

STAT540

Lecture 16: March 4th 2020

Cluster Analysis

Sara Mostafavi

Department of Statistics

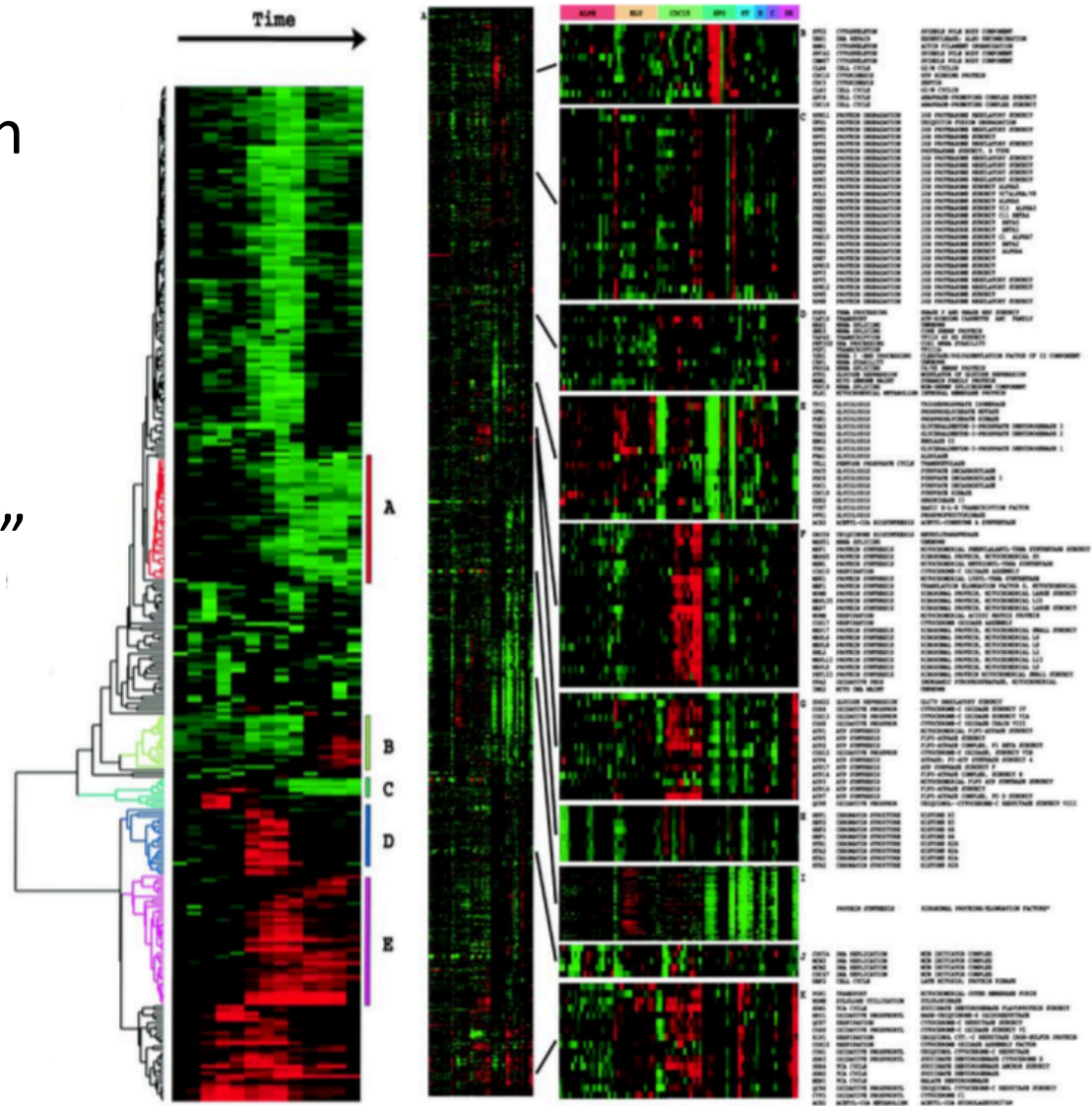
Department of Medical Genetics

Center for Molecular Medicine and Therapeutics

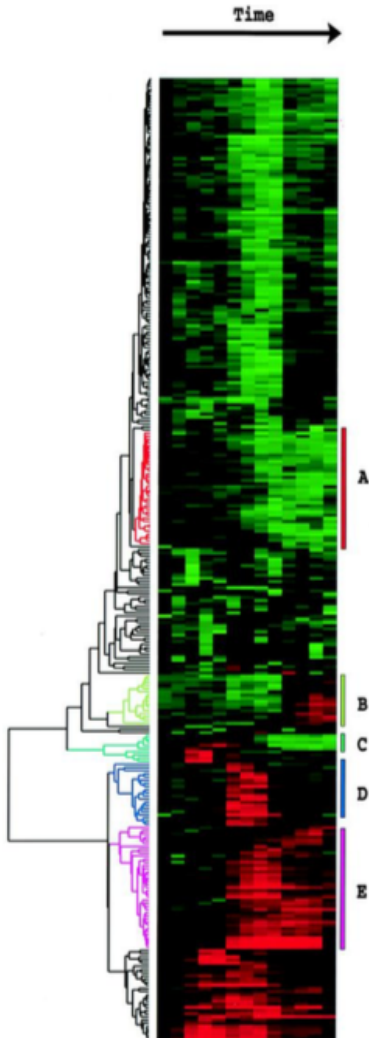
**** Slide credits:Drs. Gabriela Cohen-Freue and Jenny Bryan for lecture slides****

A familiar science in an
'omics' paper ...

Behind the scene
cluster analysis (CA)
was used to “organize”
genes into groups
(clusters) and create
dendrogram



Cluster analysis in 'omics' studies



“Cluster analysis and display of genome-wide expression patterns ” by Eisen, et al. (PNAS, Vol. 95, pp. 14863– 14868, December 1998)

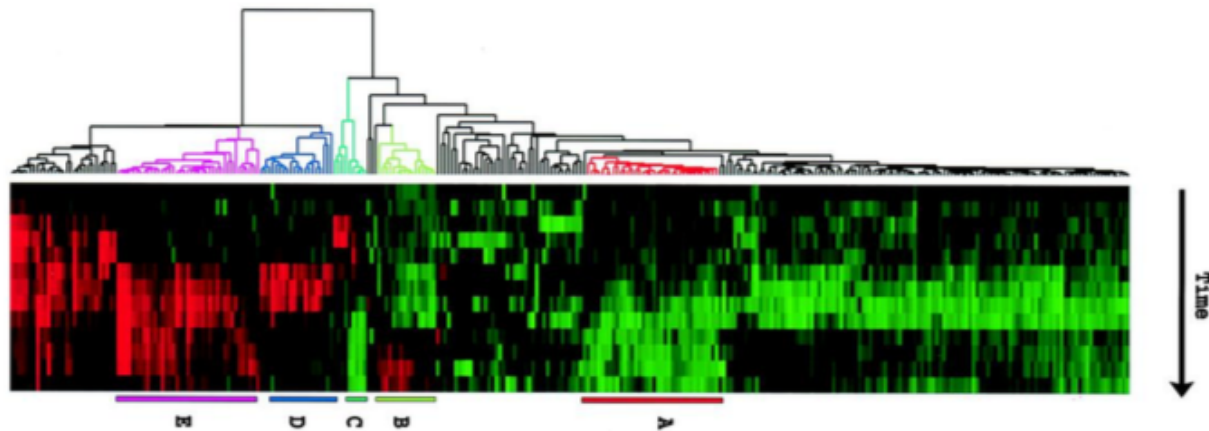
Imprinted cluster analysis (CA) on the microarray community

This precedent + explosion of array data + ease of application = widespread (over?) use of CA

Currently, CA is used similarly in many other – omics studies.

Example utility of clustering in gene expression studies

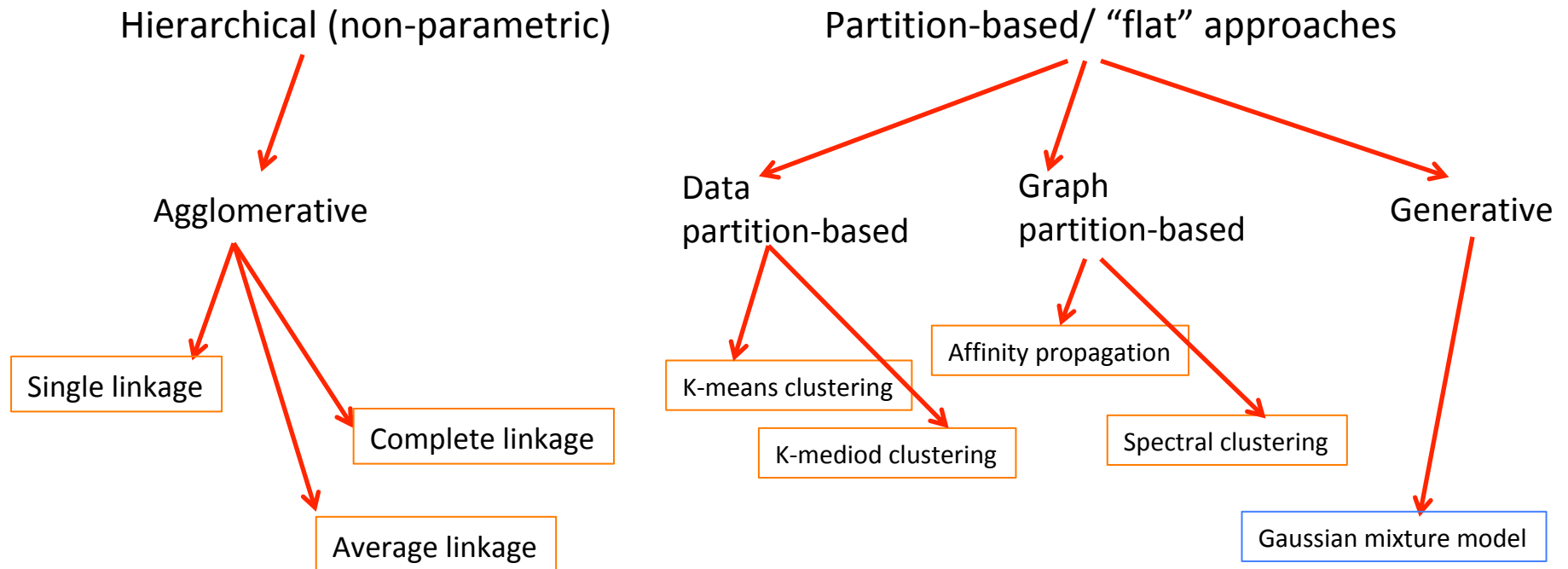
- Eisen et al., showed that clustering genes by co-expression pattern results in groups of genes that share the same function (involved in the same cellular/biological processes)



What is Clustering?

- “Clustering” Colloquially means placing/grouping a set of objects into groups/clusters.
- Clustering is a formal **problem** in Computer Science and in Statistics, with formal definitions and “solutions”.
- Clustering in bioinformatics is often used as a tool for visualization, hypothesis generation, selection of genes for further analysis.
 - Keep in mind, with typical use of clustering in bioinformatics: there is no measure of “strength of evidence” or “strength of clustering structure” provided.
- Rigorous application of clustering is very powerful for making sense of data (e.g., what are the main patterns/trends in data)

Some existing clustering approaches



- Discrete clustering assignment
- Probabilistic cluster assignment

Clustering problem: definition

- **Goal:** place a set of **objects** into groups or **clusters** in a way that **similar** objects are in the same cluster.

Cluster some rocks:



Clustering problem: definition

- **Goal:** place a set of **objects** into groups or **clusters** in a way that **similar** objects are in the same cluster.

Cluster some rocks:



Rocks were clustered according to their color and texture.

Clustering problem: definition

- **Goal:** place a set of **objects** into groups or **clusters** in a way that **similar** objects are in the same cluster.

Cluster some rocks:



Other ways you can cluster these rocks?

Clustering problem: definition

- **Goal:** place a set of **objects** into groups or **clusters** in a way that **similar** objects are in the same cluster.

Cluster some rocks:



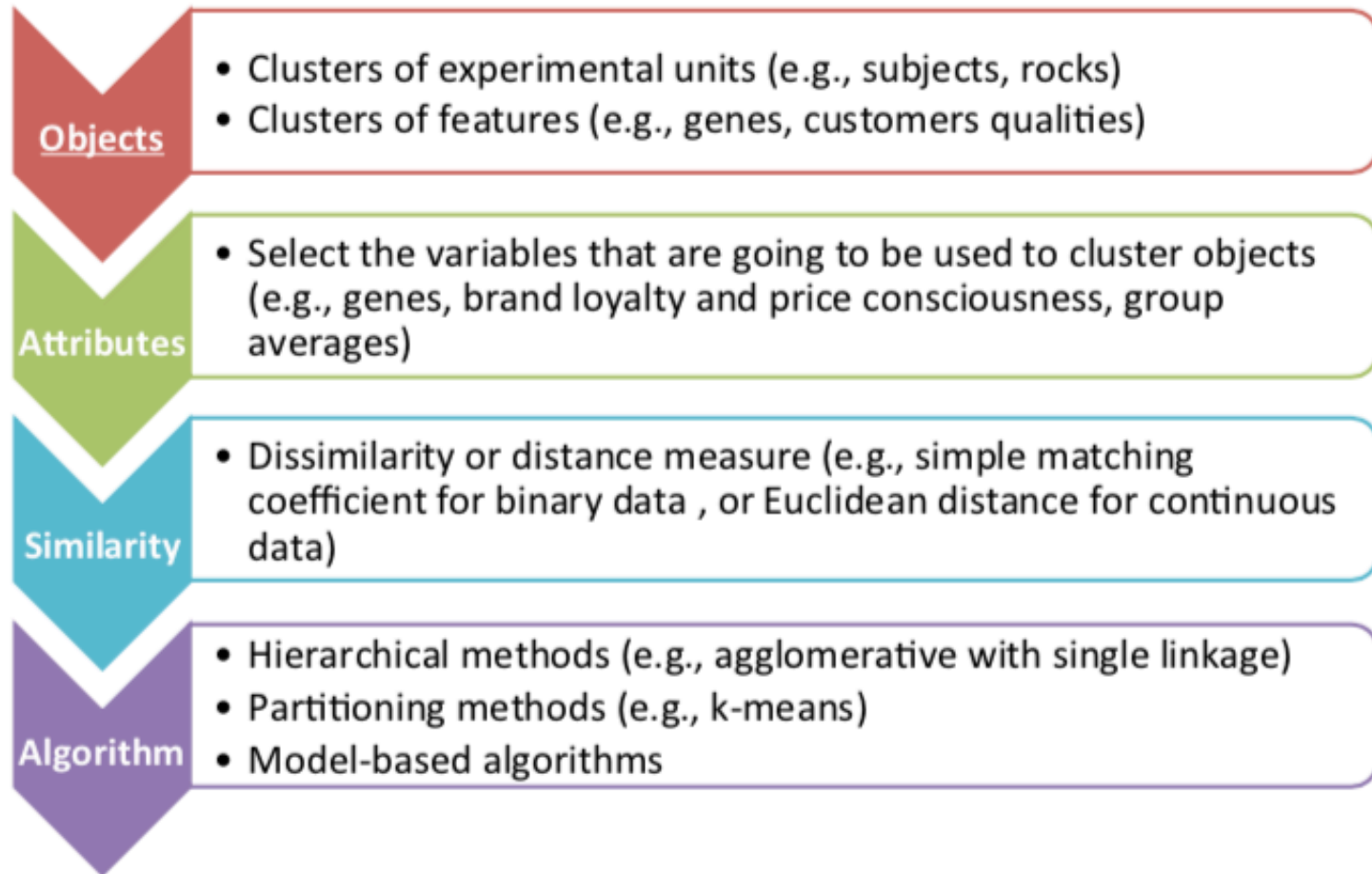
Note that you could have also considered a 2-cluster solution.

Clustering: definitions

- **Goal:** place a set of **objects** into groups or **clusters**.
- **How** do we do this?
 - gather a set of **attributes** (e.g., texture, size, shape) for each object.
 - Place objects in clusters so that objects within each cluster are more **similar** to each other compared to objects that outside their group/cluster.



Key elements



Number of clusters???

Example: photoreceptor data

- Gene expression of purified photoreceptors at distinct developmental stages and from different genetic background
- Almost 30K genes and 39 samples (mice)
- 5 developmental stages: day 16 of embryonic development (E16), postnatal days 2, 6 and 10 (P2, P7, and P10) as well as 4_weeks
- 2 genetic backgrounds: wild type (WT) and knockout Nrl mice (NrlKO)

A peak at the data...

	Wild Type			Knock-out			...			Wild Type			Knock-out		
	E16									4_weeks					
	Sample_20	...	Sample_23	Sample_16	...	Sample_17	...	Sample_i	...	Sample_36	...	Sample_39	Sample_11	...	Sample_9
1415670_at	7.24	...	7.07	7.38	...	7.34	7.25	...	7.13	7.42	...	7.32
1415671_at	9.48	...	10.13	7.64	...	10.03	9.66	...	8.73	9.83	...	9.80
1415672_at	10.01	...	9.91	8.42	...	10.24	9.51	...	9.53	10.00	...	9.85
1415673_at	8.36	...	8.49	8.36	...	8.37	8.49	...	8.65	8.60	...	8.40
1415674_a_at	8.59	...	8.64	8.51	...	8.89	8.42	...	8.28	8.43	...	8.46
1415675_at	9.59	...	9.70	9.66	...	9.61	9.67	...	9.45	9.60	...	9.51
1415676_a_at	9.68	...	10.19	8.05	...	10.02	9.95	...	8.70	9.23	...	9.82
1415677_at	7.24	...	7.49	7.34	...	7.34	7.28	...	6.84	7.33	...	7.45
1415678_at	11.71	...	11.57	10.46	...	11.75	11.56	...	11.80	12.04	...	11.81
1415679_at	9.21	...	9.92	8.22	...	9.60	9.13	...	8.08	9.06	...	9.29
...
gene g	$X_{g,i}$
...
1460746_at	6.37	...	6.12	7.25	...	6.15	6.34	...	6.52	6.36	...	6.35

Column-driven analysis

	Wild Type			Knock-out			...			Wild Type			Knock-out		
	E16									4_weeks					
	Sample_20	...	Sample_23	Sample_16	...	Sample_17	...	Sample_i	...	Sample_36	...	Sample_39	Sample_11	...	Sample_9
1415670_at	7.24	...	7.07	7.38	...	7.34	7.25	...	7.13	7.42	...	7.32
1415671_at	9.48	...	10.13	7.64	...	10.03	9.66	...	8.73	9.83	...	9.80
1415672_at	10.01	...	9.91	8.42	...	10.24	9.51	...	9.53	10.00	...	9.85
1415673_at	8.36	...	8.49	8.36	...	8.37	8.49	...	8.65	8.60	...	8.40
1415674_a_at	8.59	...	8.64	8.51	...	8.89	8.42	...	8.28	8.43	...	8.46
1415675_at	9.59	...	9.70	9.66	...	9.61	9.67	...	9.45	9.60	...	9.51
1415676_a_at	9.68	...	10.19	8.05	...	10.02	9.95	...	8.70	9.23	...	9.82
1415677_at	7.24	...	7.49	7.34	...	7.34	7.28	...	6.84	7.33	...	7.45
1415678_at	11.71	...	11.57	10.46	...	11.75	11.56	...	11.80	12.04	...	11.81
1415679_at	9.21	...	9.92	8.22	...	9.60	9.13	...	8.08	9.06	...	9.29
...
gene g	X_{gi}
...
1460746_at	6.37	...	6.12	7.25	...	6.15	6.34	...	6.52	6.36	...	6.35



Sample clustering: based on expression of g genes (attributes) how do the samples (mice) cluster?

Row-driven analysis

	Wild Type			Knock-out			***			Wild Type			Knock-out		
	E16									4_weeks					
	Sample_20	...	Sample_23	Sample_16	...	Sample_17	...	Sample_i	...	Sample_36	...	Sample_39	Sample_11	...	Sample_9
1415670_at	7.24	...	7.07	7.38	...	7.34	7.25	...	7.13	7.42	...	7.32
1415671_at	9.48	...	10.13	7.64	...	10.03	9.66	...	8.73	9.83	...	9.80
1415672_at	10.01	...	9.91	8.42	...	10.24	9.51	...	9.53	10.00	...	9.85
1415673_at	8.36	...	8.49	8.36	...	8.37	8.49	...	8.65	8.60	...	8.40
1415674_a_at	8.59	...	8.64	8.51	...	8.89	8.42	...	8.28	8.43	...	8.46
1415675_at	9.59	...	9.70	9.66	...	9.61	9.67	...	9.45	9.60	...	9.51
1415676_a_at	9.68	...	10.19	8.05	...	10.02	9.95	...	8.70	9.23	...	9.82
1415677_at	7.24	...	7.49	7.34	...	7.34	7.28	...	6.84	7.33	...	7.45
1415678_at	11.71	...	11.57	10.46	...	11.75	11.56	...	11.80	12.04	...	11.81
1415679_at	9.21	...	9.92	8.22	...	9.60	9.13	...	8.08	9.06	...	9.29
...
gene g	X_{gi}
...
1460746_at	6.37	...	6.12	7.25	...	6.15	6.34	...	6.52	6.36	...	6.35

Sample clustering: based on expression of ?? (attributes) how do the ?? cluster?

Defining attribute/feature vector for each object

- We need to numerically define a attribute or feature vector that describes the relevant properties of each object

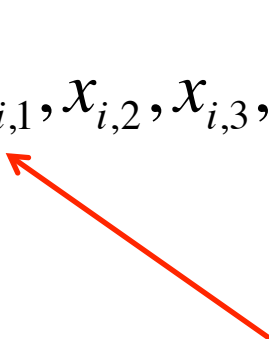
Set of n objects $\{ \vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_n \}$

Each object is represented by a numerical vector: \vec{x}_i

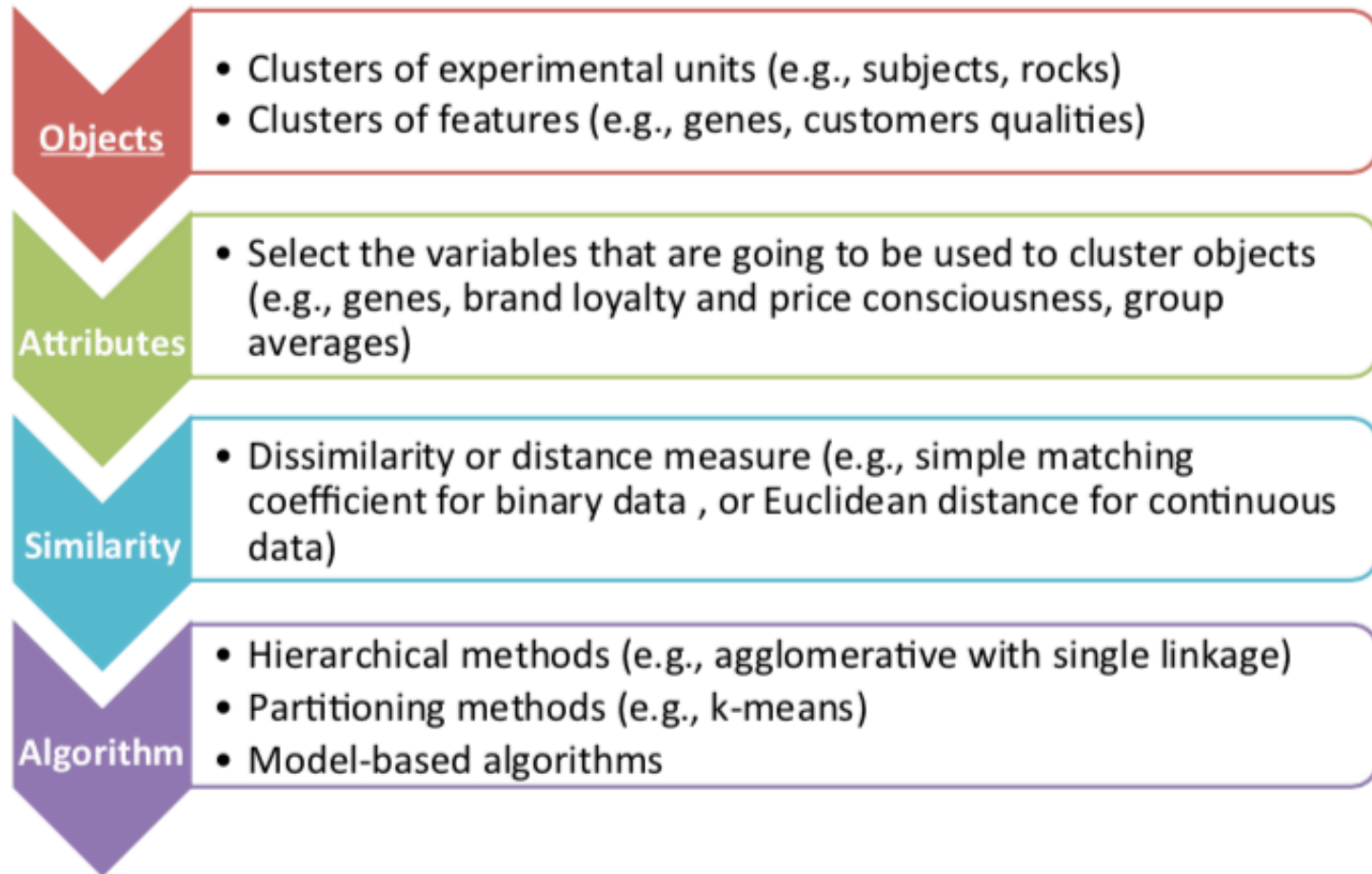
sample i: $\vec{x}_i = (x_{i,1}, x_{i,2}, x_{i,3}, \dots, x_{i,p})$

Attribute/feature p for object i

Numerical value representing expression level of each gene



Key elements



Number of clusters???

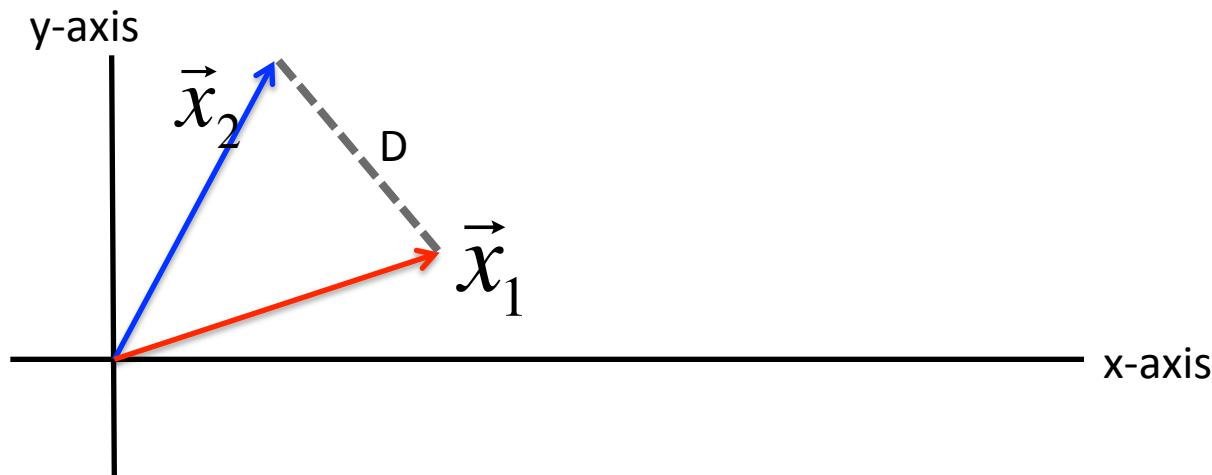
Commonly Used Measures of Similarity and Distance

- Every clustering method is based on the measure of distance or similarity.
- We need to compute pairwise similarities between all objects.
- Typical distance/similarity measures:
 - Distance:
 - Euclidean
 - Manhattan
 - Similarity: Correlation
 - Spearman
 - Pearson

Commonly used measures of similarity and distance

- Euclidian distance between two feature vectors: \vec{x}_1 and \vec{x}_2

$$D = \| \vec{x}_1 - \vec{x}_2 \|_2 = \sqrt{\sum_{j=1}^p (x_{1,j} - x_{2,j})^2}$$



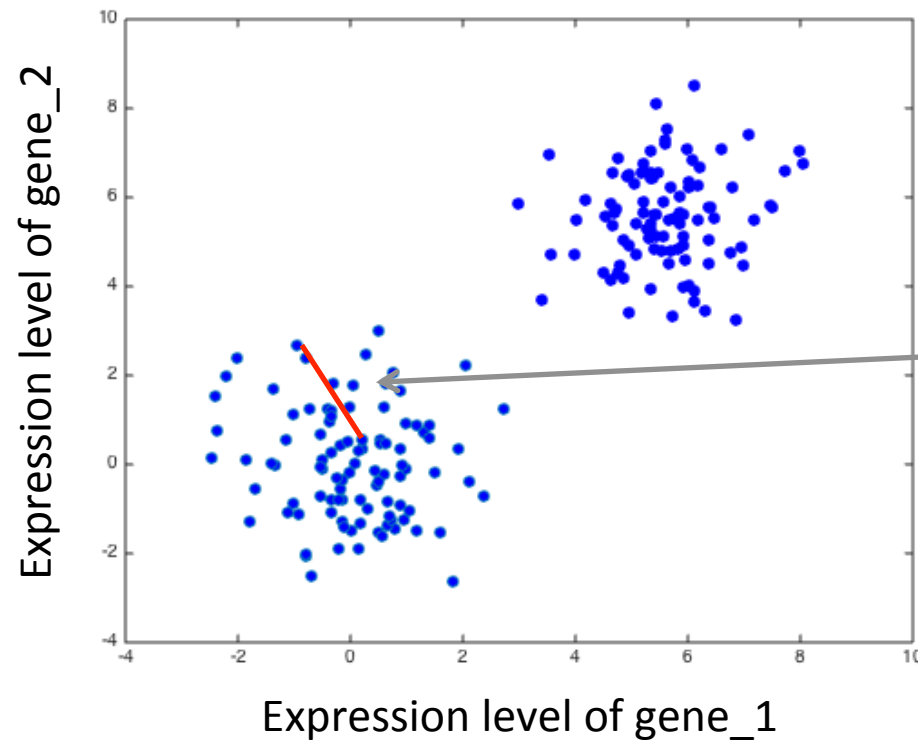
Commonly used measure of similarity and distance in genomics

- 1-Pearson correlation coefficient: \vec{x}_1 and \vec{x}_2

$$r = \frac{(\vec{x}_1 - \mu_1) \bullet (\vec{x}_2 - \mu_2)}{\sigma_x \sigma_y}$$

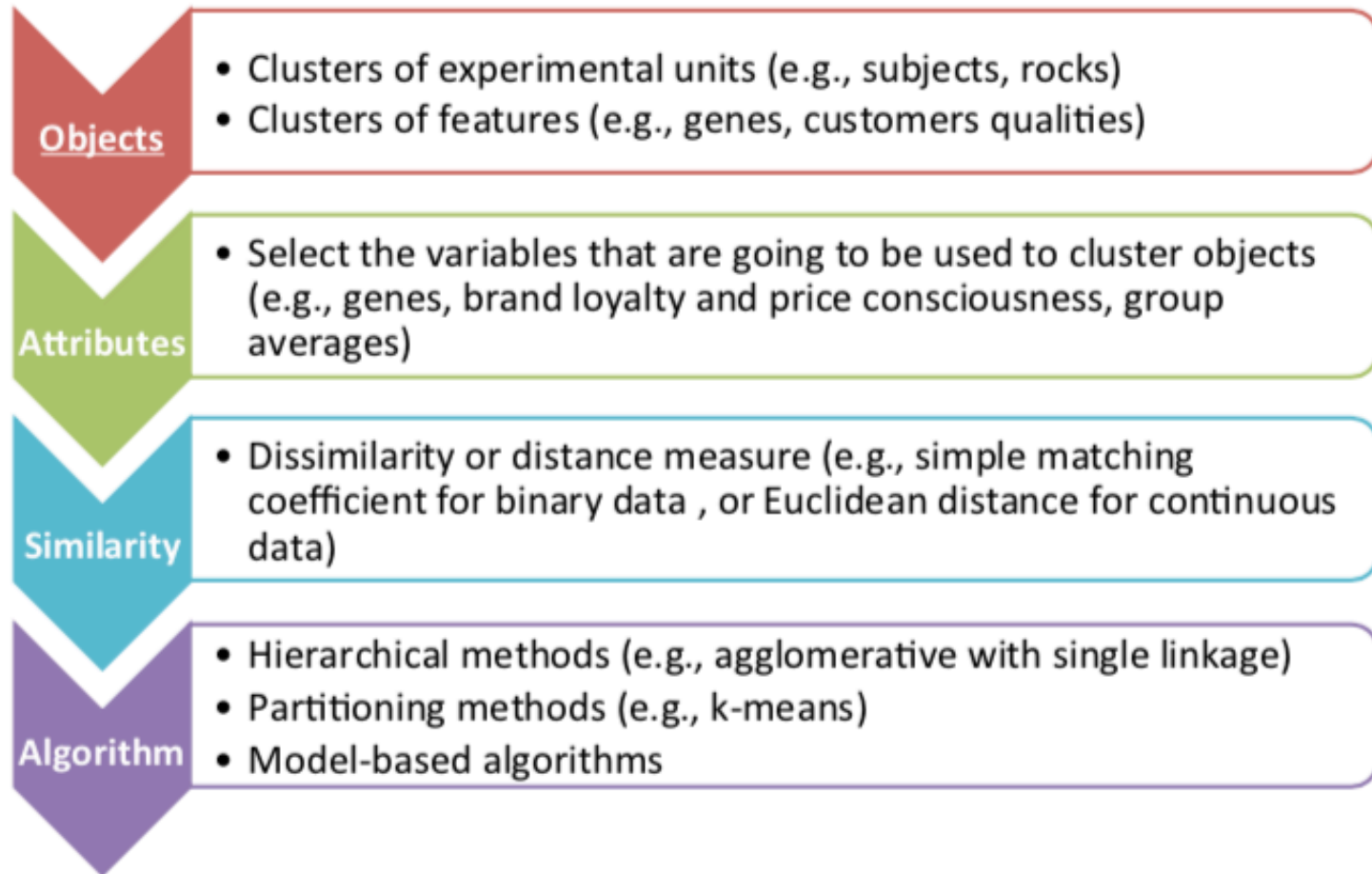
Suppose you measured gene expression level for 2 genes (A and B) for 200 individuals

Suppose you measured expression levels for 2 genes (gene A and gene B) for 200 individuals



How close are these individuals?

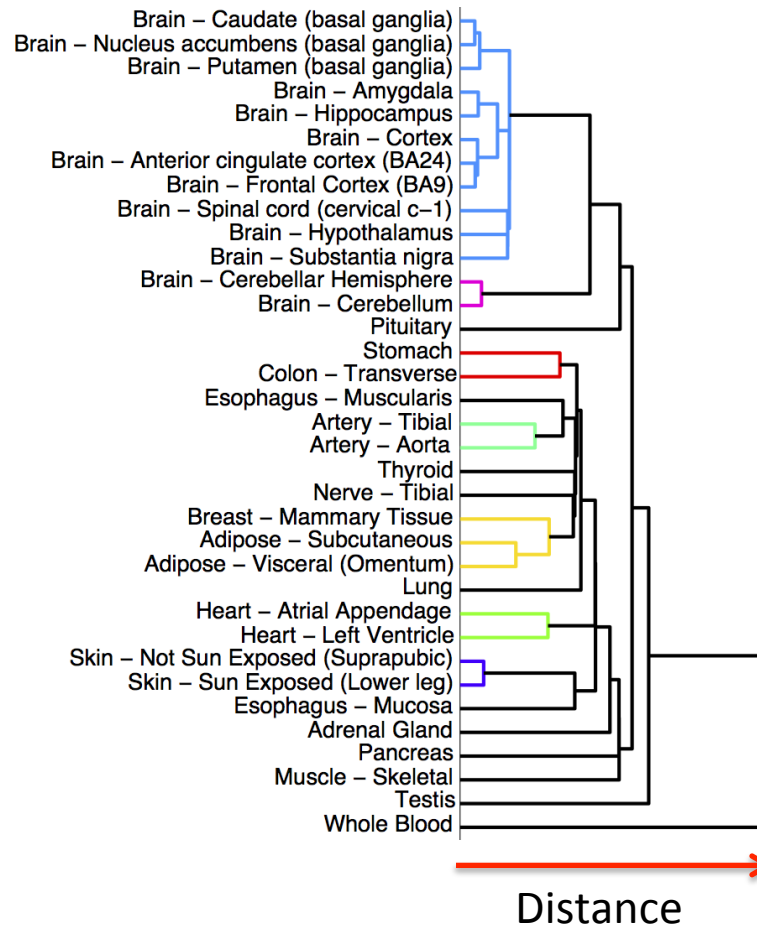
Key elements



Number of clusters???

Hierarchical Agglomerative clustering

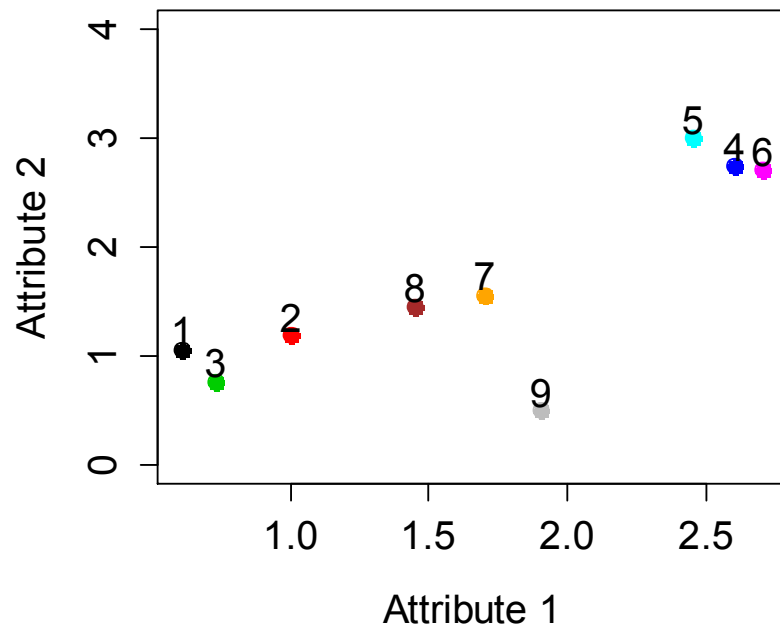
A clustering approach for revealing hierarchical relationships between objects



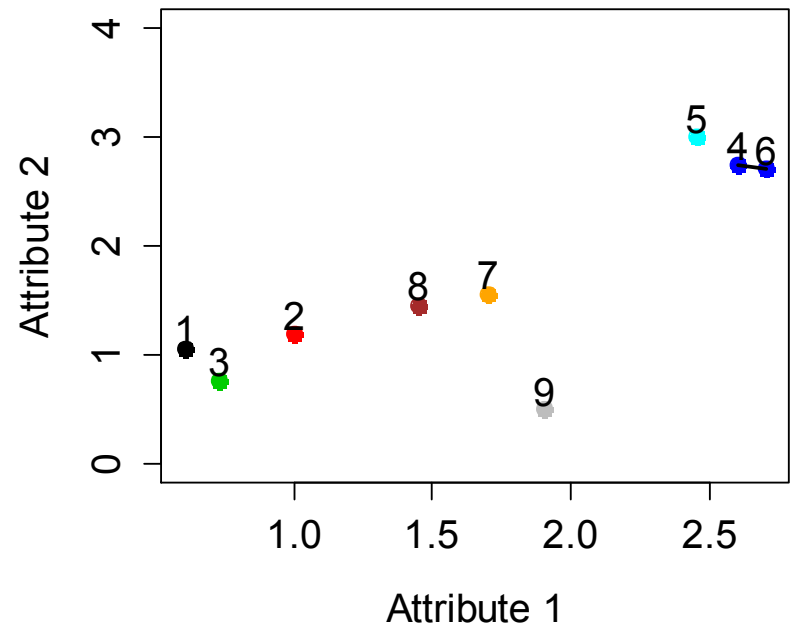
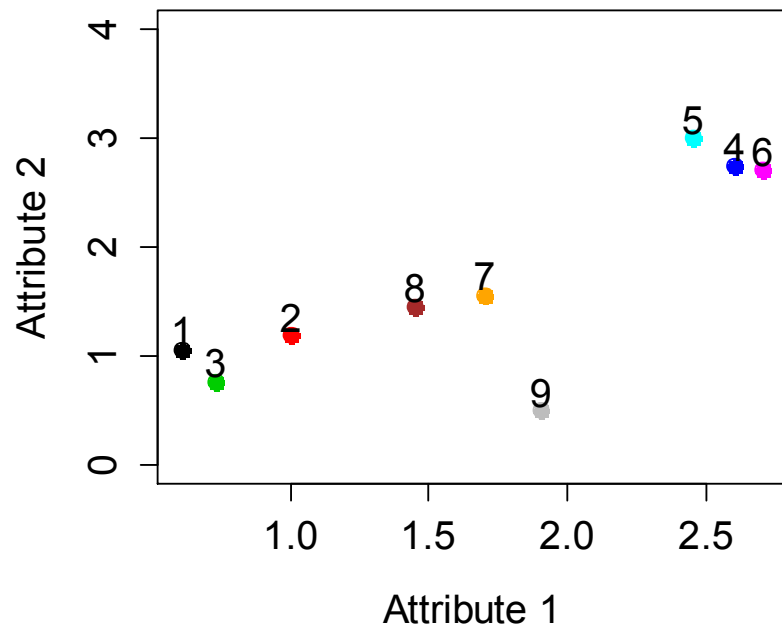
Algorithms: Hierarchical

Given *N objects* with *H attributes* and a *distance metric*:

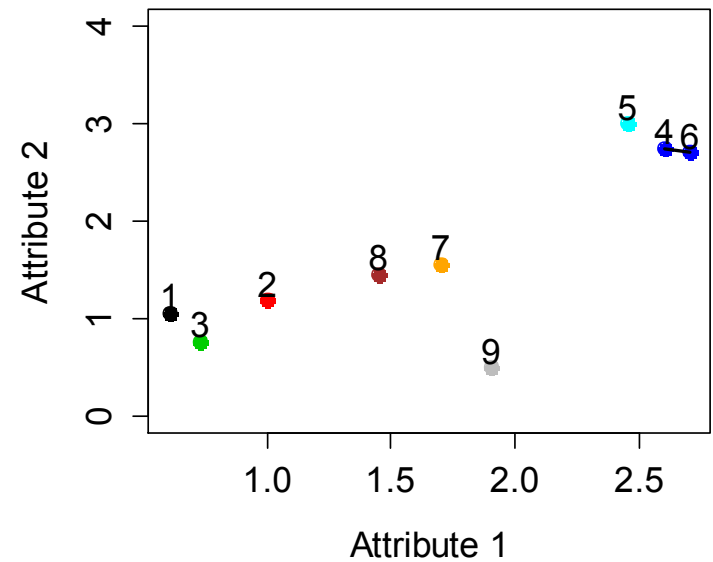
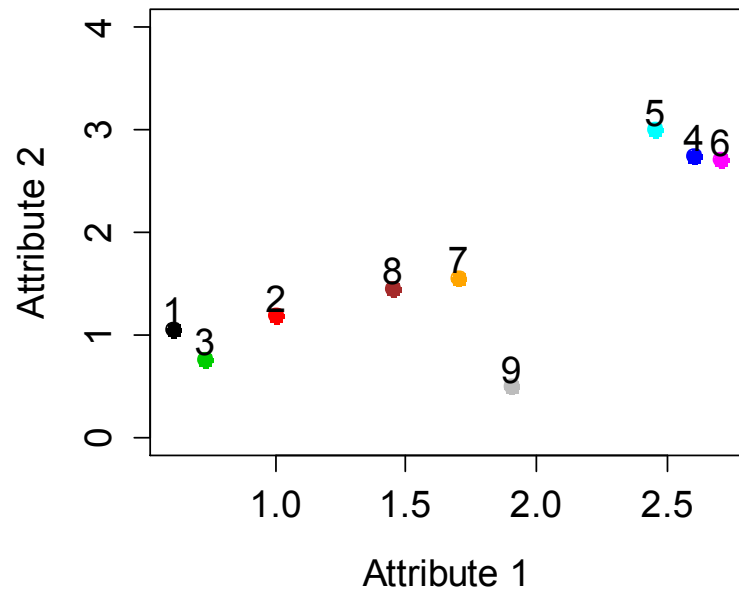
1. Assign each object to a cluster and compute the pairwise distances between all clusters
2. Find the “closest” pair of *clusters* and *merge them* into a single cluster
3. Compute new distances between clusters
4. Repeat steps 2 and 3 until all objects belong to a single cluster.



```
> round(dist(a, method='euclidean'),2)
      1      2      3      4      5      6      7      8
2 0.41
3 0.32 0.50
4 2.61 2.23 2.72
5 2.67 2.32 2.81 0.29
6 2.66 2.28 2.76 0.11 0.39
7 1.20 0.79 1.25 1.49 1.62 1.52
8 0.93 0.52 0.99 1.73 1.84 1.77 0.27
9 1.41 1.13 1.20 2.35 2.55 2.34 1.07 1.05
```



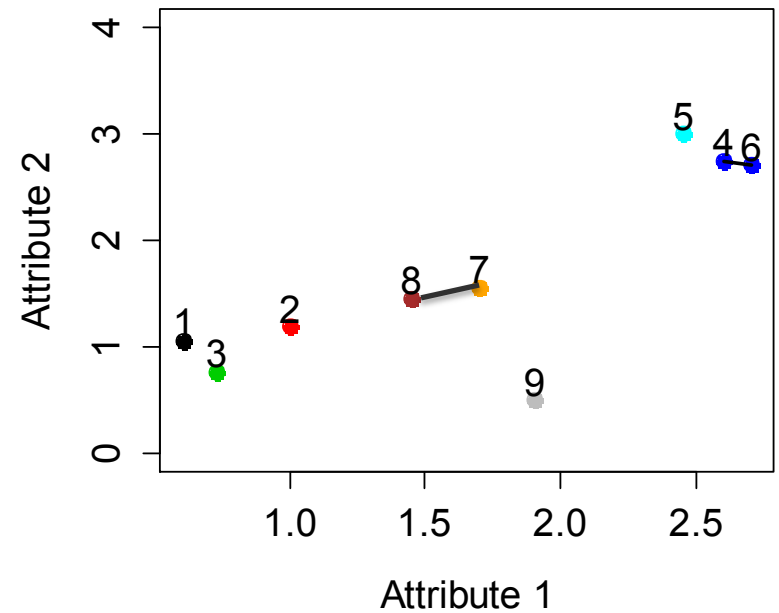
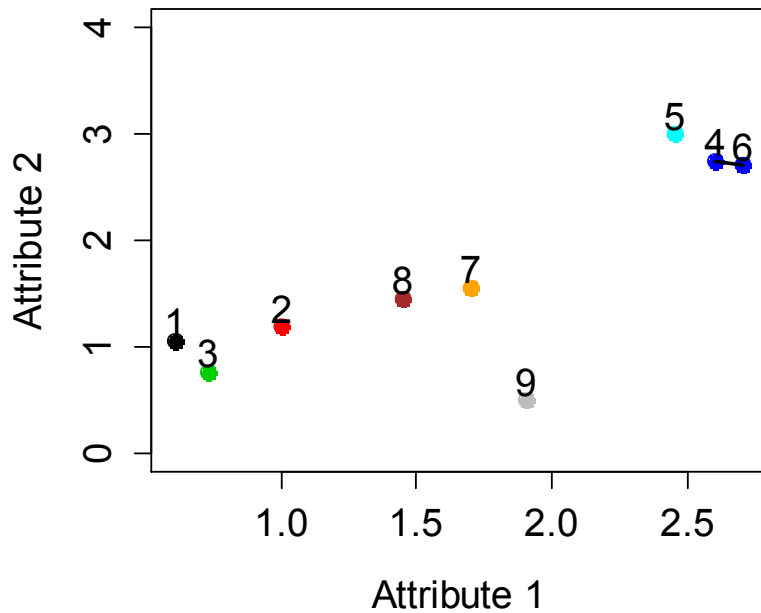
```
> round(dist(a, method='euclidean'),2)
      1      2      3      4      5      6      7      8
2 0.41
3 0.32 0.50
4 2.61 2.23 2.72
5 2.67 2.32 2.81 0.29
6 2.66 2.28 2.76 0.11 0.39
7 1.20 0.79 1.25 1.49 1.62 1.52
8 0.93 0.52 0.99 1.73 1.84 1.77 0.27
9 1.41 1.13 1.20 2.35 2.55 2.34 1.07 1.05
```



```
> round(dist(a, method='euclidean'),2)
      1      2      3      4      5      6      7      8
2 0.41
3 0.32 0.50
4 2.61 2.23 2.72
5 2.67 2.32 2.81 0.29
6 2.66 2.28 2.76 0.11 0.39
7 1.20 0.79 1.25 1.49 1.62 1.52
8 0.93 0.52 0.99 1.73 1.84 1.77 0.27
9 1.41 1.13 1.20 2.35 2.55 2.34 1.07 1.05
```

→ You can define the cluster “centroids” using:

- Single linkage
- Average linkage
- Complete linkage

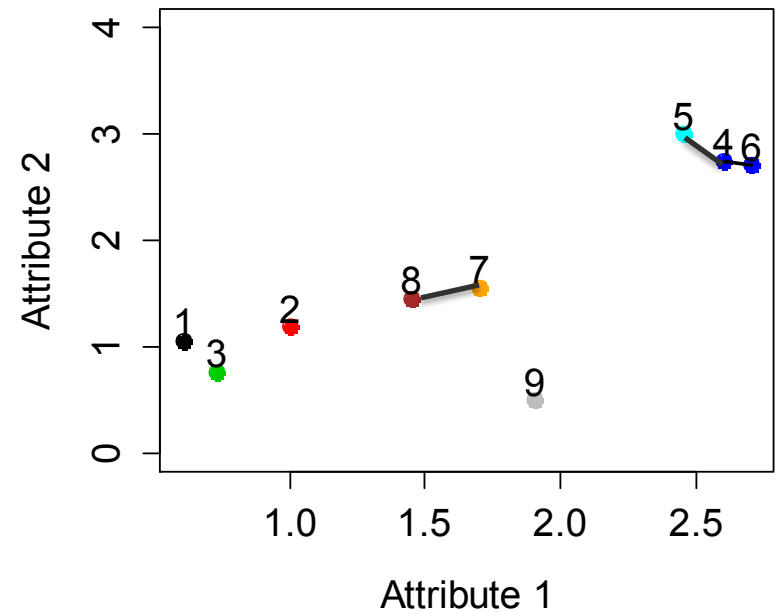
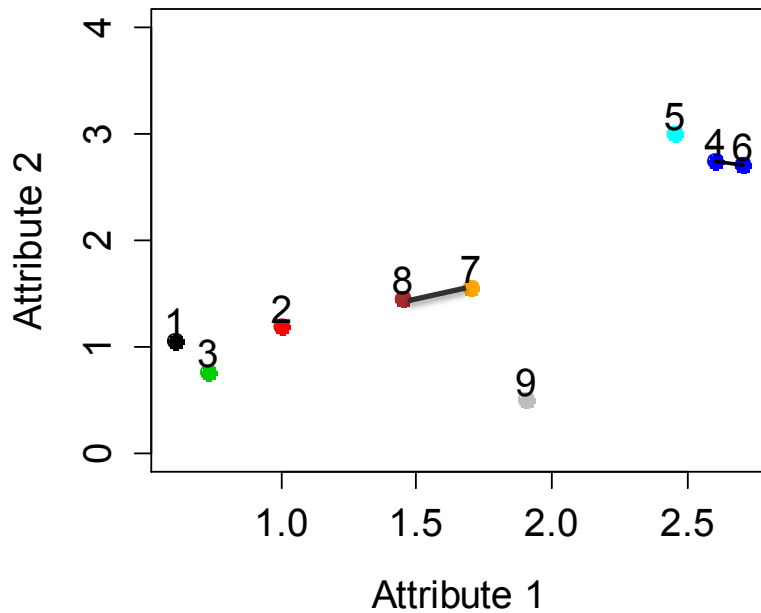


```
> round(dist(a, method='euclidean'),2)
```

	1	2	3	4	5	6	7	8
2	0.41							
3	0.32	0.50						
4	2.61	2.23	2.72					
5	2.67	2.32	2.81	0.29				
6	2.66	2.28	2.76	0.11	0.39			
7	1.20	0.79	1.25	1.49	1.62	1.52		
8	0.93	0.52	0.99	1.73	1.84	1.77	0.27	
9	1.41	1.13	1.20	2.35	2.55	2.34	1.07	1.05

→ You can define the cluster “centroids” using:

- Single linkage
- Average linkage
- Complete linkage

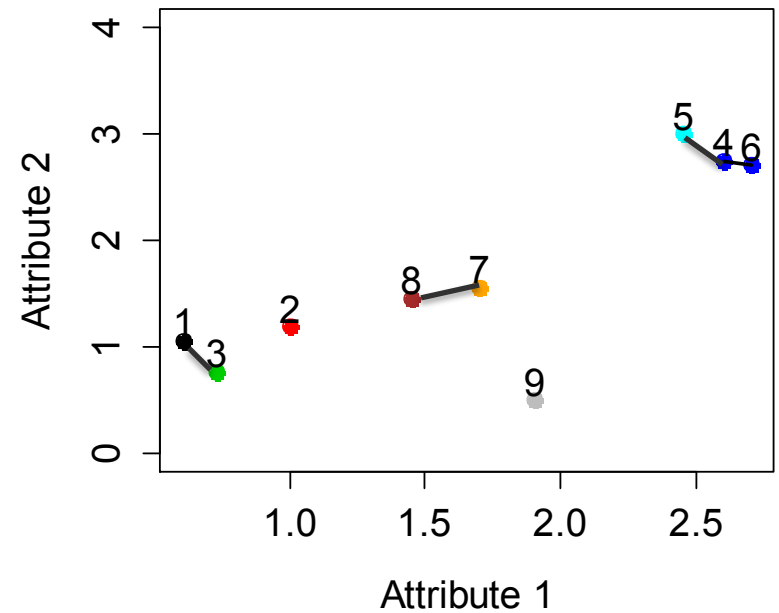
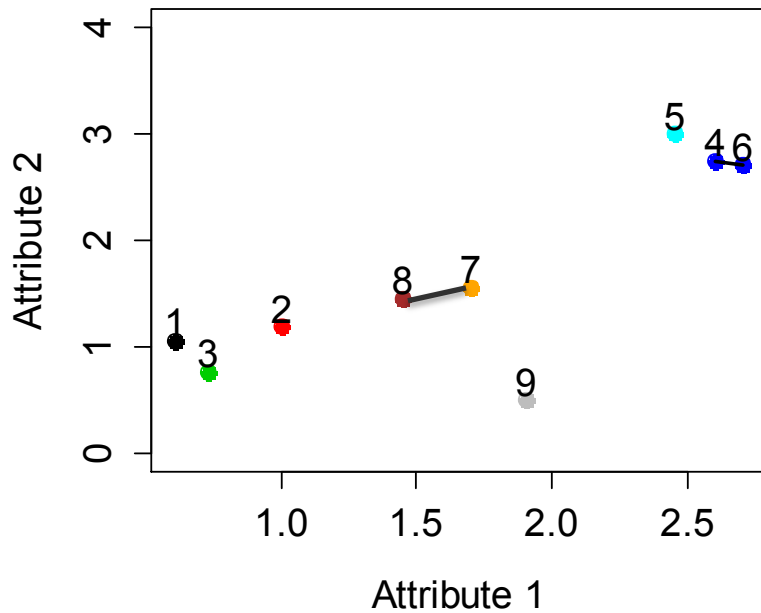


```
> round(dist(a, method='euclidean'),2)
```

	1	2	3	4	5	6	7	8
2	0.41							
3	0.32	0.50						
4	2.61	2.23	2.72					
5	2.67	2.32	2.81	0.29				
6	2.66	2.28	2.76	0.11	0.39			
7	1.20	0.79	1.25	1.49	1.62	1.52		
8	0.93	0.52	0.99	1.73	1.84	1.77	0.27	
9	1.41	1.13	1.20	2.35	2.55	2.34	1.07	1.05

→ You can define the cluster “centroids” using:

- Single linkage
- Average linkage
- Complete linkage



```
> round(dist(a, method='euclidean'),2)
```

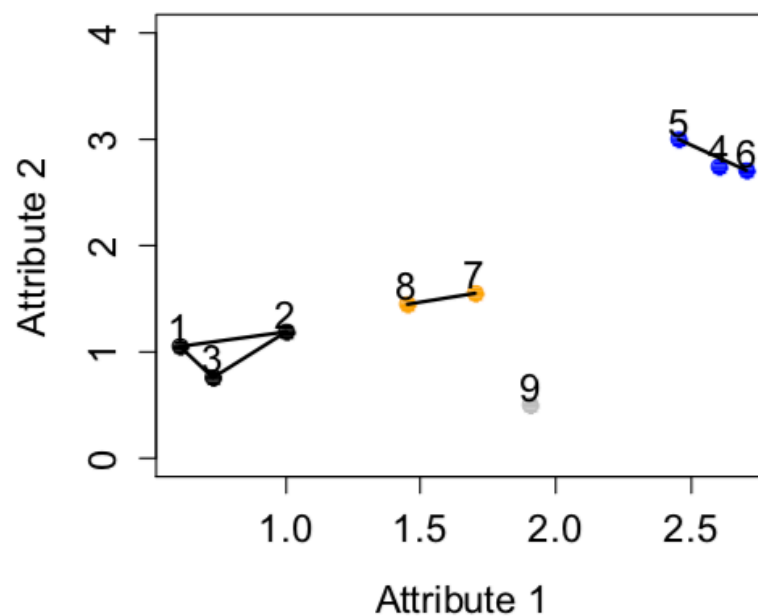
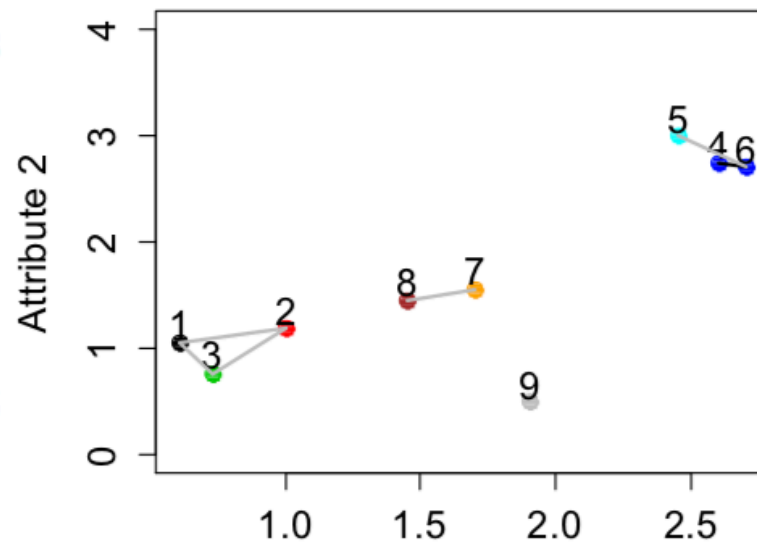
	1	2	3	4	5	6	7	8
2	0.41							
3	0.32	0.50						
4	2.61	2.23	2.72					
5	2.67	2.32	2.81	0.29				
6	2.66	2.28	2.76	0.11	0.39			
7	1.20	0.79	1.25	1.49	1.62	1.52		
8	0.93	0.52	0.99	1.73	1.84	1.77	0.27	
9	1.41	1.13	1.20	2.35	2.55	2.34	1.07	1.05

→ You can define the cluster “centroids” using:

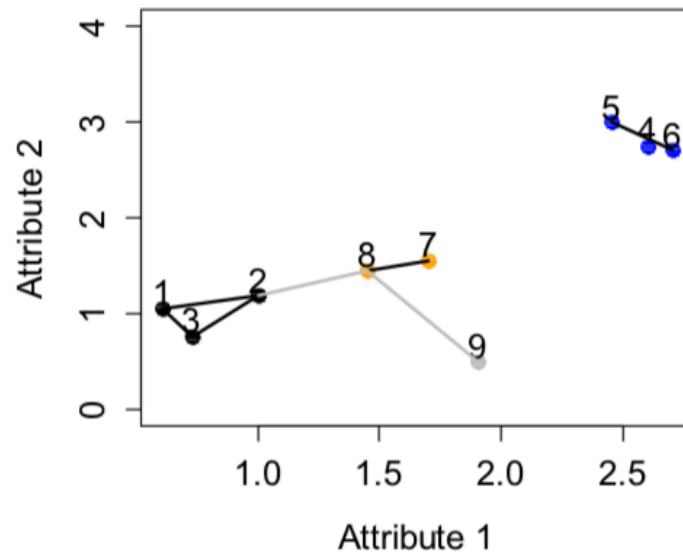
- Single linkage
- Average linkage
- Complete linkage

```
> round(dist(a, method='euclidean'),2)
```

	1	2	3	4	5	6	7	8
2	0.41							
3	0.32	0.50						
4	2.61	2.23	2.72					
5	2.67	2.32	2.81	0.29				
6	2.66	2.28	2.76	0.11	0.39			
7	1.20	0.79	1.25	1.49	1.62	1.52		
8	0.93	0.52	0.99	1.73	1.84	1.77	0.27	
9	1.41	1.13	1.20	2.35	2.55	2.34	1.07	1.05




```
> round(dist(a, method='euclidean'),2)
      1      2      3      4      5      6      7      8
2 0.41
3 0.32 0.50
4 2.61 2.23 2.72
5 2.67 2.32 2.81 0.29
6 2.66 2.28 2.76 0.11 0.39
7 1.20 0.79 1.25 1.49 1.62 1.52
8 0.93 0.52 0.99 1.73 1.84 1.77 0.27
9 1.41 1.13 1.20 2.35 2.55 2.34 1.07 1.05
```



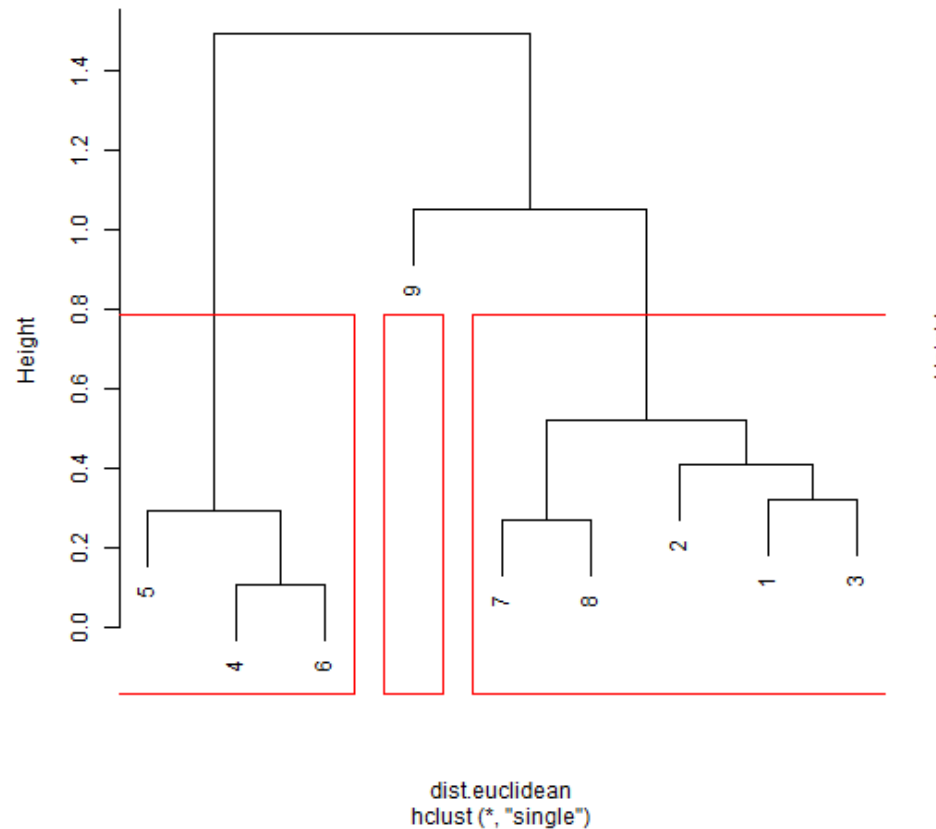
Single Linkage

```
# Dendrogram
dist.euclidean = dist(a, method = "euclidean")

# Single
ex1.hc5 <- hclust(dist.euclidean, method = "single")
plot(ex1.hc5)

# identify 3 clusters
ex1.hc5.3 <- rect.hclust(ex1.hc5, k = 3)
```

Cluster Dendrogram



Agglomerative clustering

- **Single linkage:** The distance between two clusters is the *minimum* distance between any two elements.
- **Complete linkage:** The distance between two clusters is the *maximum* distance between any two elements.
- **Average linkage:** The distance between two clusters is the *average* of all pairwise distances between any two objects.

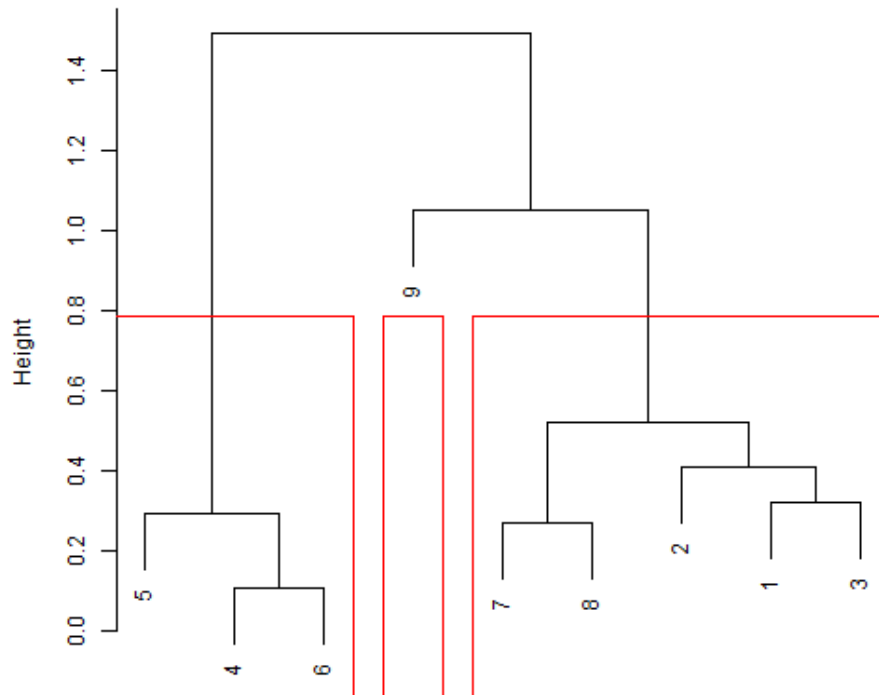
Single Linkage

```
# Dendrogram
dist.euclidean = dist(a, method = "euclidean")

# Single
ex1.hcS <- hclust(dist.euclidean, method = "single")
plot(ex1.hcS)

# identify 3 clusters
ex1.hcS.3 <- rect.hclust(ex1.hcS, k = 3)
```

Cluster Dendrogram

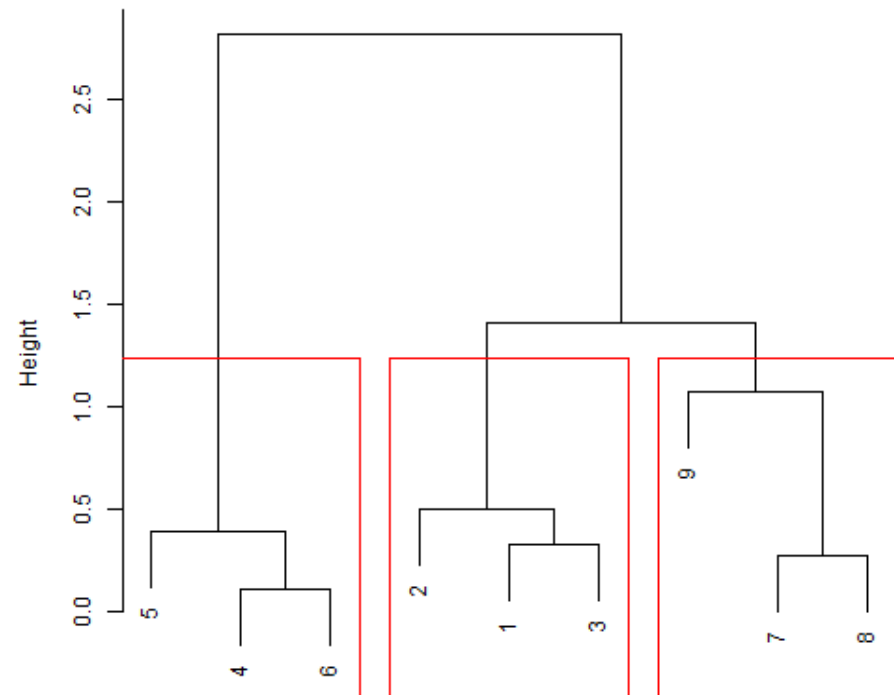


Complete Linkage

```
# Complete
ex1.hcC <- hclust(dist.euclidean, method = "complete")
plot(ex1.hcC)

# identify 3 clusters
ex1.hcC.3 <- rect.hclust(ex1.hcC, k = 3)
```

Cluster Dendrogram



Is there an intermediate solution?

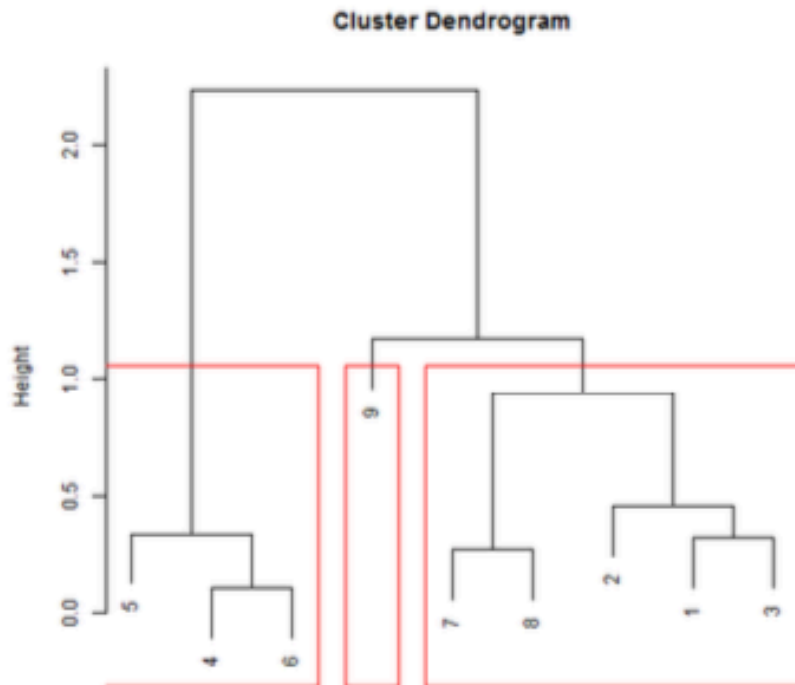
Average Linkage: the distance between two clusters is the *average* of all pairwise distances between any two objects

```
> round(dist(a, method='euclidean'),2)
```

	1	2	3	4	5	6	7	8
2	0.41							
3	0.32	0.50						
4	2.61	2.23	2.72					
5	2.67	2.32	2.81	0.29				
6	2.66	2.28	2.76	0.11	0.39			
7	1.20	0.79	1.25	1.49	1.62	1.52		
8	0.93	0.52	0.99	1.73	1.84	1.77	0.27	
9	1.41	1.13	1.20	2.35	2.55	2.34	1.07	1.05

Avg=0.95

Avg=1.06



Ward's Criterion: The distance between two clusters is the **sum** of all pairwise distances between any two objects normalized for cluster size

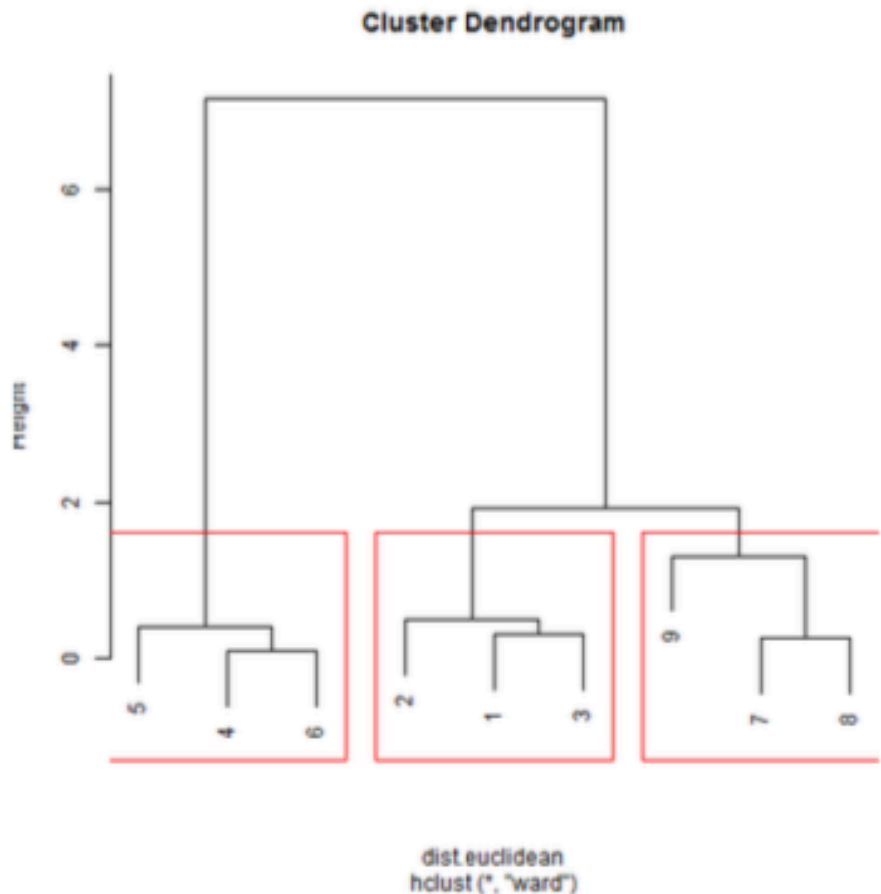
```
> round(dist(a, method='euclidean'),2)
```

	1	2	3	4	5	6	7	8
2	0.41							
3	0.32	0.50						
4	2.61	2.23	2.72					
5	2.67	2.32	2.81	0.29				
6	2.66	2.28	2.76	0.11	0.39			
7	1.20	0.79	1.25	1.49	1.62	1.52		
8	0.93	0.52	0.99	1.73	1.84	1.77	0.27	
9	1.41	1.13	1.20	2.35	2.55	2.34	1.07	1.05

Sum=5.68 Sum=2.12

```
# Ward's
ex1.hcw <- hclust(dist.euclidean, method = "ward")
plot(ex1.hcw)

# identify 3 clusters
ex1.hcw.3 <- rect.hclust(ex1.hcw, k = 3)
```

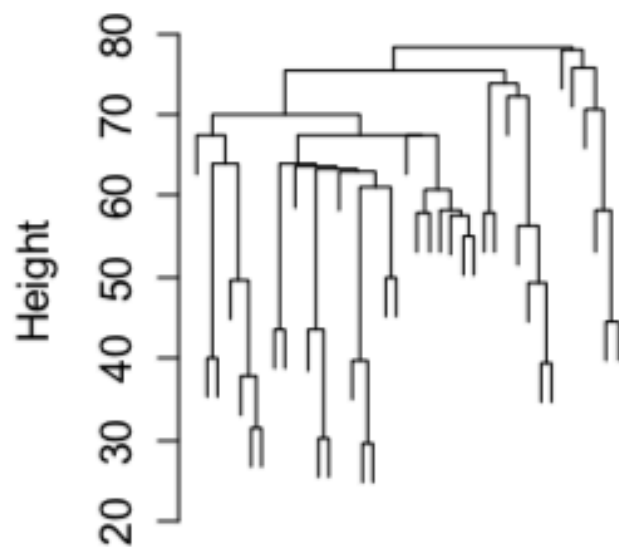


Note: in this trivial example, Ward's criterion gives the same result as complete linkage. In general, this is not true.

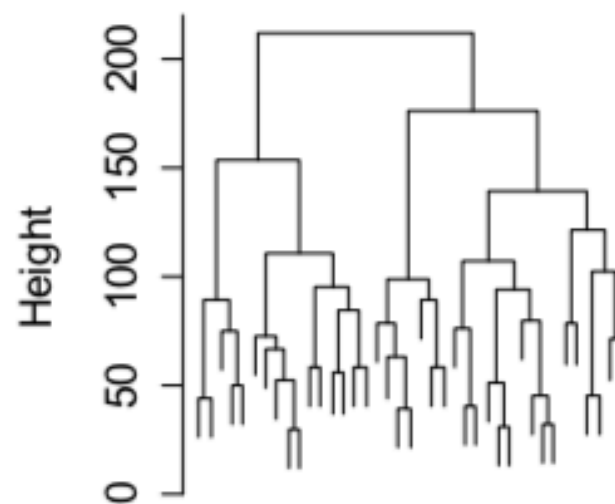
Photoreceptor data

- Data: each gene has been measured on 3 or 4 biological replicates of WT and KO mice, for each of 5 time points (~30K genes and 39 samples)
- Objects: 39 mice samples
- Attributes: ~30K genes

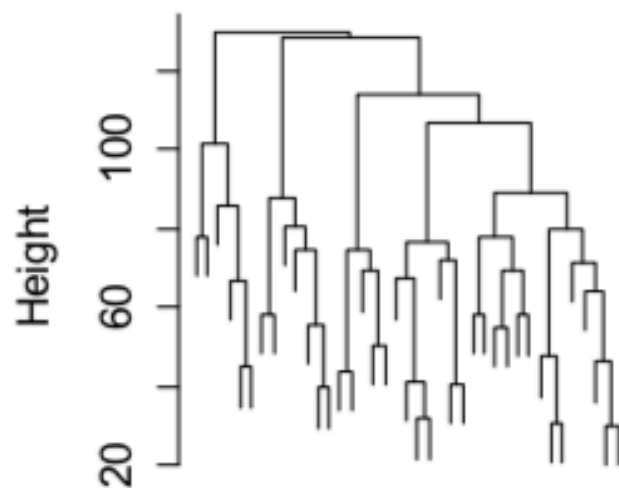
Single



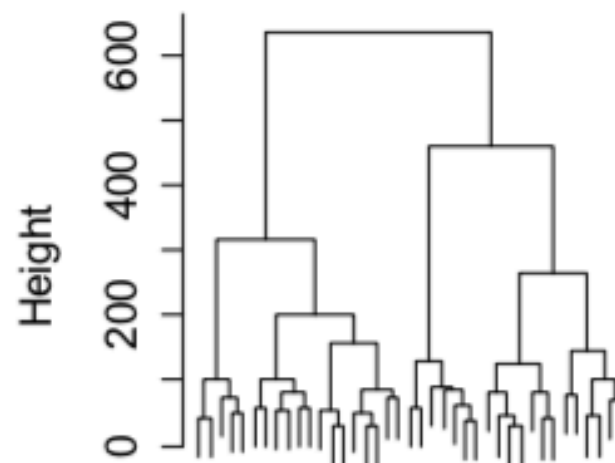
Complete



Average



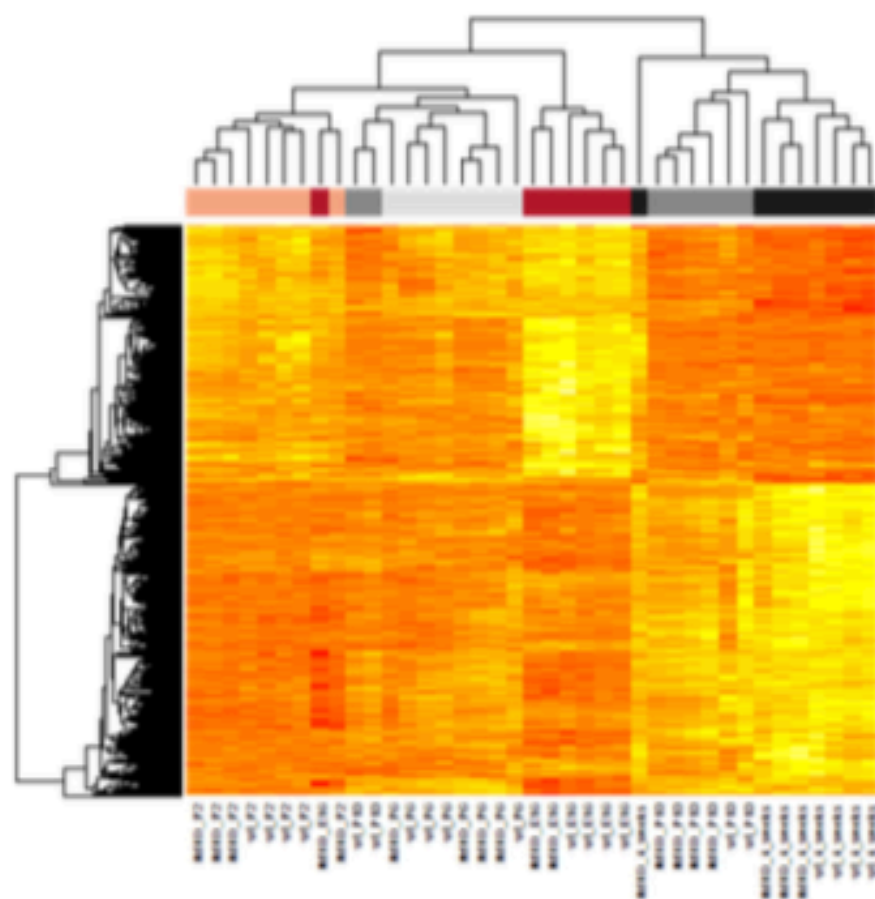
Ward



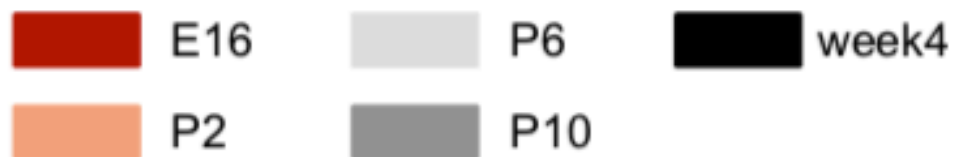
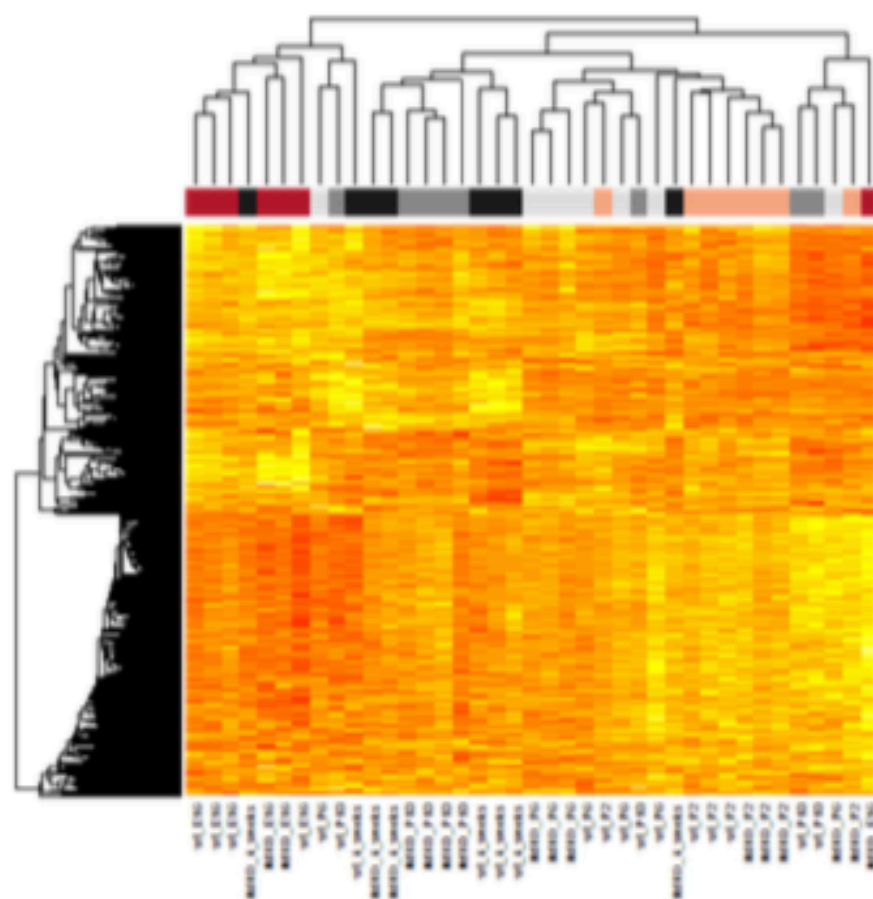
Photoreceptor data

- Data: each gene has been measured on 3 or 4 biological replicates of WT and KO mice, for each of 5 time points (~30K genes and 39 samples)
- Attributes: 39 mice samples
- Objects: ~30K genes

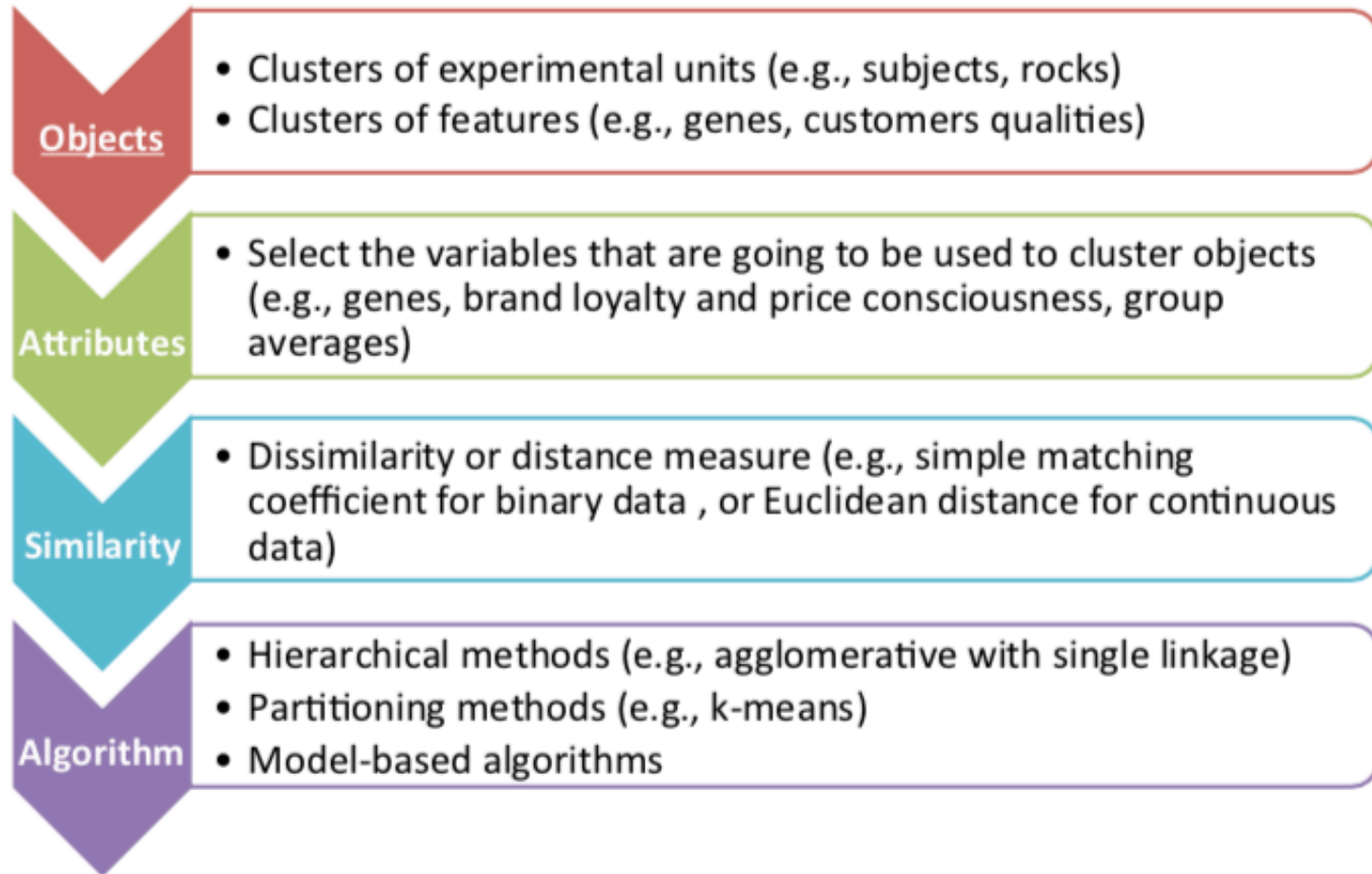
Top-972 from LIMMA



972 randomly selected



Key elements



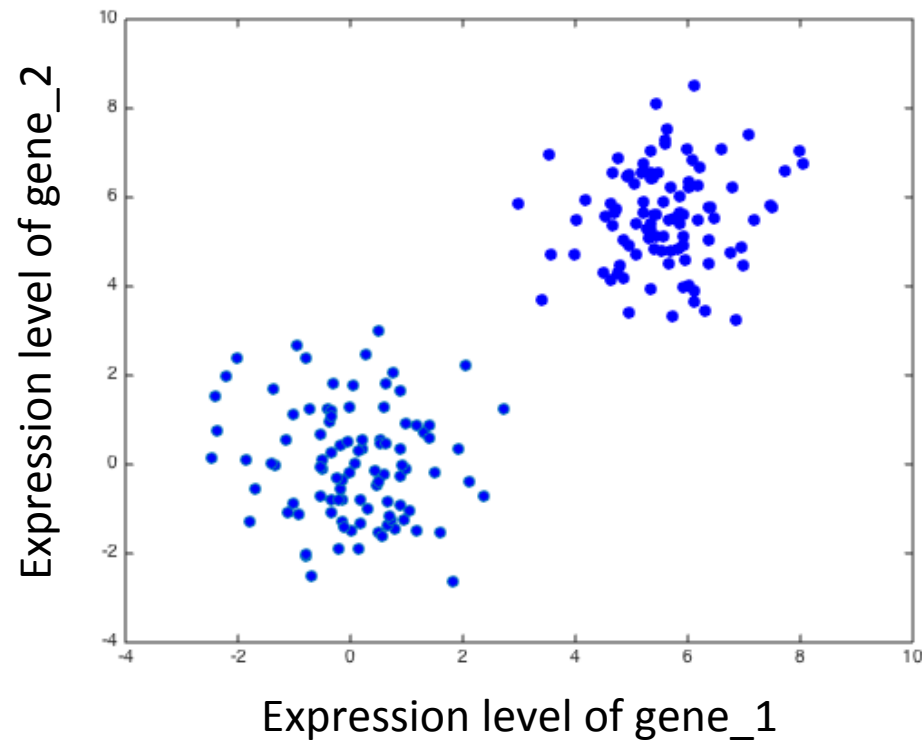
Number of clusters???

Algorithm: partitioning

- These algorithms partition the objects into K groups
- Often motivated by an objective function that attains an extreme for the “correct” or “best” partition of the objects
- K needs to be set *a priori*
- Most typical algorithms: k-means and partitioning around the medoids (PAM)

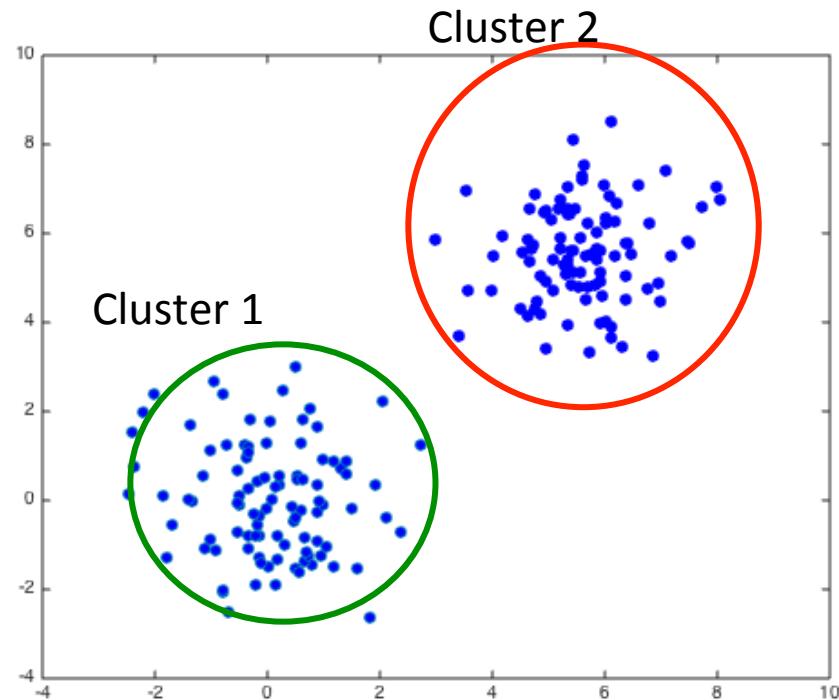
How many clusters are there?

Suppose you measured expression levels for 2 genes (gene A and gene B) for 200 individuals



How many clusters are there?

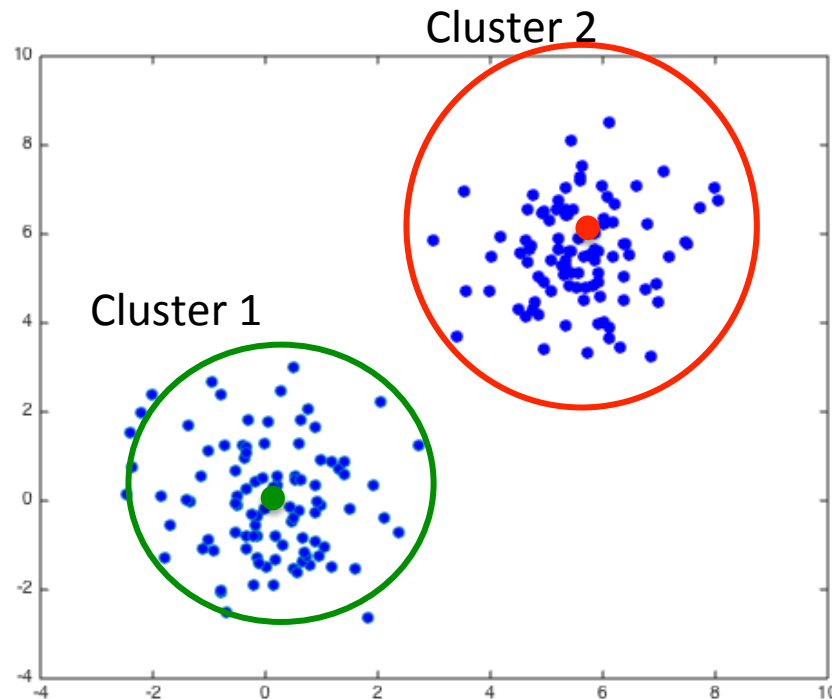
Suppose you measured expression levels for 2 genes (gene A and gene B) for 100 individuals



K-means objective function

Objective function: minimize the average squared **Euclidean distance** of objects from their assigned cluster centers. A **cluster center** (or centroid) is defined as the mean of objects in the given cluster.

Computing the “mean/centroid” for each cluster:

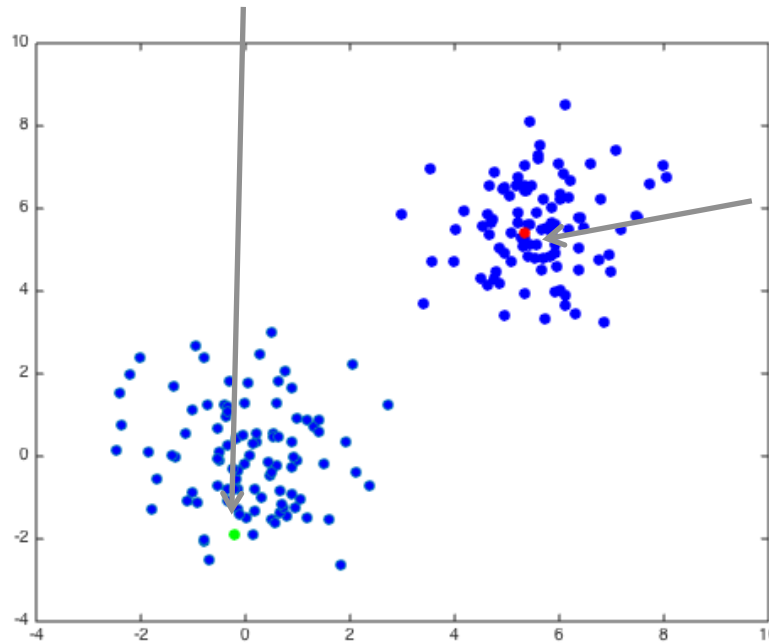


Let's go through the k-means algorithm first

Algorithm: iterative procedure

1) Pick k random points as initial cluster centers

Randomly selected point 1



Randomly selected point 2

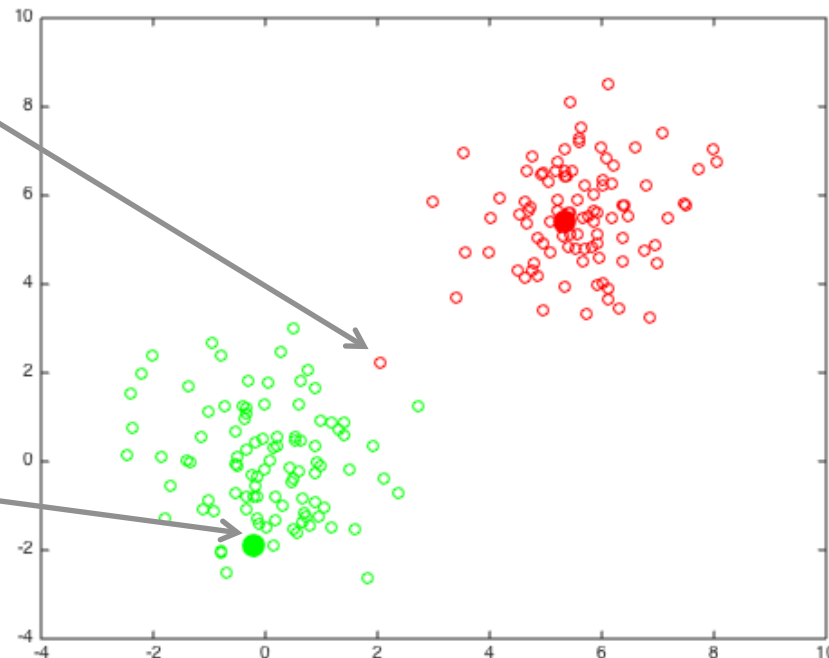
Let's go through the k-means algorithm first

Algorithm: iterative procedure

- 1) Pick k random points as initial cluster centers
- 2) Measure distance between all points and the cluster centers – assign points to nearest cluster

Doesn't look right

"Initialization point"

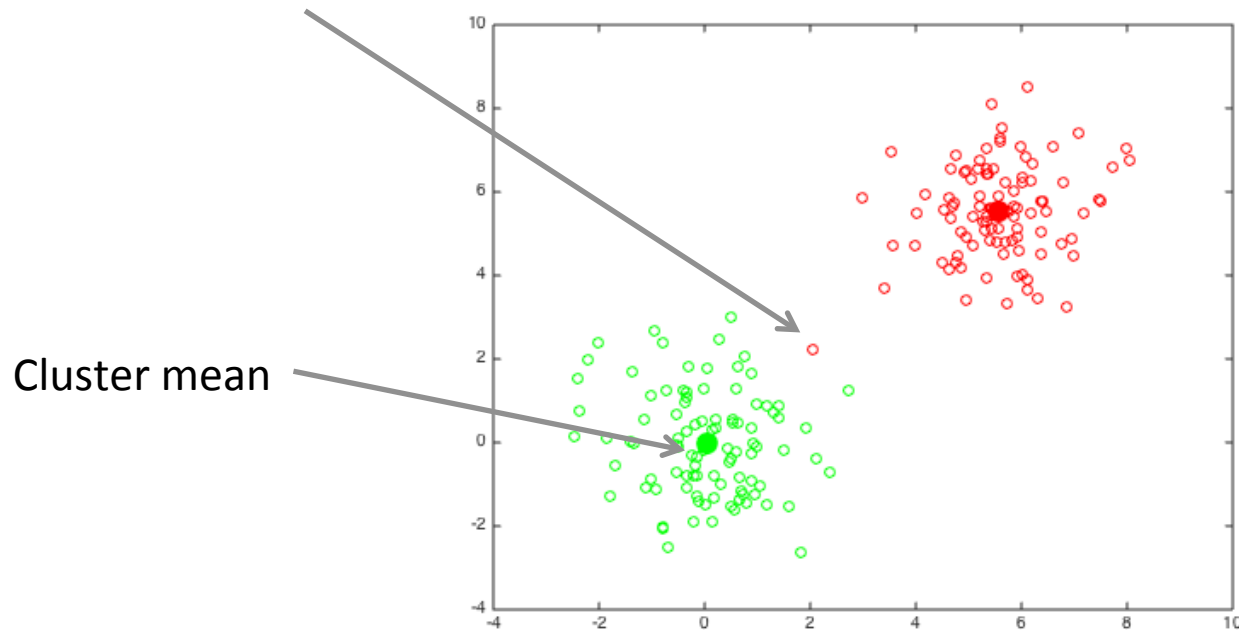


Let's go through the k-means algorithm first

Algorithm: iterative procedure

- 1) Pick k random points as initial cluster centers
- 2) Measure distance between all points and the cluster centers – assign points to nearest cluster
- 3) Compute cluster means

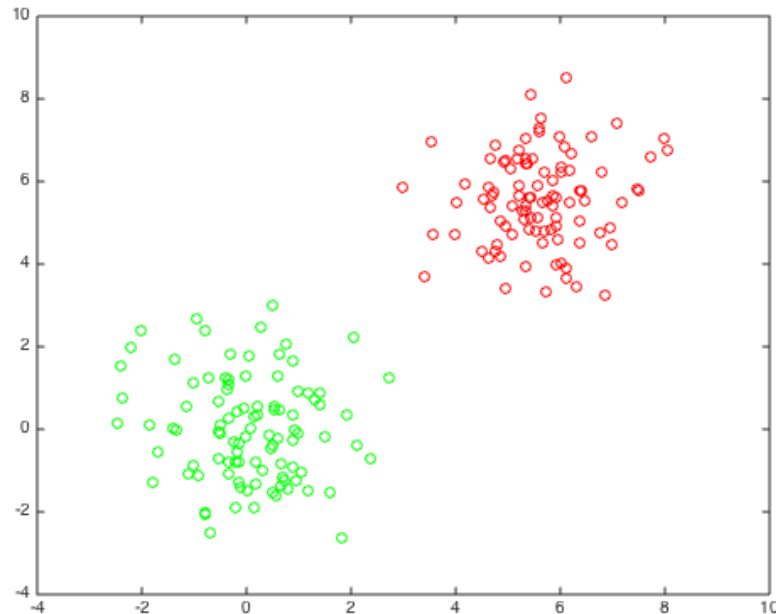
Doesn't look right



Let's go through the k-means algorithm first

Algorithm: iterative procedure

- 1) Pick k random points as initial cluster centers
- 2) Measure distance between all points and the cluster centers – assign points to nearest cluster
- 3) Compute cluster means
- 4) Reassign points to cluster based on distance
 - If not change from previous assignment, stop, else to go step 3



K-means objective function (formula/equation)

Objective function: minimize the average squared **Euclidean distance** of objects from their assigned cluster centers. A **cluster center** (or centroid) is defined as the mean of objects in the given cluster.

* N objects, each have p attributes: $\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_n\}$

* k-means objective function:

$$J = \sum_{i=1}^n \sum_{k=1, i \in k}^K \underbrace{\|\vec{x}_i - \vec{\mu}_k\|}_{\text{Euclidian distance between } x_i \text{ and } u_k}$$

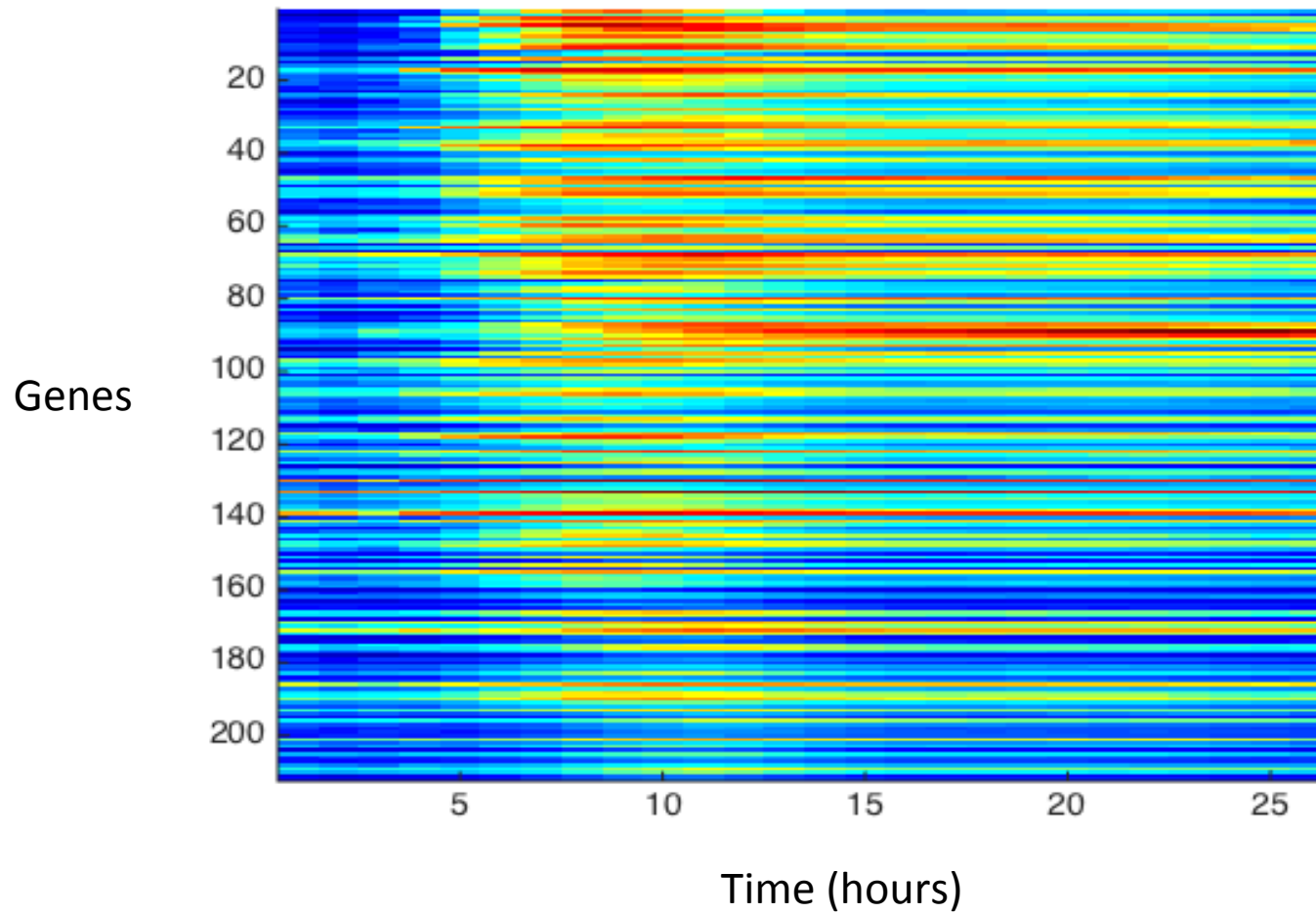
Cluster center for cluster k

Timing patterns for IFN induced genes in CD19⁺ Bcells



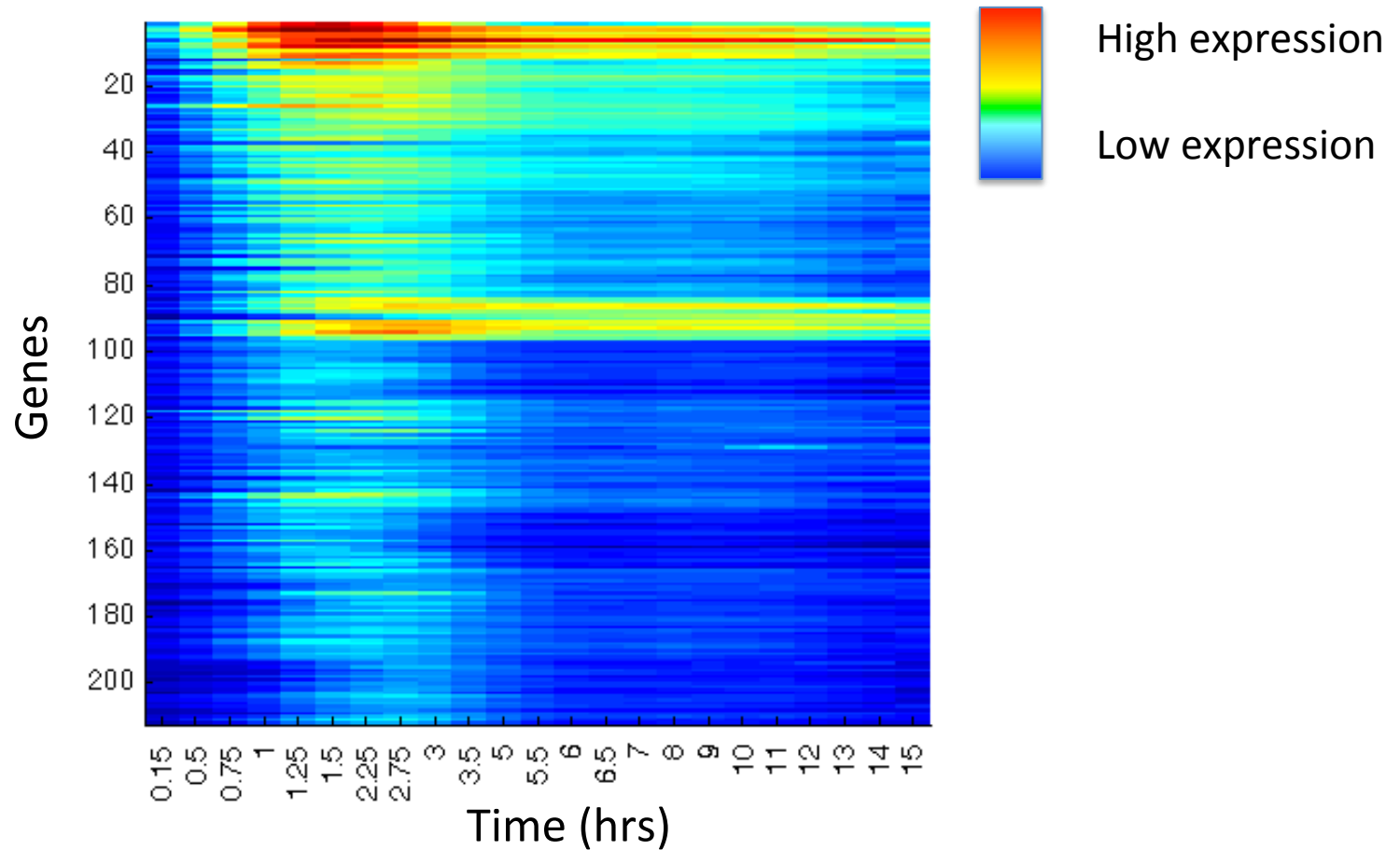
High expression

Low expression



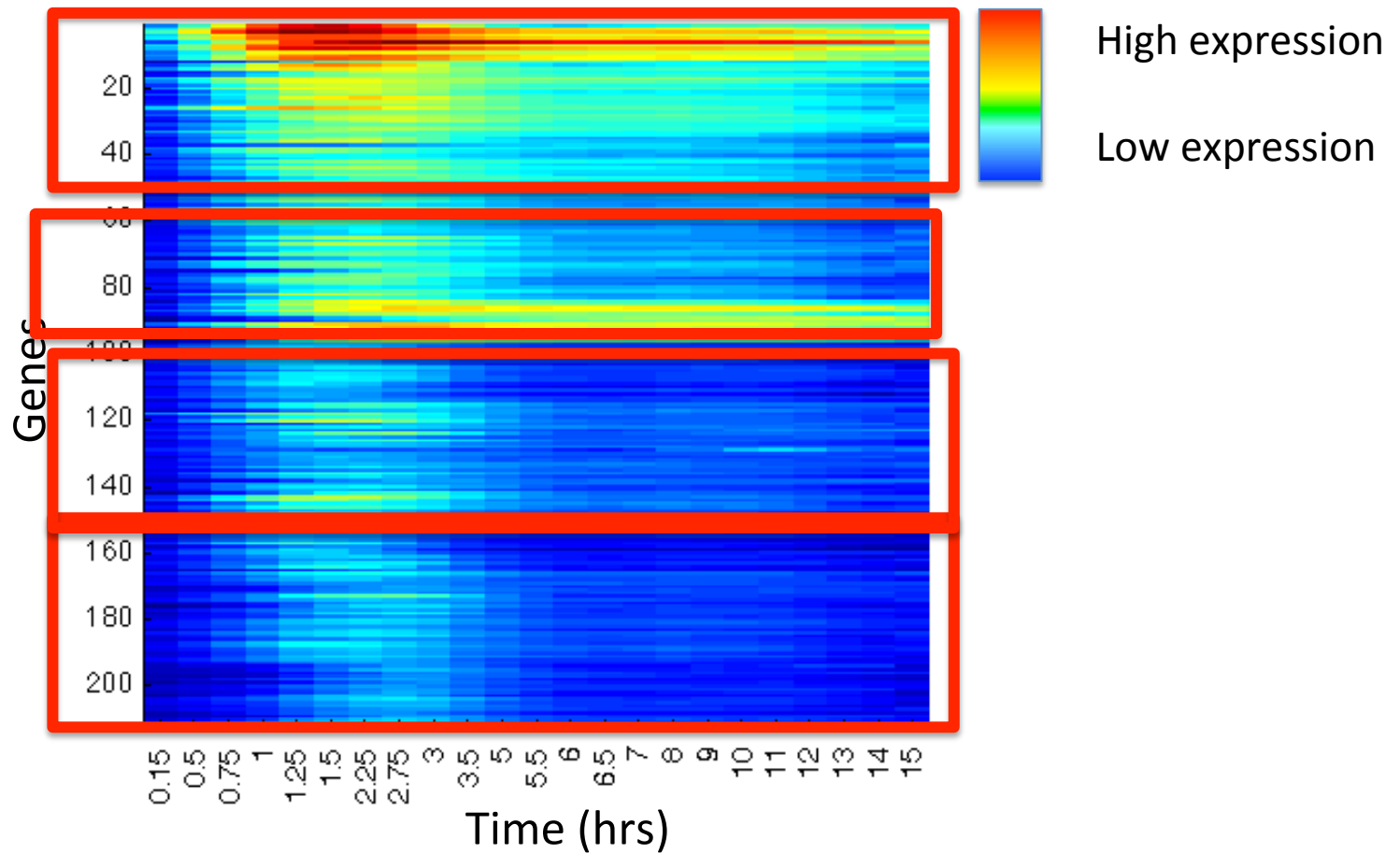
Timing patterns for IFN induced genes in CD19⁺ Bcells

Application of k-means clustering with k=4



Timing patterns for IFN induced genes in CD19⁺ Bcells

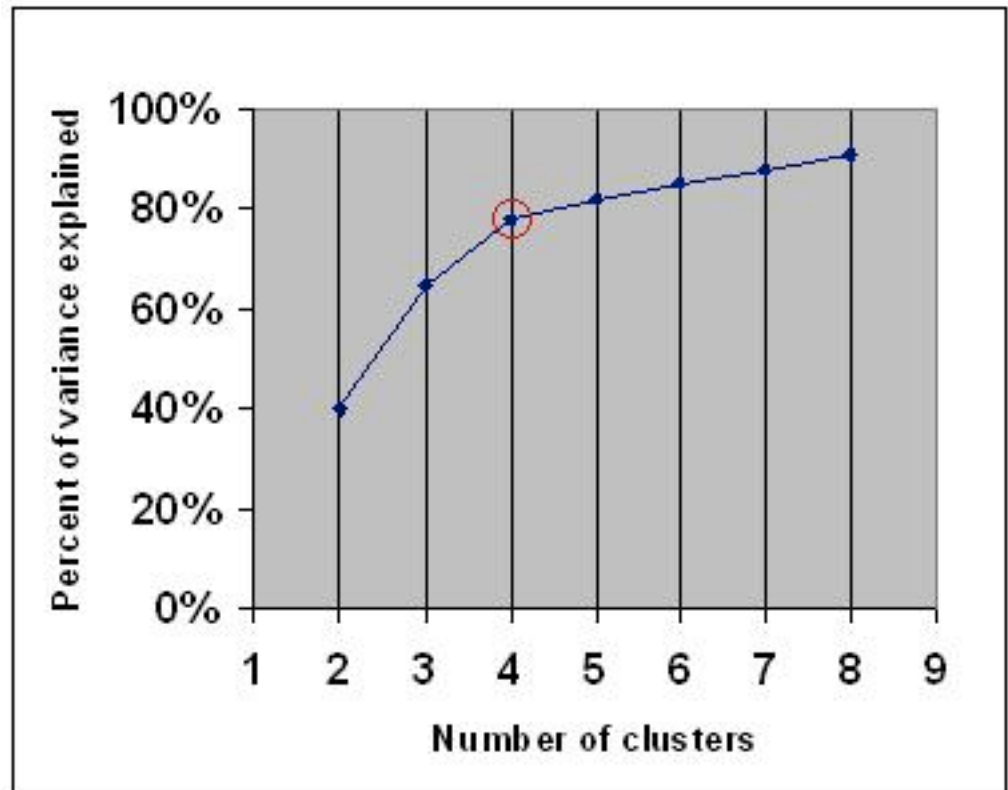
Application of k-means clustering with k=4



How do you determine k (number of clusters)?

Note: maximizing the clustering likelihood/objective will not be informative → each object should be in its own cluster. Therefore, need an algorithm that takes into account the “cost” of additional clusters.

- Prior knowledge
- The “elbow method”



How do you determine k (number of clusters)?

Note: maximizing the clustering likelihood/objective will not be informative → each object should be in its own cluster. Therefore, need an algorithm that takes into account the “cost” of additional clusters.

- Prior knowledge
- The “elbow method”
- Information Criteria Approach: AIC or BIC
- Silhouette method
- The Gap Statistics
- Cross-validation

Summary & conclusions

- Many choices to make when you want to cluster a set of objects:
 - Objective, algorithm, **attributes/features**, distance metric, number of clusters.
- Not possible to say which method is the best. It all depends on data and goal.
- Clustering is very powerful, but reckless application leads to misguided conclusions.
- CA is still a good way to explore the data and summarize results.

