n(V_2)]$$

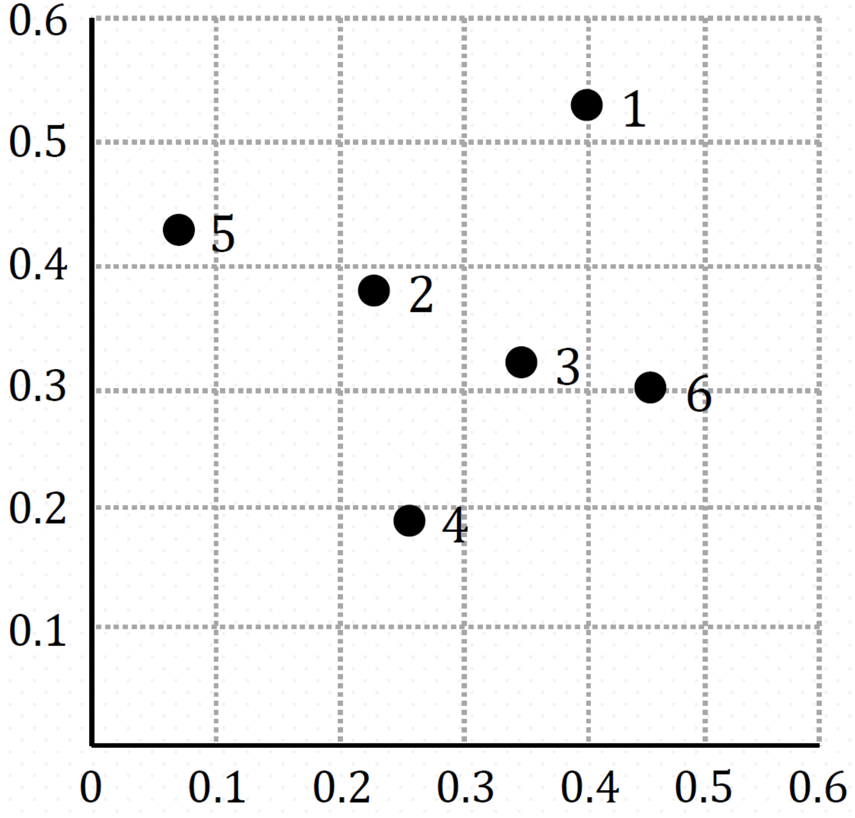


Basic Agglomerative Clustering

```
Compute the proximity matrix, if necessary.
repeat
        - Merge the closest two clusters.
        - Update the proximity matrix to reflect the
                proximity between the new cluster
                and the original clusters.
until Only one cluster remains.
```

Example: clustering 6 data points

Set of 6 Two-Dimensional Points



xy Coordinates of 6 Points

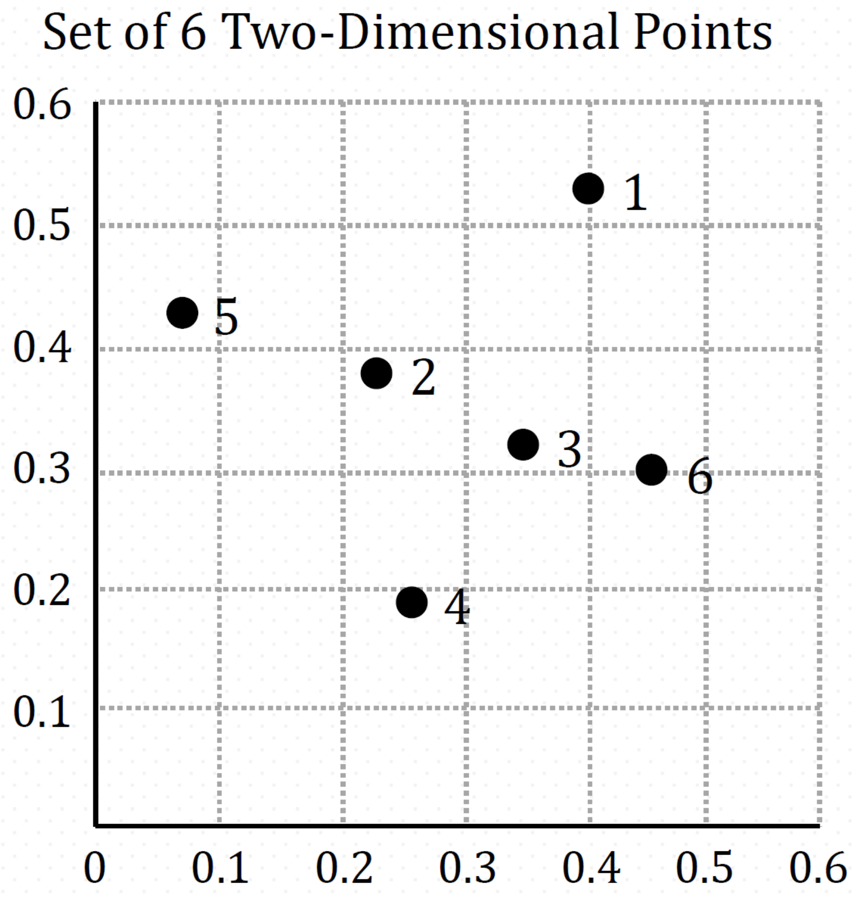
Point	*x* Coordinate	*y* Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Sample name

	P_1	P_2	P_3	P_4	P_5	P_6
x	0.40	0.22	0.35	0.26	0.08	0.45
y	0.53	0.38	0.32	0.19	0.41	0.30

Feature name

Euclidean distance matrix

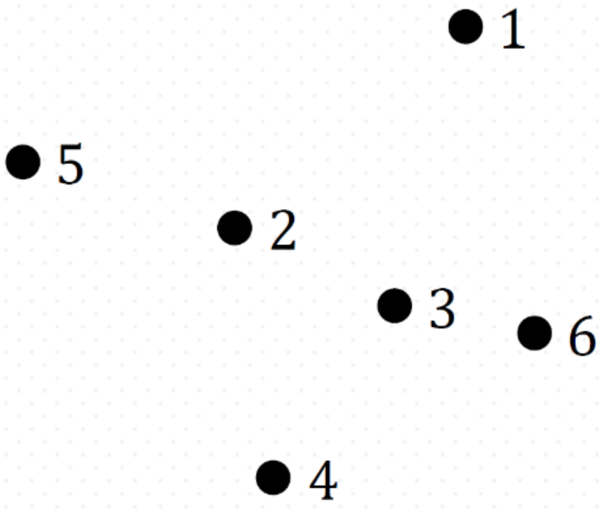

Set of 6 Two-Dimensional Points

Euclidean Distance Matrix for 6 Points

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Single link (min) clustering

Nested Cluster Diagram

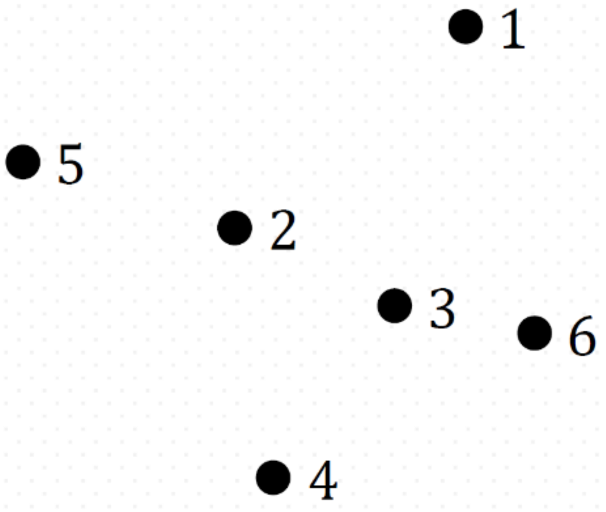


Single Link Distance Matrix

	1	2	3	4	5	6
1	0	0.24	0.22	0.37	0.34	0.23
2		0	0.15	0.20	0.14	0.25
3			0	0.15	0.28	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Single link (min) clustering

Nested Cluster Diagram



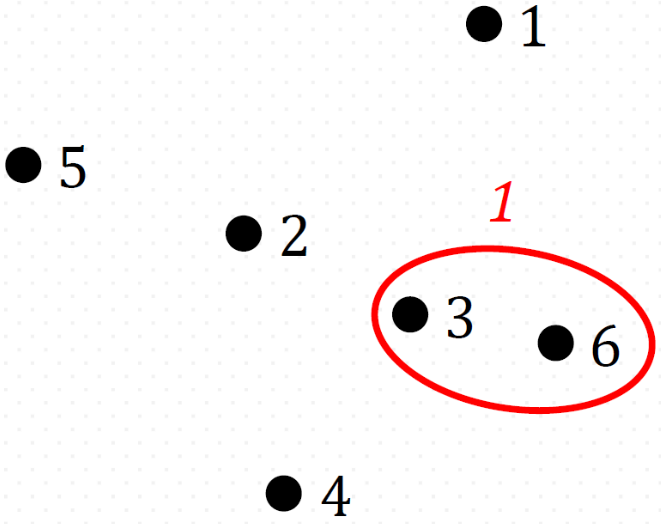
Single Link Distance Matrix

	1	2	3	4	5	6
1	0	0.24	0.22	0.37	0.34	0.23
2		0	0.15	0.20	0.14	0.25
3			0	0.15	0.28	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Which data points are merged at first glance?

Single link clustering

Nested Cluster Diagram



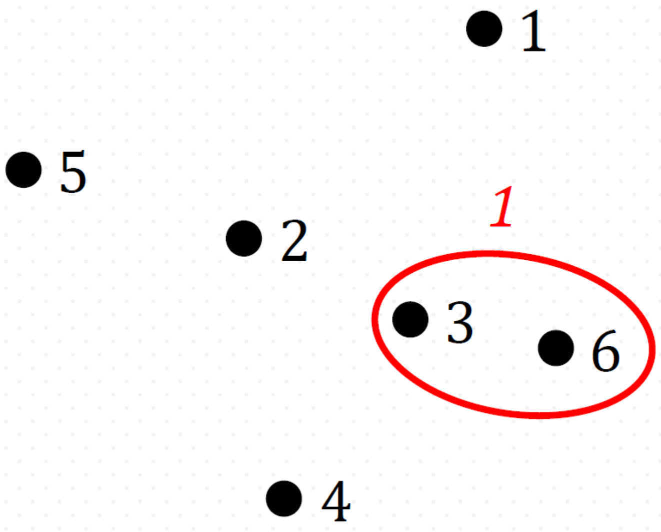
Single Link Distance Matrix

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Data points 3 and 6 have the **smallest (minimum) single link proximity distance**. These data points are merged into one cluster and update the distances to this new cluster.

Single link clustering

Nested Cluster Diagram



Single Link Distance Matrix

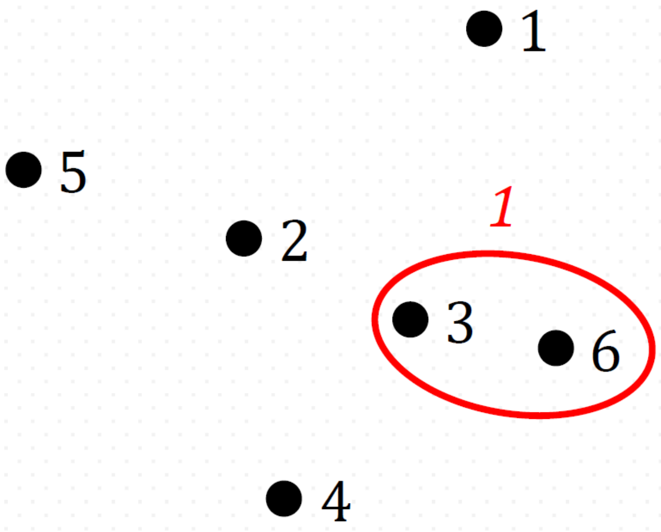
	1	2	4	5	3,6
1	0	?	?	?	?
2		0	?	?	?
4			0	?	?
5				0	?
3,6					0

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Now, update the proximity matrix

Single link clustering

Nested Cluster Diagram



Single Link Distance Matrix

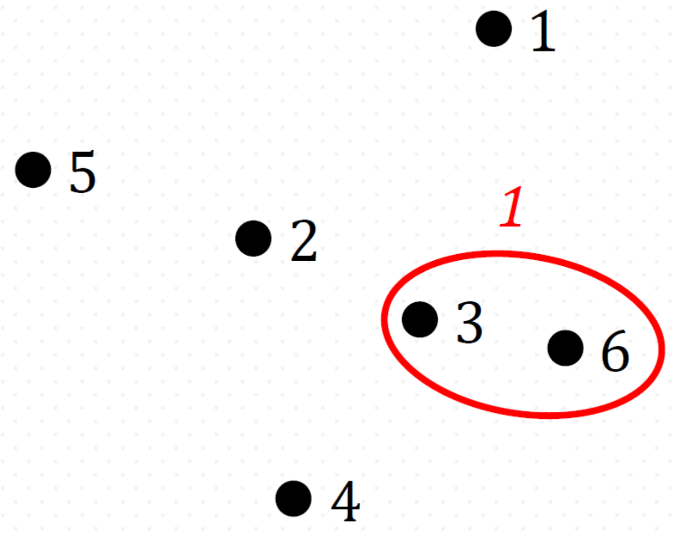
	1	2	4	5	3,6
1	0	0.24	?	?	?
2		0	?	?	?
4			0	?	?
5				0	?
3,6					0

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Now, update the proximity matrix

Single link clustering

Nested Cluster Diagram



Single Link Distance Matrix

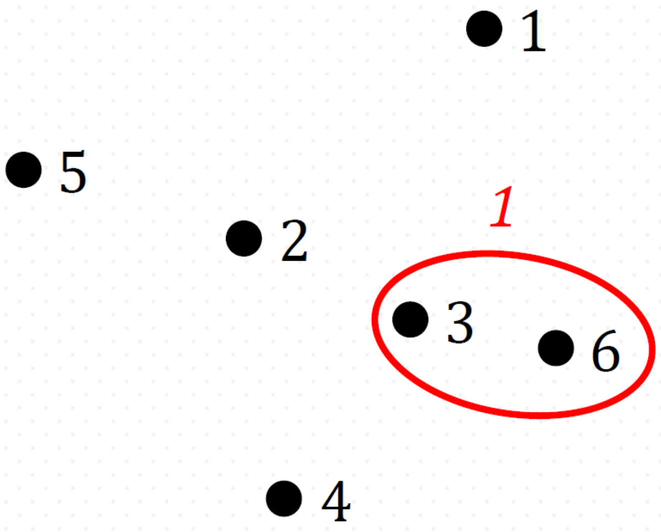
	1	2	4	5	3,6
1	0	0.24	0.37	?	?
2		0	?	?	?
4			0	?	?
5				0	?
3,6					0

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Now, update the proximity matrix

Single link clustering

Nested Cluster Diagram



Single Link Distance Matrix

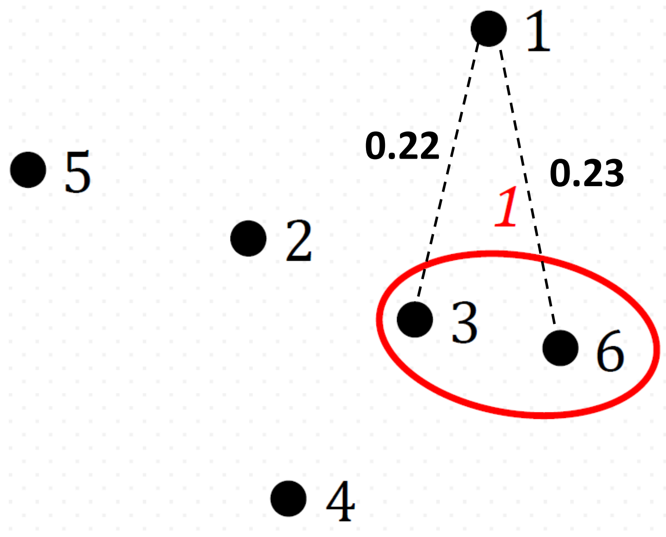
	1	2	4	5	3,6
1	0	0.24	0.37	0.34	?
2		0	?	?	?
4			0	?	?
5				0	?
3,6					0

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Now, update the proximity matrix

Single link clustering

Nested Cluster Diagram



Single Link Distance Matrix

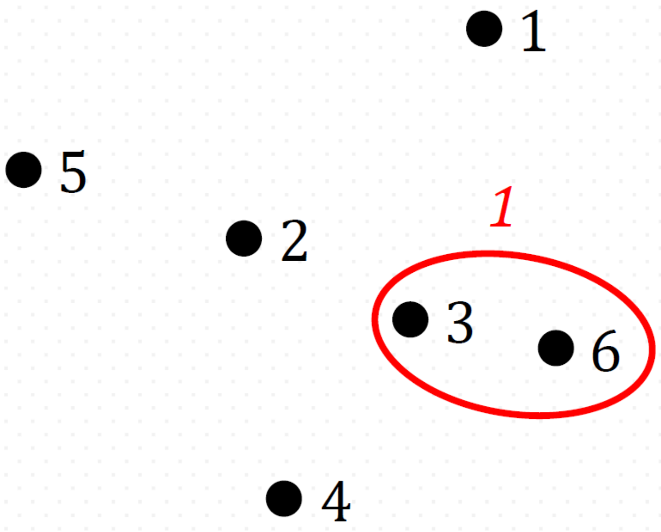
	1	2	4	5	3,6
1	0	0.24	0.37	0.34	0.22
2		0	?	?	?
4			0	?	?
5				0	?
3,6					0

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

Now, update the proximity matrix

Single link clustering

Nested Cluster Diagram



Single Link Distance Matrix

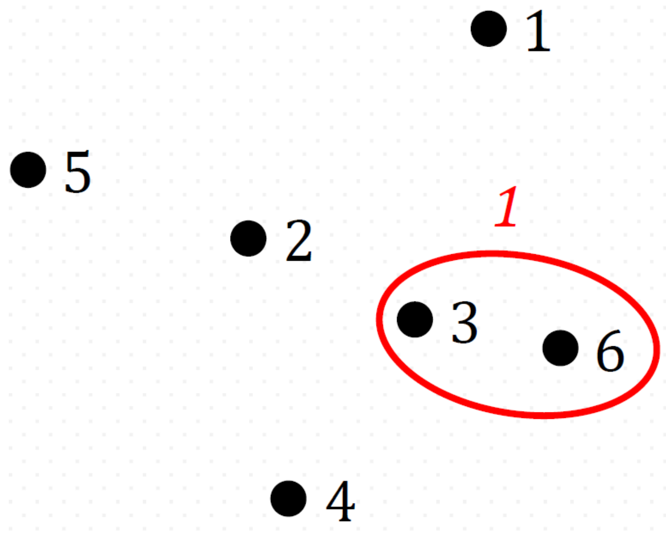
	1	2	4	5	3,6
1	0	0.24	0.37	0.34	0.22
2		0	0.20	0.14	0.15
4			0	0.29	0.15
5				0	0.28
3,6					0

	1	2	3	4	5	6
1	0	0.24	*0.22*	0.37	0.34	0.23
2		0	*0.15*	0.20	0.14	0.25
3			0	*0.15*	*0.28*	0.11
4				0	0.29	0.22
5					0	0.39
6						0

The proximity matrix was **Updated**

Single link clustering

Nested Cluster Diagram



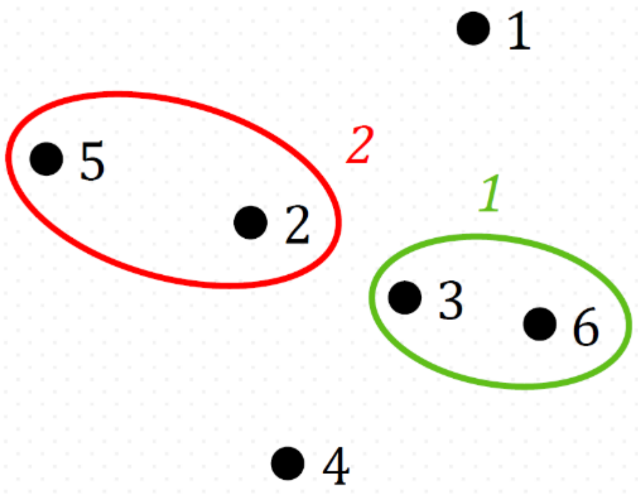
Single Link Distance Matrix

	1	2	4	5	3,6
1	0	0.24	0.37	0.34	0.22
2		0	0.20	0.14	0.15
4			0	0.29	0.15
5				0	0.28
3,6					0

Which data points are merged next?

Single link clustering

Nested Cluster Diagram

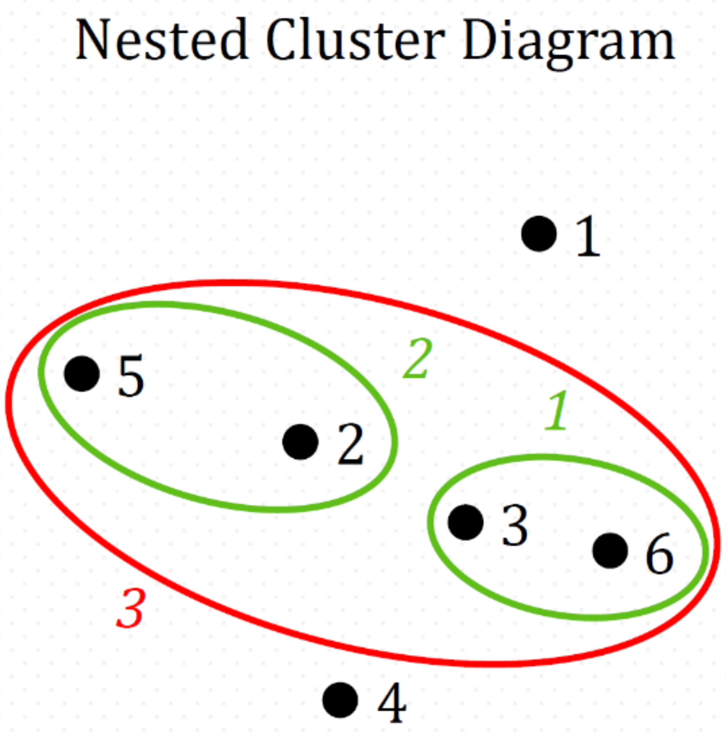


Single Link Distance Matrix

	1	2	4	5	3,6
1	0	*0.24*	0.37	0.34	0.22
2		0	*0.20*	0.14	0.15
4			0	0.29	0.15
5				0	0.28
3,6					0

Data points 2 and 5 have the smallest single link proximity distance. These data points are merged into one cluster and update the distances to this new cluster.

Single link clustering



Nested Cluster Diagram

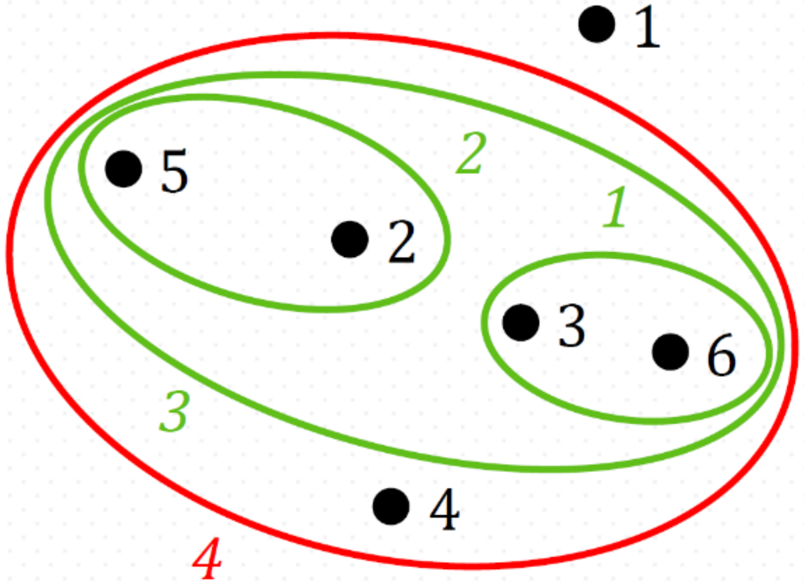
Single Link Distance Matrix

	1	4	2,5	3,6
1	0	0.37	0.24	*0.22*
4		0	0.20	*0.15*
2,5			0	0.15
3,6				0

Iterate …

Single link clustering

Nested Cluster Diagram



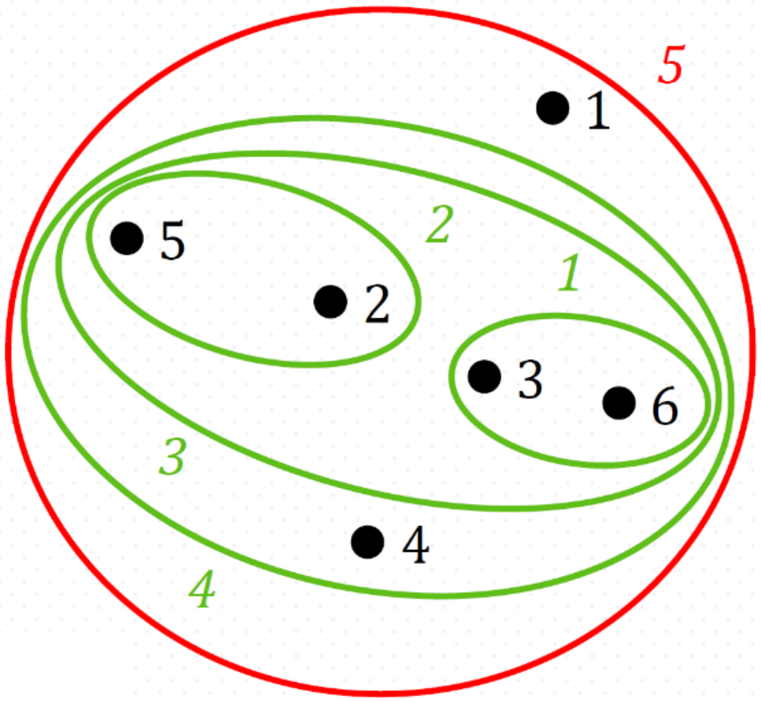
Single Link Distance Matrix

$$\begin{array}{c|ccc} & 1 & 4 & 2,5,3,6 \\ \hline 1 & 0 & \underline{0.37} & \underline{\mathit{0.22}} \\ 4 & & 0 & \color{red}{0.15} \\ 2,5,3,6 & & & 0 \end{array}$$

Iterate …

Single link clustering

Nested Cluster Diagram

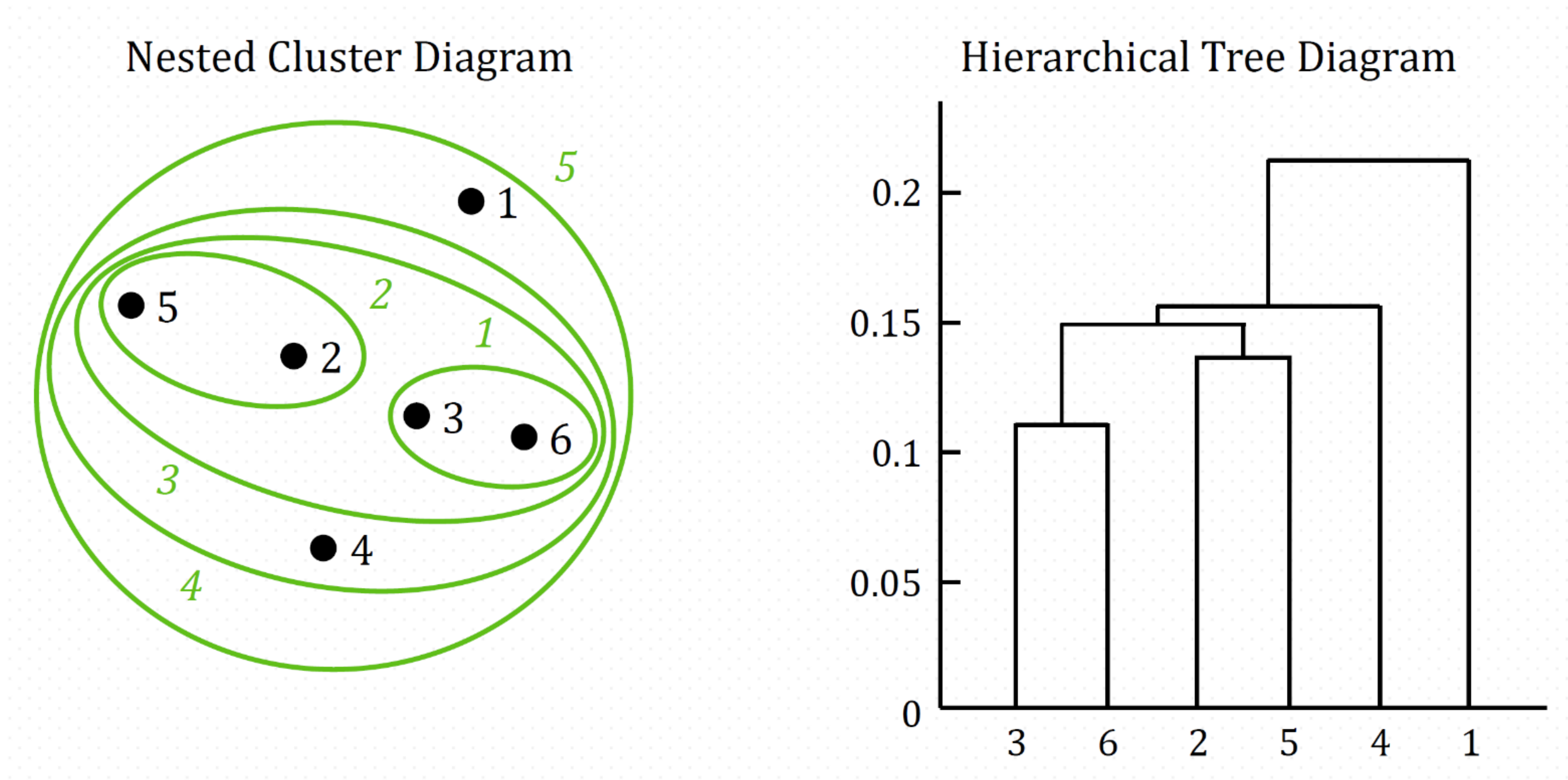


Single Link Distance Matrix

	1	4,2,5,3,6
1	0	0.22
2,5,3,6		0

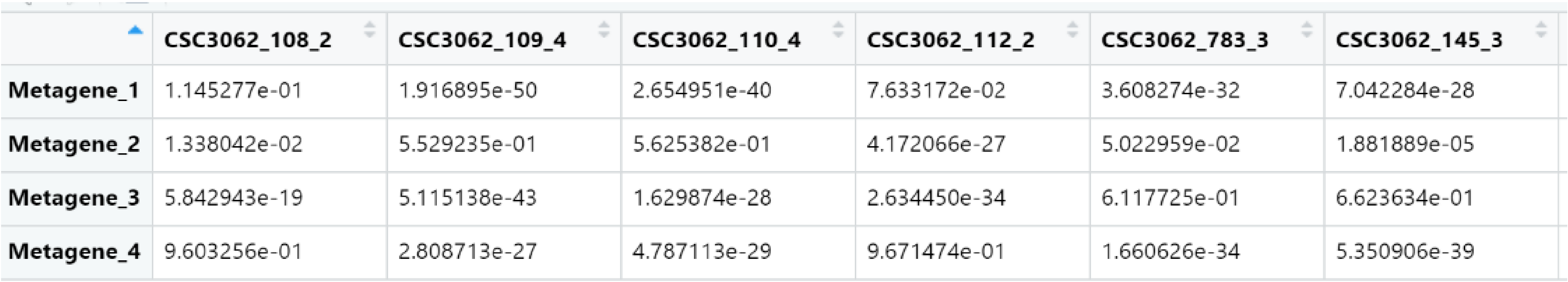
Iterate until there would be only one all-inclusive cluster.

Single link clustering

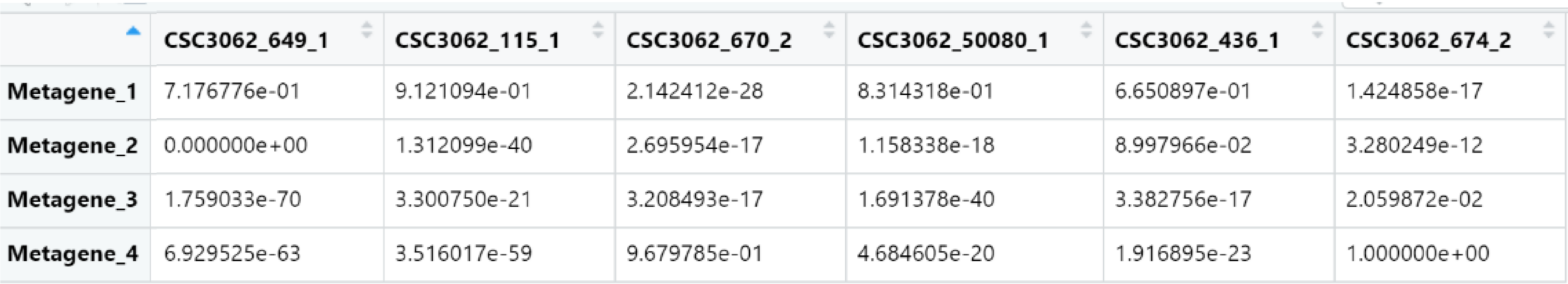


Tree-like diagram which is called a **dendrogram**

n=12 samples with 4 subgroups

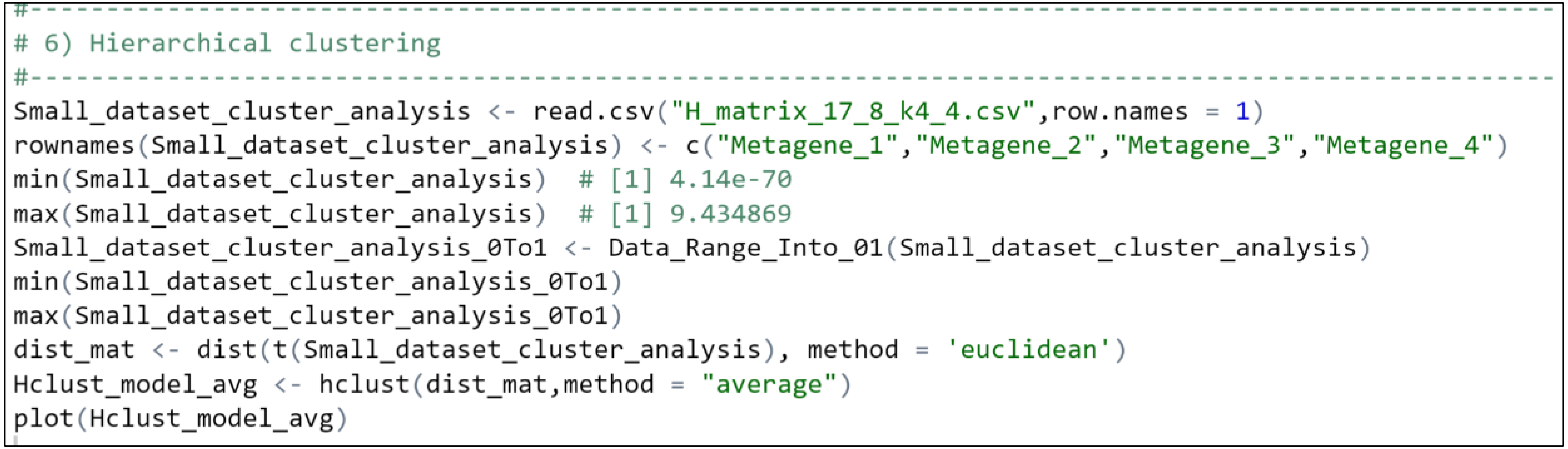


	CSC3062_108_2	CSC3062_109_4	CSC3062_110_4	CSC3062_112_2	CSC3062_783_3	CSC3062_145_3
Metagene_1	1.145277e-01	1.916895e-50	2.654951e-40	7.633172e-02	3.608274e-32	7.042284e-28
Metagene_2	1.338042e-02	5.529235e-01	5.625382e-01	4.172066e-27	5.022959e-02	1.881889e-05
Metagene_3	5.842943e-19	5.115138e-43	1.629874e-28	2.634450e-34	6.117725e-01	6.623634e-01
Metagene_4	9.603256e-01	2.808713e-27	4.787113e-29	9.671474e-01	1.660626e-34	5.350906e-39



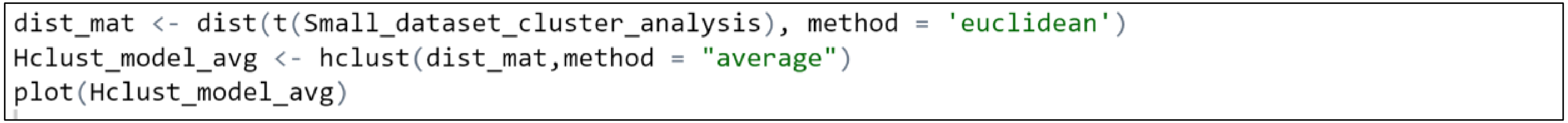
	CSC3062_649_1	CSC3062_115_1	CSC3062_670_2	CSC3062_50080_1	CSC3062_436_1	CSC3062_674_2
Metagene_1	7.176776e-01	9.121094e-01	2.142412e-28	8.314318e-01	6.650897e-01	1.424858e-17
Metagene_2	0.000000e+00	1.312099e-40	2.695954e-17	1.158338e-18	8.997966e-02	3.280249e-12
Metagene_3	1.759033e-70	3.300750e-21	3.208493e-17	1.691378e-40	3.382756e-17	2.059872e-02
Metagene_4	6.929525e-63	3.516017e-59	9.679785e-01	4.684605e-20	1.916895e-23	1.000000e+00

Running hclust() on H matrix (n=12)

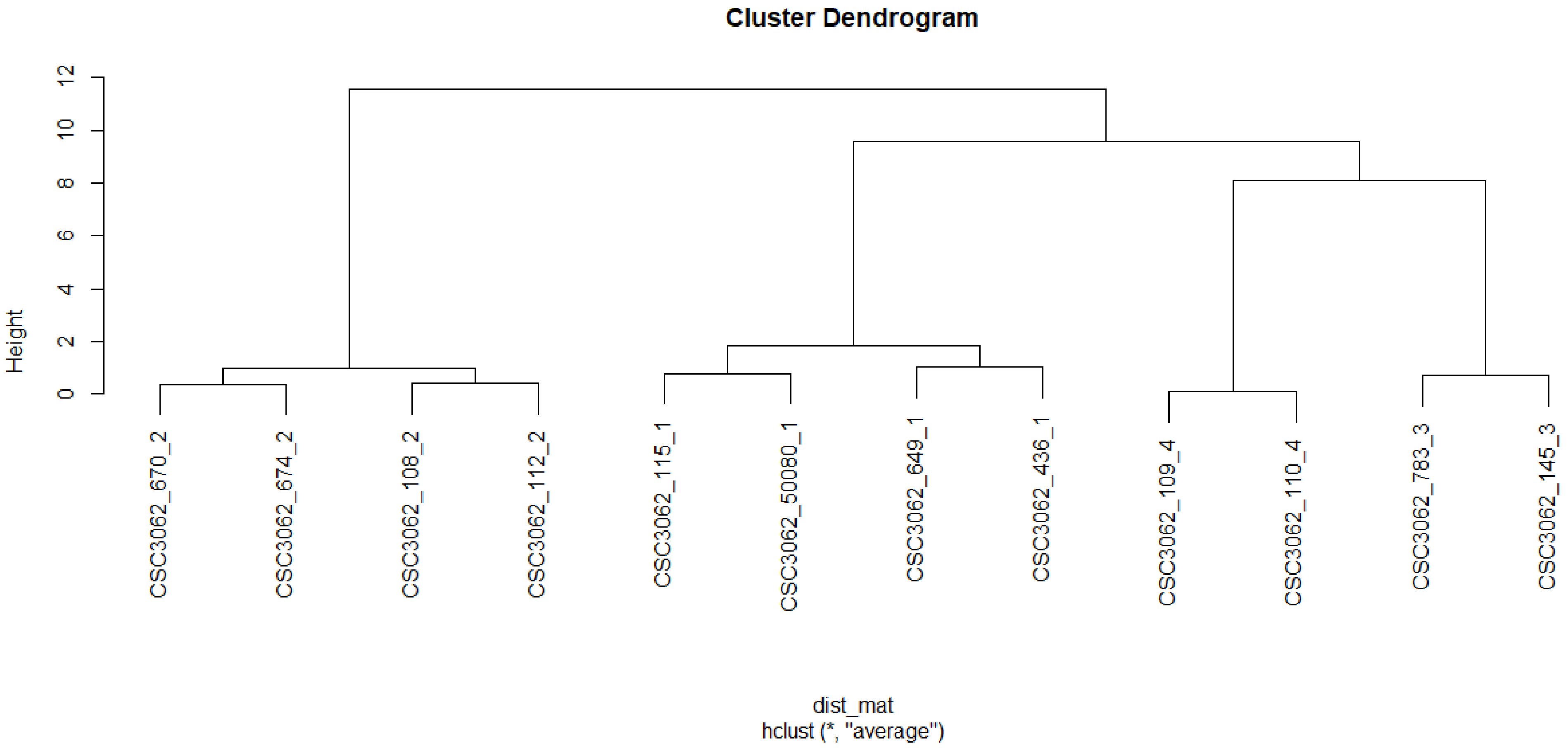


```
#-----------------------------------------------------------------------------------
# 6) Hierarchical clustering
#-----------------------------------------------------------------------------------
Small_dataset_cluster_analysis <- read.csv("H_matrix_17_8_k4_4.csv",row.names = 1)
rownames(Small_dataset_cluster_analysis) <- c("Metagene_1","Metagene_2","Metagene_3","Metagene_4")
min(Small_dataset_cluster_analysis)  # [1] 4.14e-70
max(Small_dataset_cluster_analysis)  # [1] 9.434869
Small_dataset_cluster_analysis_0To1 <- Data_Range_Into_01(Small_dataset_cluster_analysis)
min(Small_dataset_cluster_analysis_0To1)
max(Small_dataset_cluster_analysis_0To1)
dist_mat <- dist(t(Small_dataset_cluster_analysis), method = 'euclidean')
Hclust_model_avg <- hclust(dist_mat,method = "average")
plot(Hclust_model_avg)
```

Running hclust() on H matrix (n=12)



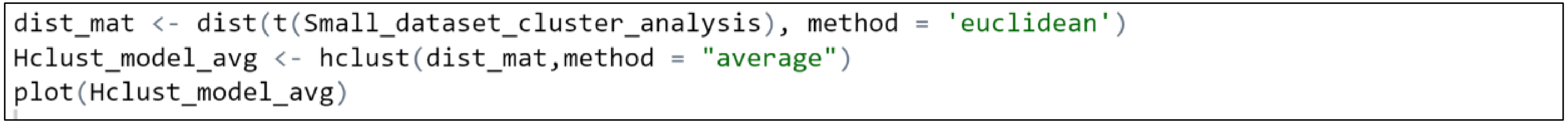
```
dist_mat <- dist(t(Small_dataset_cluster_analysis), method = 'euclidean')
Hclust_model_avg <- hclust(dist_mat,method = "average")
plot(Hclust_model_avg)
```



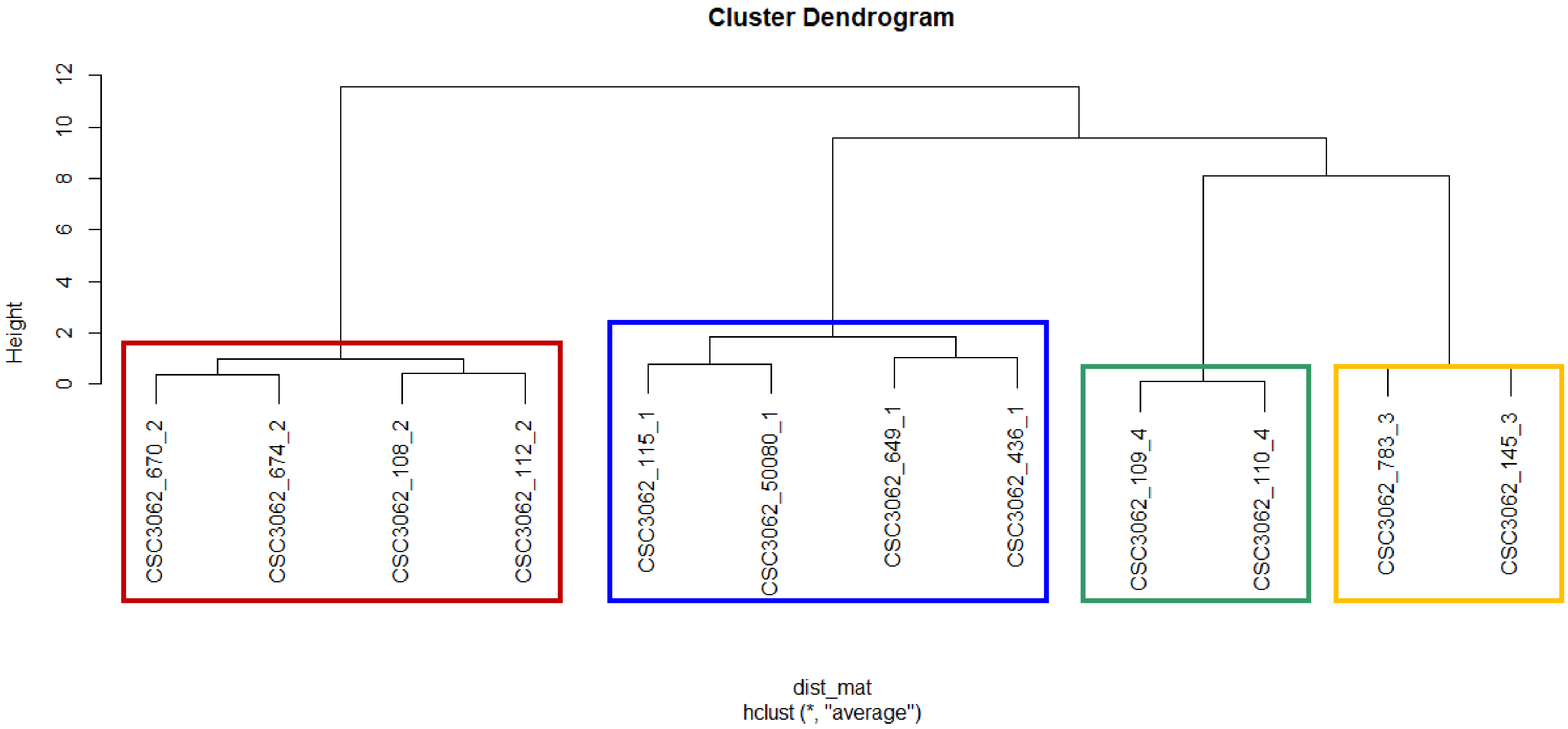
QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Running hclust() on H matrix (n=12)

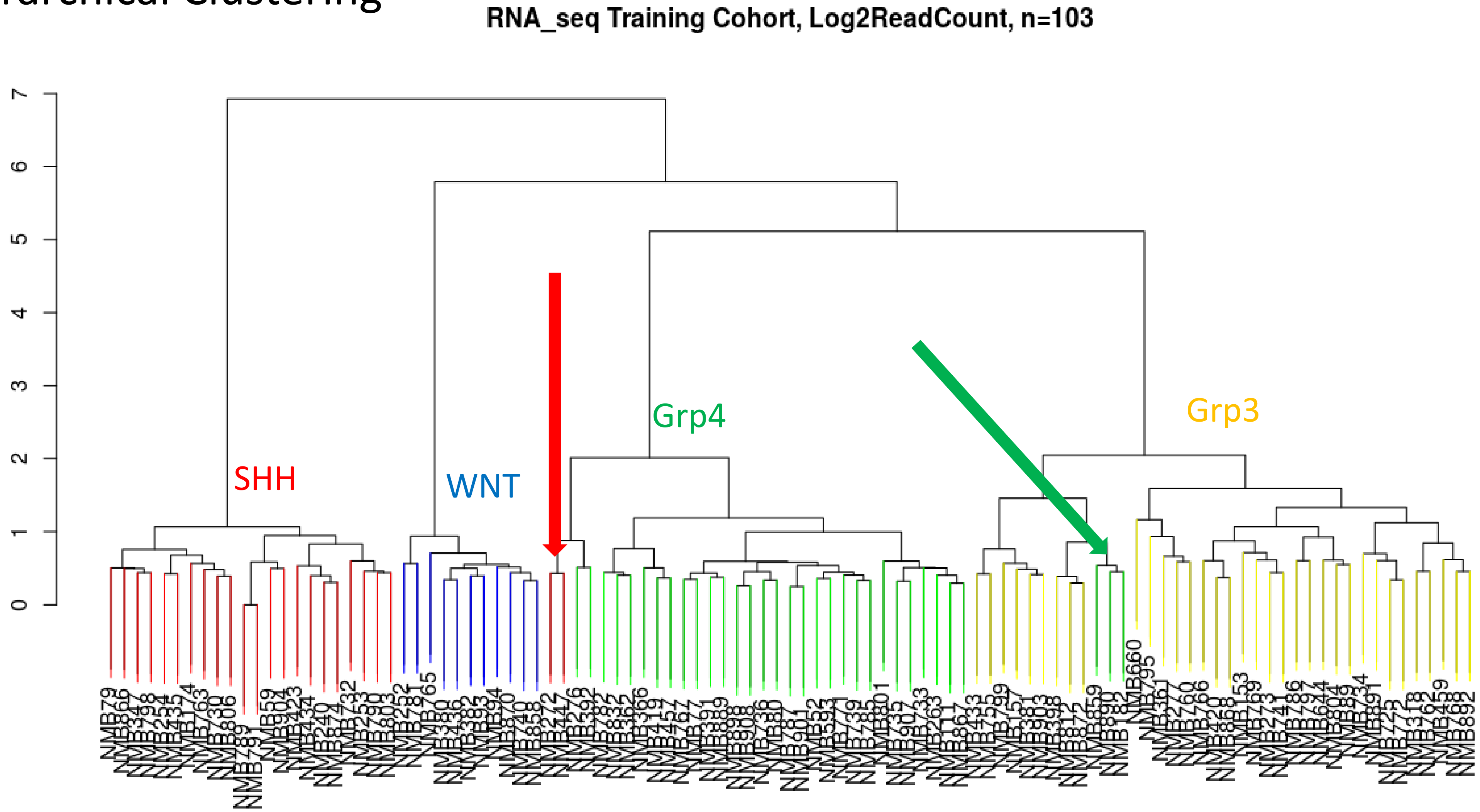


```
dist_mat <- dist(t(Small_dataset_cluster_analysis), method = 'euclidean')
Hclust_model_avg <- hclust(dist_mat,method = "average")
plot(Hclust_model_avg)
```



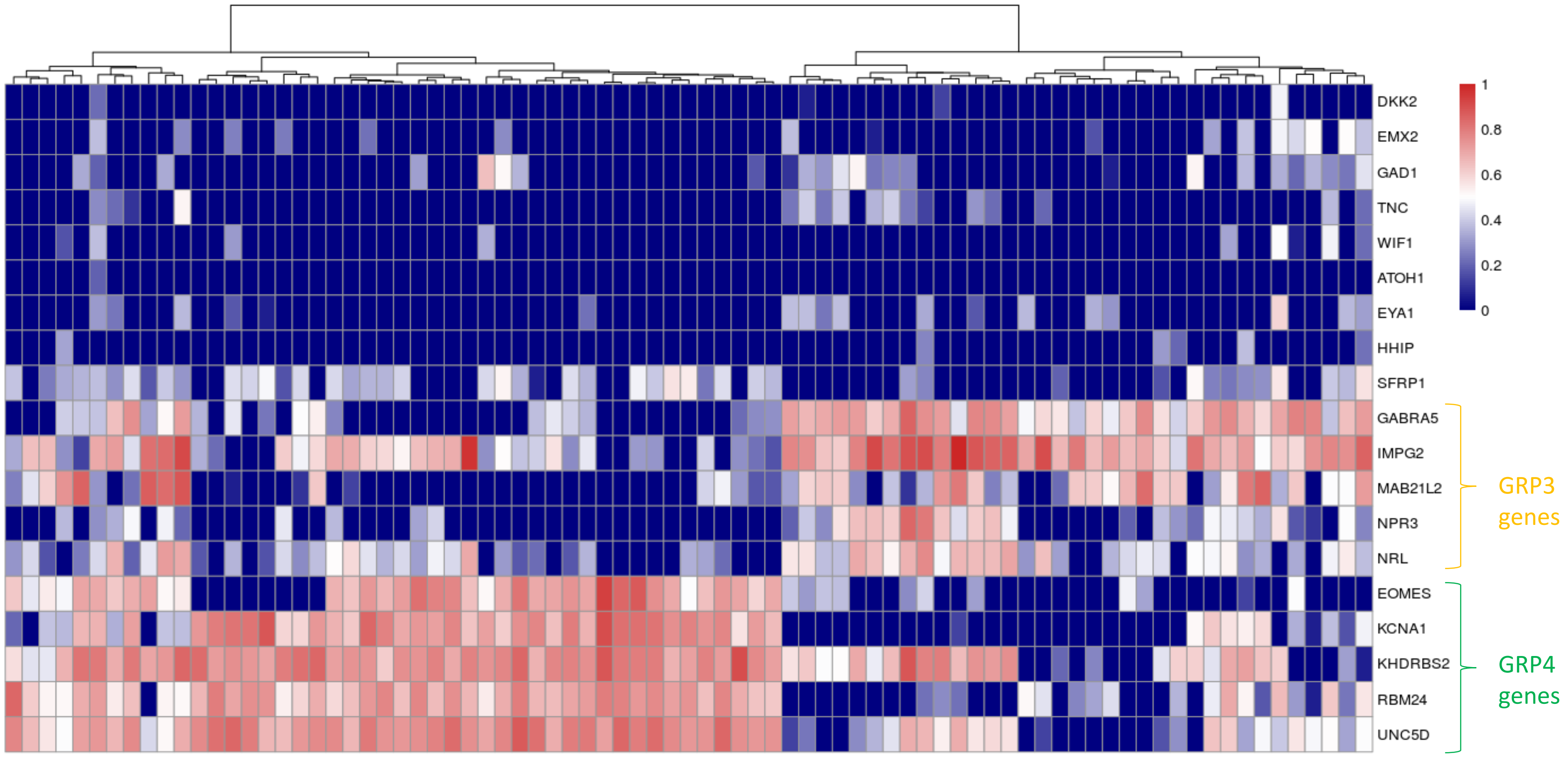
Running hclust() on n=103 samples

Ward's Hierarchical Clustering



Running hclust() on n=81 samples

Ward's Hierarchical Clustering Heatmap of raw data (n=81, 19 genes)



Hieratical clustering

Question?

Is this clustering useful for estimating the number of clusters?

Any Questions?