

Clustering

Marcel Reinders

TU Delft

Delft Bioinformatics Lab

Faculty of Electrical Engineering, Computer Science and Mathematics

Delft University of Technology



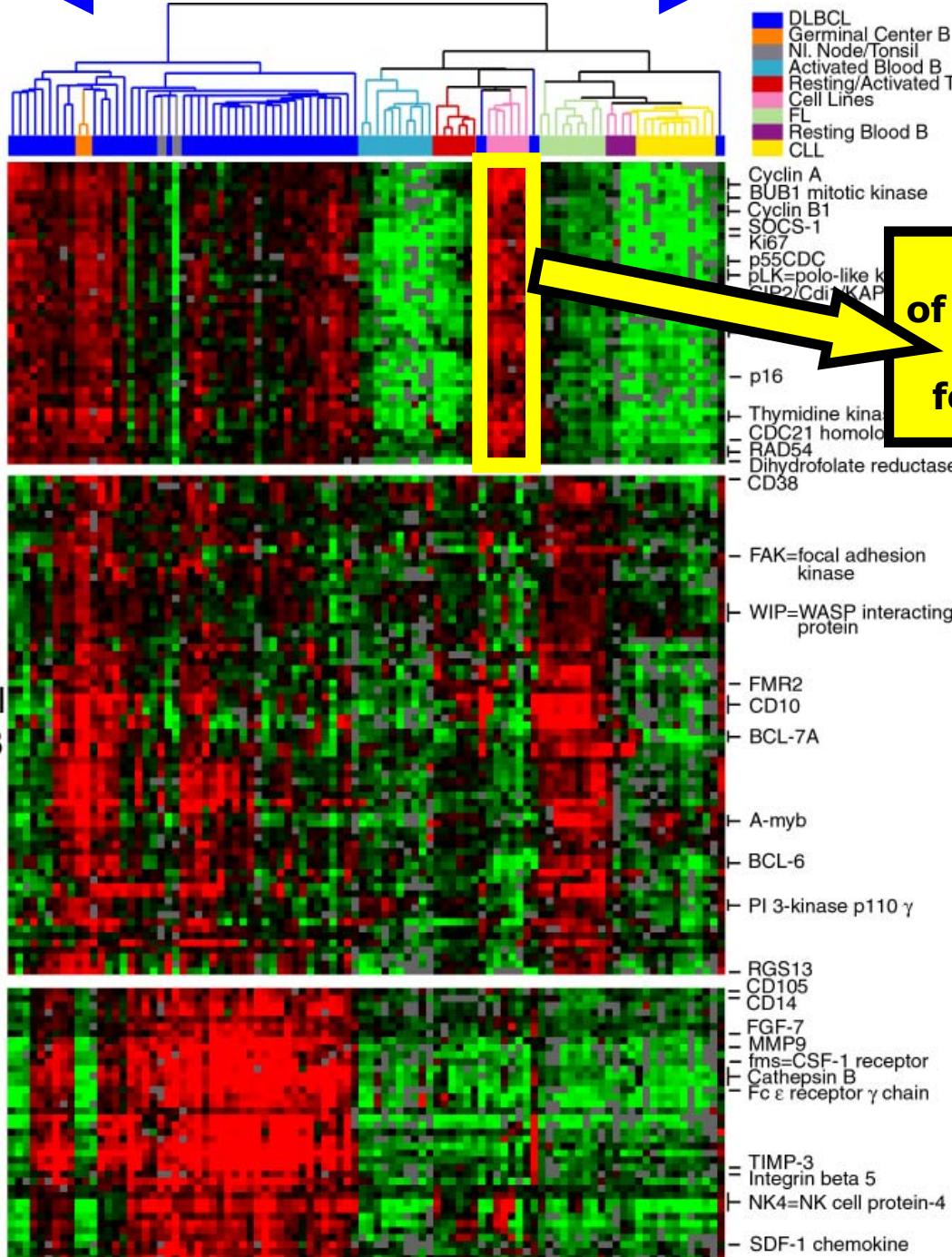
ordered on
similarity

related tumors

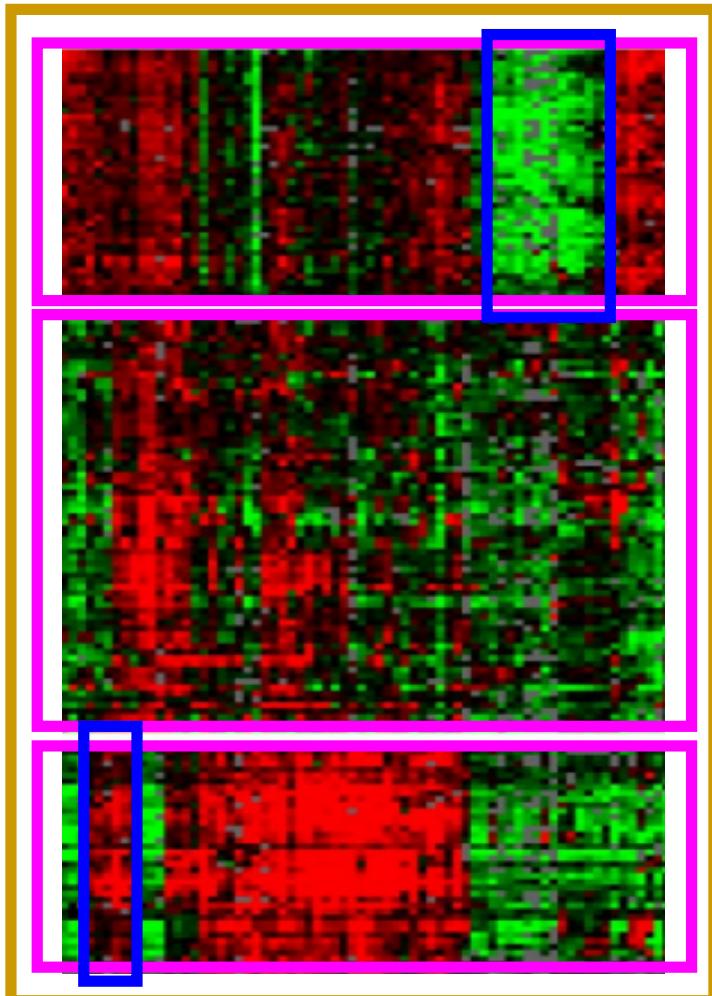
Proliferation

related
genes

Lymph
Node

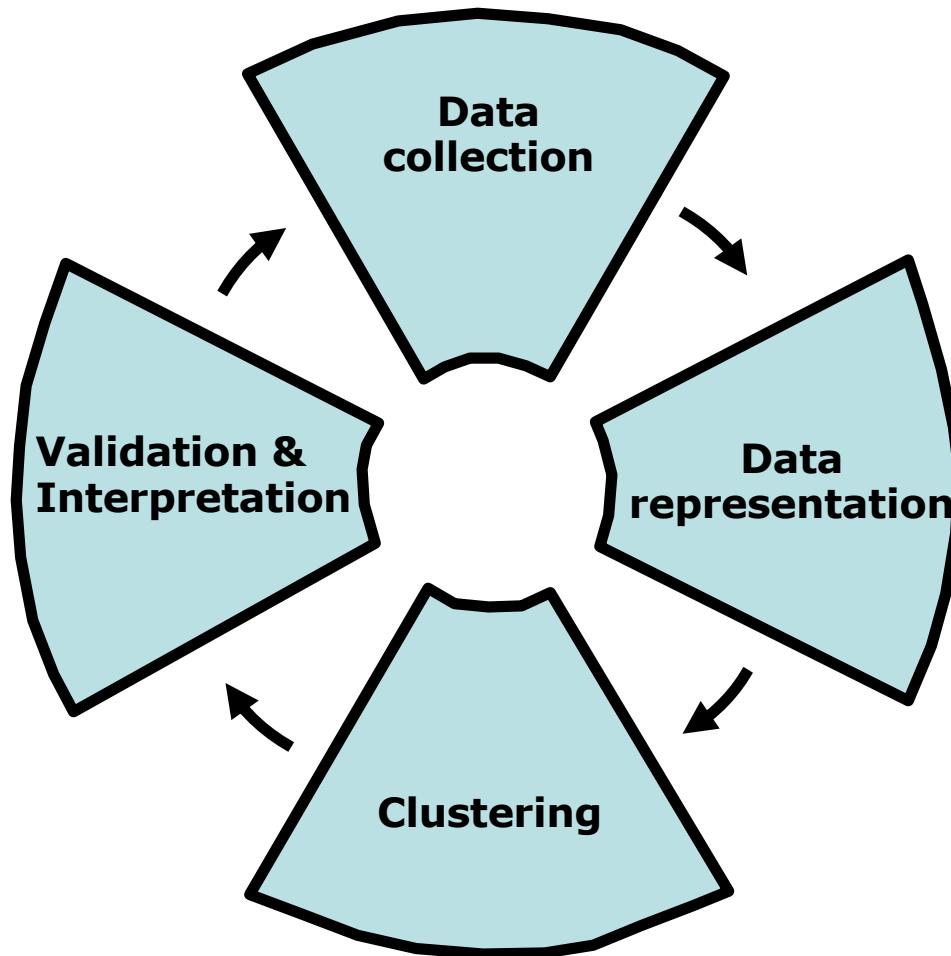


Cluster analysis

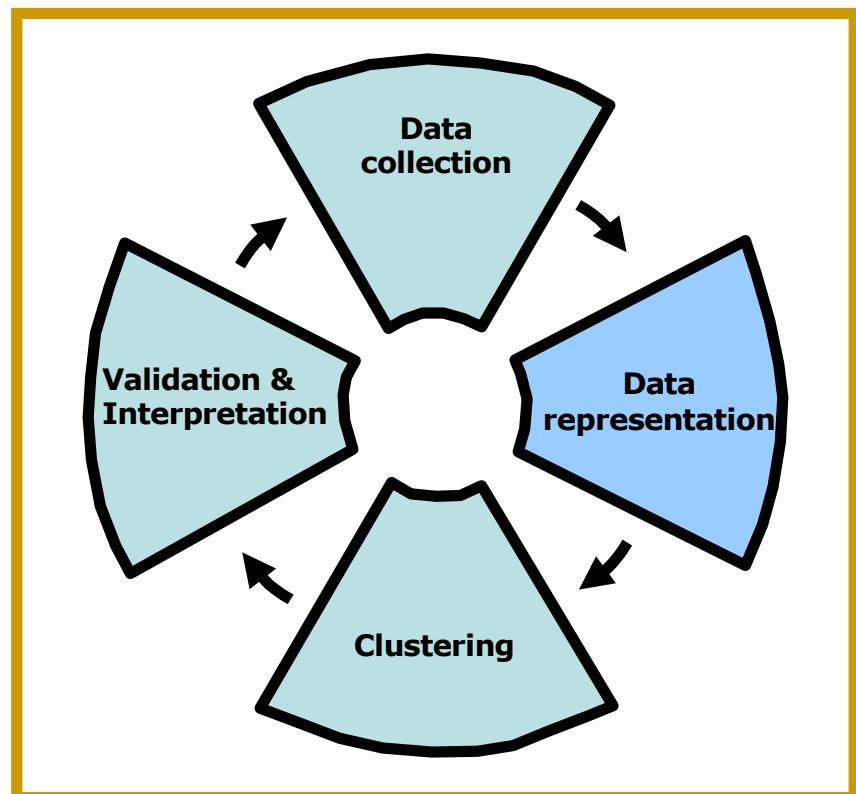


- **Group similar profiles**
- **Finding structure in the data**
- **Requisites:**
 - Measure of similarity
 - Grouping method
- **Subjective measures:**
 - Validation
 - Clustering is a process

Cluster analysis: The process



Data representation



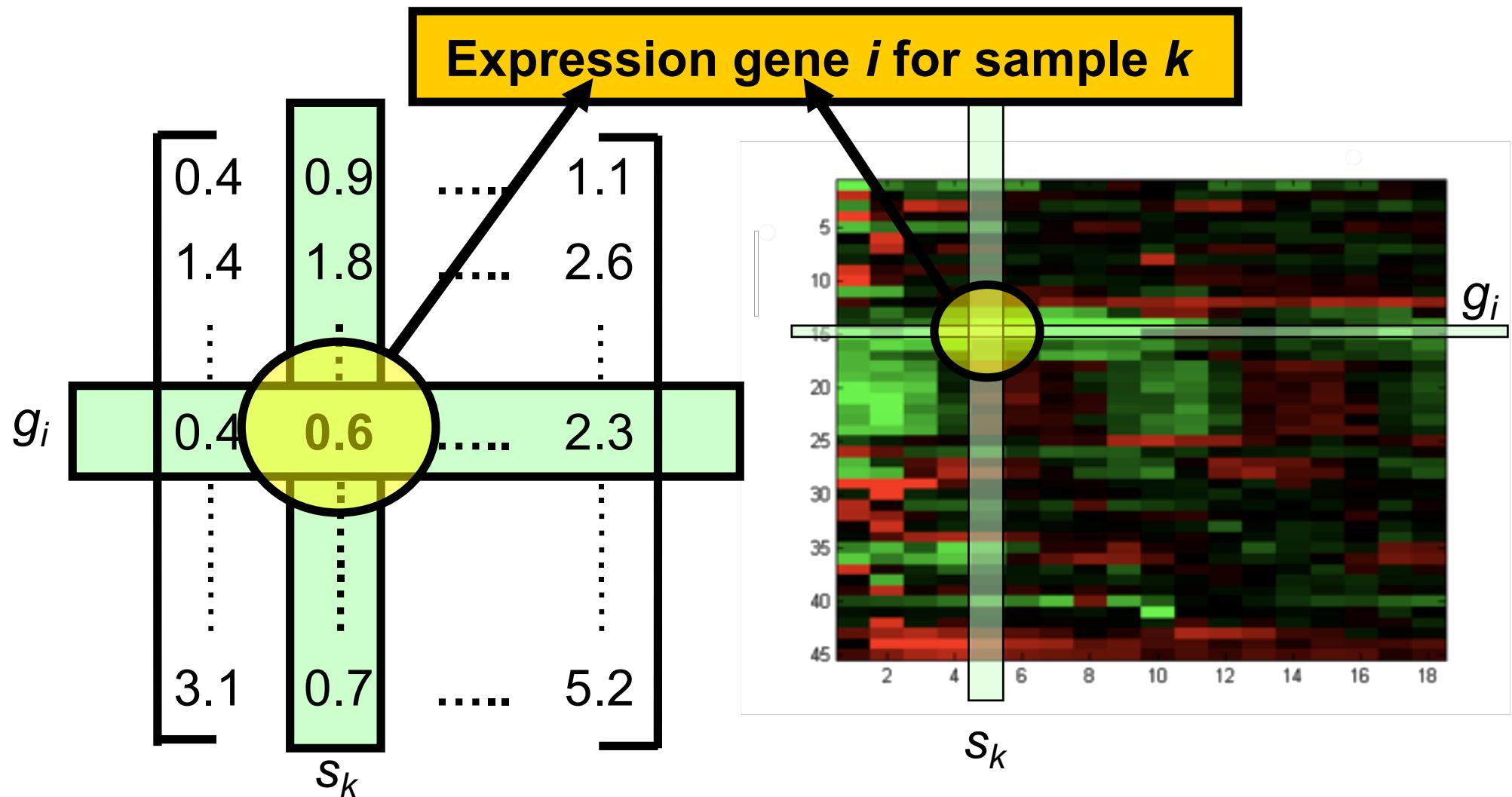
TOPICS

- **Data representation**
- **Data spaces:**
 - Sample-space
 - Gene-space
- **Clustering**

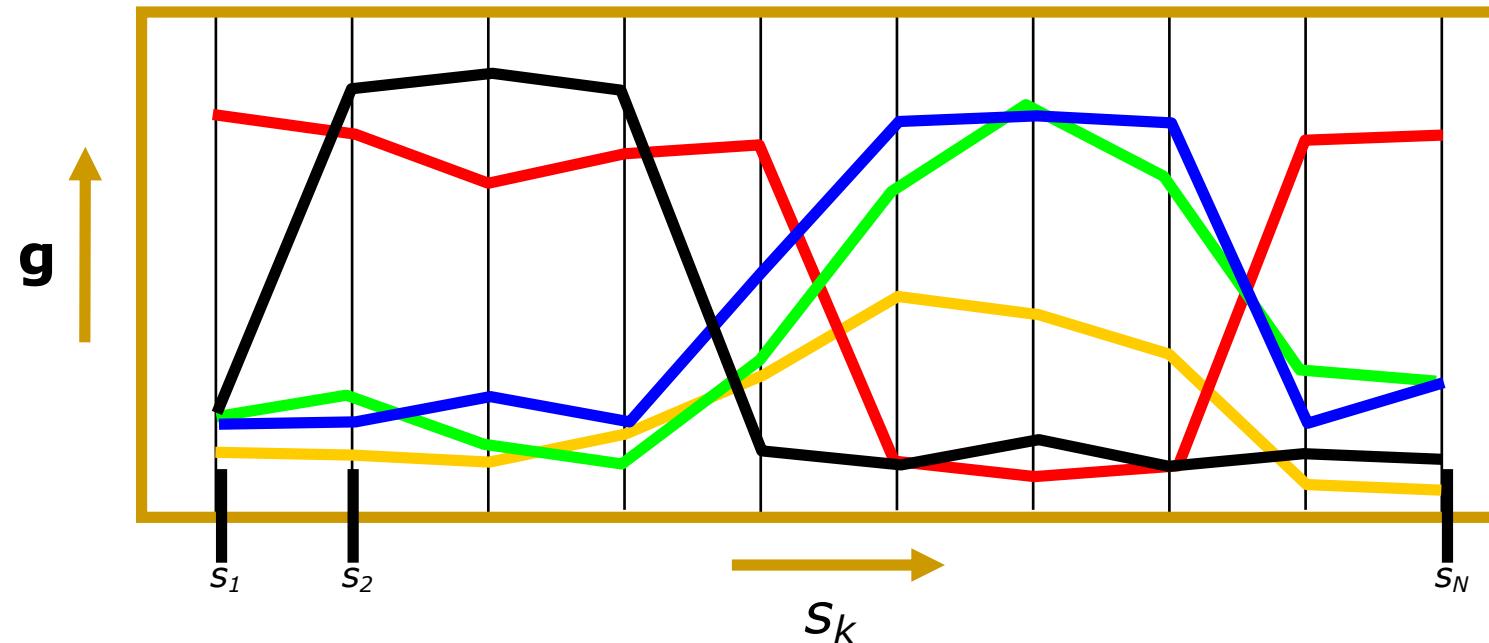
Data representation, one sample

data		
	0.4	
	1.4	gene 2: g_2
	:	
	0.4	gene i: g_i
	:	
	3.1	
vector		

Data representation, multiple samples



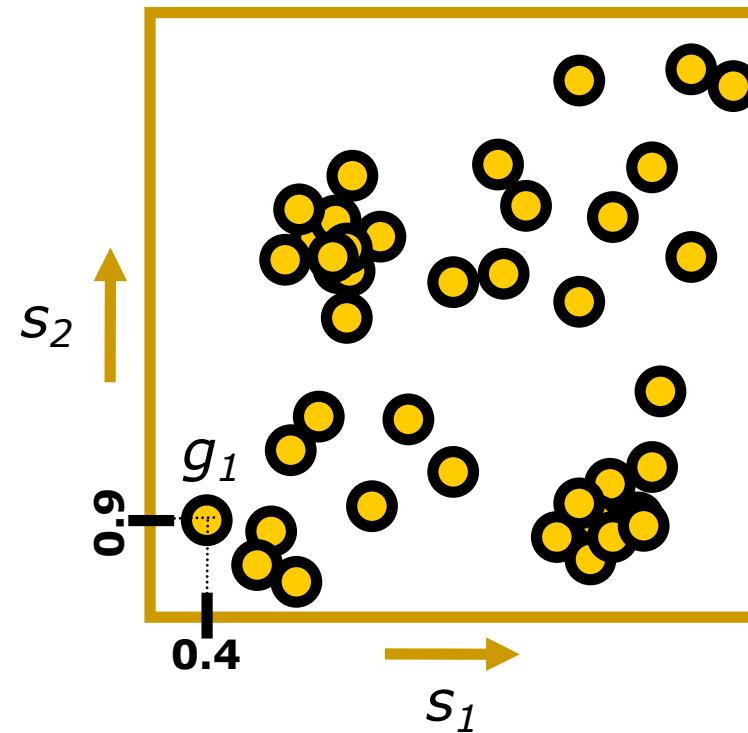
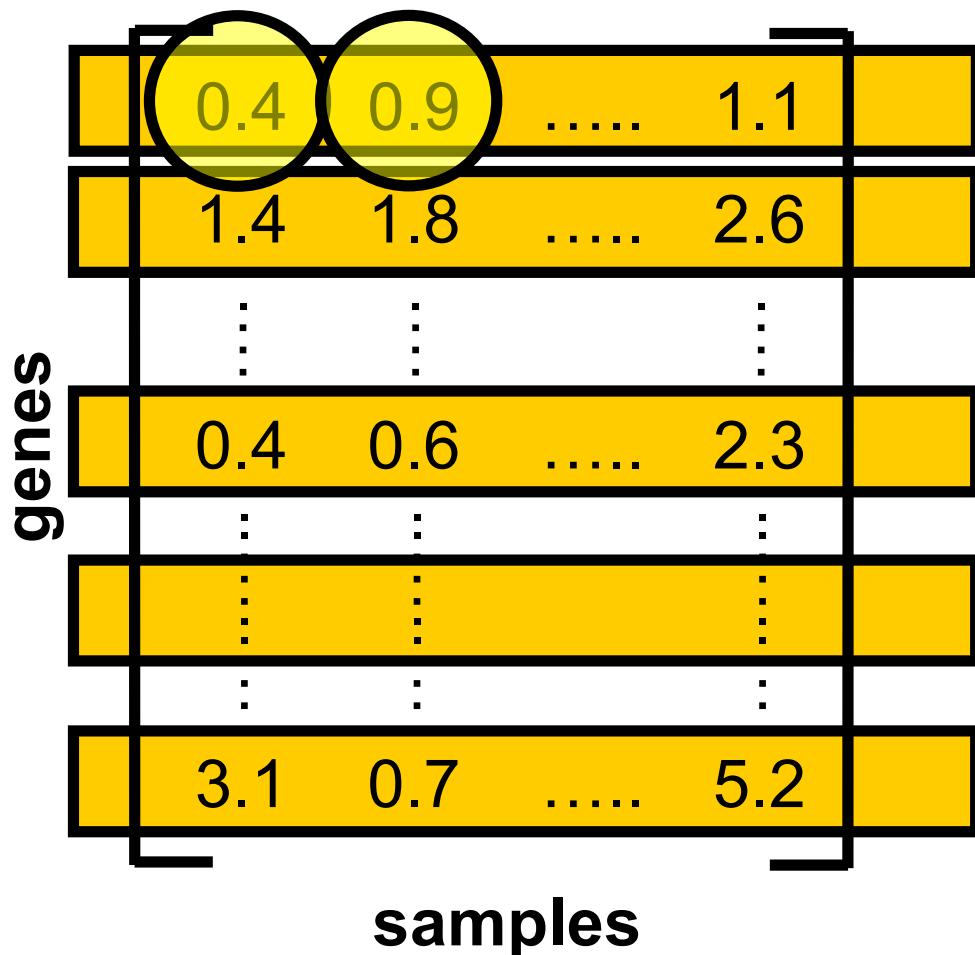
View as profiles



Profile for each gene

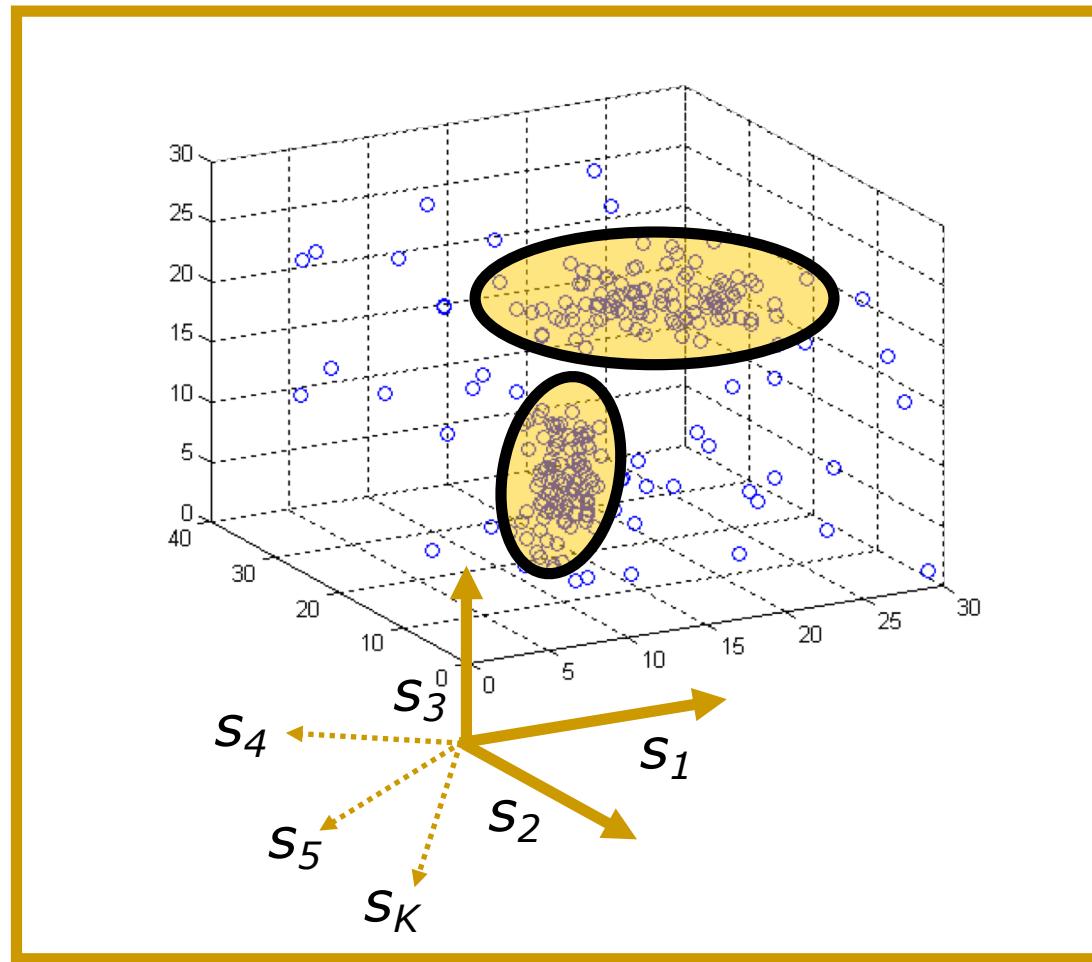
Represents gene activity across all samples

Sample-space (1/2)



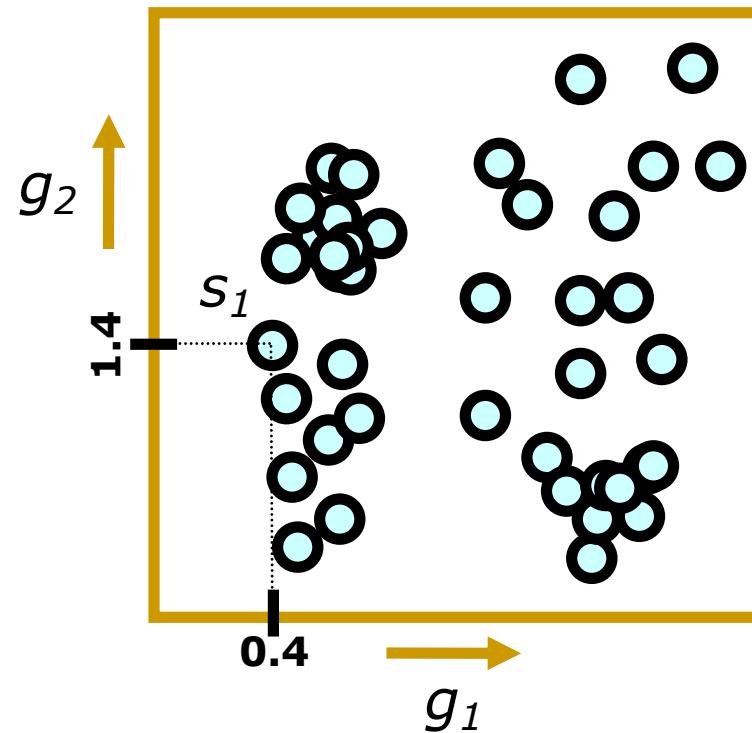
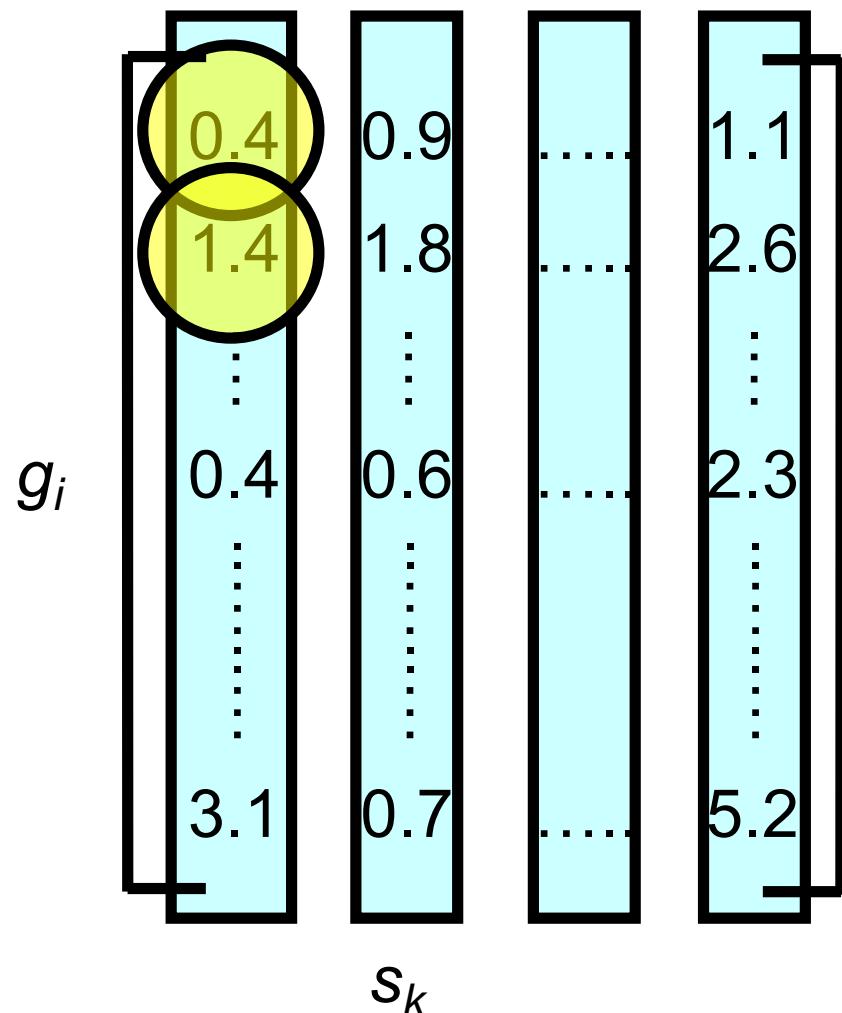
Sample-space
Genes are data points

Sample-space (2/2)



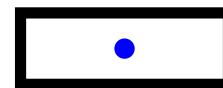
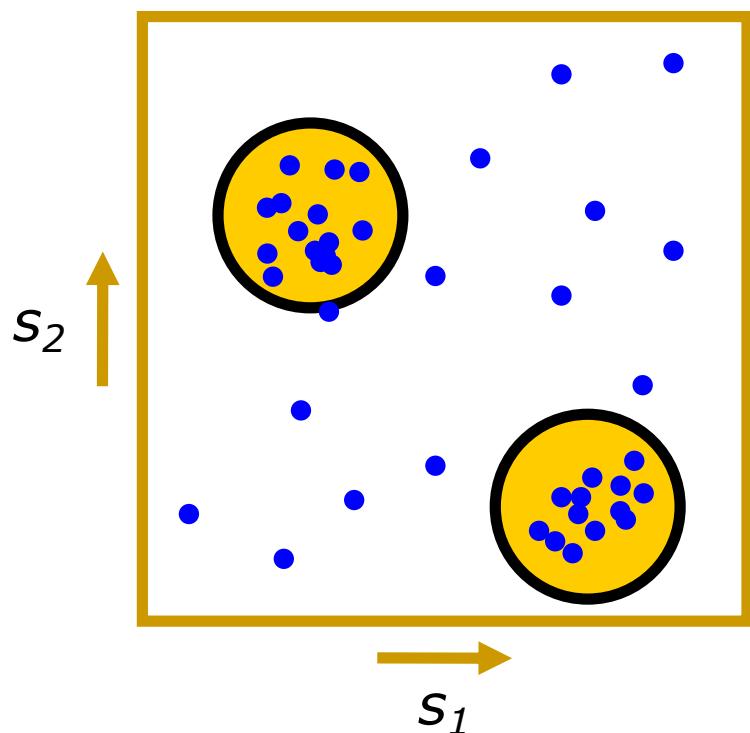
Note!
Sample-space is high dimensional

Gene-space



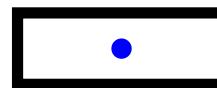
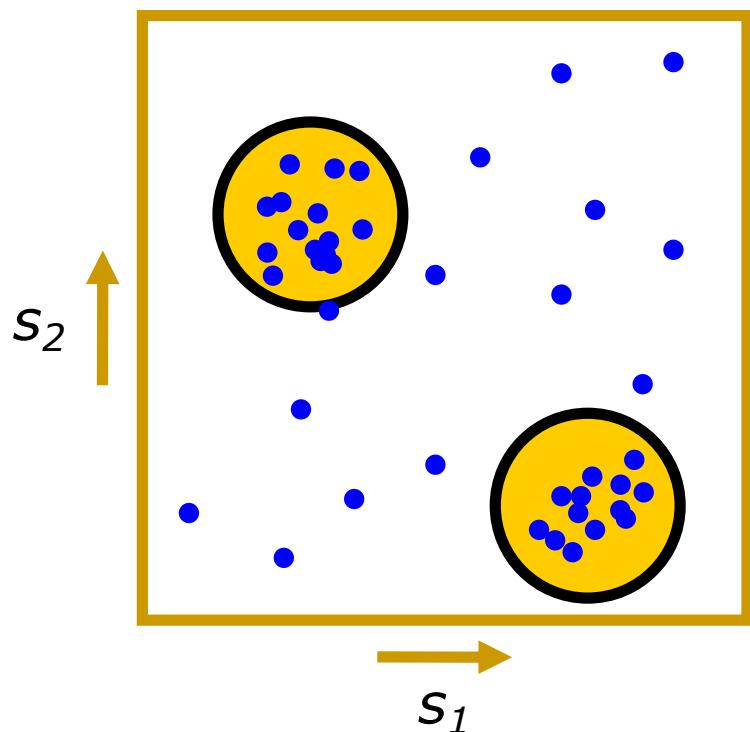
Gene-space
Samples are data points

Genes in *sample-space*



Gene g_i : activity level in Samples s_1 and s_2
(each point represents one gene)

Genes in *experiment-space*



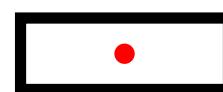
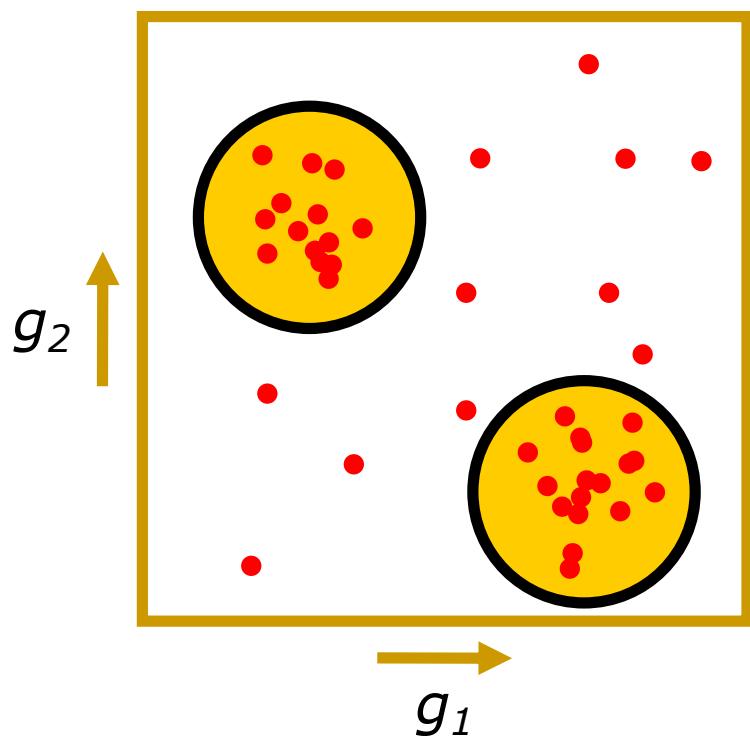
Gene g_i : activity level in samples s_1 and s_2
(each point represents one gene)



Group of genes that have the same activity levels across “all” samples

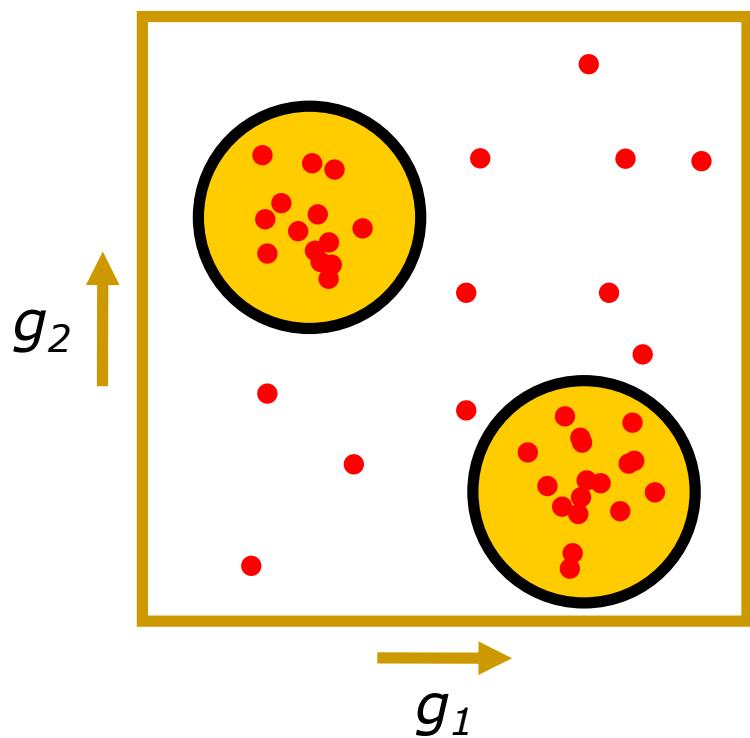
Cluster: Genes functionally related
Characterization of unknown genes
Hypothesis testing!

Samples in *gene-space*



Sample s_k : activity of genes g_1 and g_2
(each point represents one sample)

Samples in *gene-space*



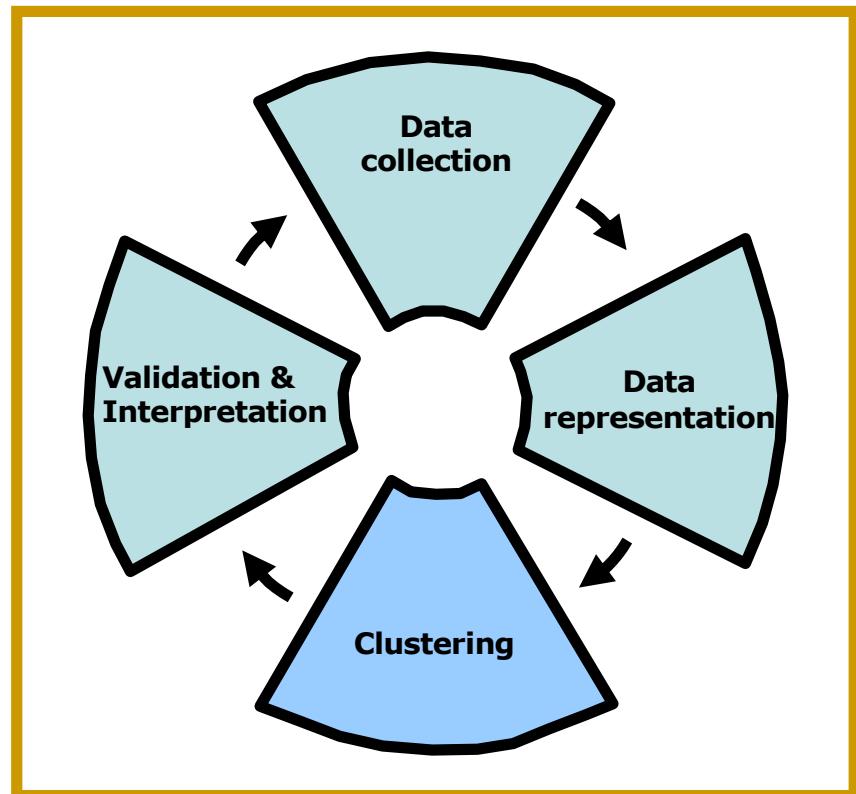
Sample s_k : activity of genes g_1 and g_2
(each point represents one sample)



Group of samples in which genes have the same activity levels across samples

**Cluster: Genetic profile for related samples
Characterization of related diseases**

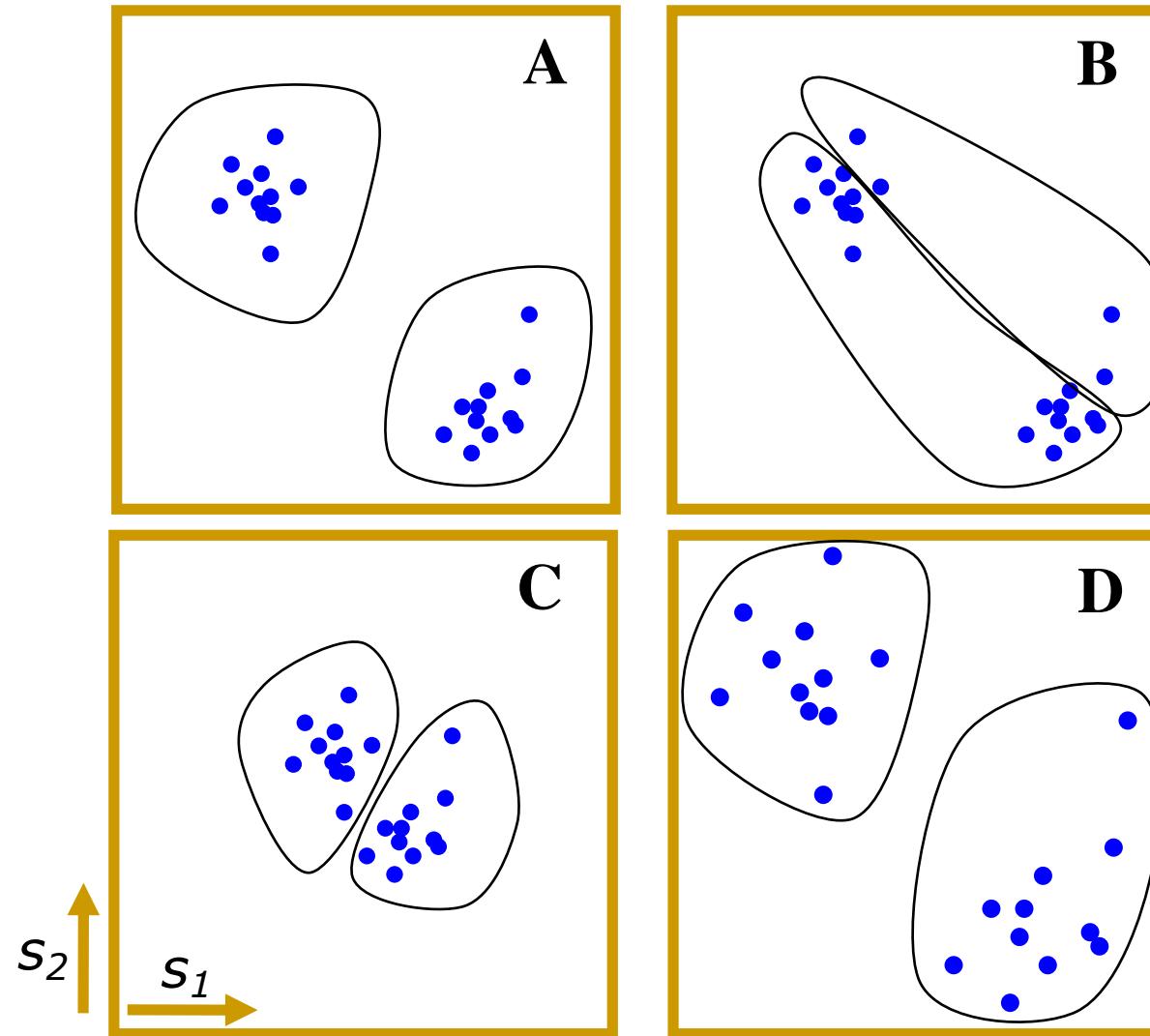
Clustering methods



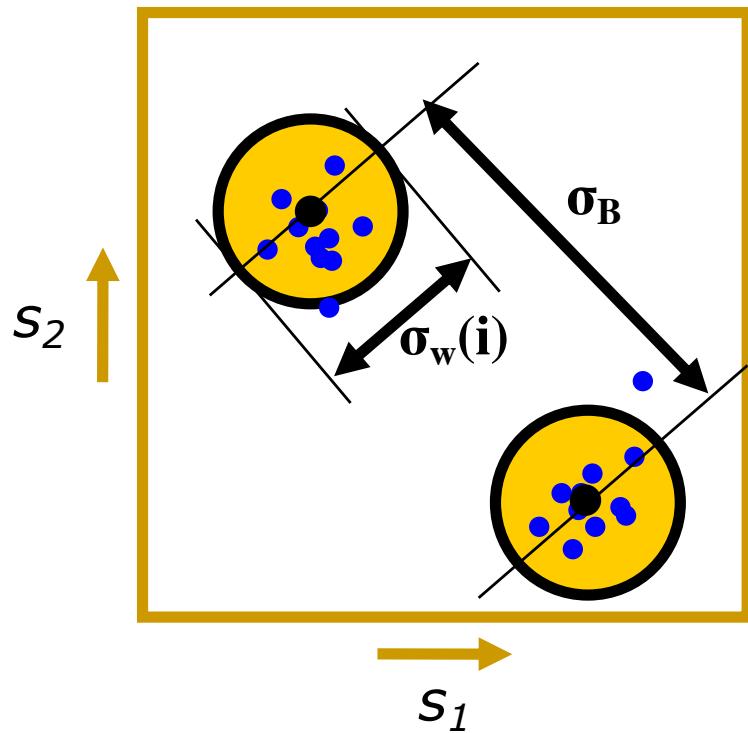
TOPICS

- **Finding structure**
- **Hierarchical clustering**
- **Similarity & Linkage**
- **K-means clustering**
- **Graph-based clustering**

Finding structure: better grouping?

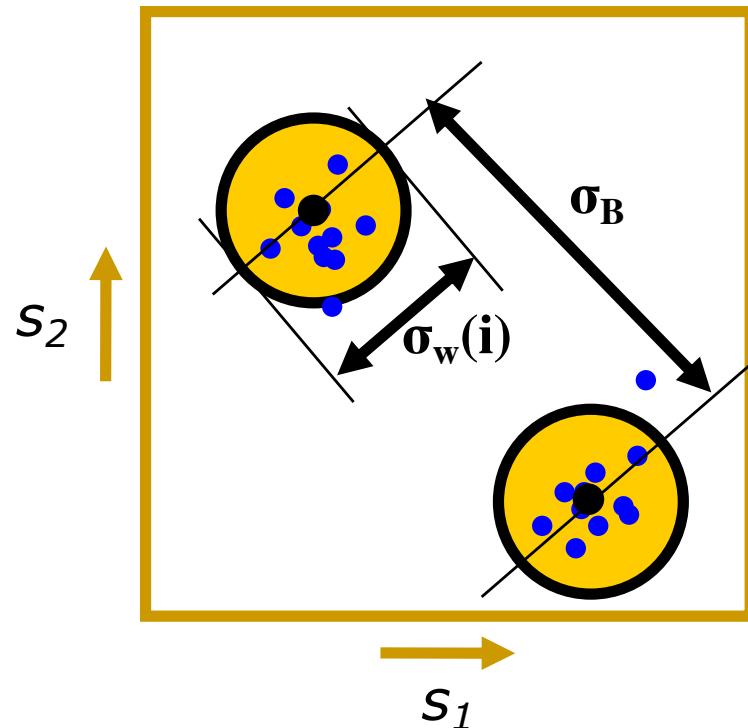


Finding structure: better grouping?



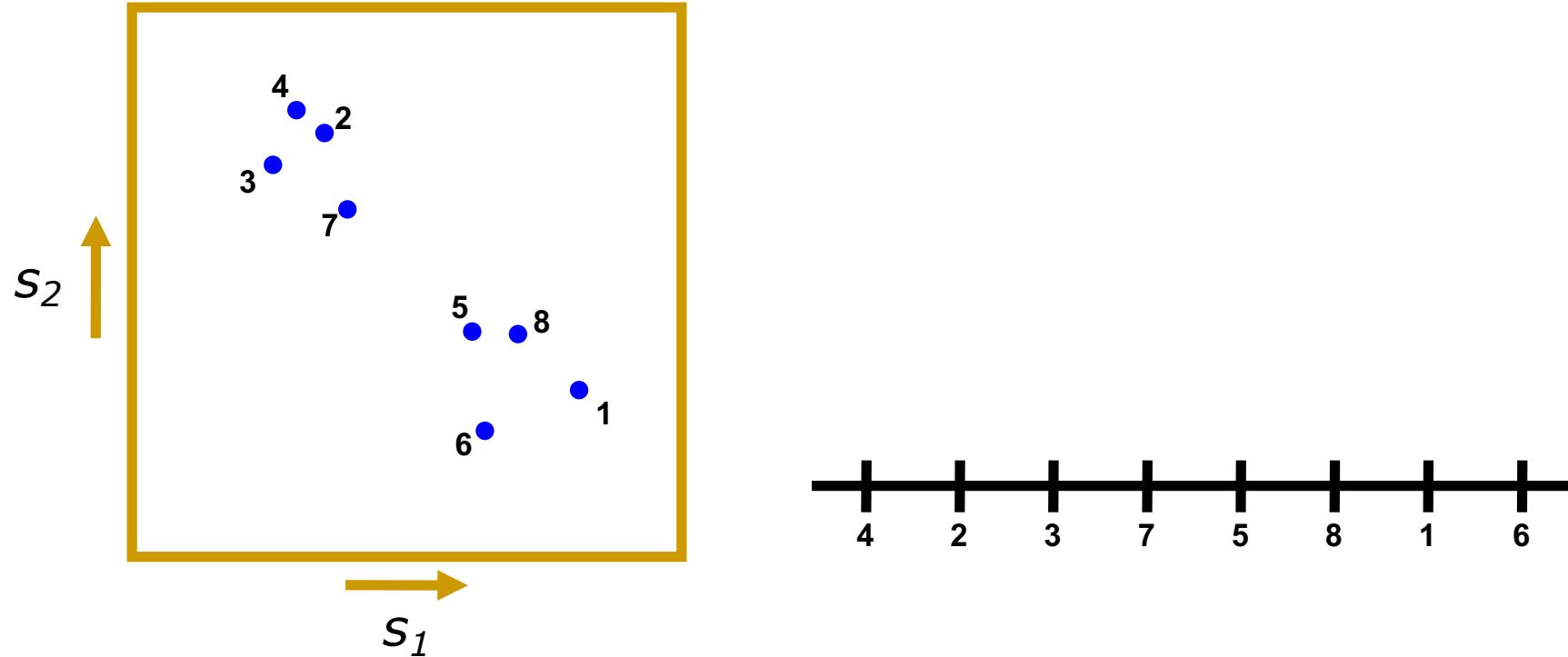
- **Structure when:**
 - 1) Points within cluster resemble each other (*within variance, $\sigma_w(i)$*)
 - 2) Clusters deviate from each other (*between variance, σ_B*)
- **Group points such that**
$$\text{MIN} \left[\frac{\sum \text{within variance}}{\text{between variance}} \right] \rightarrow \begin{array}{l} \sigma_w: \text{small \&} \\ \sigma_B: \text{large} \end{array}$$

General approaches



- **Agglomerative
(building trees)**
hierarchical clustering
(Mike Eisen's Cluster)
- **Partitional
(finding prototypes)**
**k-means, graph-based
clustering, som-mapping**
gene shaving

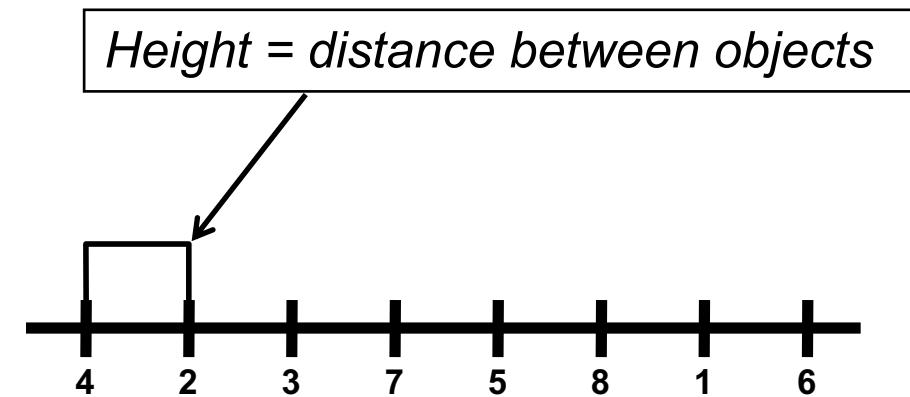
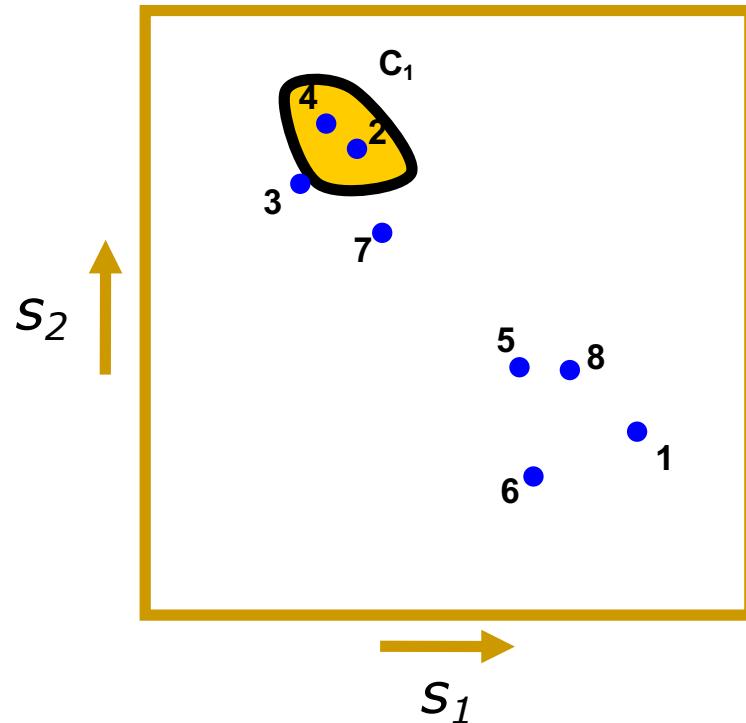
Hierarchical clustering



Find most similar objects (points/genes) and group them

Hierarchical clustering

dendrogram

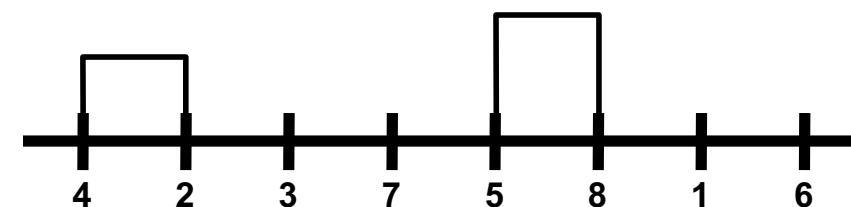
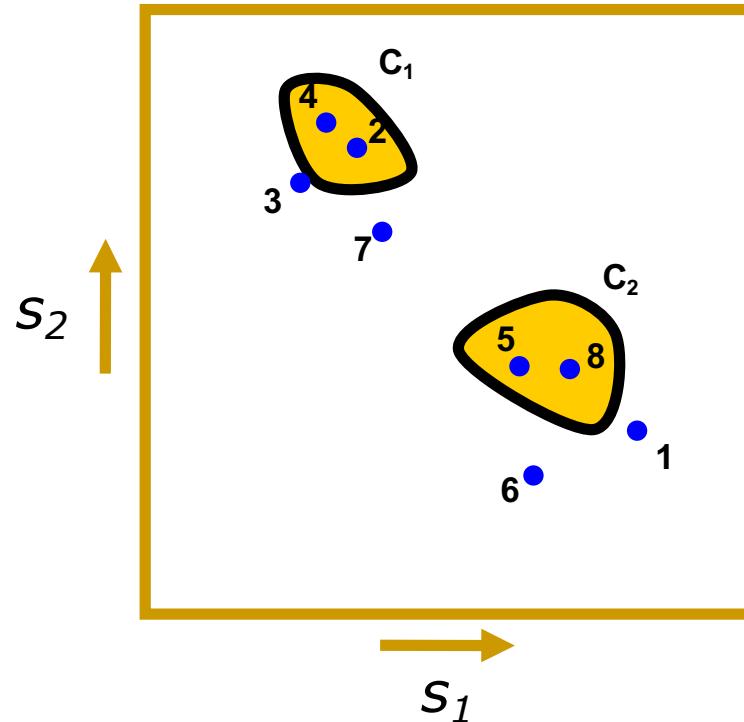


These are: objects 4 and 2

Again, find most similar objects (genes or clusters)
and group them

Hierarchical clustering

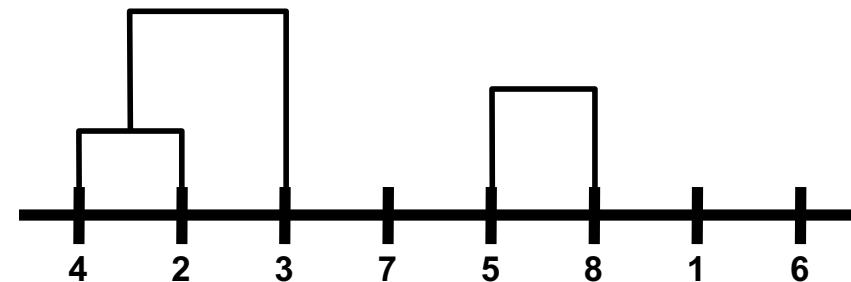
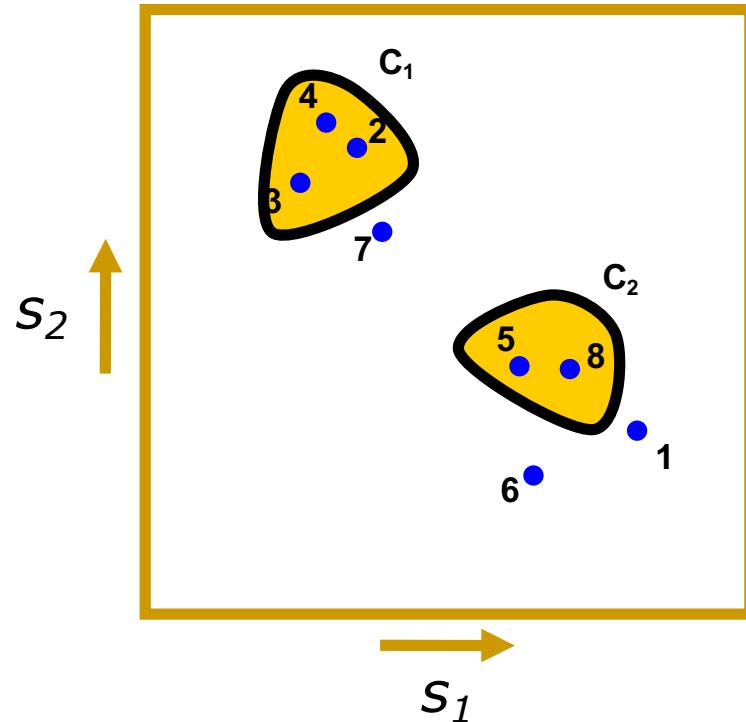
dendrogram



These are: objects 5 and 8
Repeat finding most similar objects (genes or clusters) and grouping them

Hierarchical clustering

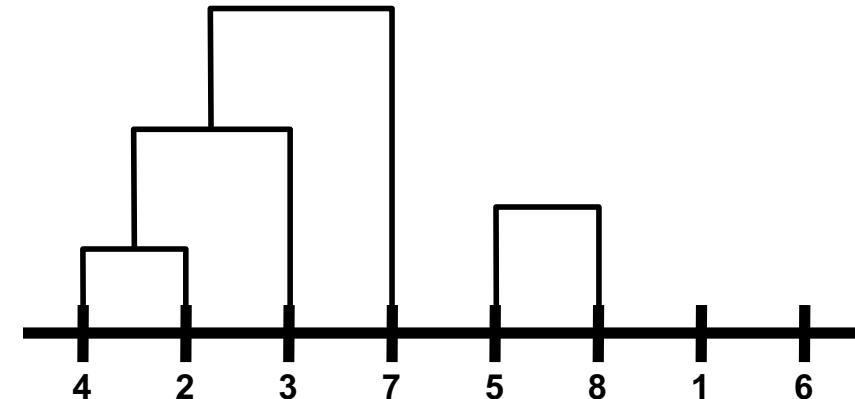
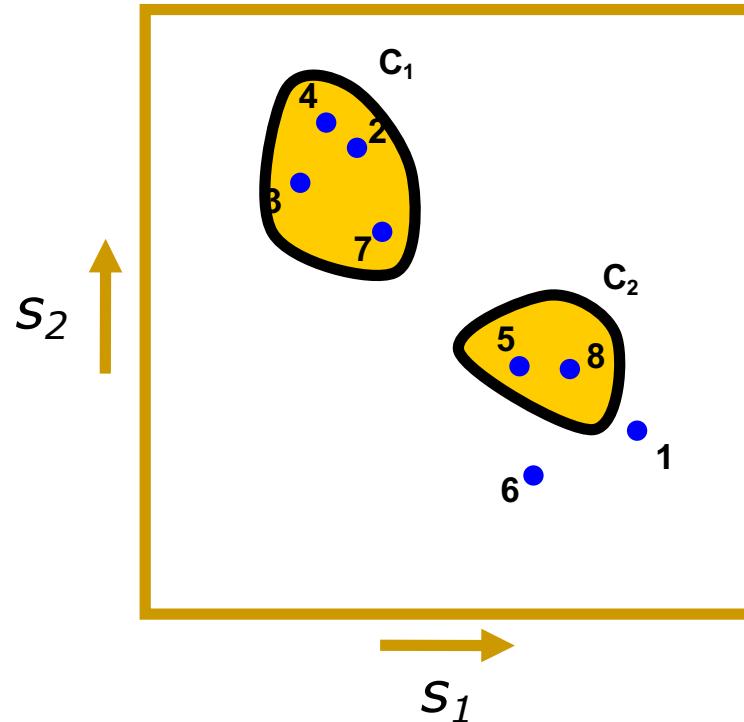
dendrogram



**Join object 3 and cluster 1
Repeat process**

Hierarchical clustering

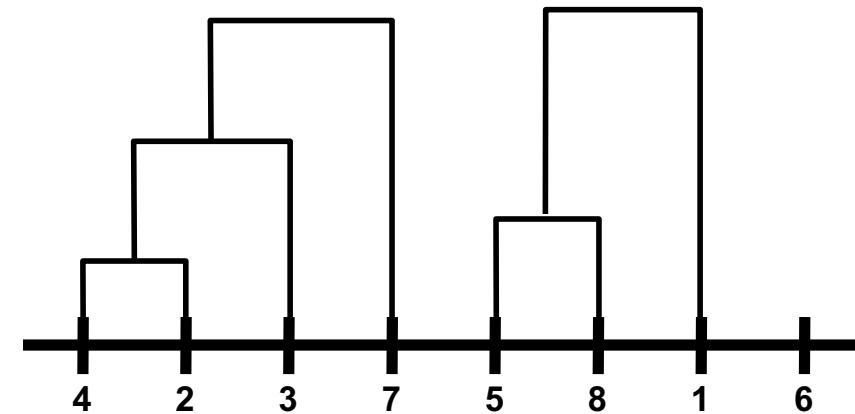
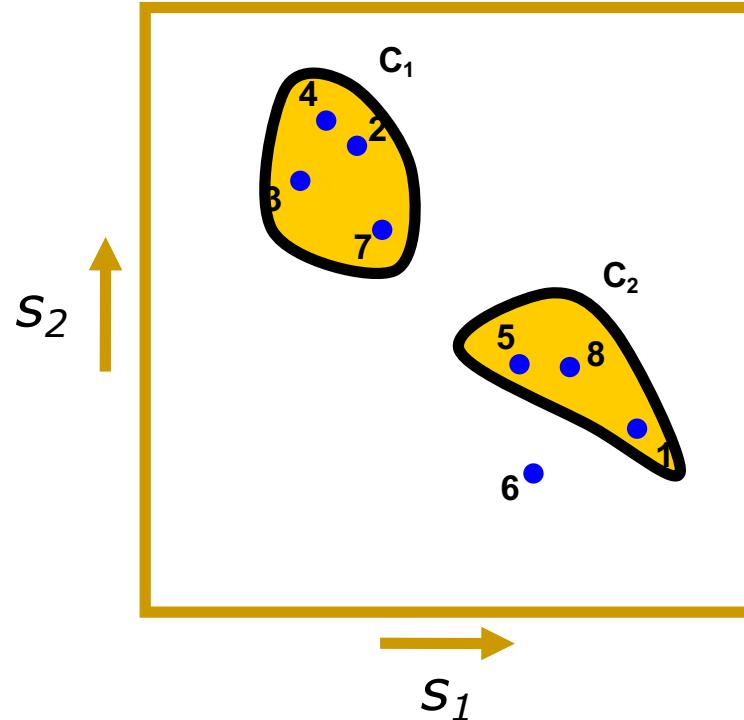
dendrogram



Join [object 7 and cluster 1] into [cluster 1]
Repeat process

Hierarchical clustering

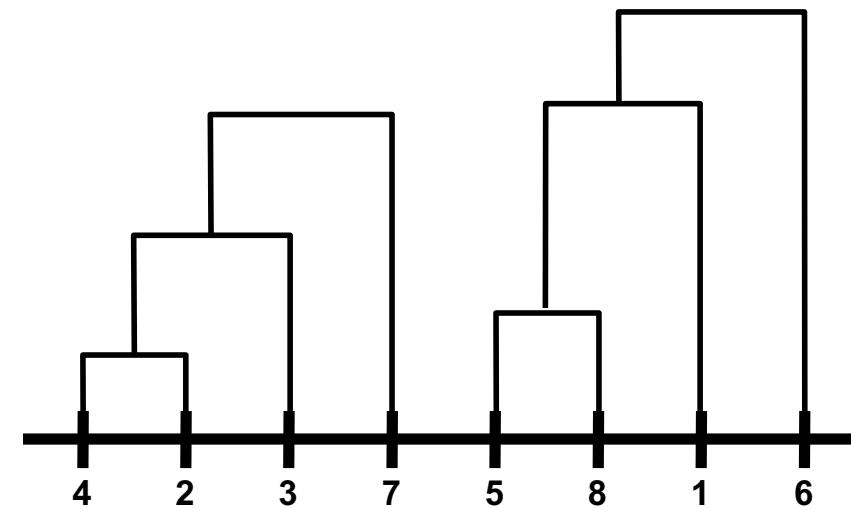
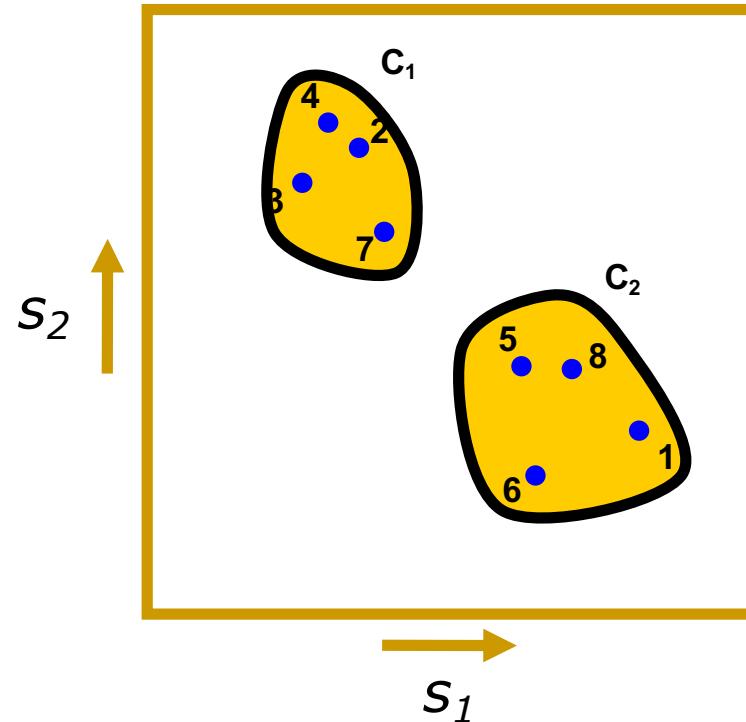
dendrogram



Join [object 1 and cluster 2] \rightarrow [cluster 2]
Repeat process

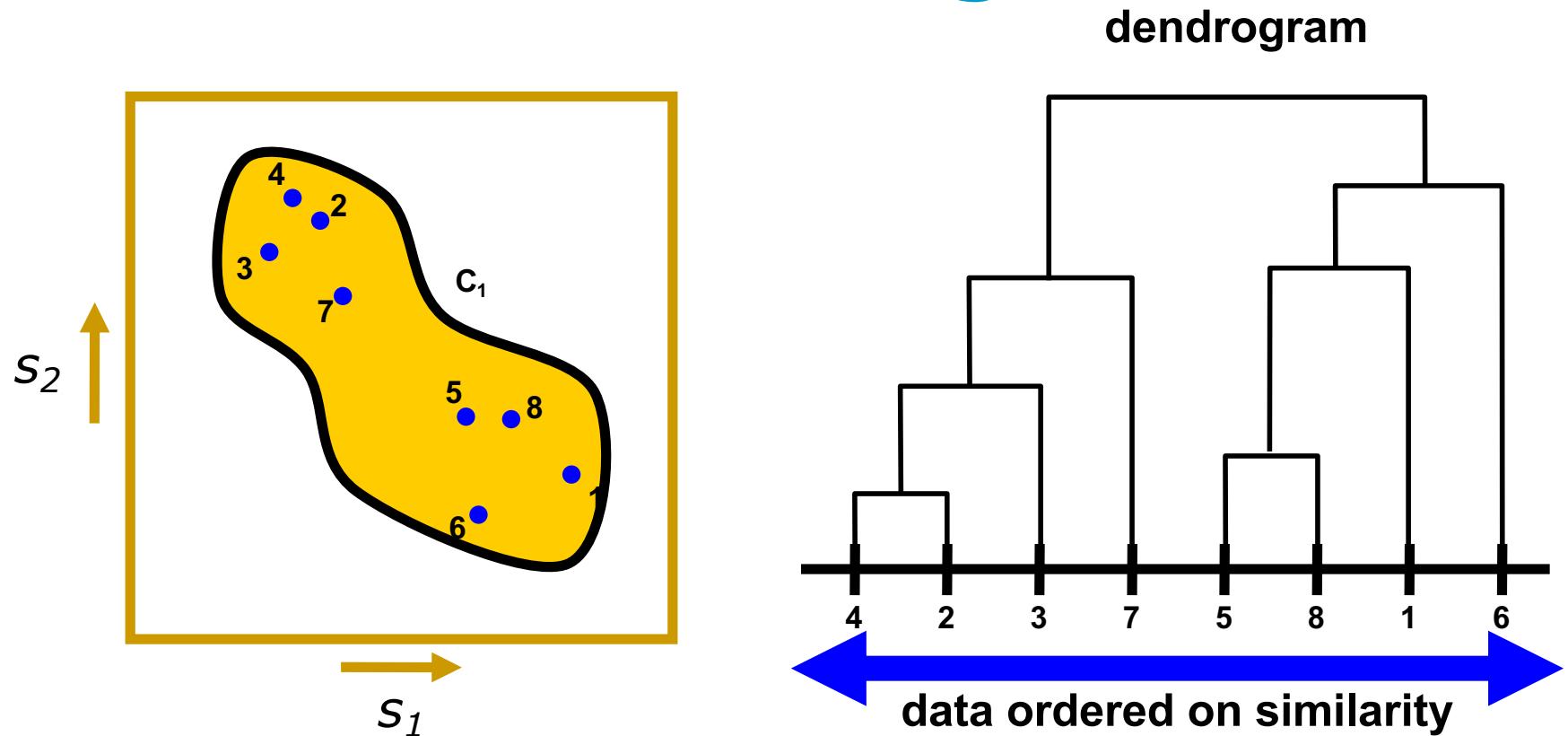
Hierarchical clustering

dendrogram



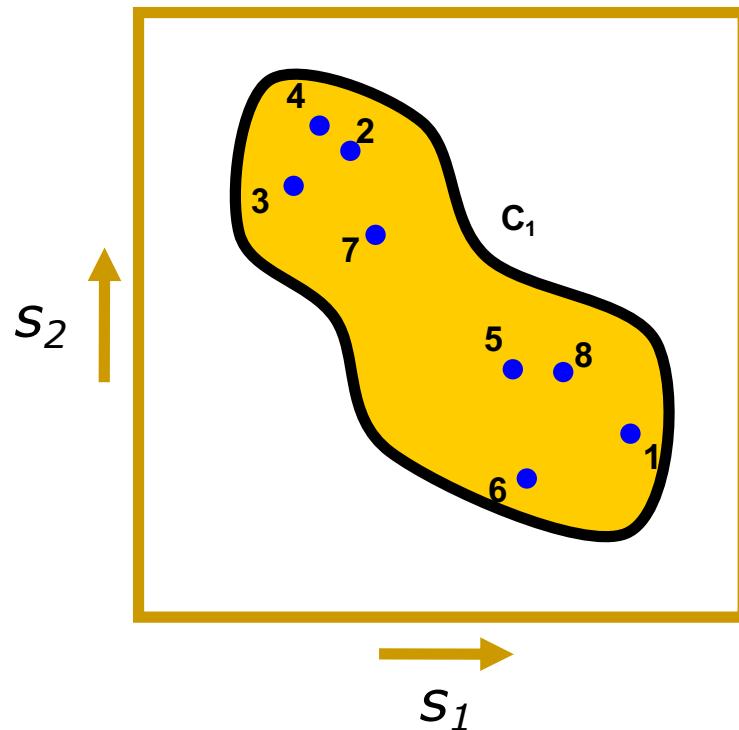
Join [object 6 and cluster 2] \rightarrow [cluster 2]
Repeat process

Hierarchical clustering



Join [cluster 1 and cluster 2] → [cluster 1]
All in one cluster: FINISHED!

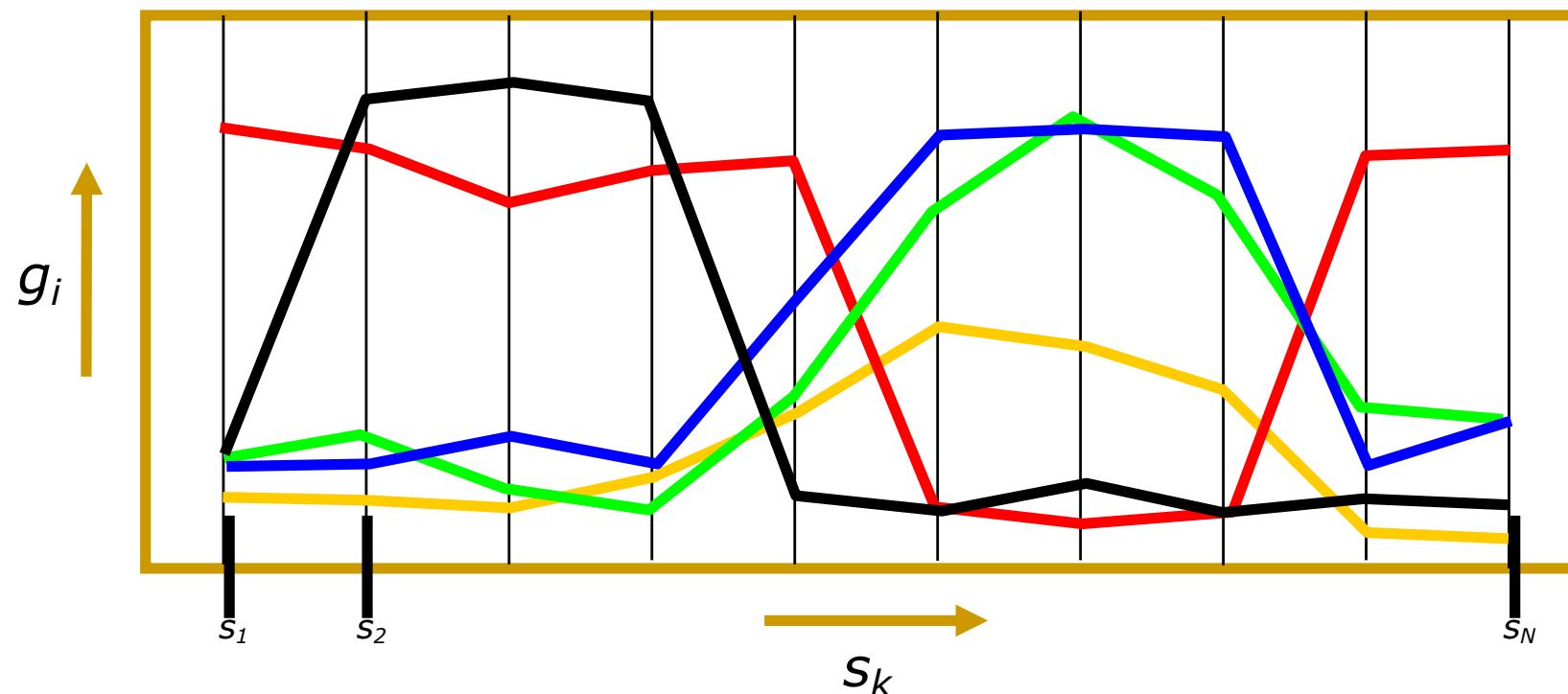
Hierarchical clustering: Choices

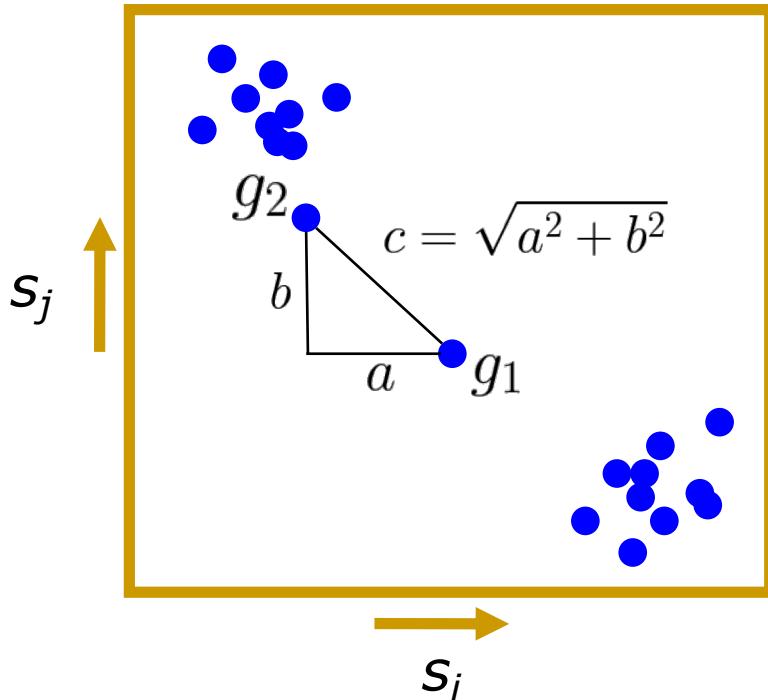


Need to know:

- **Similarity between objects**
- **Similarity between clusters**

Hierarchical clustering: Similarity between objects





Euclidean distance

$$a = g_1(s_i) - g_2(s_i)$$

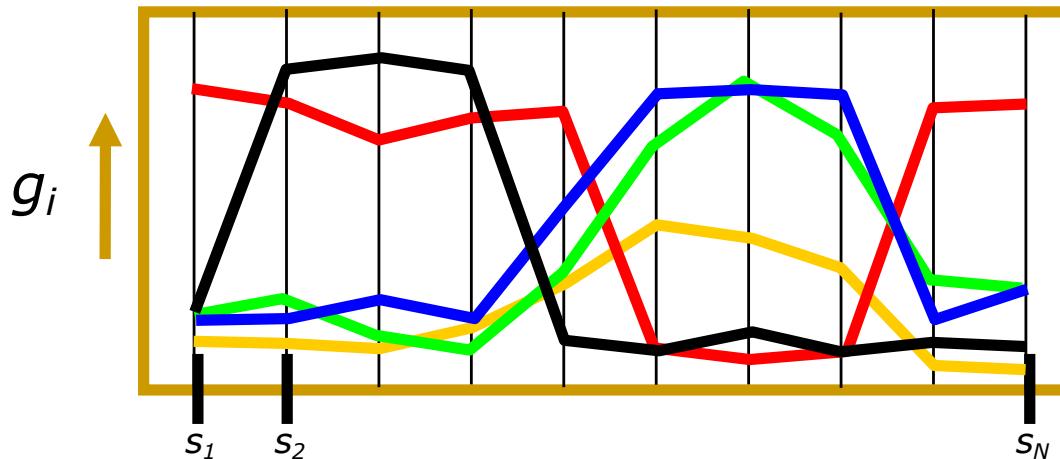
$$b = g_1(s_j) - g_2(s_j)$$

$$c = \sqrt{a^2 + b^2}$$

$$c = \sqrt{(g_1(s_i)) - g_2(s_i))^2 + (g_1(s_j) - g_2(s_j))^2}$$

$$d(g_1, g_2) = c = \sqrt{\sum_{k=1}^K (g_1(s_k) - g_2(s_k))^2}$$

Similarity between objects



Euclidean distance

$$d(g_1, g_2) = \sqrt{\sum_{k=1}^K (g_1(s_k) - g_2(s_k))^2}$$

$d(\bullet, \bullet)$ < $d(\bullet, \circ)$
 $d(\bullet, \bullet)$ << $d(\bullet, \bullet)$
 $d(\bullet, \bullet)$ << $d(\bullet, \bullet)$

Pearson correlation

$$\rho_{g_1,g_2} = \frac{\sum_{k=1}^K g_1(s_k) * g_2(s_k) - \mu_1 * \mu_2}{(\sigma_1 * \sigma_2)}$$

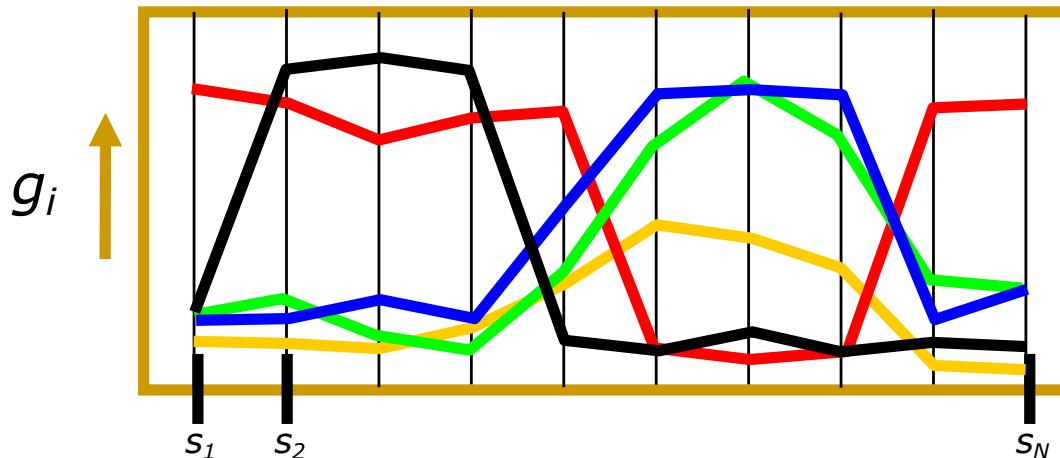
$d(\bullet, \bullet)$ \approx $d(\bullet, \circ)$
 $d(\bullet, \bullet)$ << $d(\bullet, \bullet)$
 $d(\bullet, \bullet)$ << $d(\bullet, \bullet)$

Mixed Pearson correlation

$$1 - |\rho_{g_1,g_2}|$$

$d(\bullet, \bullet)$ \approx $d(\bullet, \circ)$
 $d(\bullet, \bullet)$ \approx $d(\bullet, \bullet)$
 $d(\bullet, \bullet)$ << $d(\bullet, \bullet)$

Similarity between objects



Euclidean distance

match exact shape

$$\begin{aligned} d(\bullet, \bullet) &< d(\bullet, \circ) \\ d(\bullet, \bullet) &<< d(\bullet, \textcolor{red}{\bullet}) \\ d(\bullet, \bullet) &<< d(\bullet, \textcolor{black}{\bullet}) \end{aligned}$$

Pearson correlation

ignore amplitude

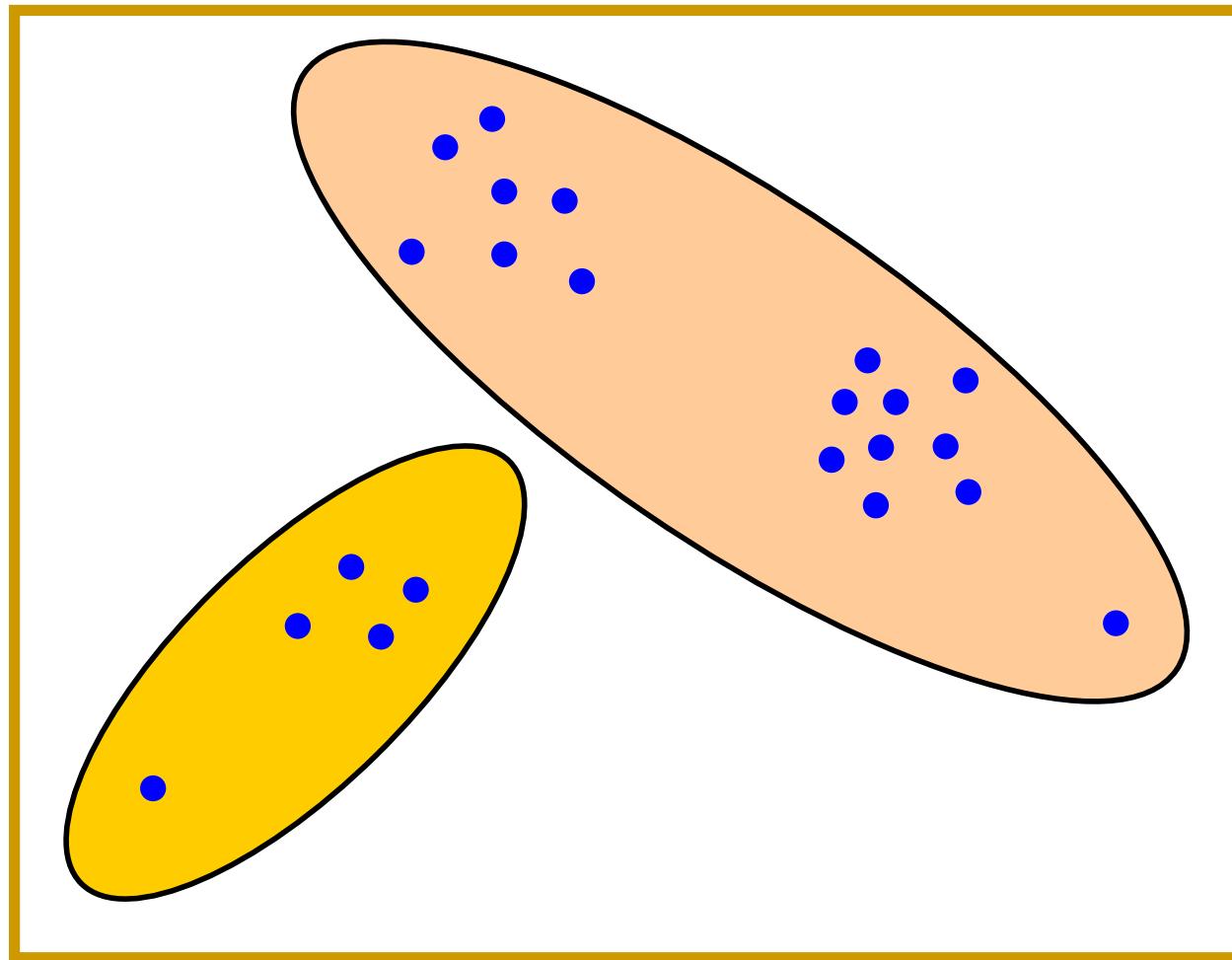
$$\begin{aligned} d(\bullet, \bullet) &\approx d(\bullet, \textcolor{yellow}{\bullet}) \\ d(\bullet, \bullet) &<< d(\bullet, \textcolor{red}{\bullet}) \\ d(\bullet, \bullet) &<< d(\bullet, \textcolor{black}{\bullet}) \end{aligned}$$

Mixed Pearson correlation

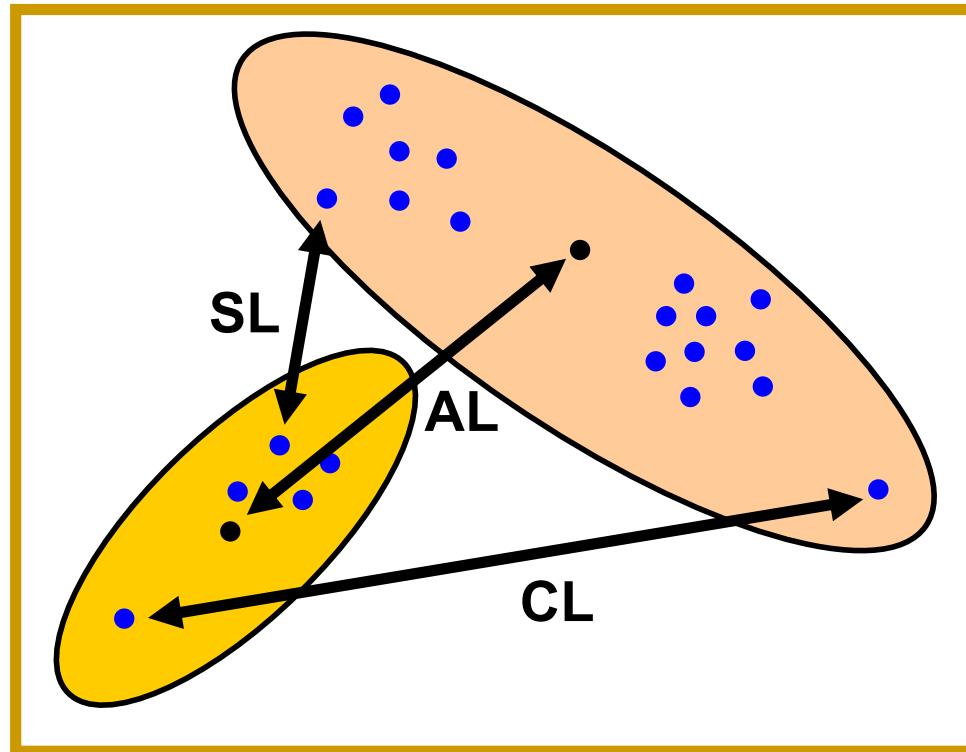
ignore amplitude & sign

$$\begin{aligned} d(\bullet, \bullet) &\approx d(\bullet, \textcolor{yellow}{\bullet}) \\ d(\bullet, \bullet) &\approx d(\bullet, \textcolor{red}{\bullet}) \\ d(\bullet, \bullet) &<< d(\bullet, \textcolor{black}{\bullet}) \end{aligned}$$

Similarity between clusters?

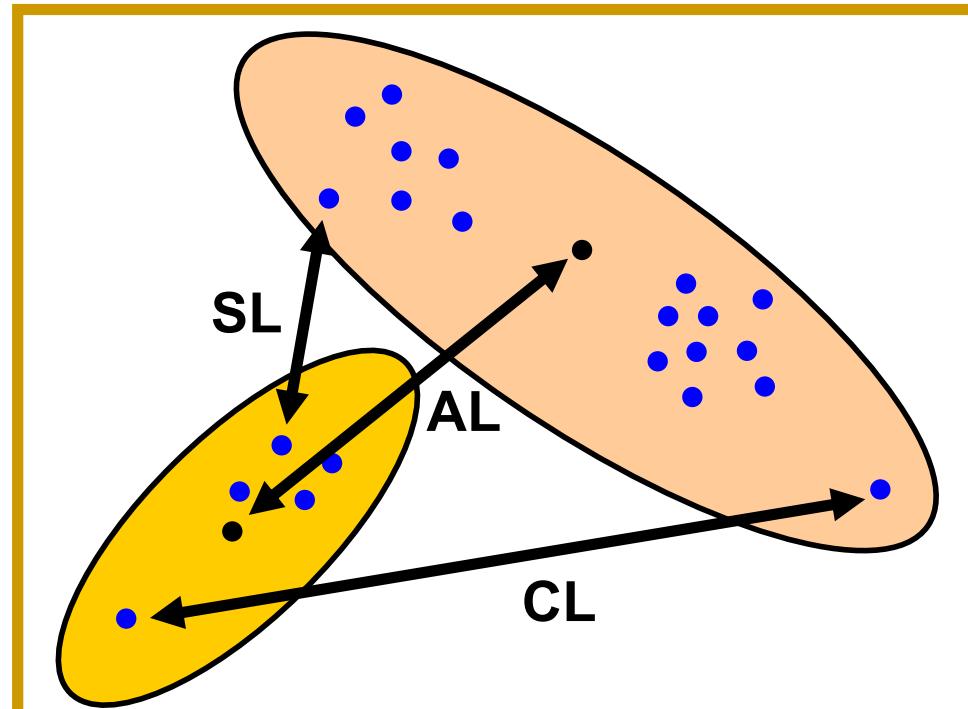


Similarity between clusters (linkage)



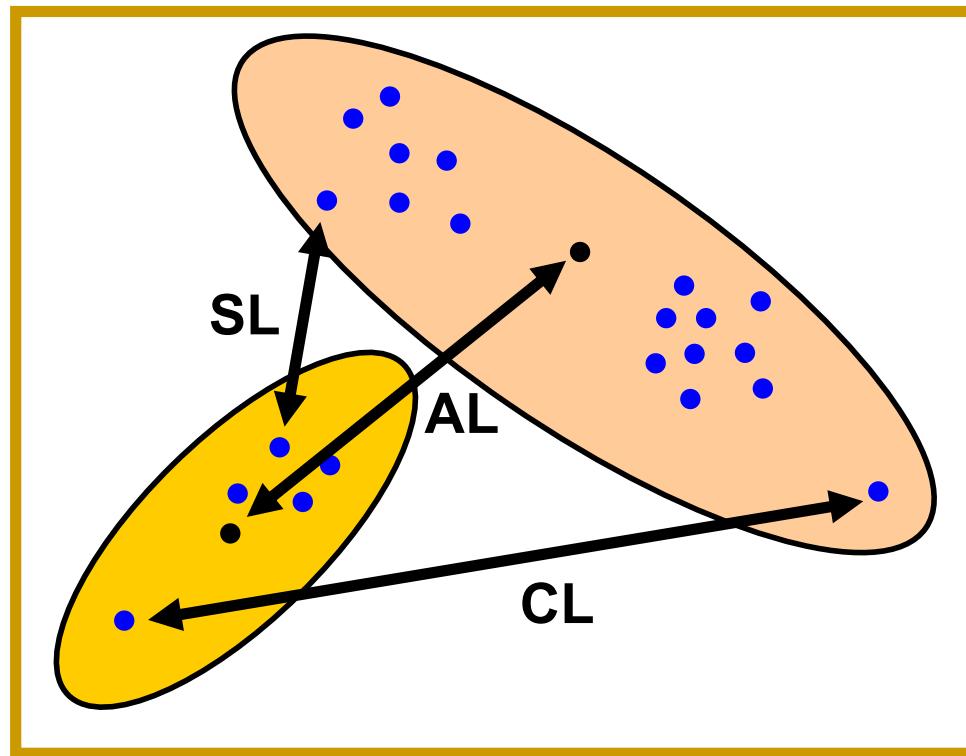
- **Single linkage:** Closest objects
- **Complete linkage:** Furthest objects
- **Average linkage:** Average dissimilarity

Similarity between clusters (linkage)



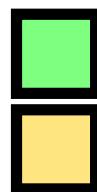
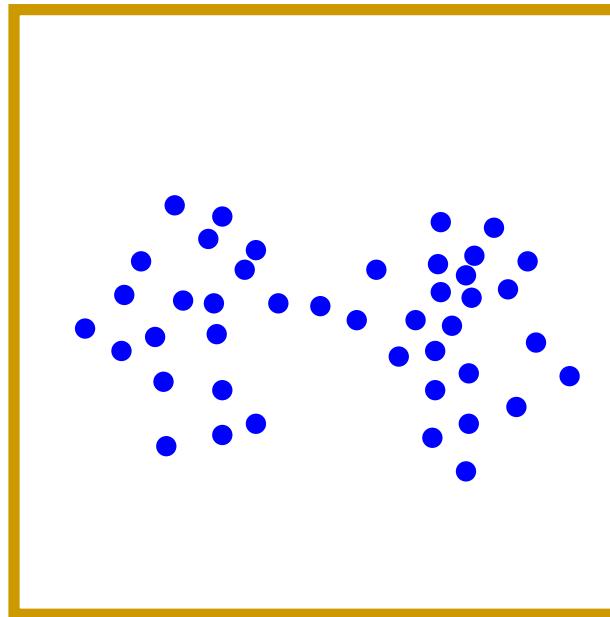
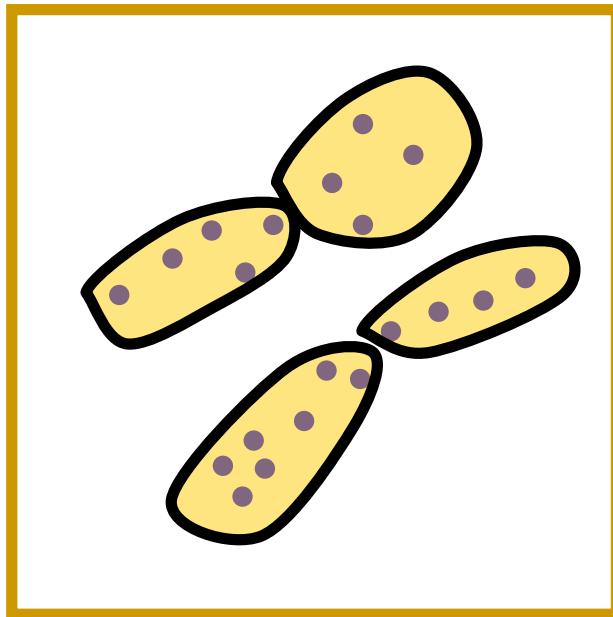
- **Single linkage:** $d(X_i, X_j) = \min(d(a, b) : a \in X_i, b \in X_j)$
- **Complete linkage:** $d(X_i, X_j) = \max(d(a, b) : a \in X_i, b \in X_j)$
- **Average linkage:** $d(X_i, X_j) = \frac{1}{|X_i||X_j|} \sum_{a \in X_i} \sum_{b \in X_j} d(a, b)$

Similarity between clusters (2)



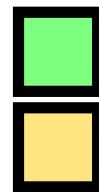
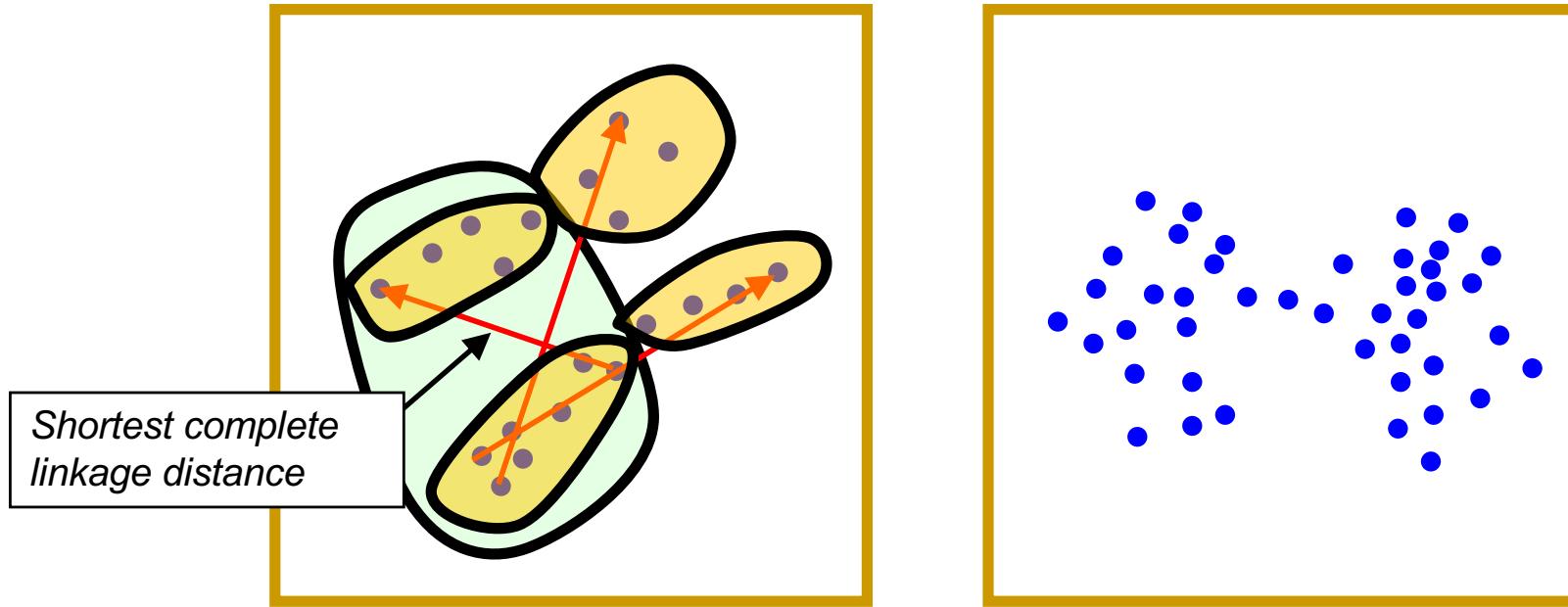
- **Ward's method:** $d(X_i, X_j) = \|X_i - X_j\|^2$
(total variance)
- **Centroid distance:** $d(X_i, X_j) = \|c_s - c_t\|$

SL vs CL: Shape influences



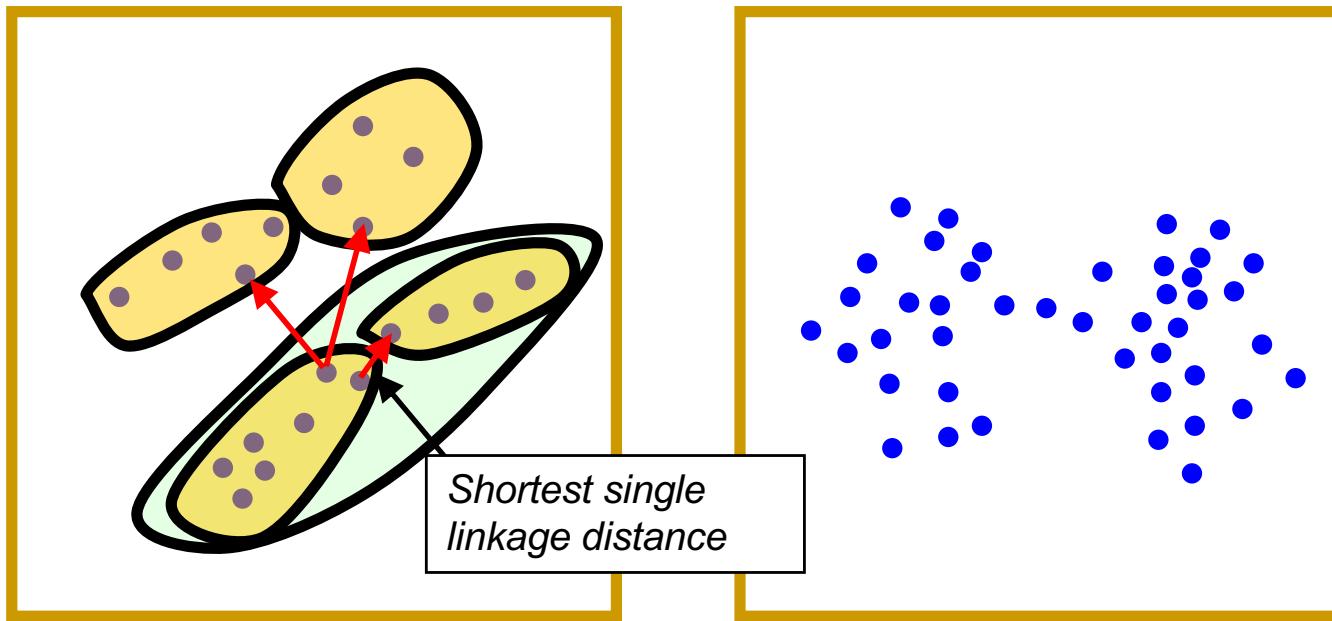
complete linkage ?
single linkage ?

SL vs CL: Shape influences



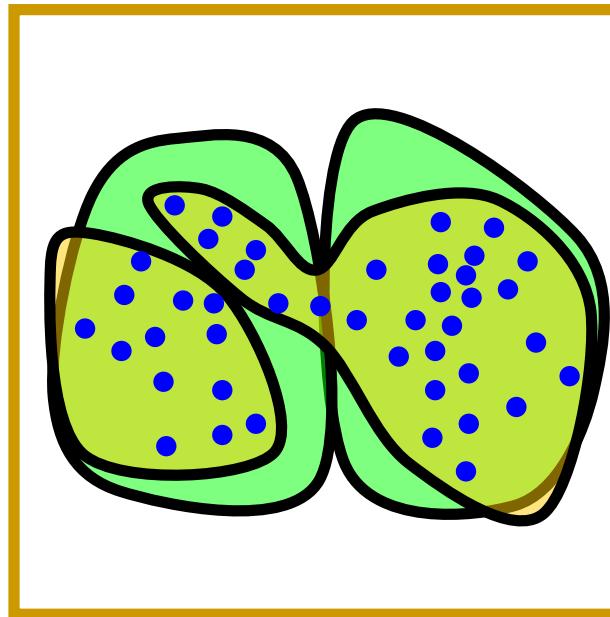
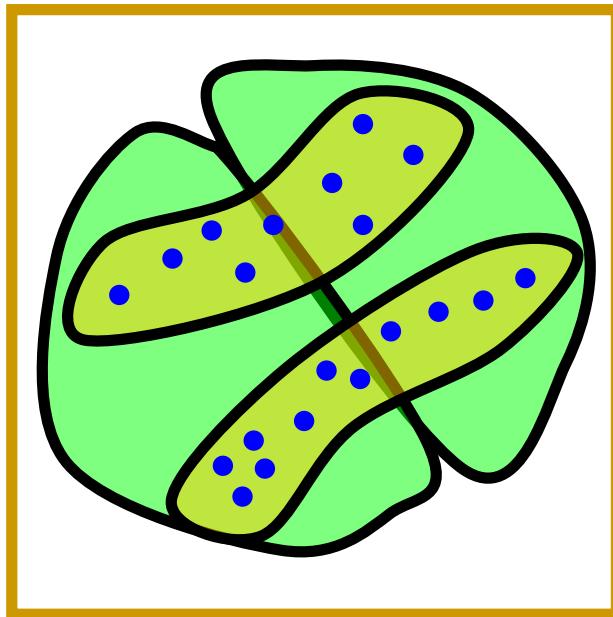
complete linkage ?
single linkage ?

SL vs CL: Shape influences



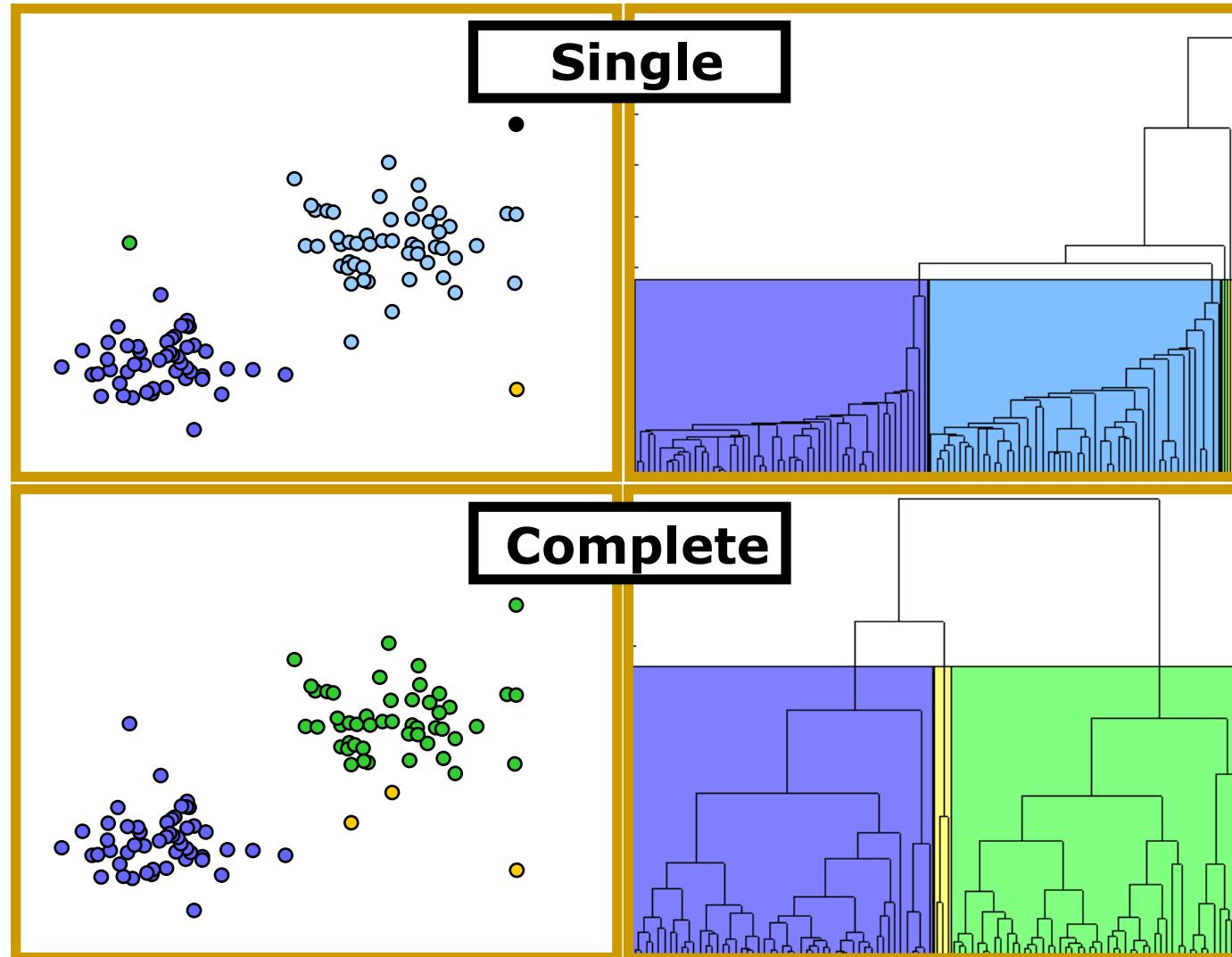
complete linkage ?
single linkage ?

SL vs CL: Shape influences



complete linkage
single linkage

SL vs CL: Outlier influences



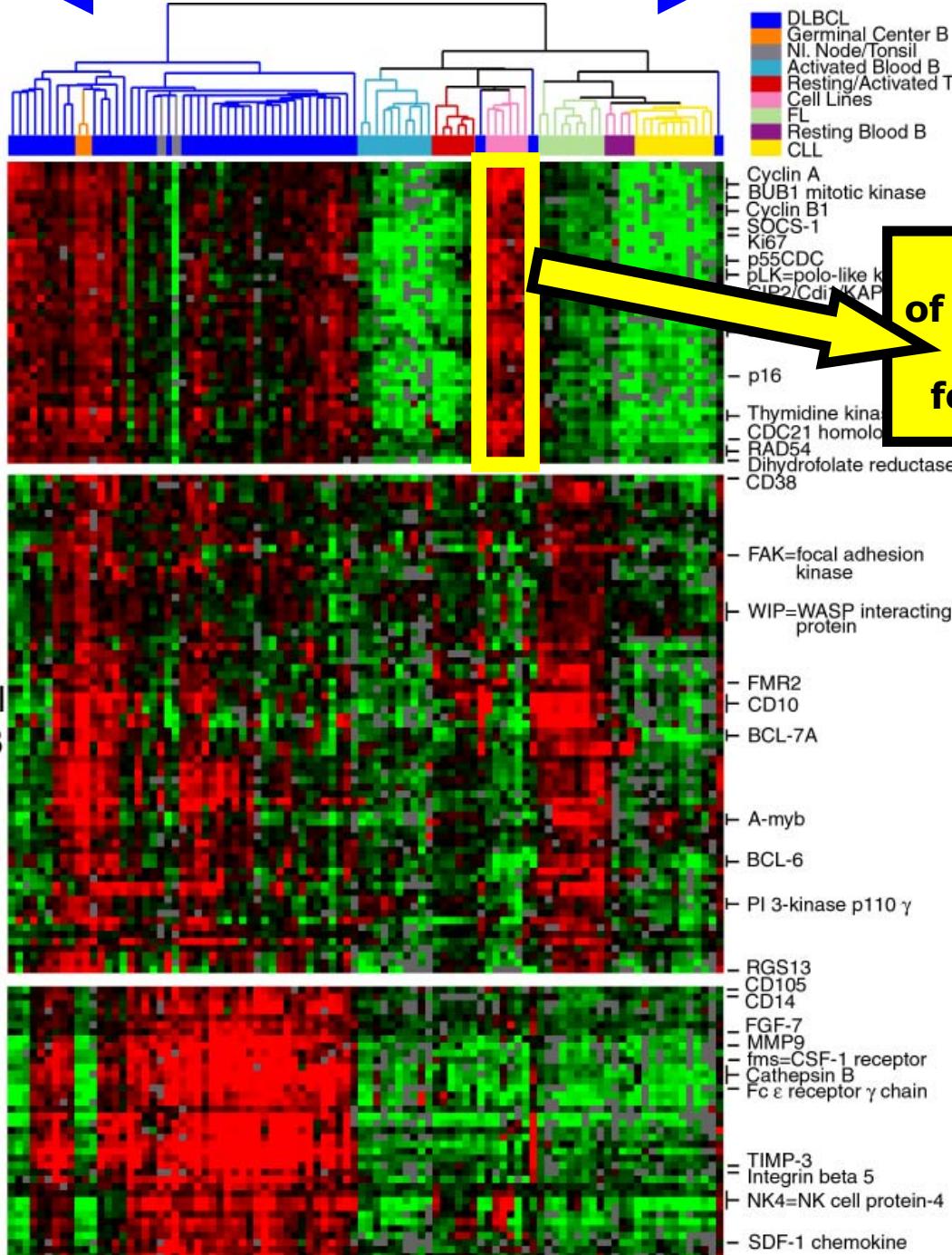
ordered on
similarity

related tumors

Proliferation

related
genes

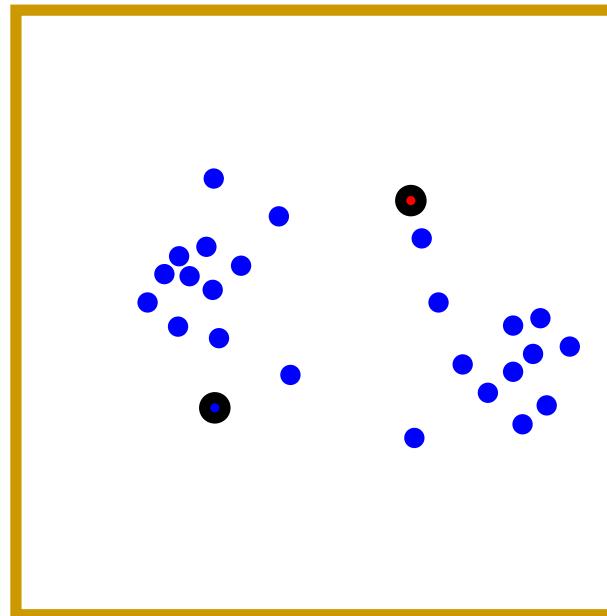
Lymph
Node



Genetic profile
of functionally related
genes
for a specific tumor

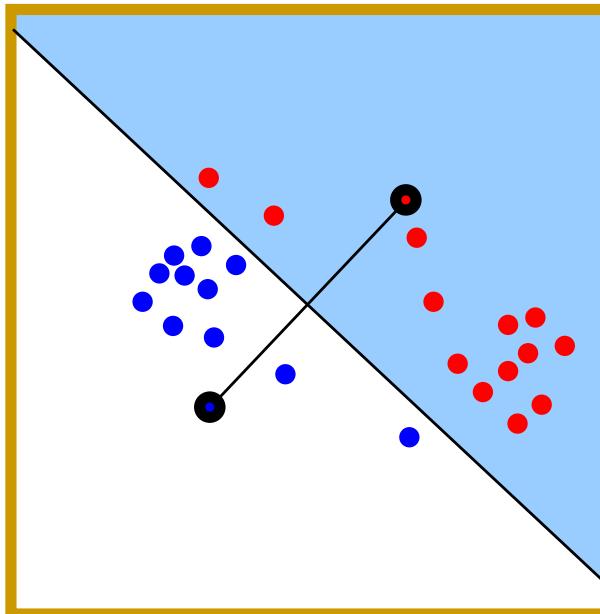
cluster

K-means clustering: Explanation by example



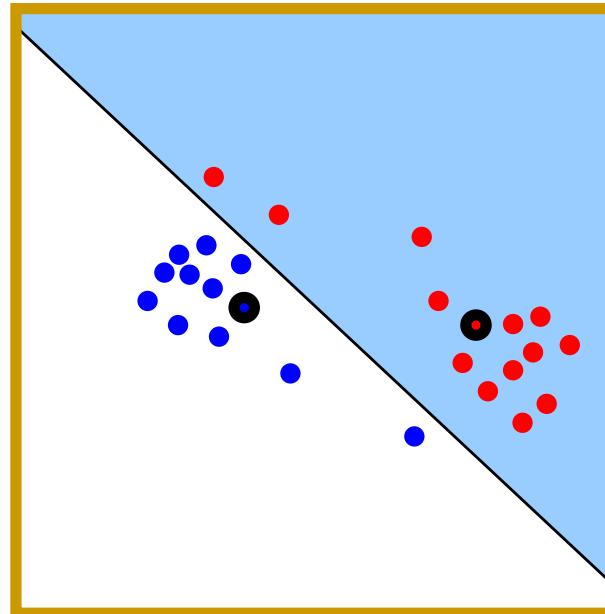
Choose randomly 2 prototypes

K-means clustering: Explanation by example



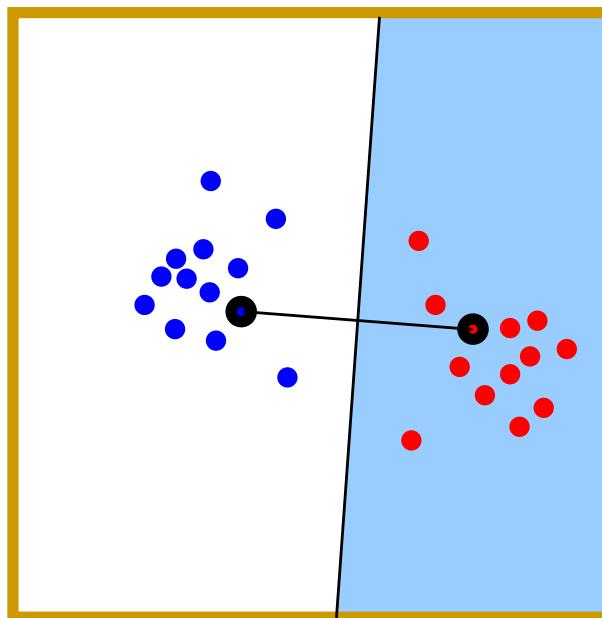
**Assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2**

K-means clustering: Explanation by example



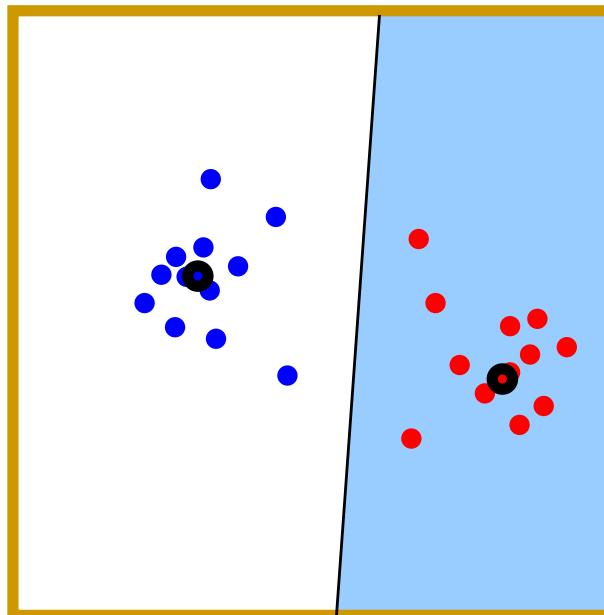
**Calculate new cluster prototypes
By averaging objects**

K-means clustering: Explanation by example



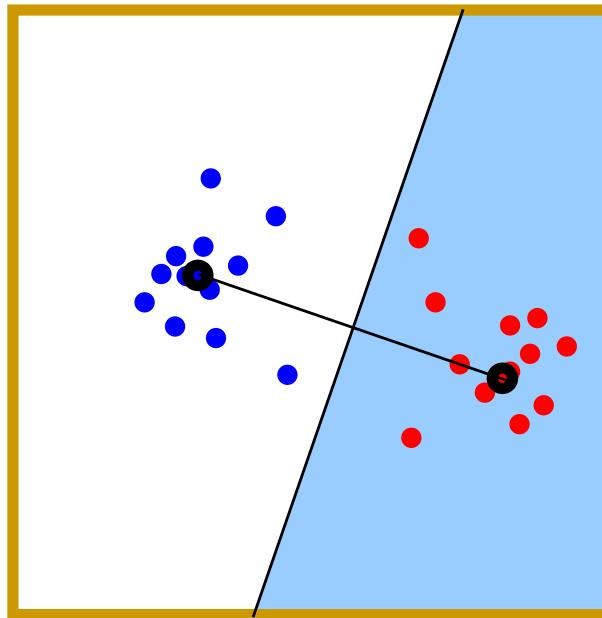
**Re-assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2**

K-means clustering: Explanation by example



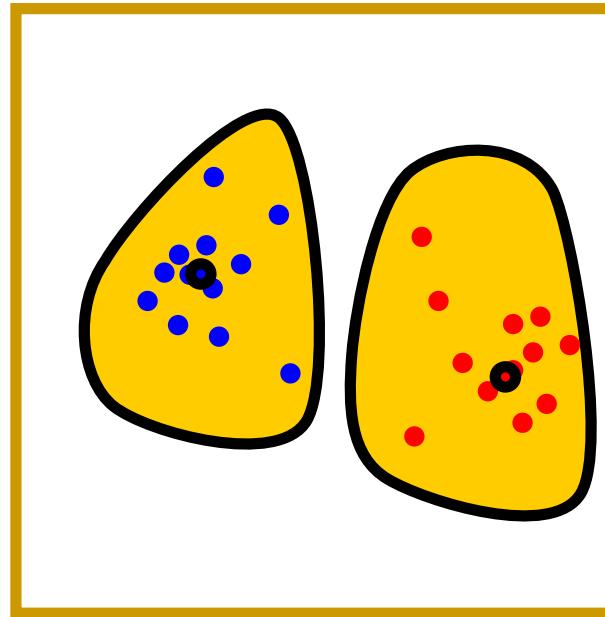
Re-calculate new cluster prototypes

K-means clustering: Explanation by example



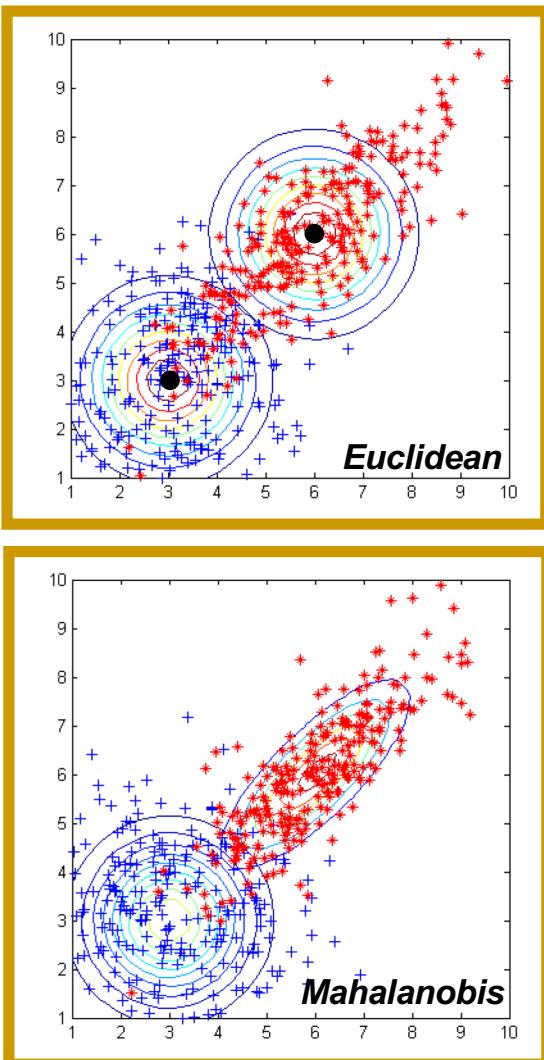
**Re-assign objects to closest prototype
If no objects change cluster then finished**

K-means clustering: Explanation by example



Establish clusters

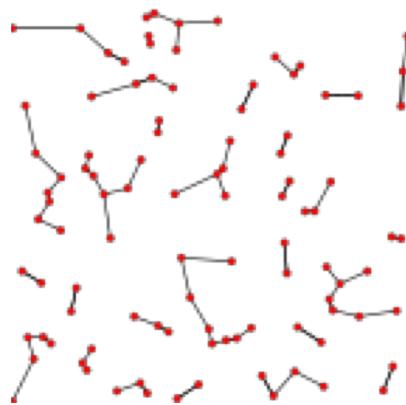
K-means clustering: Parameters



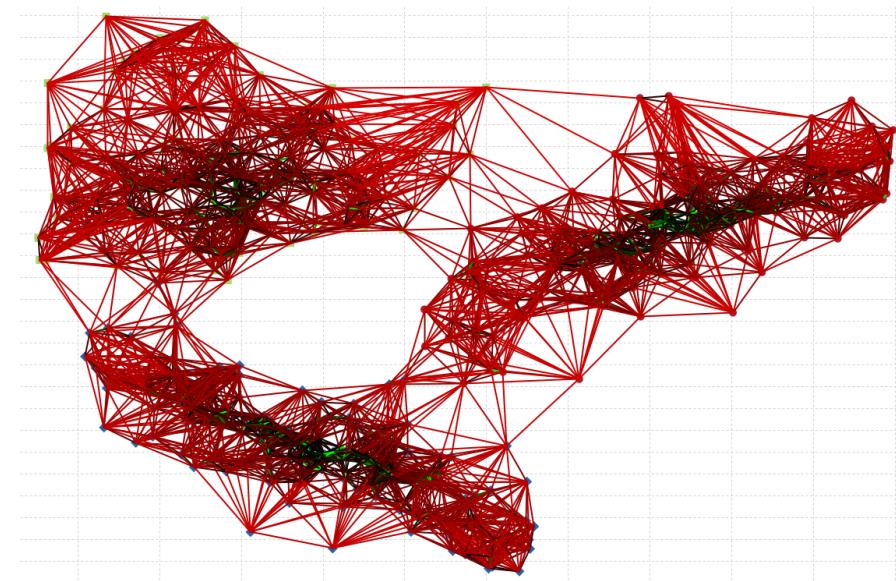
- **K-means**
 - Fixed number of clusters (need to know a priori)
 - Choice of distance measure
 - Prototype choice
- **Distance measure**
 - Euclidean: Round clusters
 - Mahalanobis: Elongated clusters
- **Prototype choice**
 - Point
 - Line etc.
- **Number of clusters**
 - Validate clustering!

Graph-based Clustering

- **K-NN graph:** Connect every node to its k-nearest nodes
- Find densely connected components



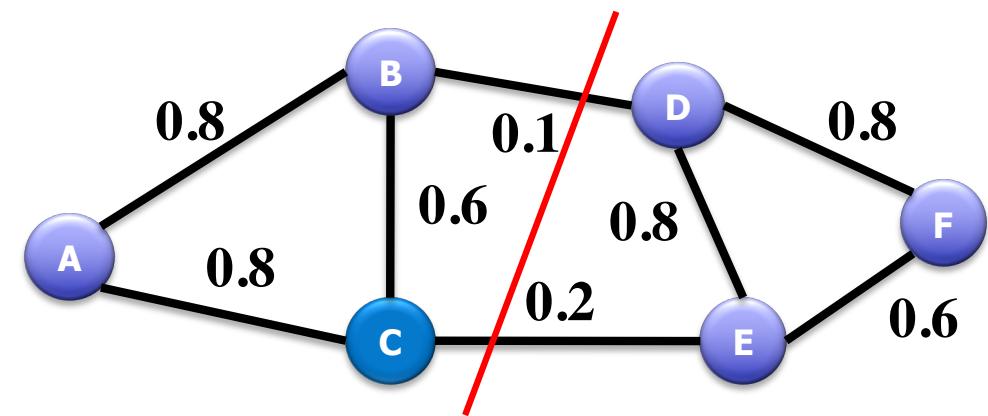
$k-1$



$k-20$

Spectral clustering (1)

- Minimise normalised cut
- Normalised cut between two clusters C_1 and C_2 :

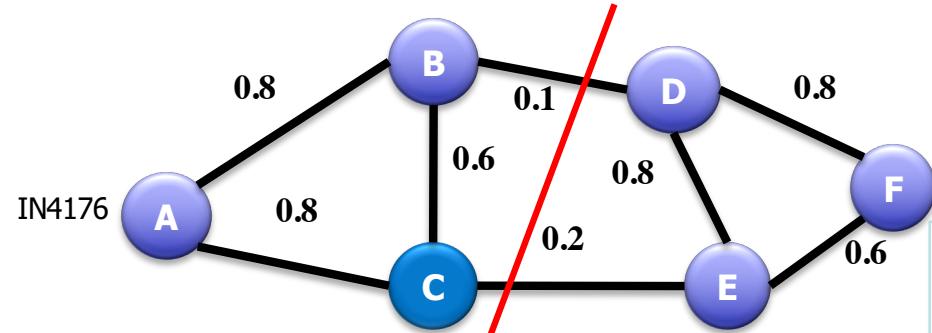
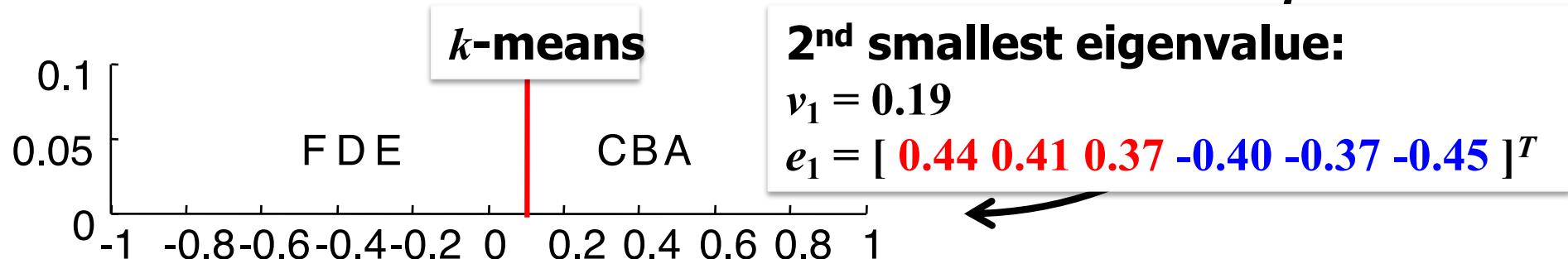


$$NC(C_1, C_2) = \frac{\text{cut}(C_1, C_2)}{\text{assoc}(C_1, V)} + \frac{\text{cut}(C_2, C_1)}{\text{assoc}(C_2, V)} = 2 - \left(\frac{\text{assoc}(C_1, C_1)}{\text{assoc}(C_1, V)} + \frac{\text{assoc}(C_2, C_2)}{\text{assoc}(C_2, V)} \right)$$

- $\text{cut}(C_1, C_2)$ = weight of links between C_1 and C_2
- $\text{cut}(C_2, C_1)$ = same
- $\text{assoc}(C_1, V)$ = total weight of links from nodes in C_1 to entire graph
- $\text{assoc}(C_2, V)$ = total weight of links from nodes in C_2 to entire graph

Spectral clustering (2) (*sum weights on diagonal – W*)

$$W = A \begin{pmatrix} 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{pmatrix} \xrightarrow{\substack{\text{Laplacian} \\ L = \text{diag}(1^T W) - W}} L = A \begin{pmatrix} 1.6 & -0.8 & -0.8 & 0 & 0 & 0 \\ -0.8 & 1.5 & -0.6 & -0.1 & 0 & 0 \\ -0.8 & -0.6 & 1.6 & 0 & -0.2 & 0 \\ 0 & -0.1 & 0 & 1.7 & -0.8 & -0.8 \\ 0 & 0 & -0.2 & -0.8 & 1.6 & -0.6 \\ 0 & 0 & 0 & -0.8 & -0.6 & 1.4 \end{pmatrix}$$



IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 8, AUGUST 2000

Normalized Cuts and Image Segmentation

Jianbo Shi and Jitendra Malik, Member, IEEE

Markov clustering (1)

- **Markov clustering**

- Random walks are Markov chains
- Start with Markov matrix: $M_{ij} = p(j \rightarrow i)$
- (optionally add loops (diagonal))

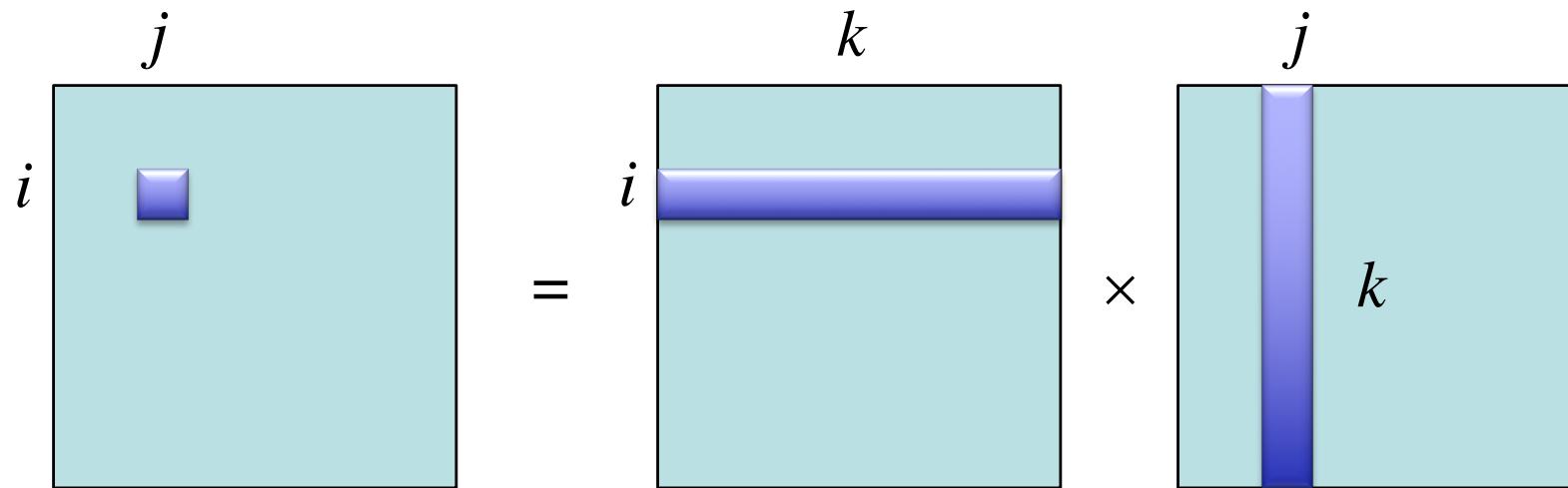
$$W = A \begin{pmatrix} A & B & C & D & E & F \\ 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ B & 0.8 & 0 & 0.6 & 0.1 & 0 \\ C & 0.8 & 0.6 & 0 & 0 & 0.2 \\ D & 0 & 0.1 & 0 & 0 & 0.8 \\ E & 0 & 0 & 0.2 & 0.8 & 0 \\ F & 0 & 0 & 0 & 0.8 & 0.6 \end{pmatrix} \xrightarrow{\text{Markov}} M = A \begin{pmatrix} A & B & C & D & E & F \\ 0 & 0.53 & 0.5 & 0 & 0 & 0 \\ B & 0.5 & 0 & 0.38 & 0.06 & 0 \\ C & 0.5 & 0.4 & 0 & 0 & 0.12 \\ D & 0 & 0.07 & 0 & 0 & 0.5 \\ E & 0 & 0 & 0.12 & 0.47 & 0 \\ F & 0 & 0 & 0 & 0.47 & 0.38 \end{pmatrix}$$

$\xrightarrow{M = W \text{diag}(1^T W)^{-1}}$

(sum columns = 1)

Markov clustering (2)

- $M_{ij} = p(j \rightarrow i)$ probability of arriving in i from j in 1 step



$$(M^2)_{ij} = \sum_k p(j \rightarrow k)p(k \rightarrow i)$$

probability of arriving in i from j in 2 steps (etc.)

Markov clustering (3)

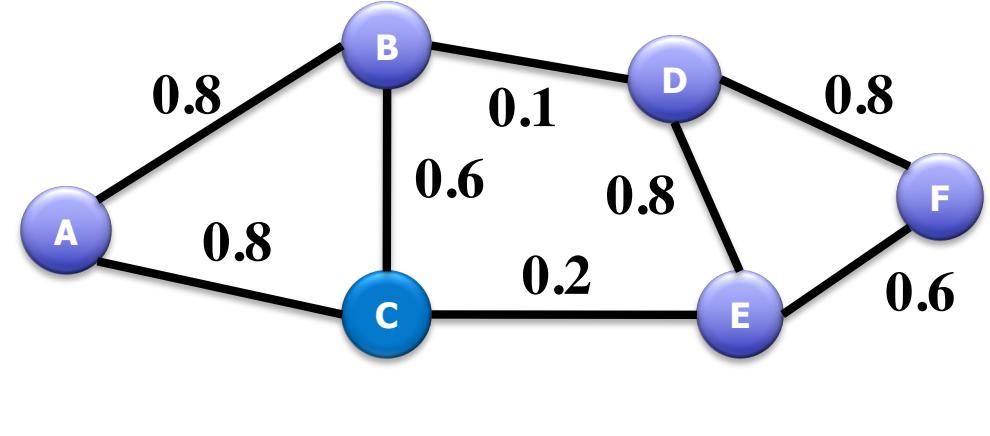
- **Markov clustering:**

- Iterate

1. $M = M^\alpha$

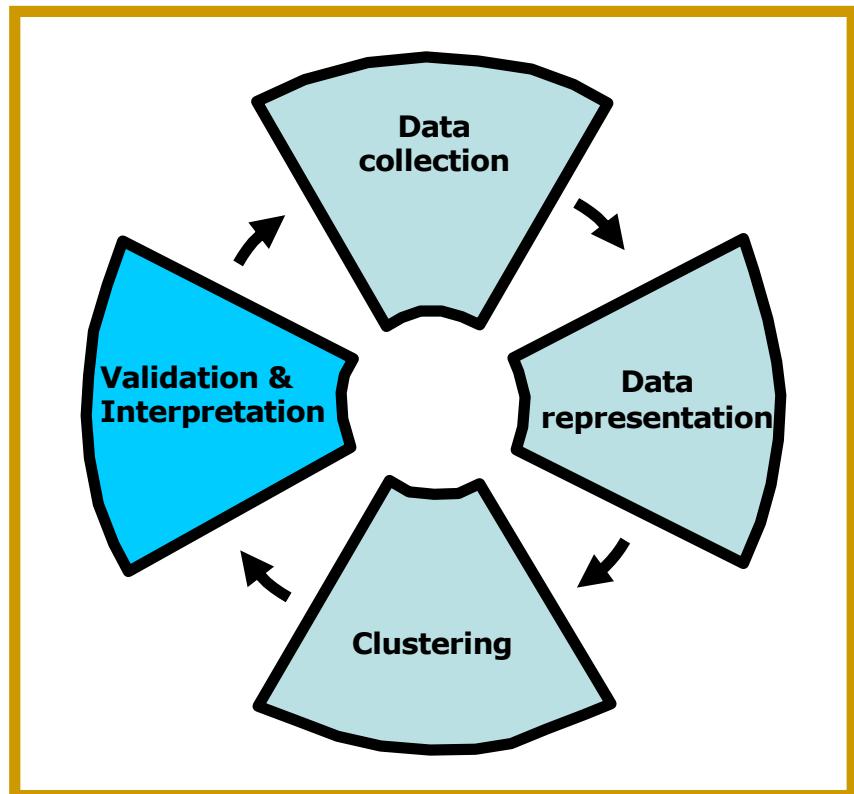
2. $M_{ij} = M_{ij}^\beta / \sum_i M_{ij}^\beta$

- Step 1 : take walk of α steps
- Step 2 : increase difference between small probabilities (<0.5) and large probabilities (>0.5) ; converges to maximum value in column
- E.g. for $\alpha = \beta = 2$



	A	B	C	D	E	F
M = A	0.00	0.50	0.50	0.00	0.00	0.00
B	0.00	0.50	0.50	0.06	0.00	0.00
C	0.00	0.50	0.50	0.00	0.00	0.00
D	0.00	0.00	0.00	0.50	0.50	0.50
E	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.50

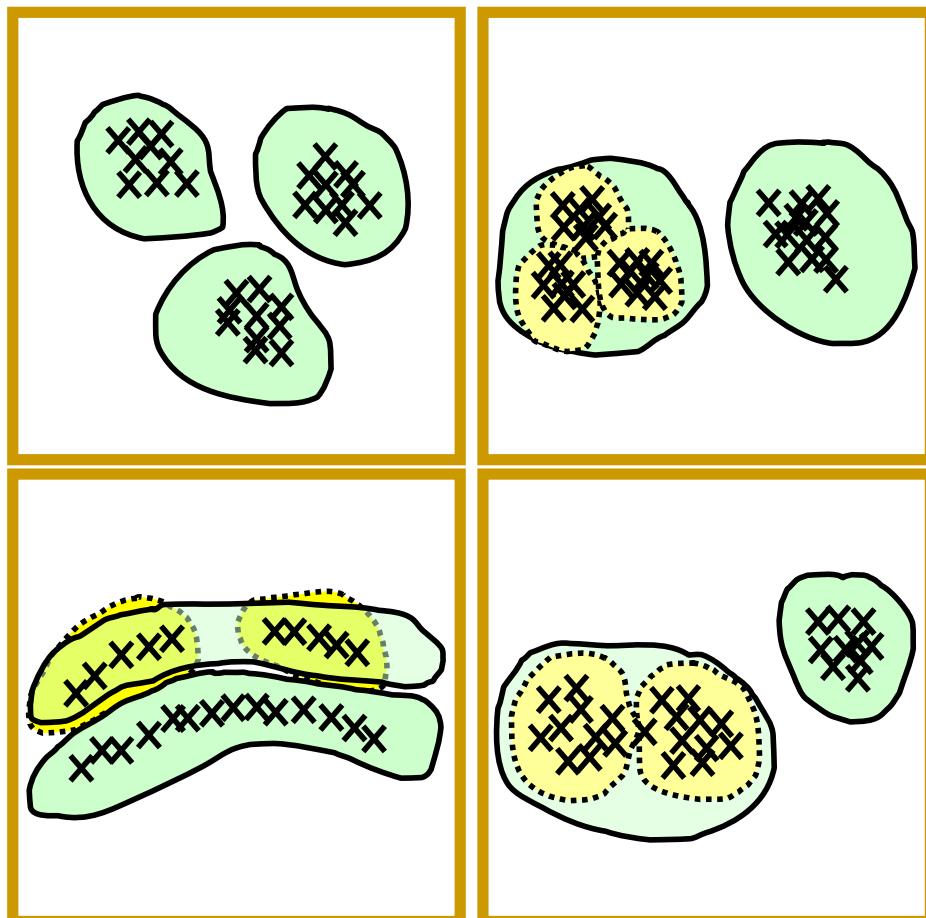
Validation



TOPICS

- **Cluster tendency**
- **Cluster validity**

Subjectivity



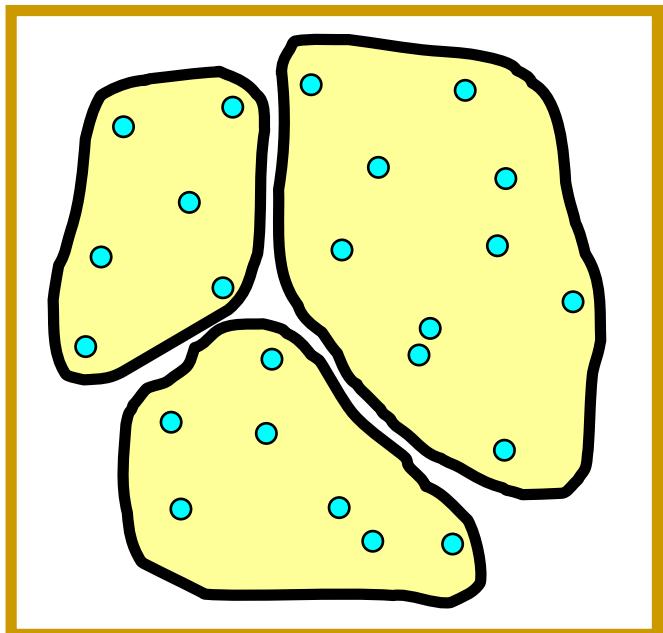
- **Principle choices:**
 - similarity
 - algorithm
- **Different choice leads to different results**
Subjectivity becomes reality
- **Cluster process**
Validate, interpret (generate hypothesis), repeat steps

Cluster Validation

- **Cluster tendency**

Clustering **IMPOSES** structure even though data may not posses it

Aim: Test whether data possesses structure

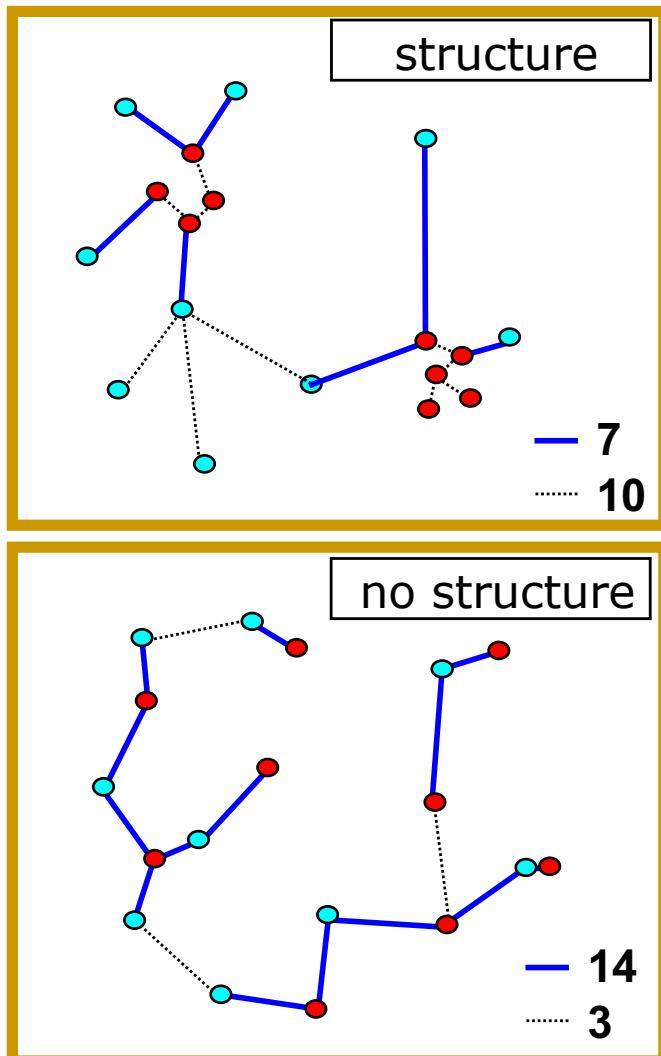


- **Cluster validity**

Choices impose restrictions on for example shape

Aim: Quantitative evaluation of the clustering results

Test for spatial randomness



- **Test**

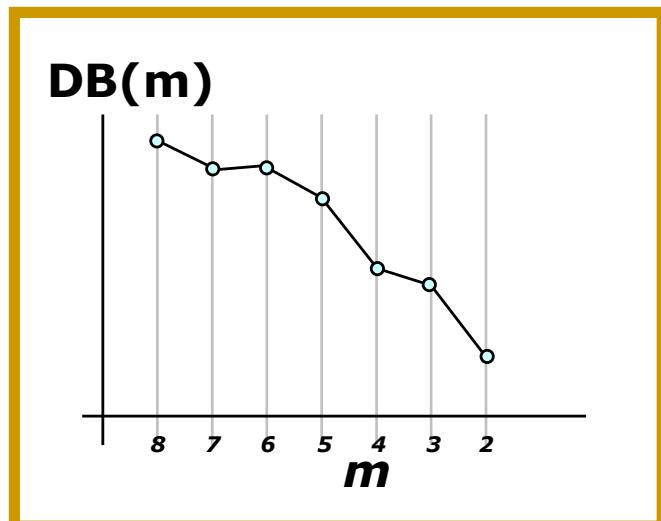
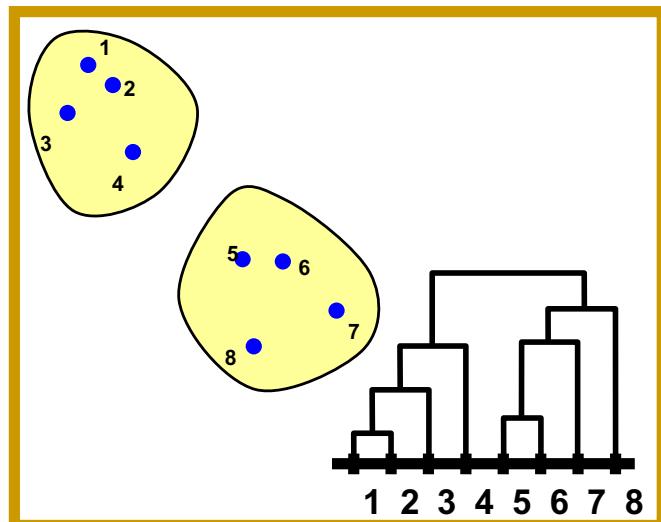
If data (\bullet) clusters frequently with random data (\circ) then data structureless

- **Approach**

- Generate random vectors (\mathbf{Y}) uniformly over observed region of data (\mathbf{X})
- Find MST (single linkage HC) of $\mathbf{X} \vee \mathbf{Y}$
- Determine number of edges q that connect vectors of \mathbf{X} with \mathbf{Y}
- If \mathbf{X} contains clusters q should be small!

(multiple random vs random measurements gives likelihood for q)

Davis-Bouldin index



- **Test**

Select specific clustering according to a criteria

For example: Davis-Bouldin index

- **DB index**

For a specific clustering m , $DB(m)$:
Average similarity of a cluster with its
most similar cluster

- **Approach**

Goal: Clusters to have minimal similarity

Seek: Clustering that minimize $DB(m)$ wrt m

Davis-Bouldin index

- **Similarity cluster C_i and C_j**

$$D_{i,j} = \frac{(\bar{d}_i + \bar{d}_j)}{\|\mu_i - \mu_j\|}$$

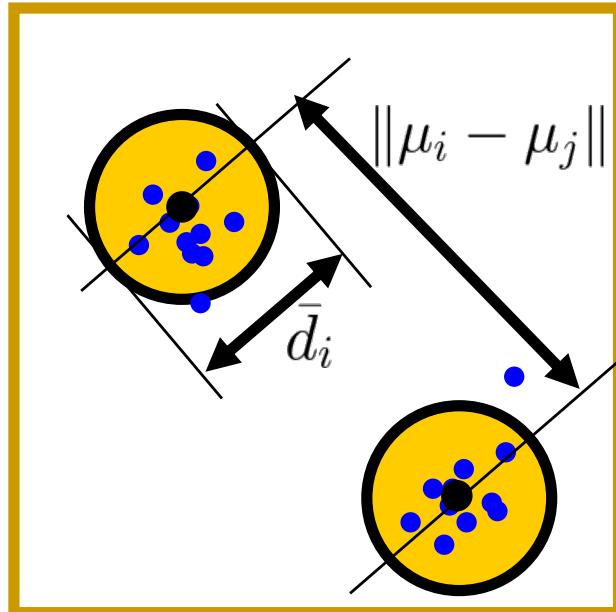
- \bar{d}_i : average distance within cluster i , μ_i : centroid of cluster i

- **Most similar cluster to C_i**

$$R_{i,j} = \max_{j \neq i} \{ D_{i,j} \}$$

- **DB index**

$$DB = \frac{1}{k} \sum_{k=1}^k R_{i,j}$$



Silhouette score

- **Measure similarity of object to its own cluster**

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

with $a(i)$ being average distance to all objects in same cluster and $b(i)$ being closest object from all other clusters:

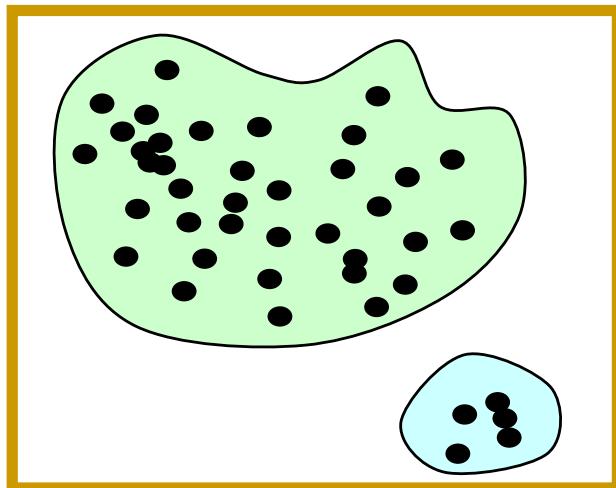
$$a(i) = \frac{1}{|C_i|} \sum_{\forall j} d(x_i, x_j) \quad b(i) = \min_{\forall j, j \notin C_i} d(x_i, x_j)$$

$-1 \leq s(i) \ll 1$; $s(i)$ is close to 1, if $a(i) \ll b(i)$; average distance within cluster much smaller than nearest objects

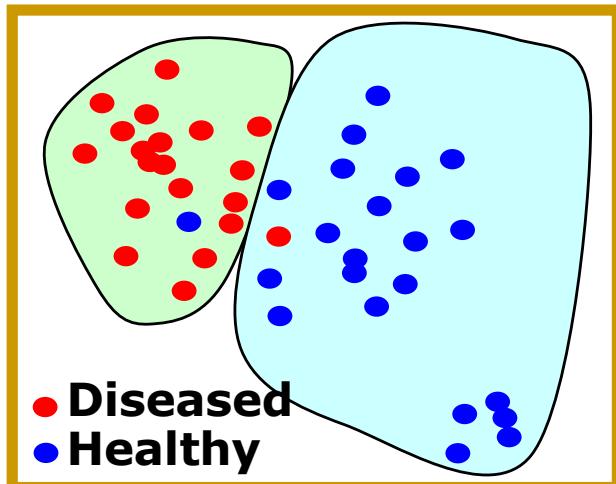
- **Silhouette score is average of all these similarities**

$$S = \frac{1}{N} \sum S(i)$$

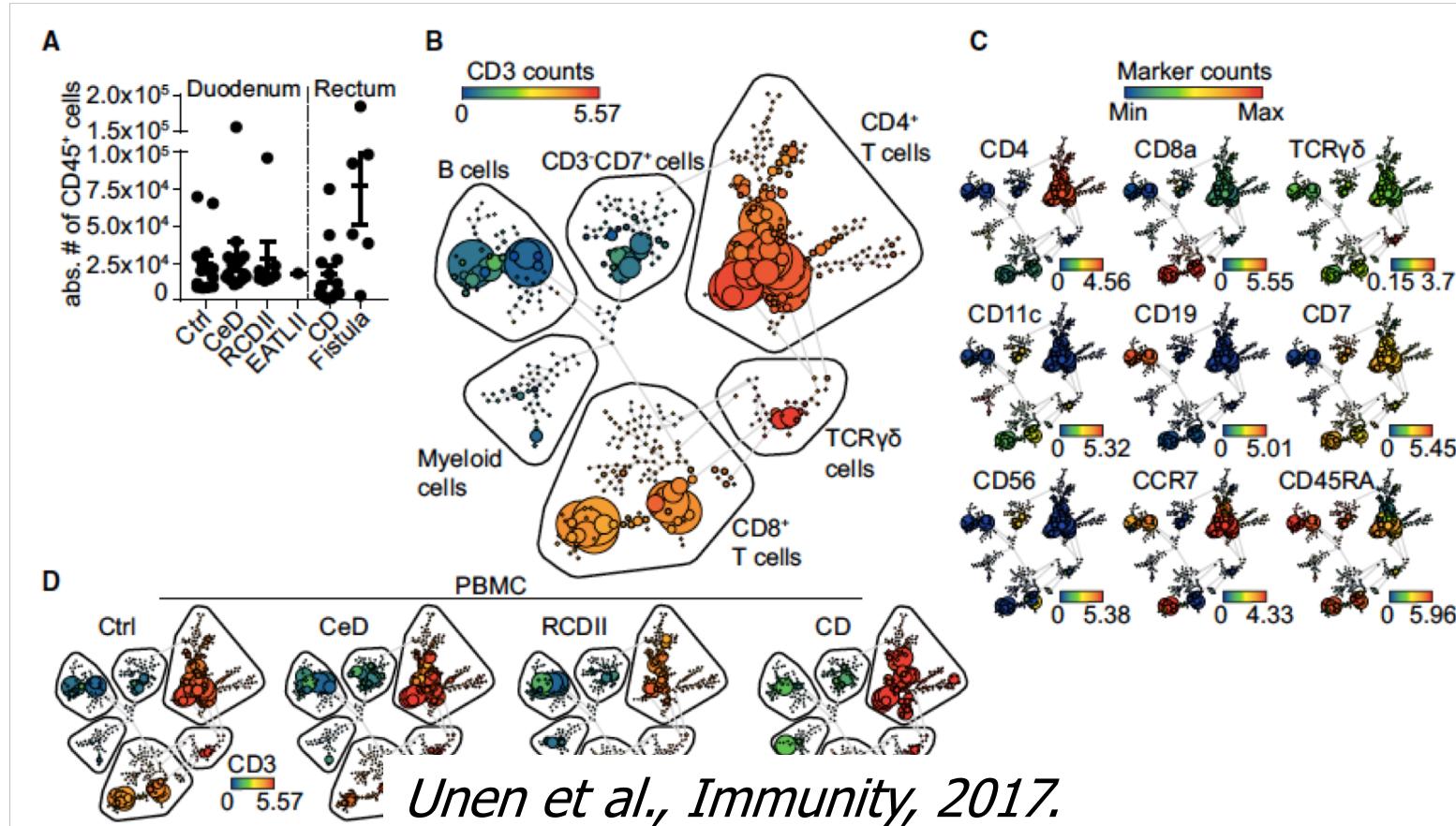
Clustering vs classification



- Machine learning
- Clustering
 - **unsupervised** learning
 - discovering structure/relations
- Classification
 - **Supervised** learning
 - Learning certain behavior
 - Prior information available about different groups



Exploration, clustering



Unen et al., Immunity, 2017.

... mass cytometry to **dissect the human mucosal immune system** ...

... **identify immune subsets** with tissue- and disease-specificity with implications for diagnostic procedures and individualized therapeutics....

KT3701



bioRxiv preprint doi: <https://doi.org/10.1101/113701>; this version posted January 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [aCC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

Clustering: Summary

- Clustering is an unsupervised, iterative process of interpreting data – not proof!
- Cluster results highly depend on the choice of cluster algorithm and dissimilarity measure
- Clustering and classification serve different purposes



Thank you!



References/Reading material

Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., Zhao, Y. *Design and Analysis of DNA Microarray Investigations*. Springer. 2003.

D'haeseleer, P. *How does gene expression clustering work?* Nature Biotechnology, 23, 1499 - 1501 (2005)
<https://www.nature.com/nbt/journal/v23/n12/full/nbt1205-1499.html>

Ringer M. *What is principal component analysis?* Nature Biotechnology, 26, 303 - 304 (2008)
<https://www.nature.com/nbt/journal/v26/n3/full/nbt0308-303.html>