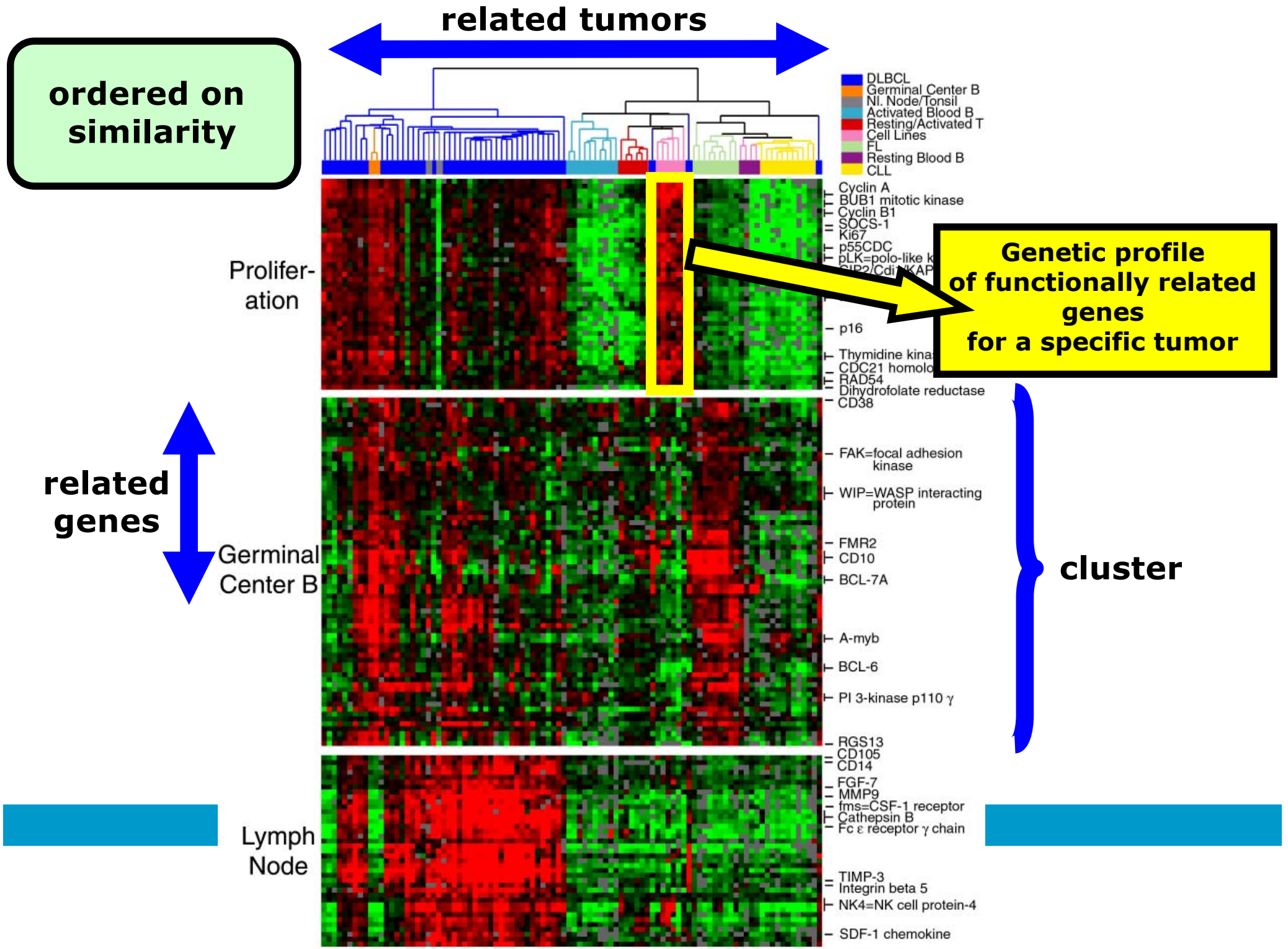
Clustering

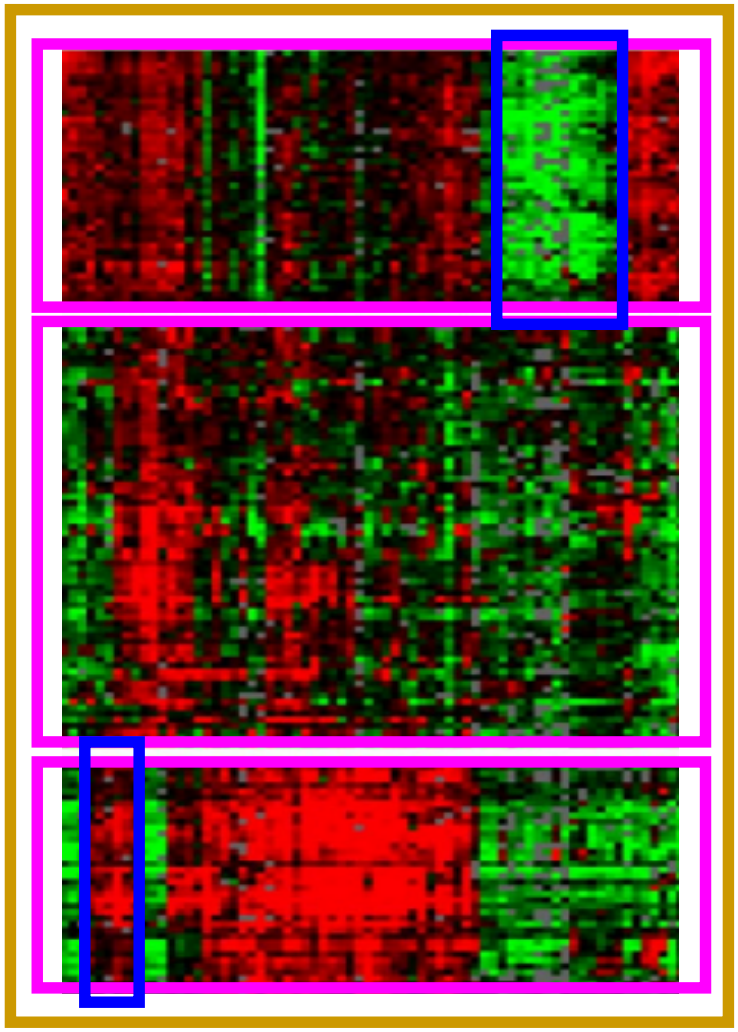
Marcel Reinders

TU Delft
Delft Bioinformatics Lab
Faculty of Electrical Engineering, Computer Science and Mathematics
Delft University of Technology

TU Delft
Delft University of Technology

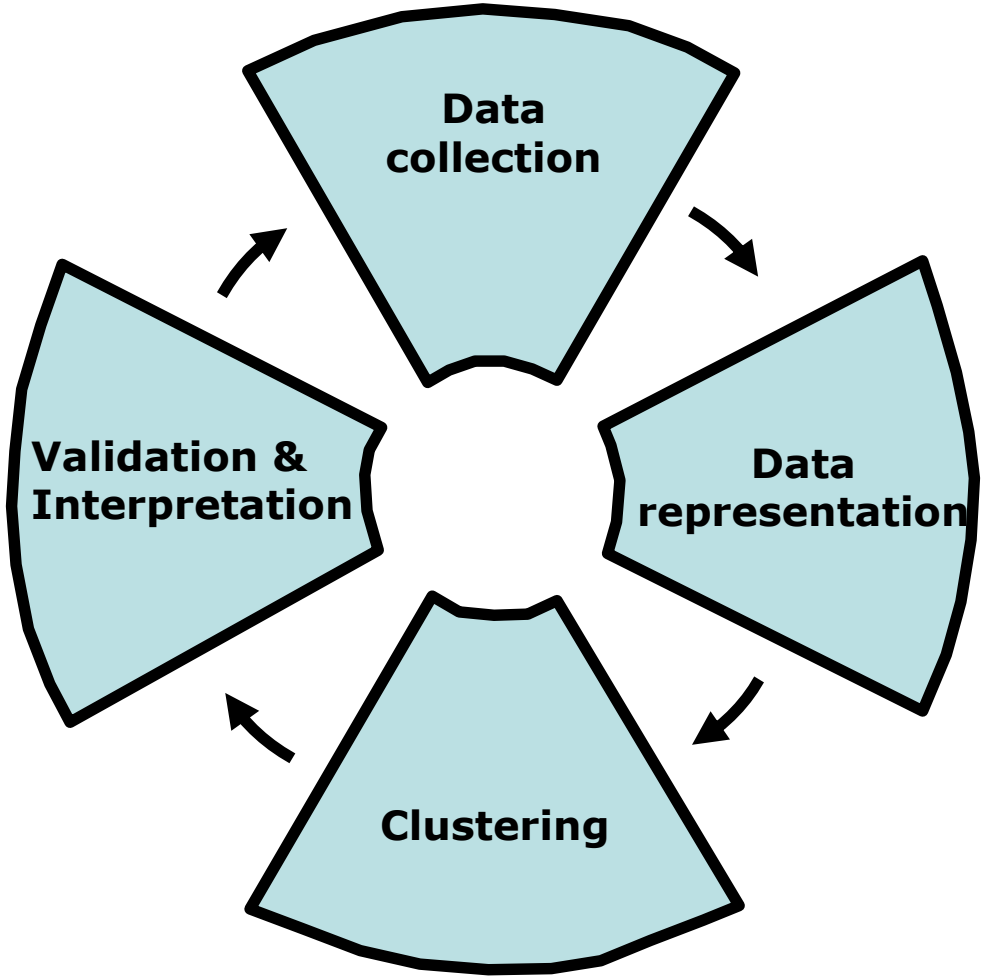

related tumors
ordered on similarity
DLBCL
Germinal Center B
NI. Node/Tonsil
Activated Blood B
Resting/Activated T
Cell Lines
FL
Resting Blood B
CLL
Cyclin A
BUB1 mitotic kinase
Cyclin B1
SOCS-1
Ki67
p55CDC
Proliferation
p16
Thymidine kina
CDC21 homolo
RAD54
Dihydrofolate reductase
CD38
Genetic profile of functionally related genes for a specific tumor
related genes
FAK=focal adhesion kinase
WIP=WASP interacting protein
FMR2
CD10
BCL-7A
Germinal Center B
A-myb
BCL-6
PI 3-kinase p110 γ
cluster
RGS13
CD105
CD14
FGF-7
MMP9
fms=CSF-1 receptor
Cathepsin B
Fc ε receptor γ chain
Lymph Node
TIMP-3
Integrin beta 5
NK4=NK cell protein-4
SDF-1 chemokine

Cluster analysis

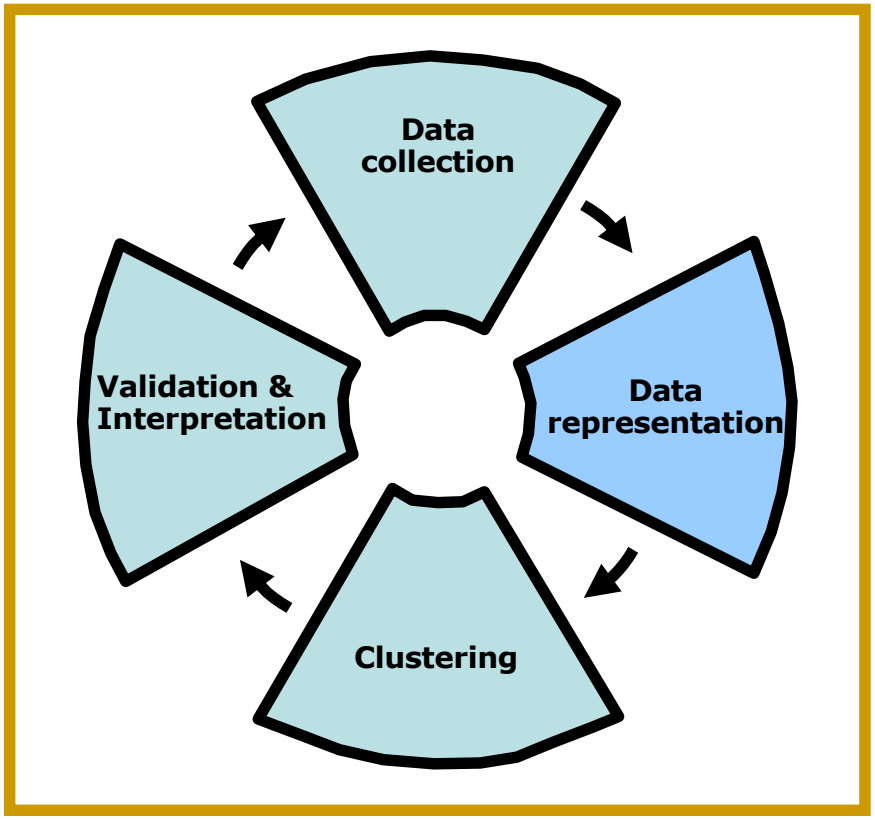


- **Group similar profiles**
- **Finding structure in the data**
- **Requisites:**
  - Measure of similarity
  - Grouping method
- **Subjective measures:**
  - Validation
  - Clustering is a process

Cluster analysis: The process



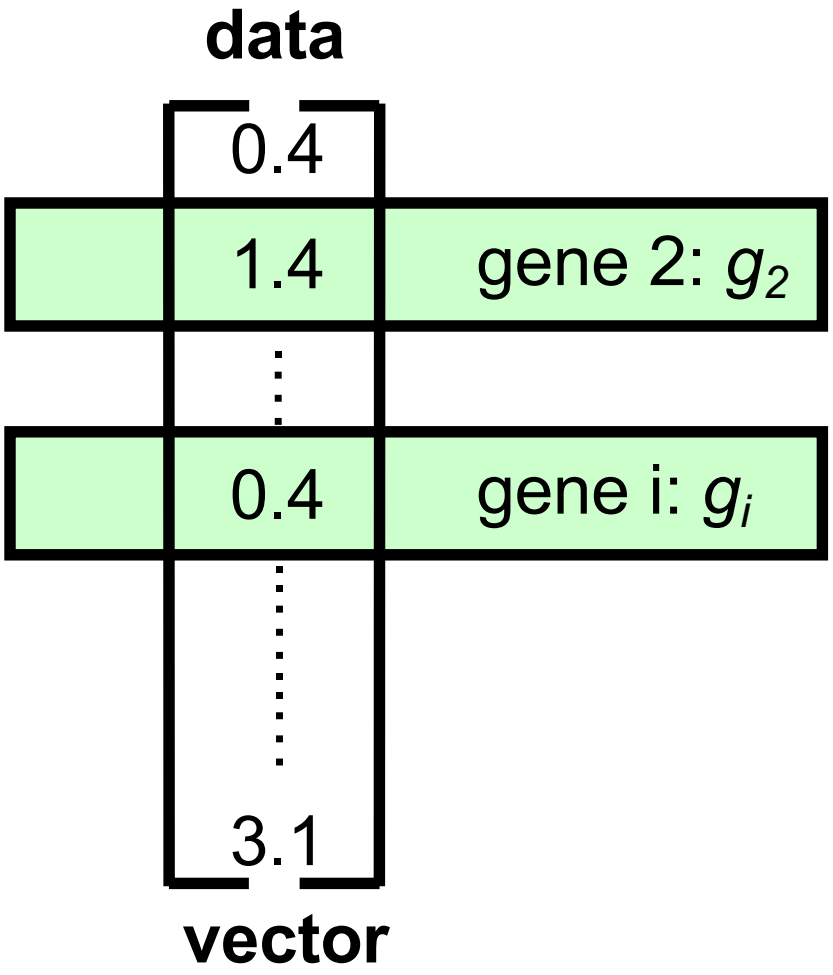
Data representation



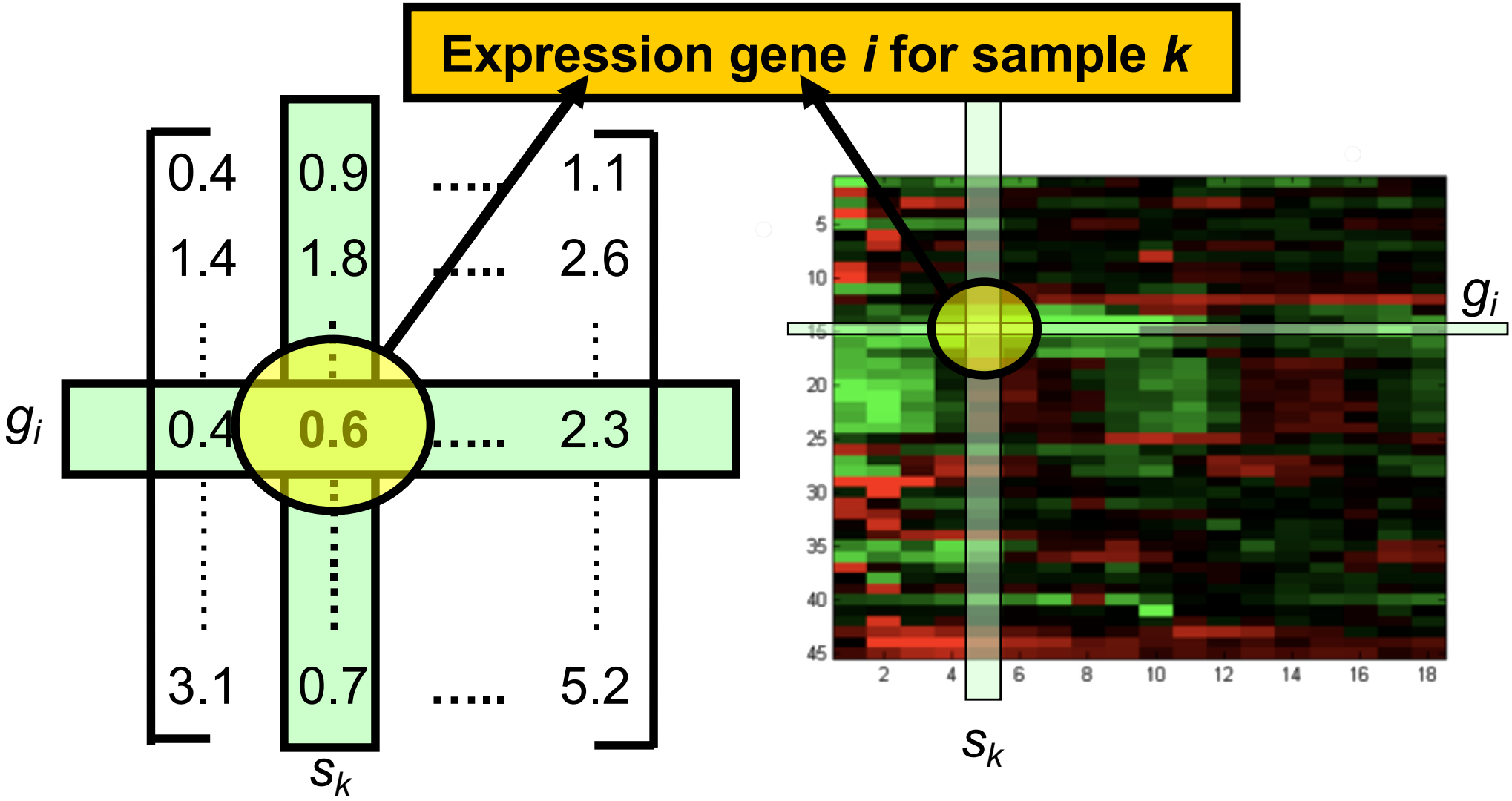
TOPICS

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

Data representation, one sample



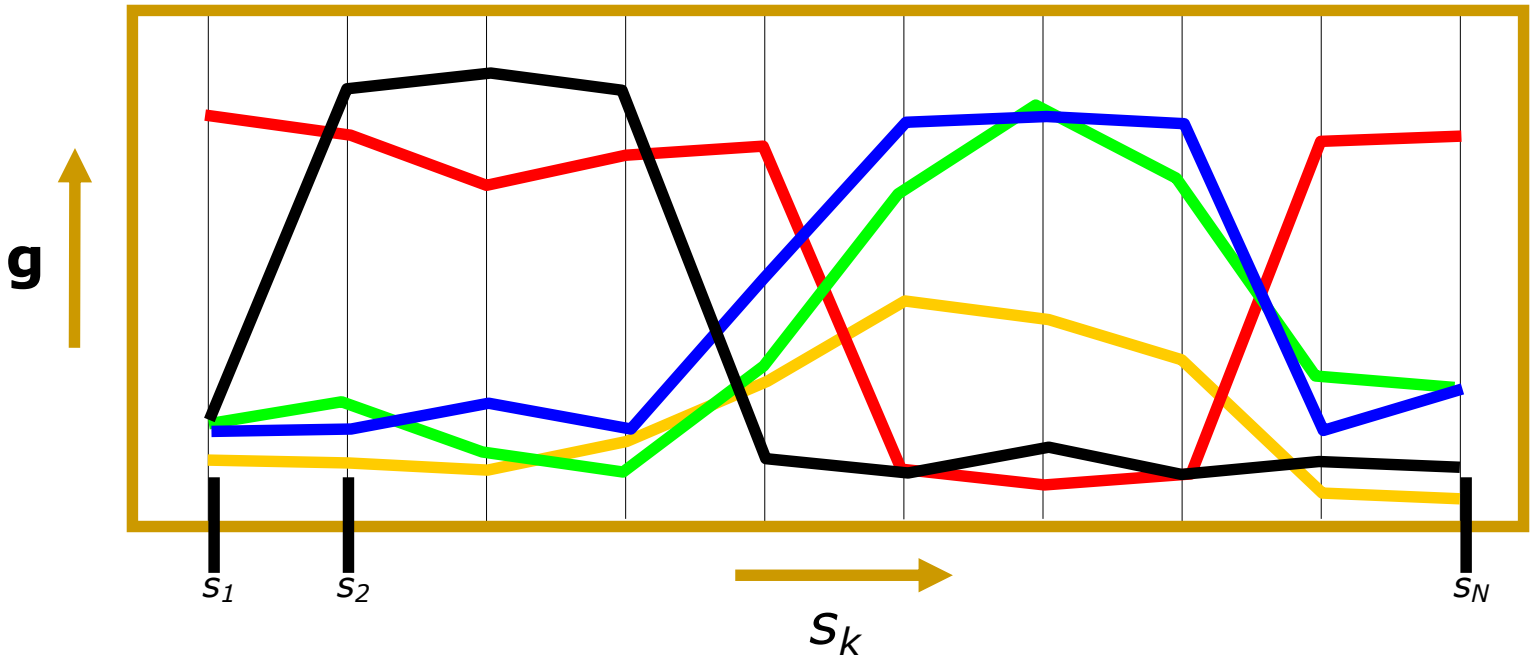
Data representation, multiple samples



The Delft Bioinformatics Lab

TU Delft

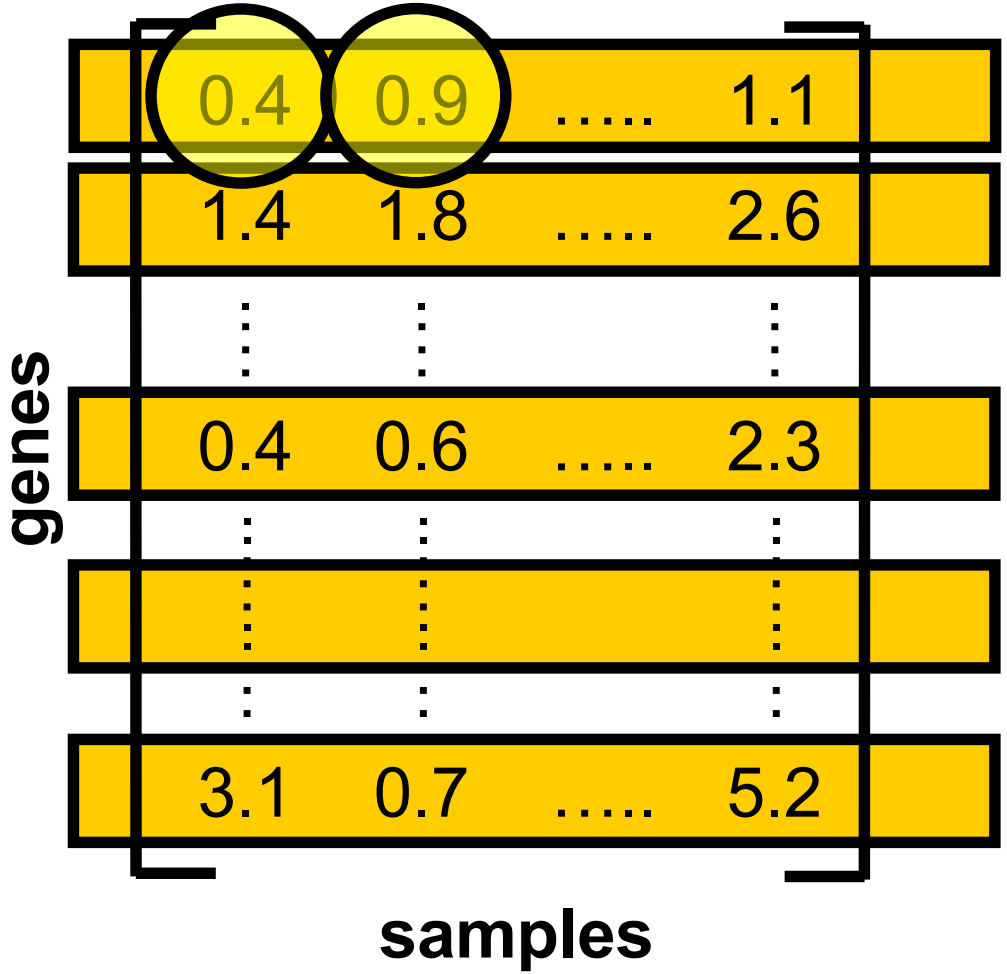
View as profiles

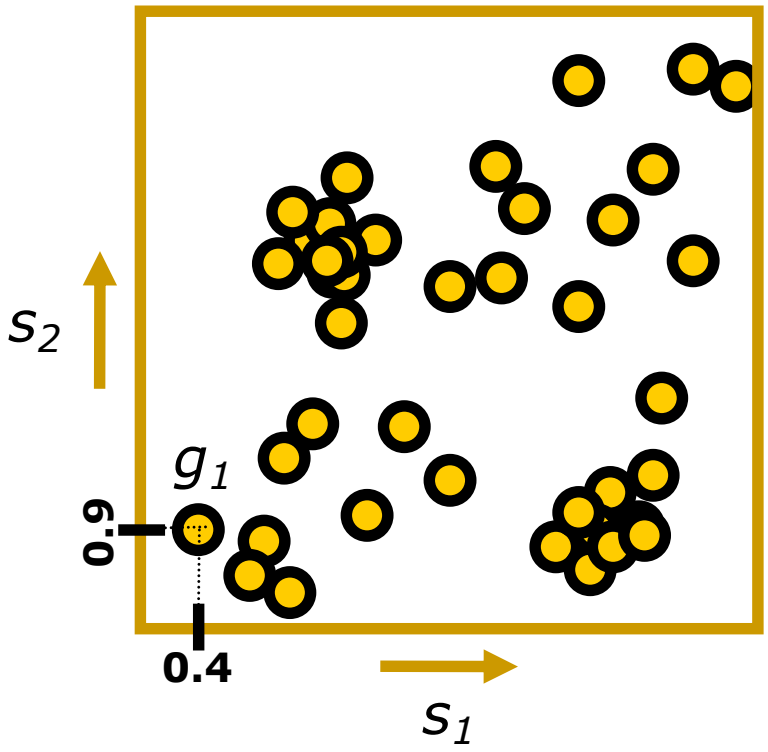


Profile for each gene

Represents gene activity across all samples

Sample-space (1/2)

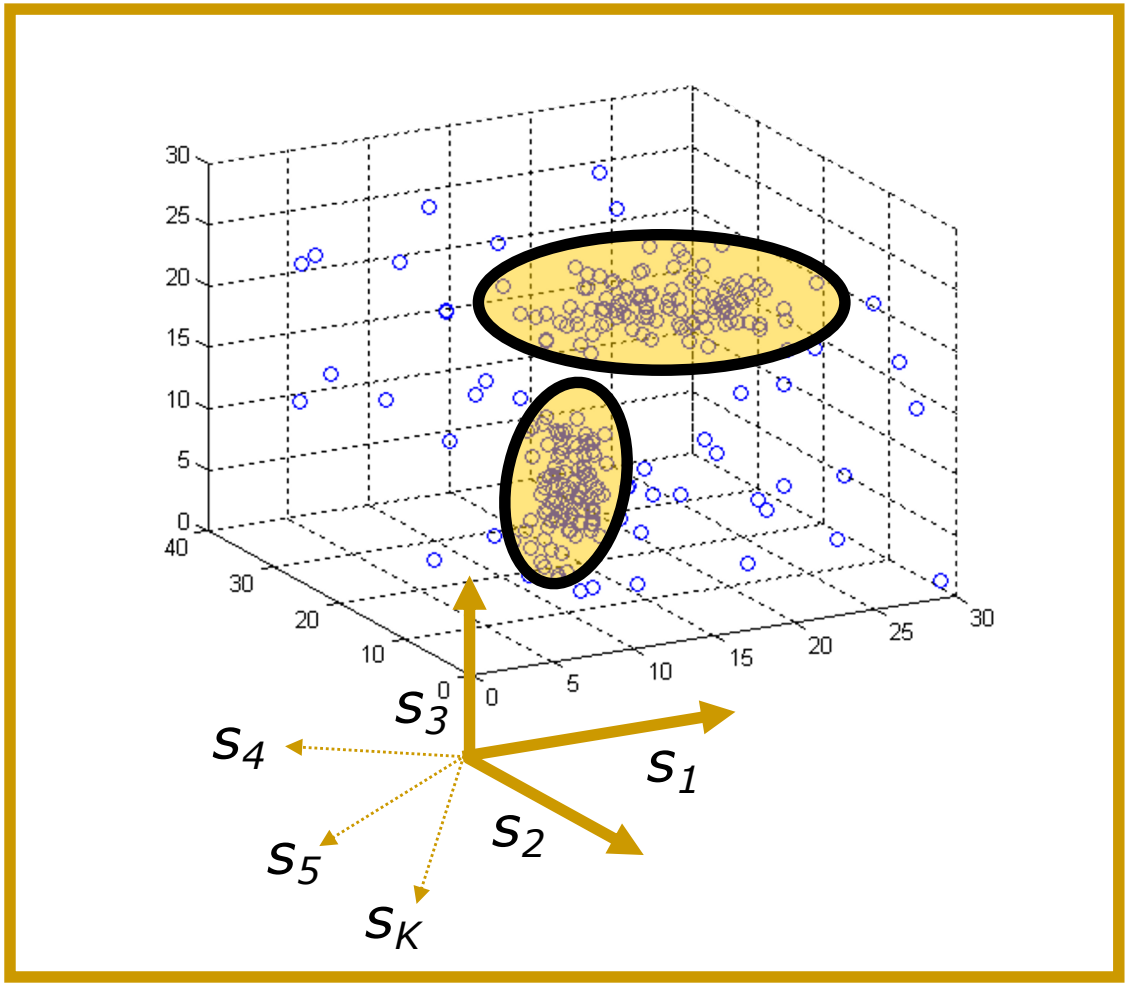




Sample-space

Genes are data points

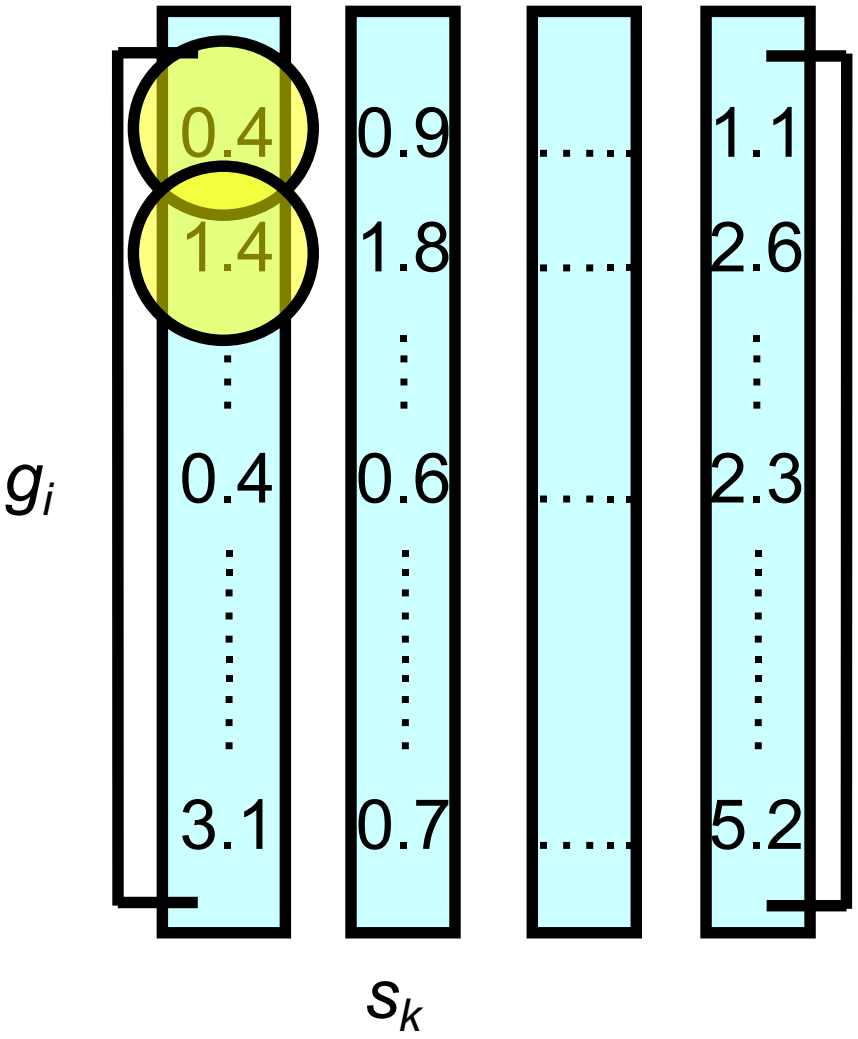
Sample-space (2/2)

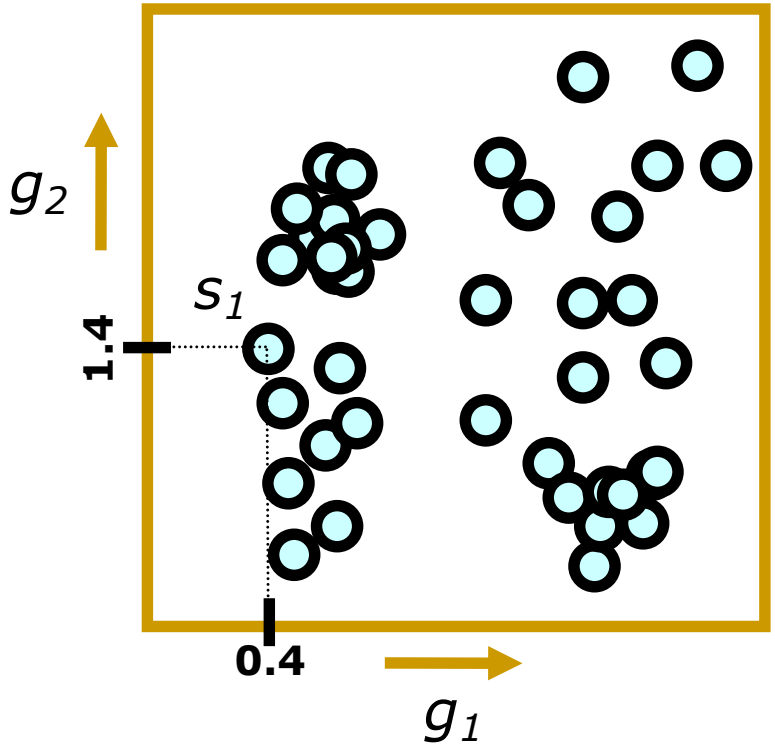


Note!

Sample-space is high dimensional

Gene-space

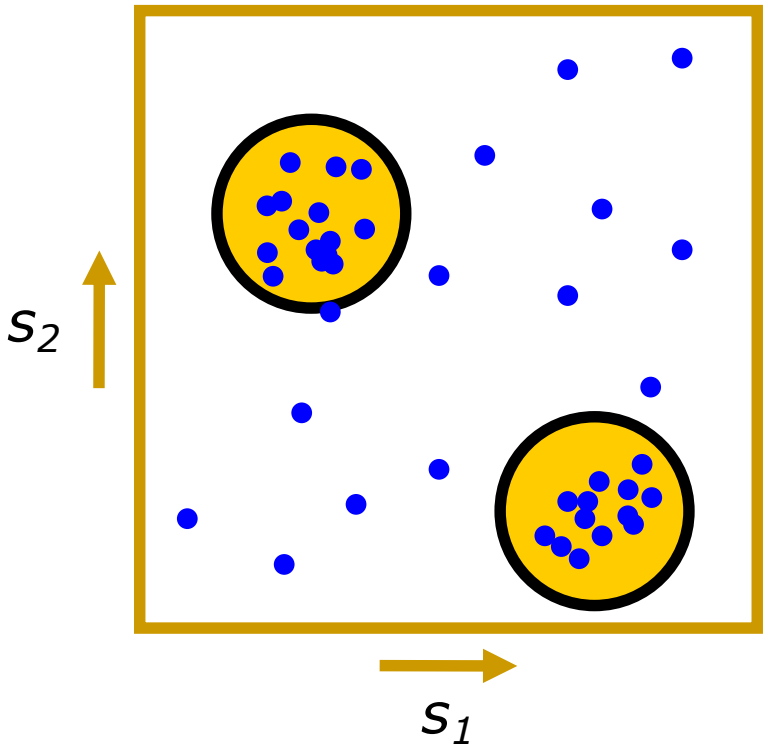




Gene-space

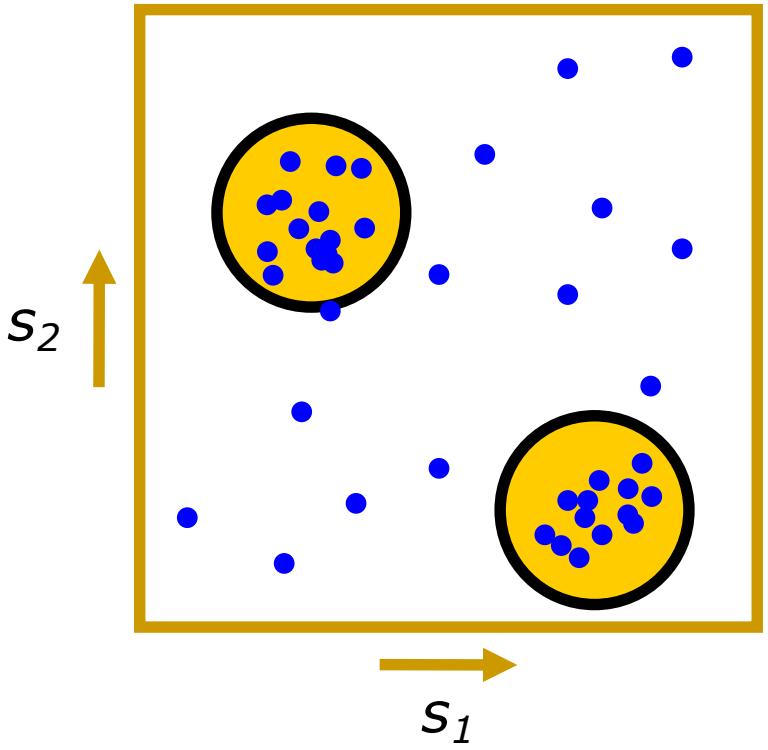
Samples are data points

Genes in *sample-space*

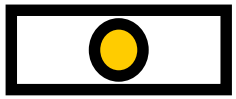


Gene g_i: activity level in Samples s_1 and s_2
(each point represents one gene)

Genes in *experiment-space*



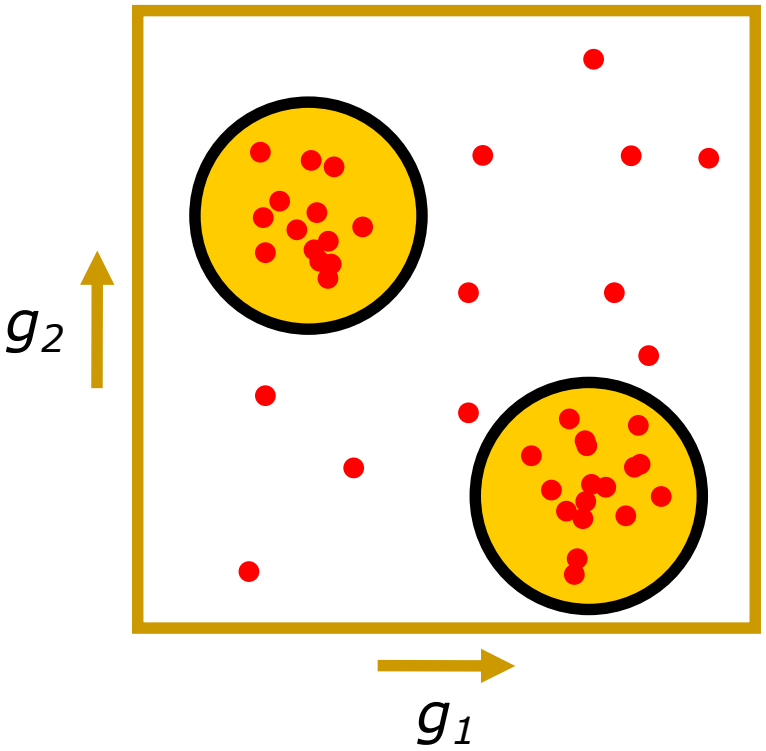
Gene g_i: activity level in samples s_1 and s_2
(each point represents one gene)

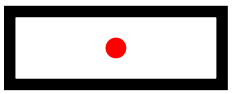


Group of genes that have the same activity levels across"all" samples

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

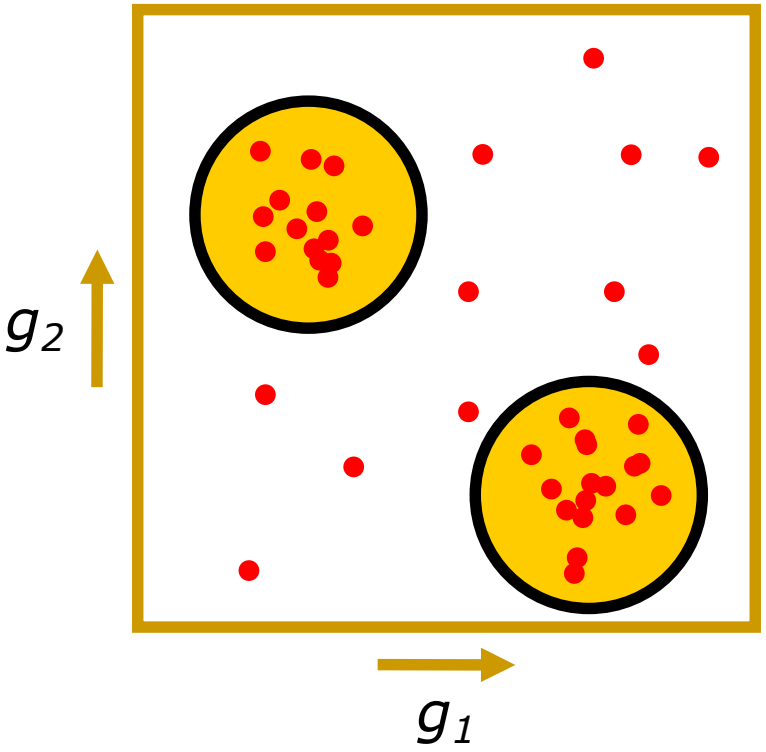
Samples in *gene-space*

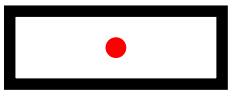




Sample s_k: activity of genes g_1 and g_2
(each point represents one sample)

Samples in *gene-space*



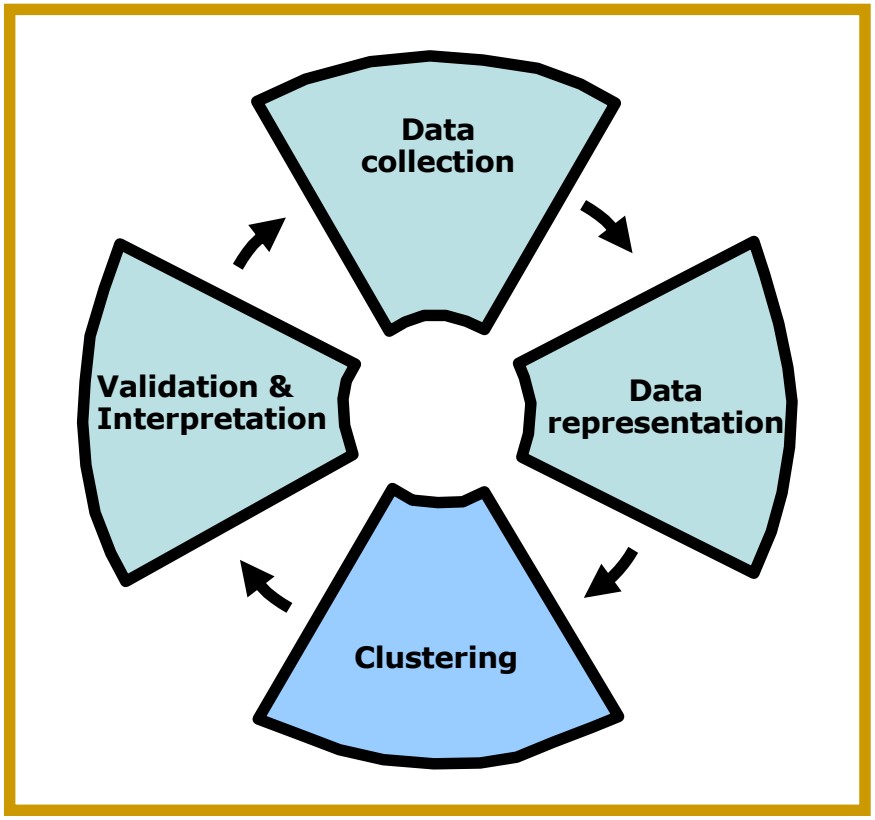


Sample s_k: activity of genes g_1 and g_2
(each point represents one sample)

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

Cluster: Genetic profile for related samples
Characterization of related diseases

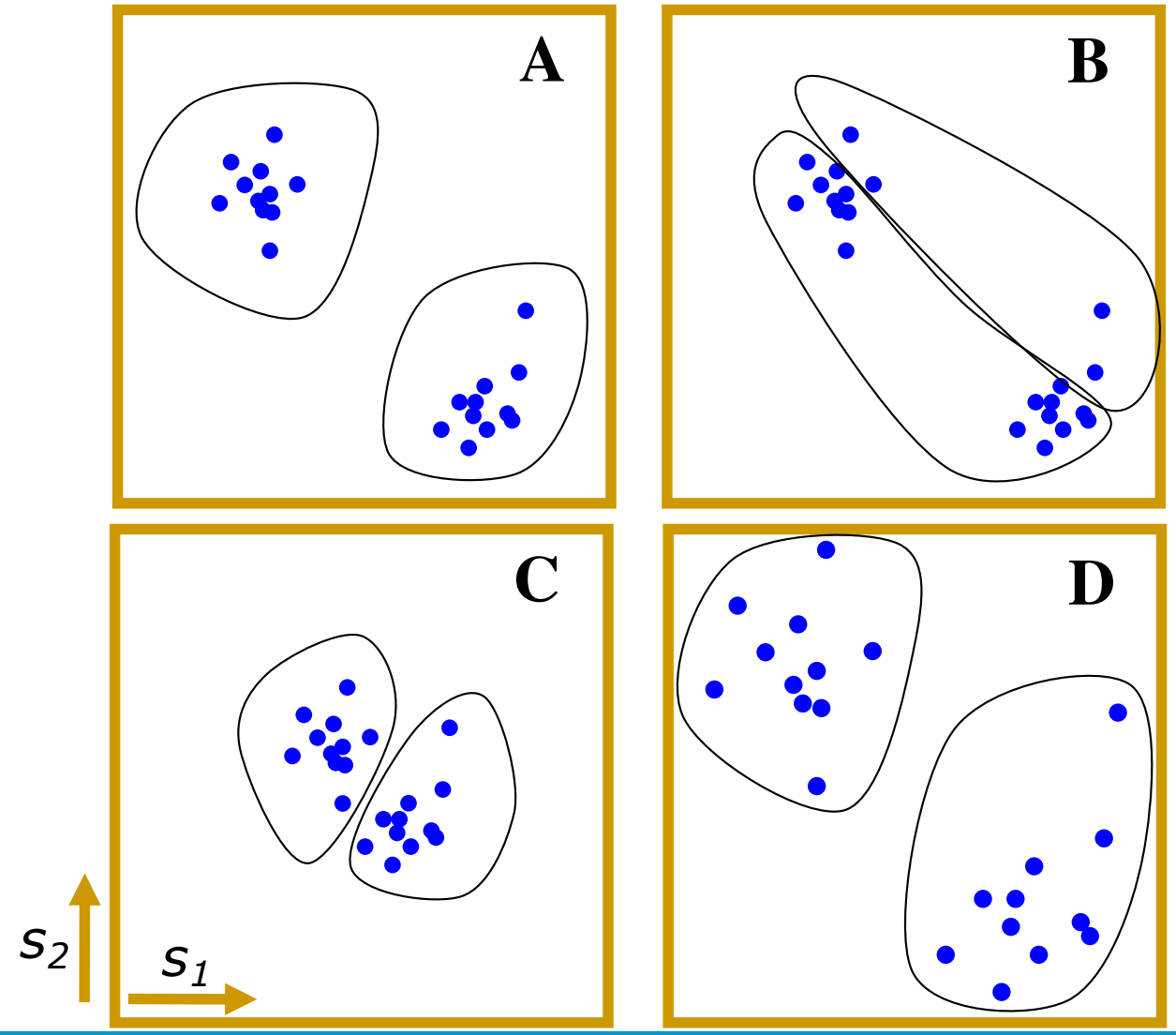
Clustering methods



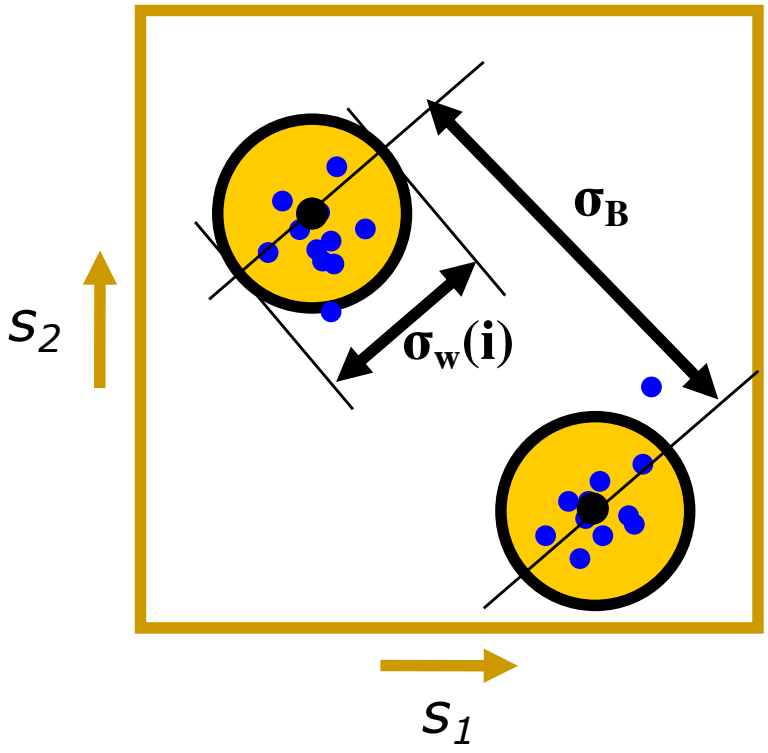
TOPICS

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

Finding structure: better grouping?



Finding structure: better grouping?



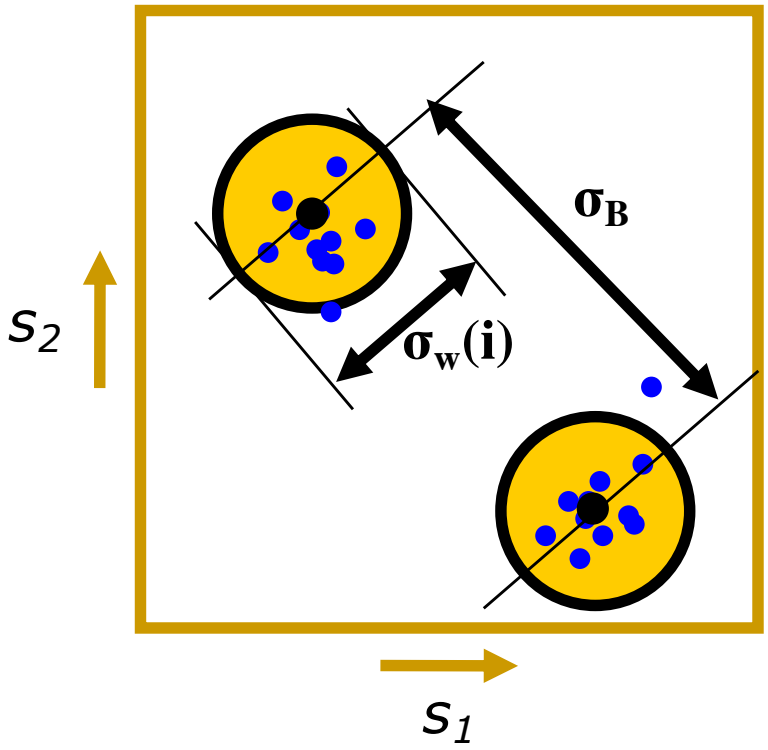
- **Structure when:**
  1) **Points within cluster resemble each other**

     (within variance, $\sigma_W(i)$)
  2) **Clusters deviate from each other *(between variance, σ_B)***

- **Group points such that**

$$\textbf{MIN}\left[\frac{\sum \textbf{within variance}}{\textbf{between variance}}\right] \rightarrow \sigma_W\textbf{: small \&}\ \sigma_B\textbf{ : large}$$

General approaches



- **Agglomerative**

  (building trees)

  hierarchical clustering (Mike Eisen's Cluster), Louvain graph clustering

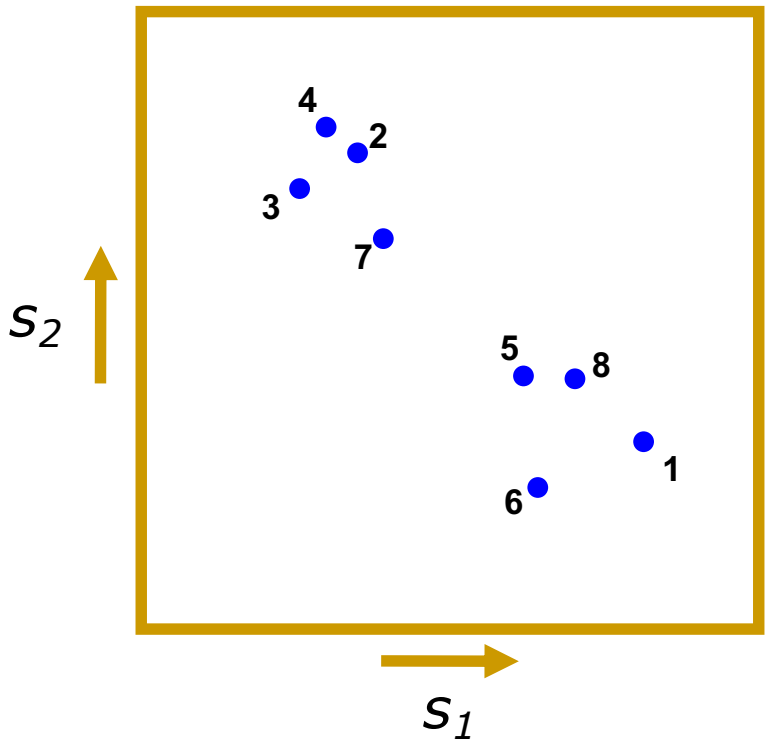
- **Partitional**

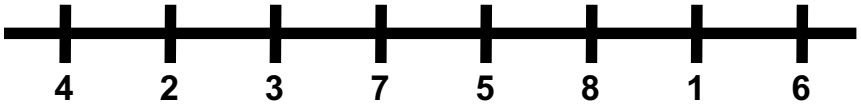
  (finding prototypes)

  k-means, spectral-based graph-based clustering, som-mapping

  gene shaving

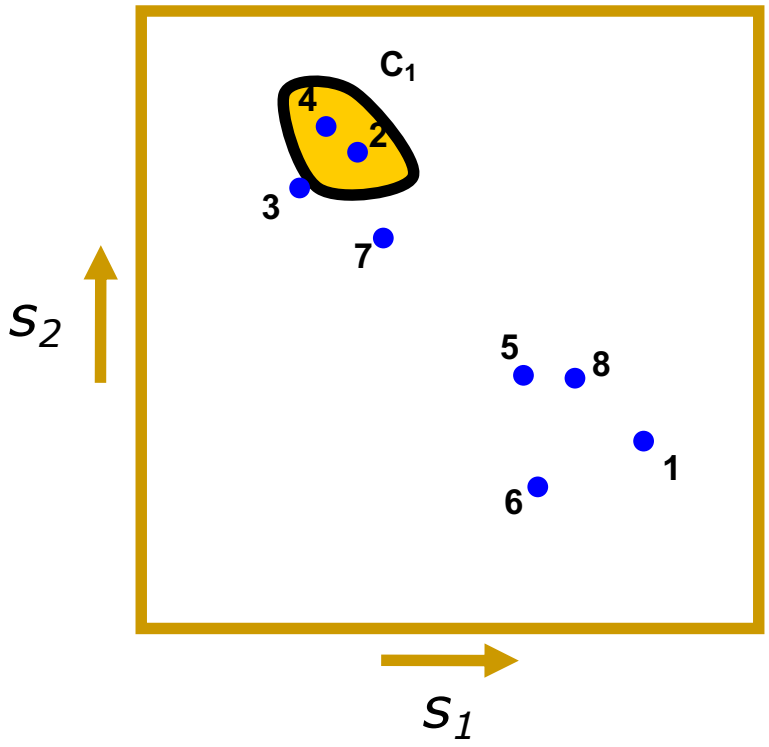
Hierarchical clustering



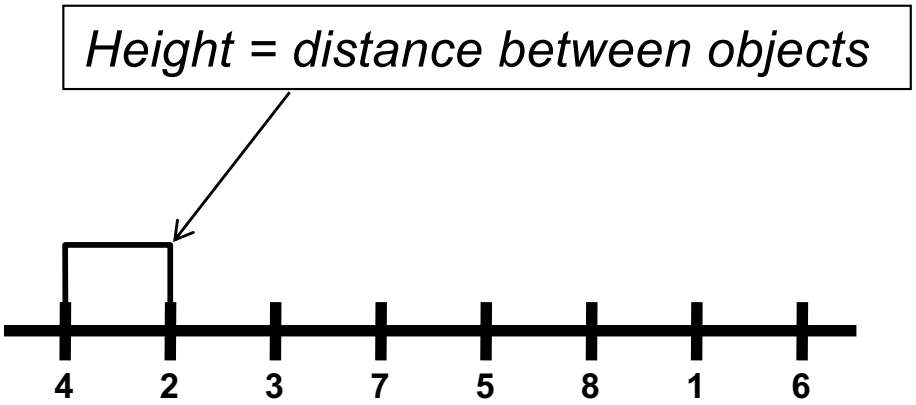


Find most similar objects (points/genes) and group them

Hierarchical clustering



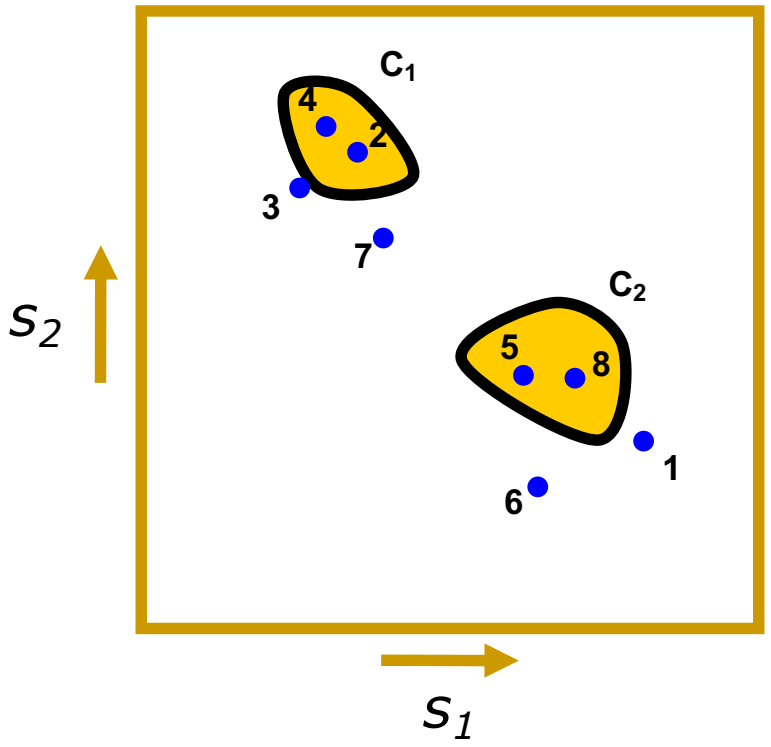
dendrogram

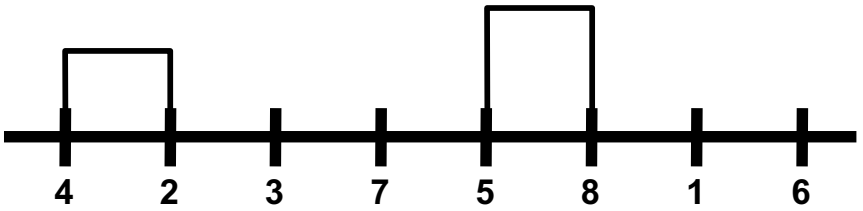


These are: objects 4 and 2
Again, find most similar objects (genes or clusters) and group them

Hierarchical clustering

dendrogram



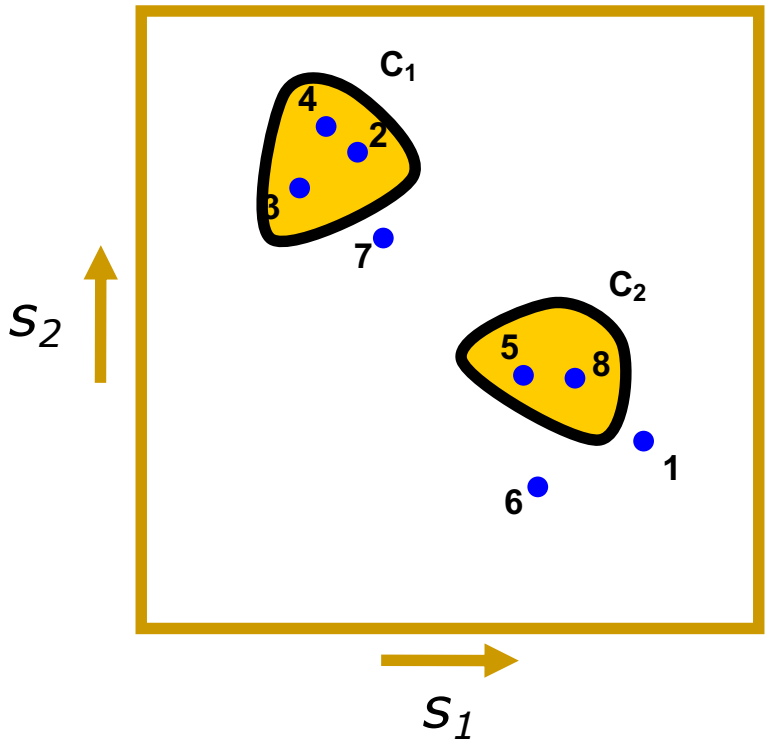


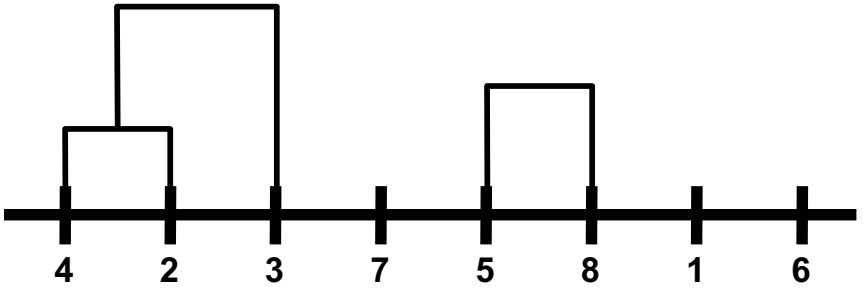
These are: objects 5 and 8

Repeat finding most similar objects (genes or clusters) and grouping them

Hierarchical clustering

dendrogram

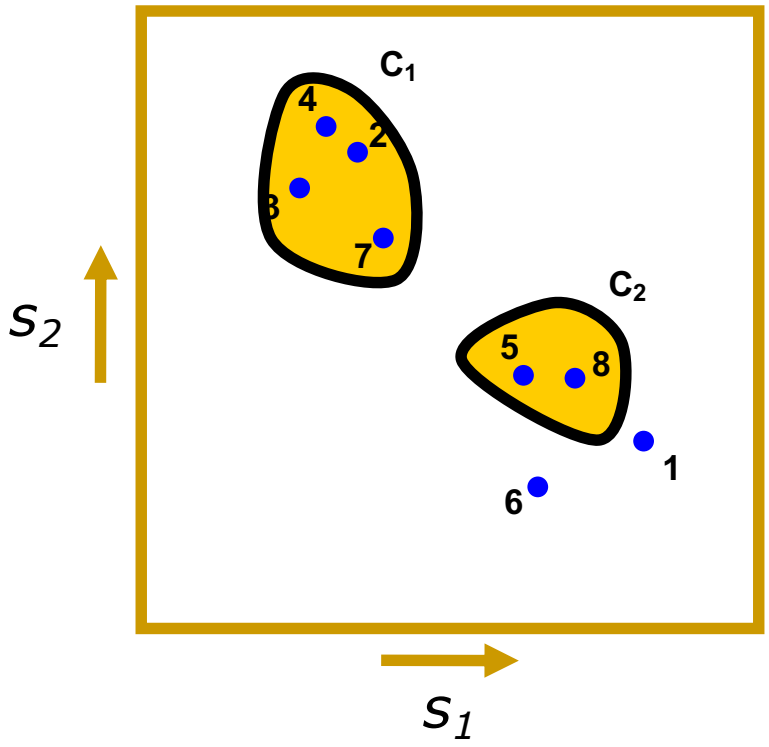


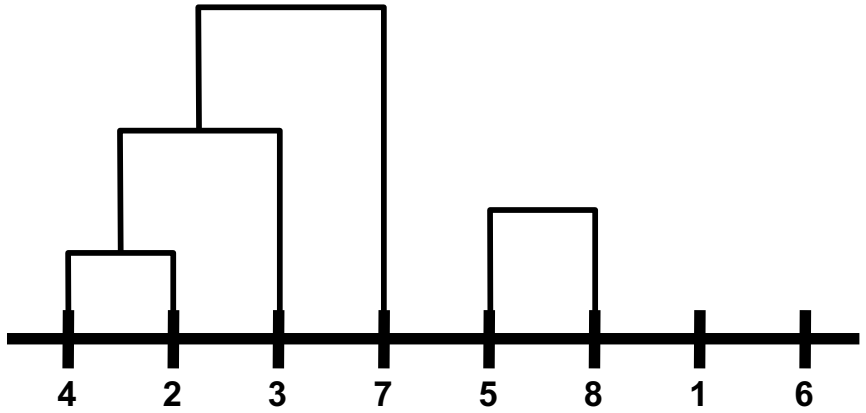


Join object 3 and cluster 1
Repeat process

Hierarchical clustering

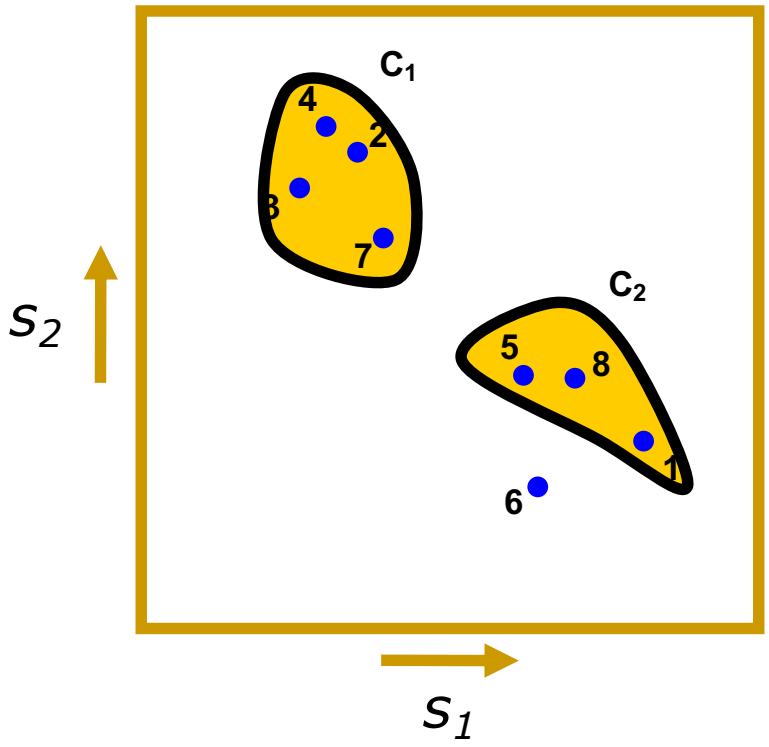
dendrogram



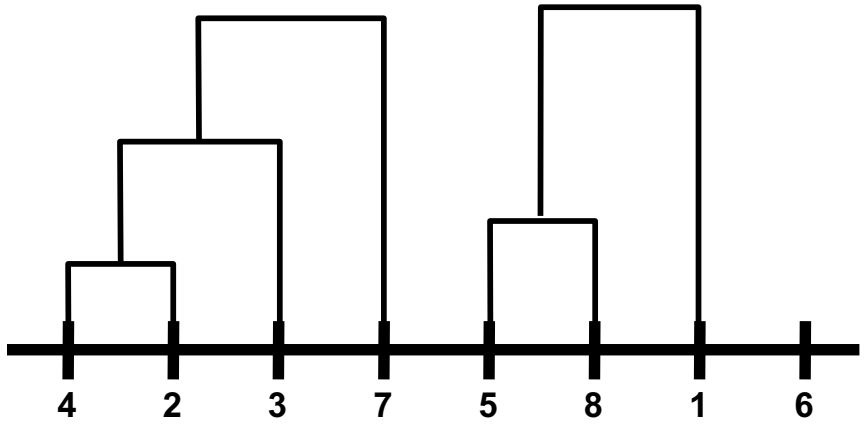


Join [object 7 and cluster 1] into [cluster 1]
Repeat process

Hierarchical clustering

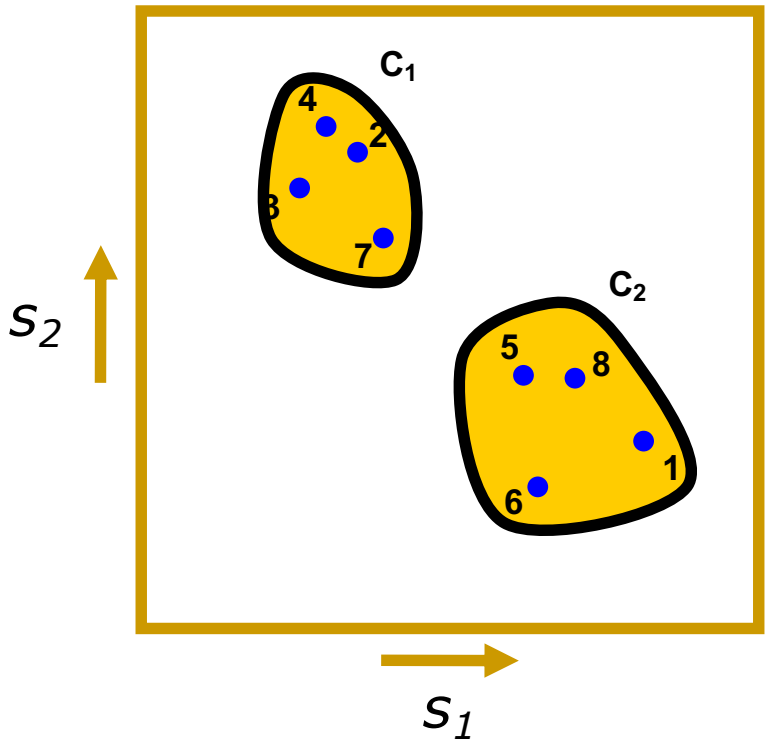


dendrogram

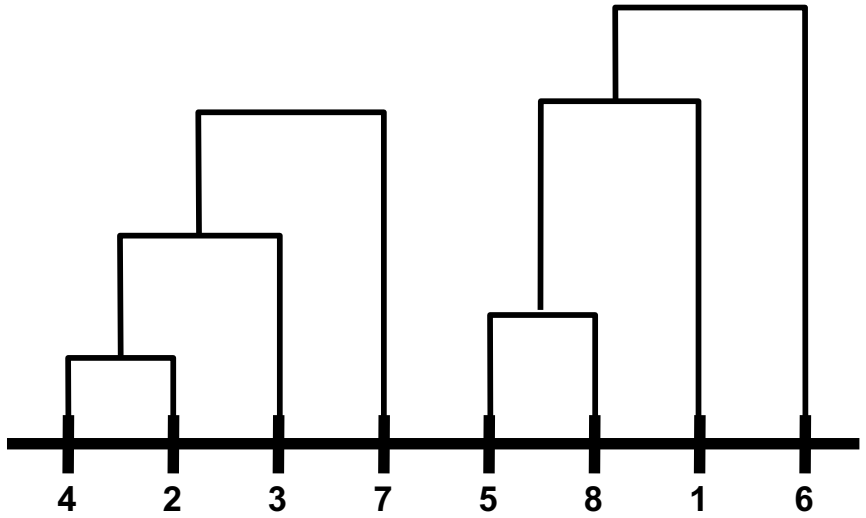


Join [object 1 and cluster 2] ➔ [cluster 2]
Repeat process

Hierarchical clustering

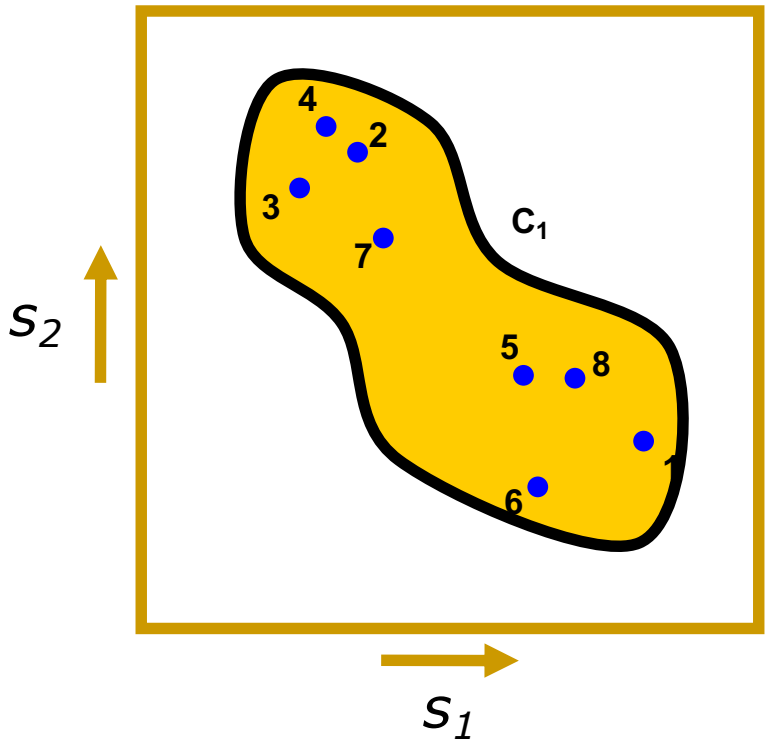


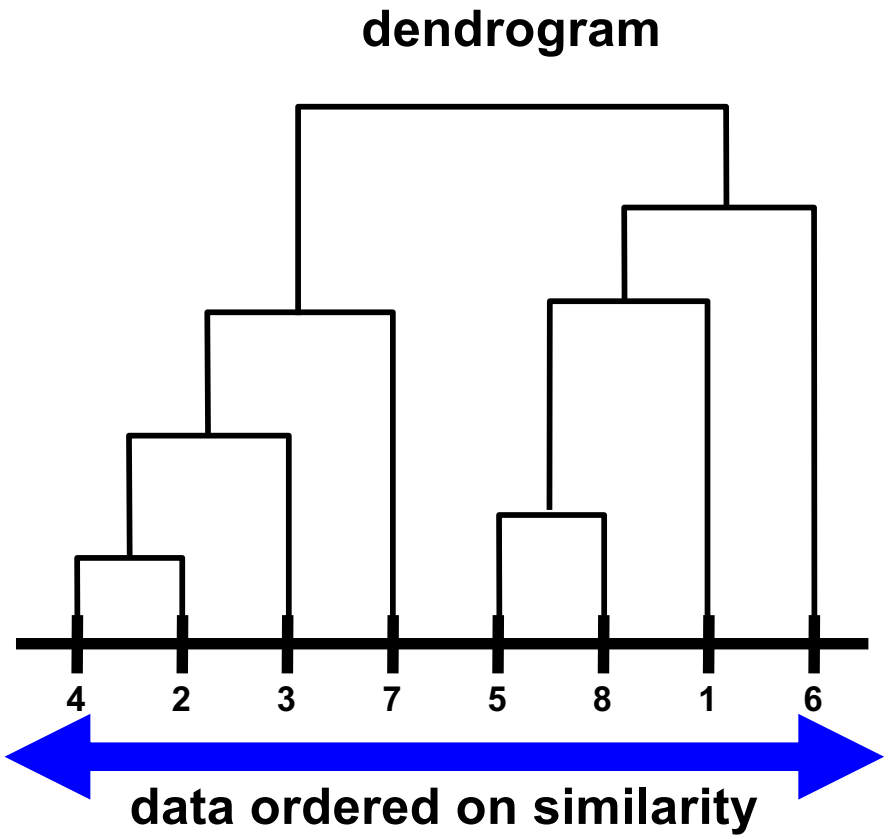
dendrogram



Join [object 6 and cluster 2] ➔ [cluster 2]
Repeat process

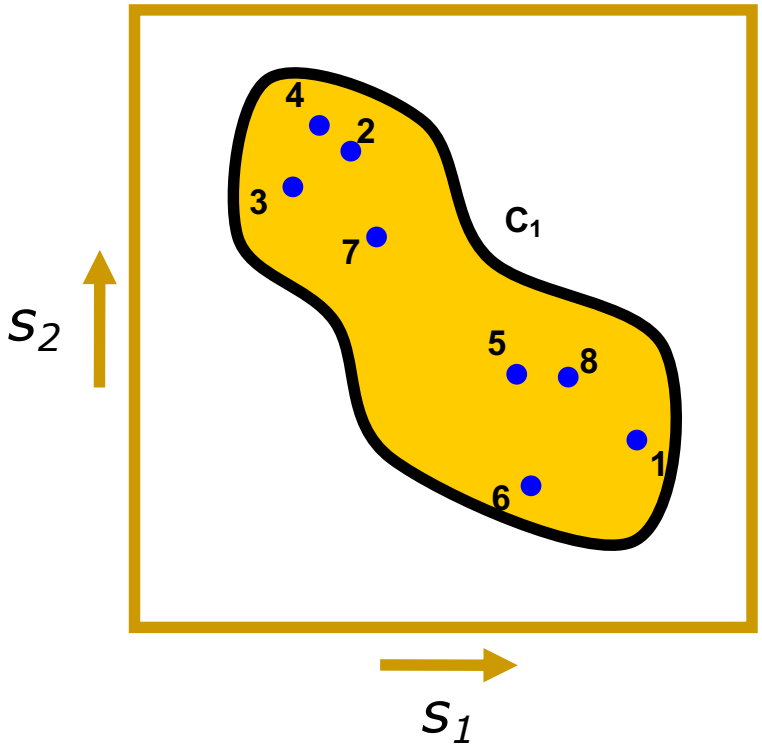
Hierarchical clustering





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

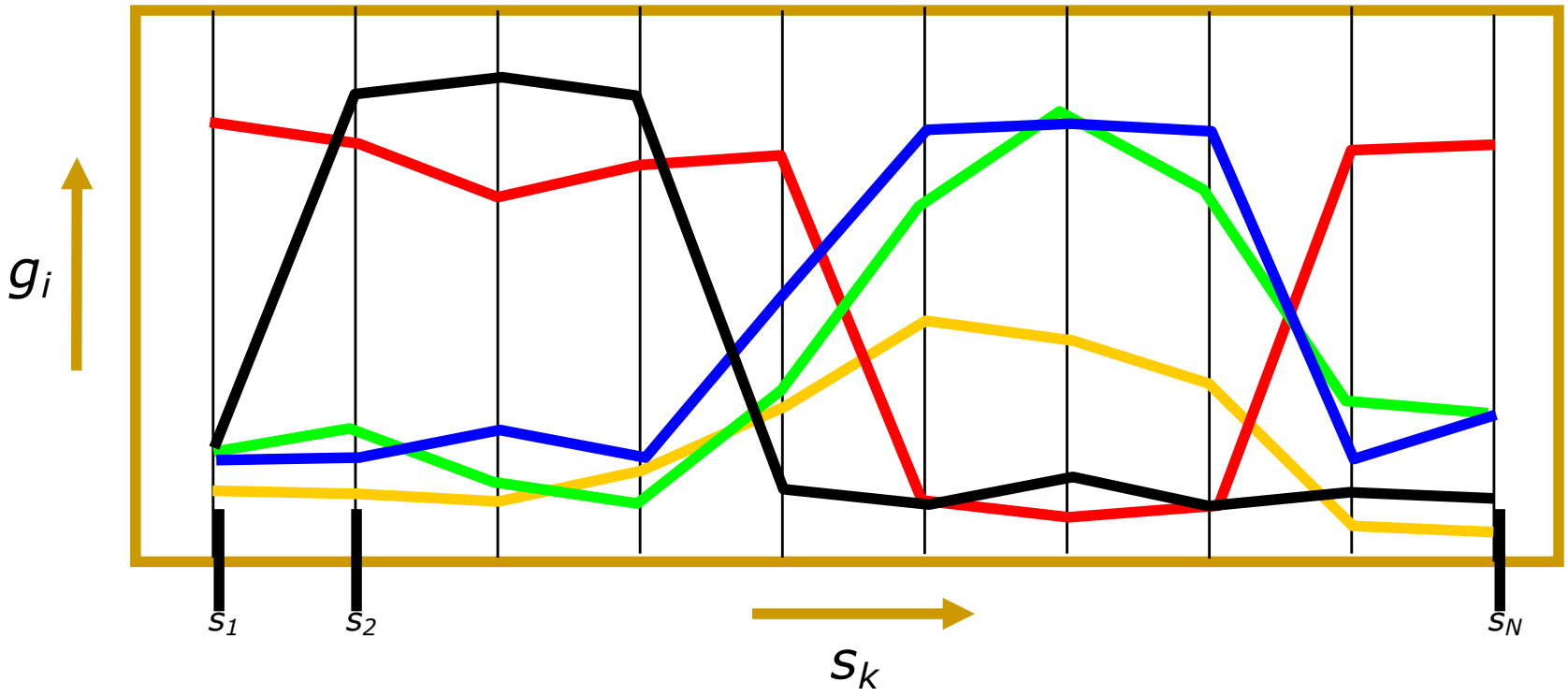
Hierarchical clustering: Choices

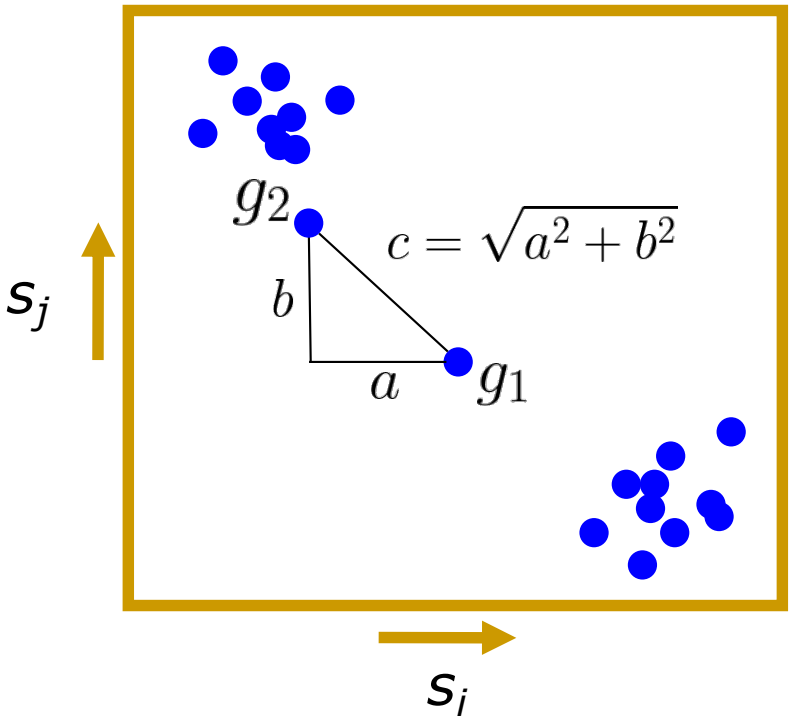


Need to know:

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

Hierarchical clustering: Similarity between objects





Euclidean distance

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

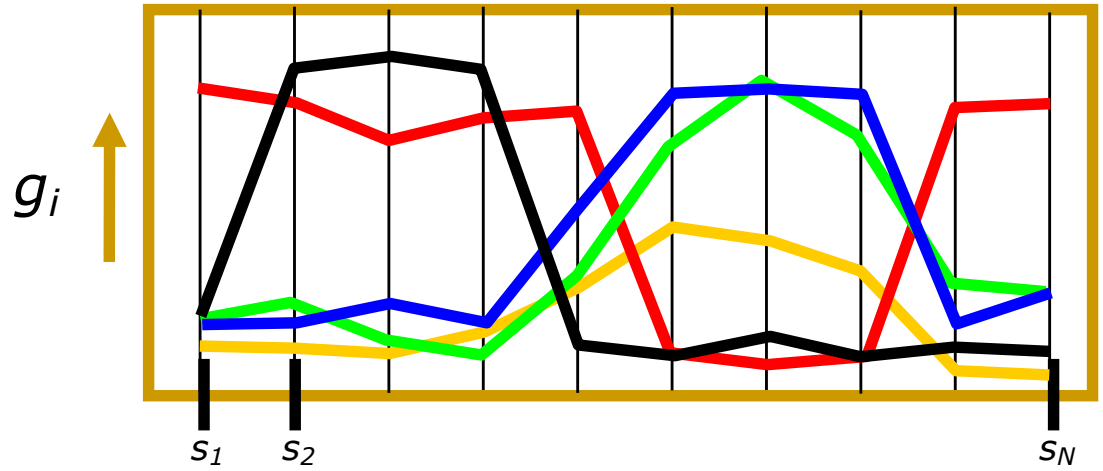
$$b = g_1(s_j) - g_2(s_j)$$

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

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

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

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(●,●) < d(●,●)
d(●,●) << d(●,●)
d(●,●) << d(●,●)

Pearson correlation

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

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

d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)
d(●,●) << d(●,●)

Mixed Pearson correlation

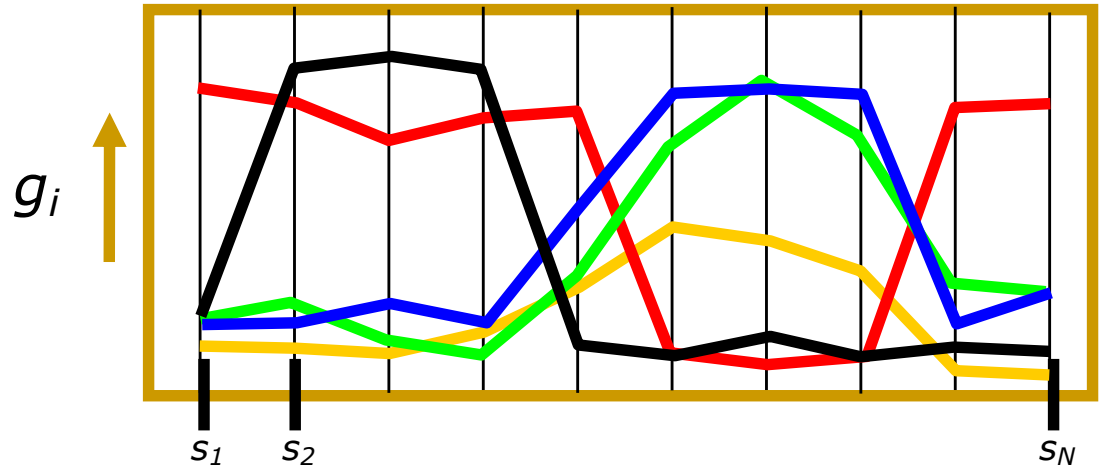
$$1 - |\rho_{g_1,g_2}|$$

d(●,●) ≈ d(●,●)
d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)

The Delft Bioinformatics Lab

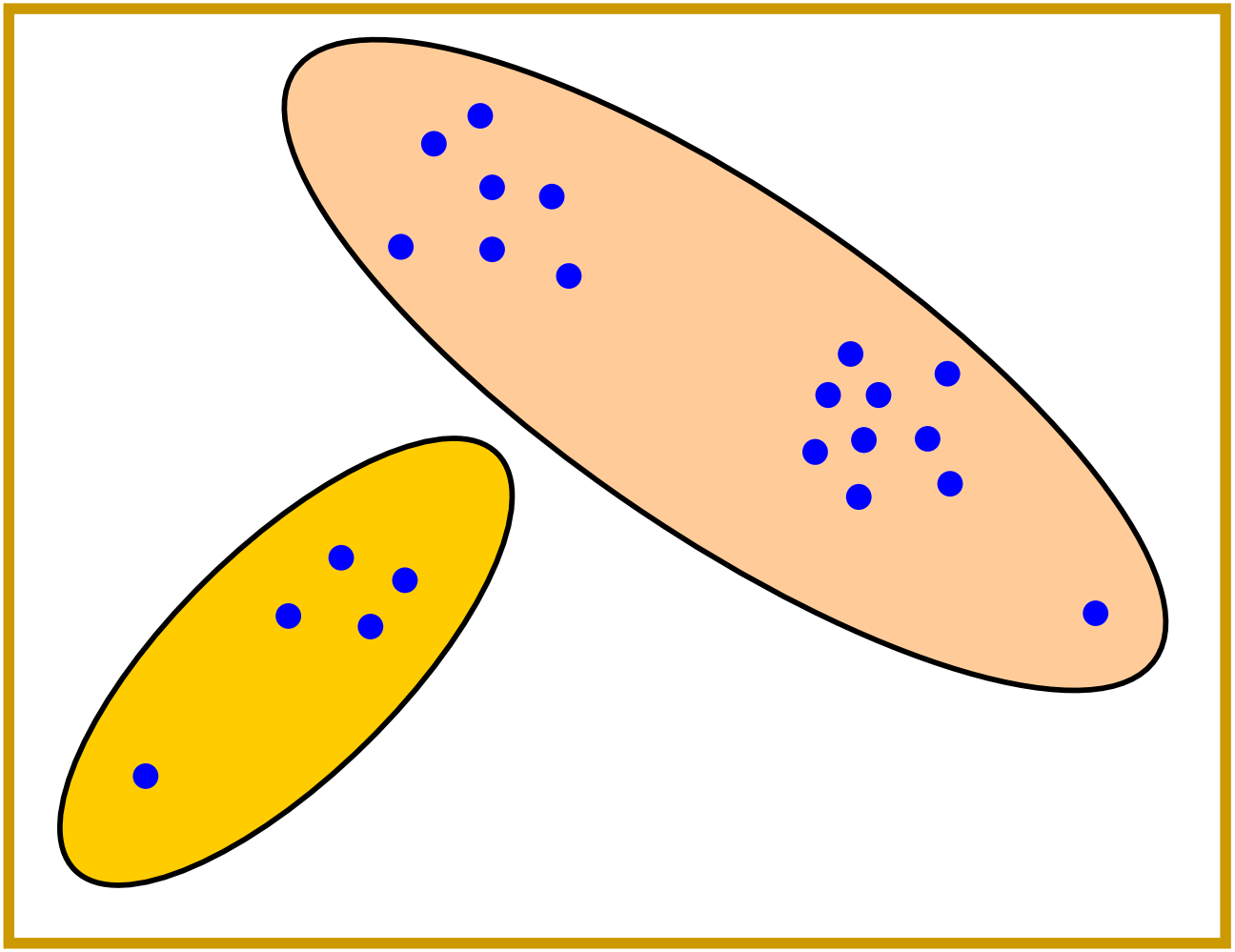
TU Delft

Similarity between objects

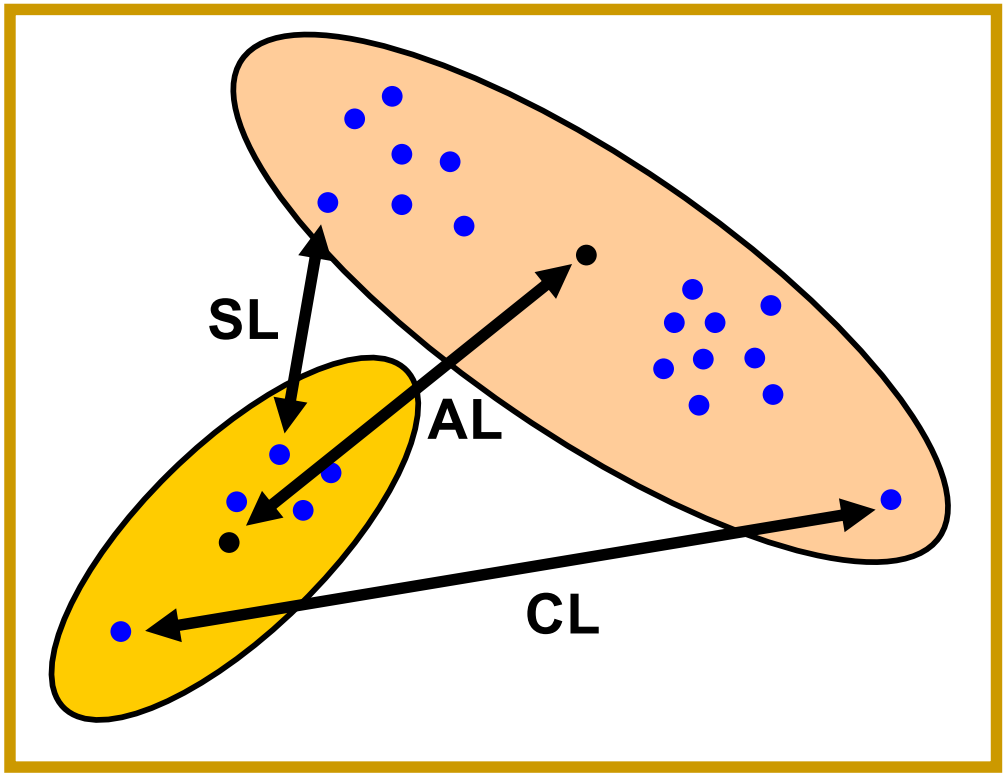


Euclidean distance	Pearson correlation	Mixed Pearson correlation
match exact shape	ignore amplitude	ignore amplitude & sign
d(●,●) < d(●,●)	d(●,●) ≈ d(●,●)	d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)	d(●,●) << d(●,●)	d(●,●) ≈ d(●,●)
d(●,●) << d(●,●)	d(●,●) << d(●,●)	d(●,●) << d(●,●)

Similarity between clusters?

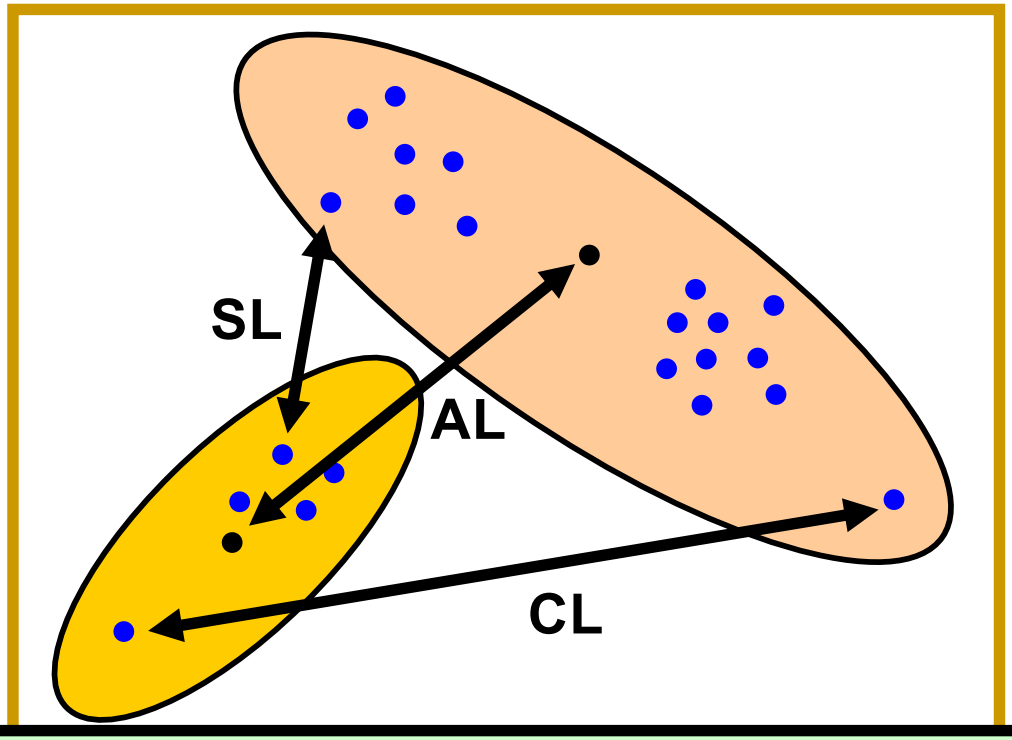


Similarity between clusters (linkage)



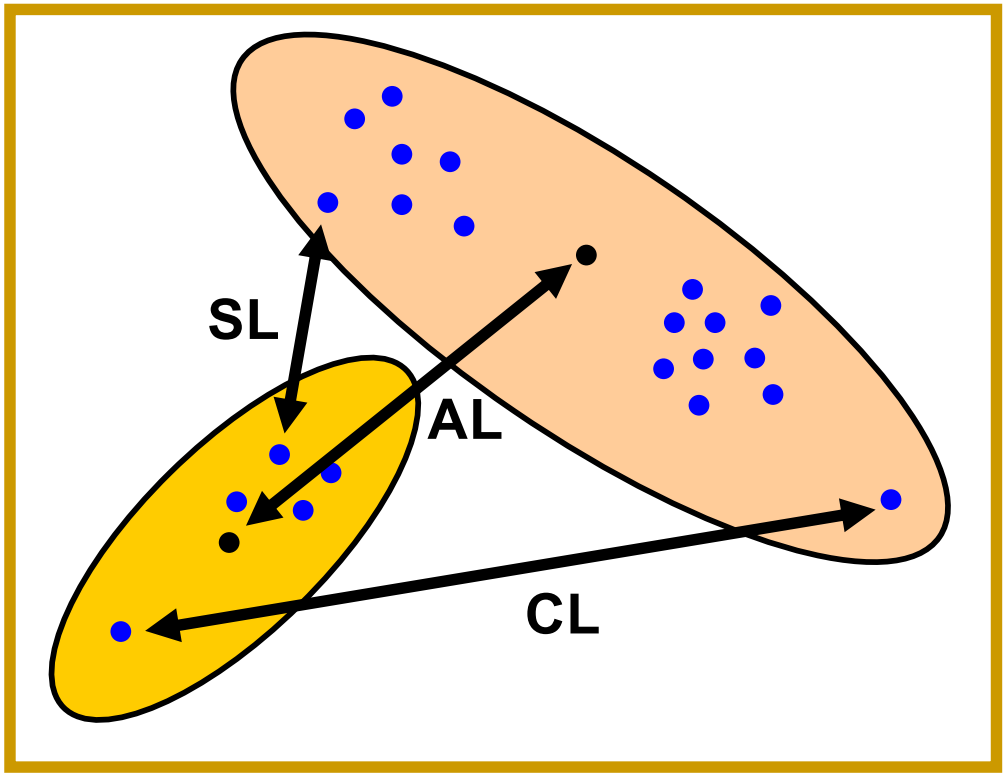
- **Single linkage:** Closest objects
- **Complete linkage:** Furthest objects
- **Average linkage:** Average dissimilarity

Similarity between clusters (linkage)



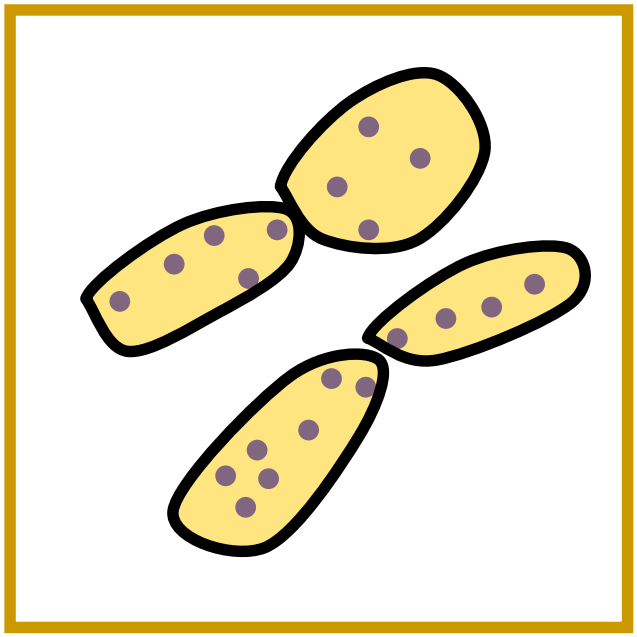
- **Single linkage:** $d(X_i, X_j) = min(d(a,b) : a \in X_i, b \in X_j)$
- **Complete linkage:** $d(X_i, X_j) = max(d(a,b) : a \in X_i, b \in X_j)$
- **Average linkage:** $d(X_i, X_j) = \frac{1}{|X_i||X_j|} \sum_{a \in X_i} \sum_{b \in X_j} d(a,b)$

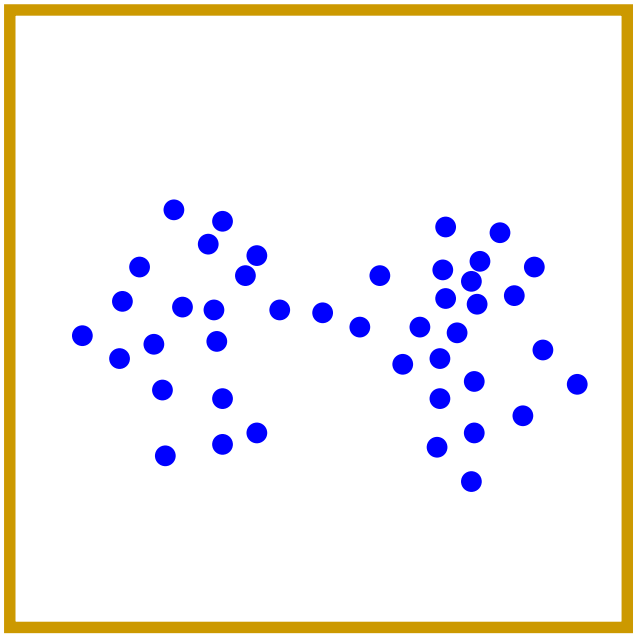
Similarity between clusters (2)

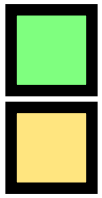


- **Ward's method:** $d(X_i, X_j) = \|X_i - X_j\|^2$ (total variance)
- **Centroid distance:** $d(X_i, X_j) = \|c_s - c_t\|$

SL vs CL: Shape influences



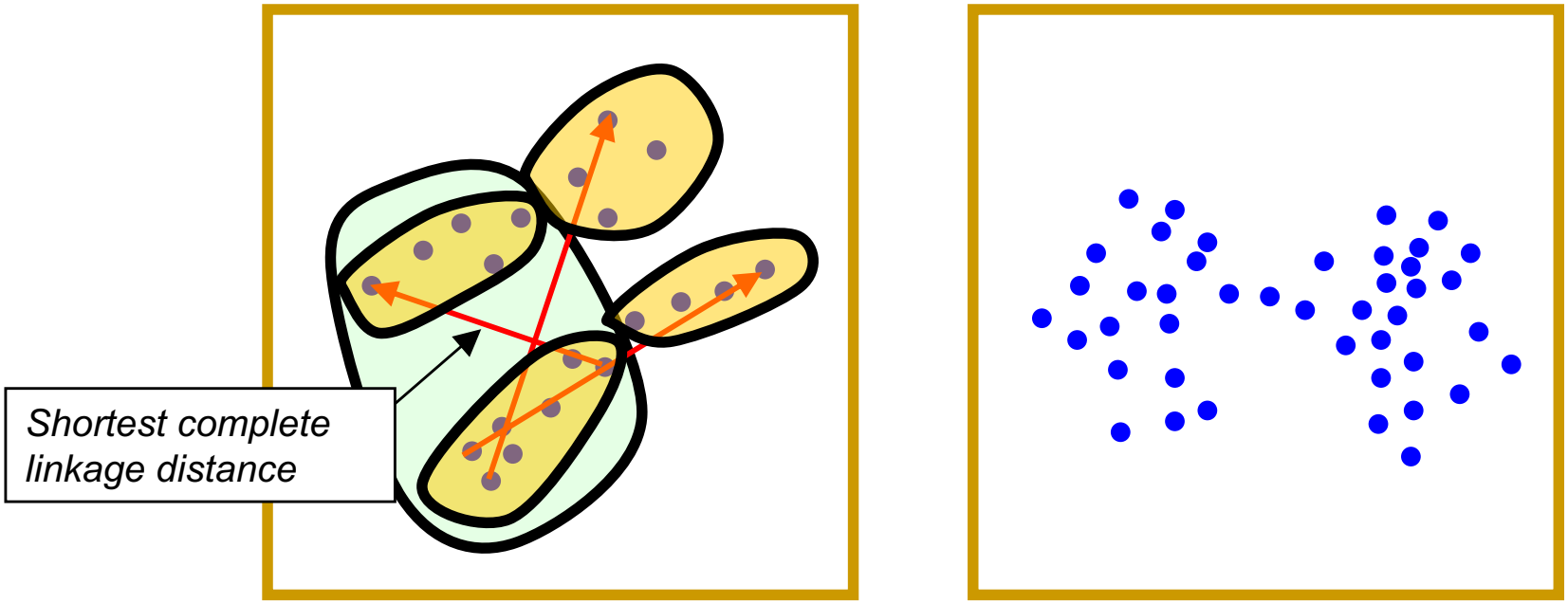




complete linkage ?

single linkage ?

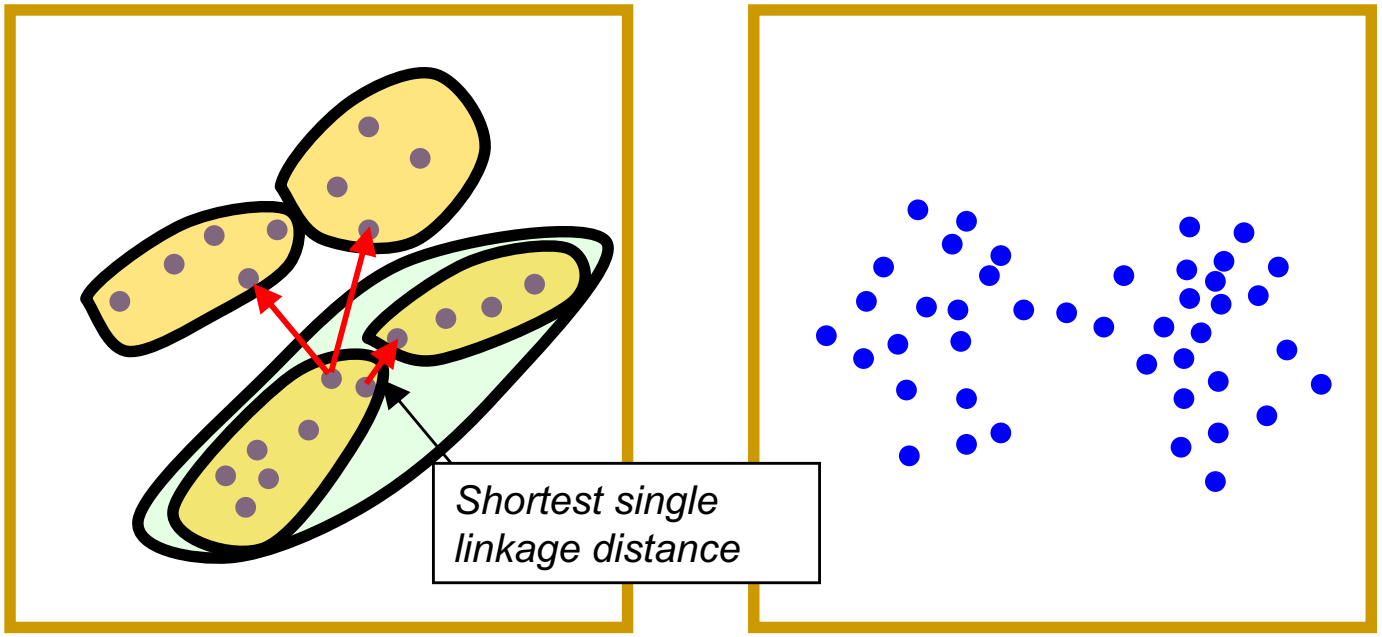
SL vs CL: Shape influences

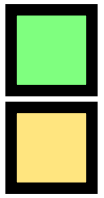


complete linkage ?

single linkage ?

SL vs CL: Shape influences

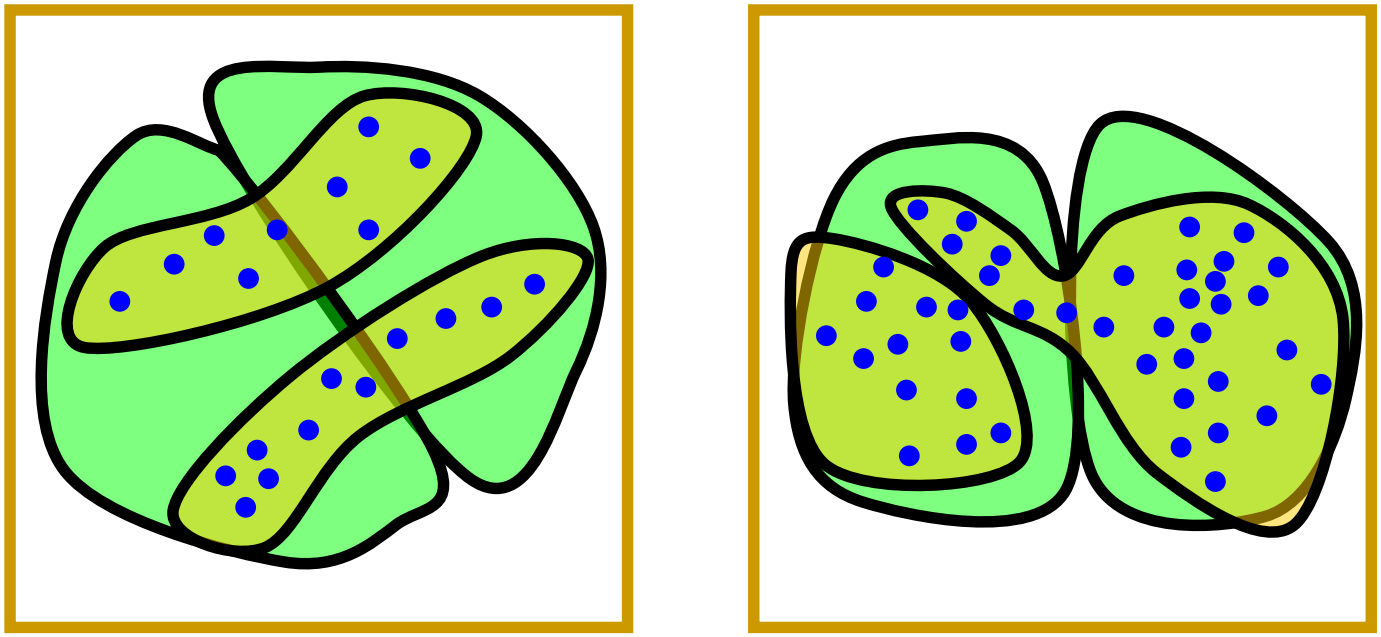




complete linkage ?

single linkage ?

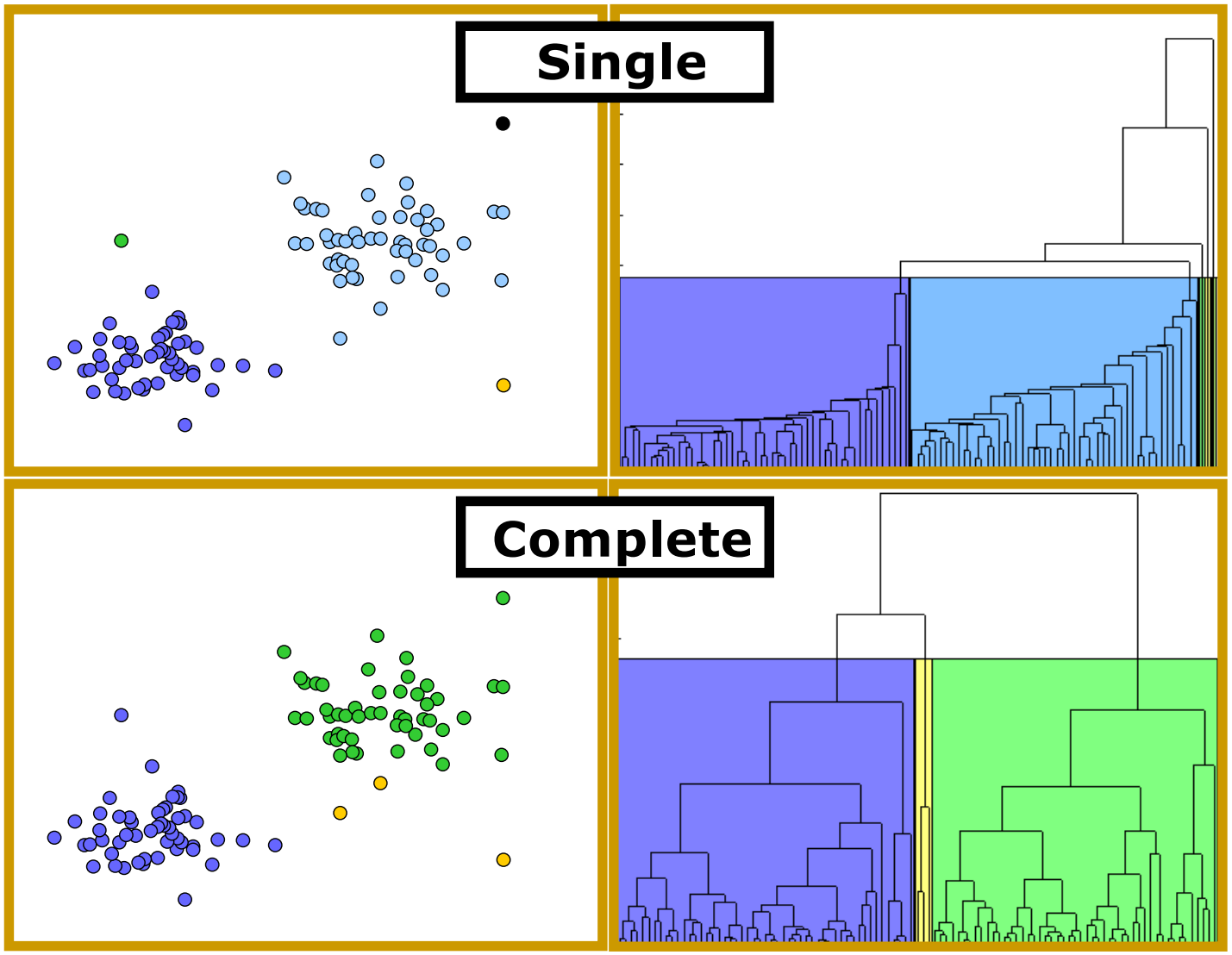
SL vs CL: Shape influences

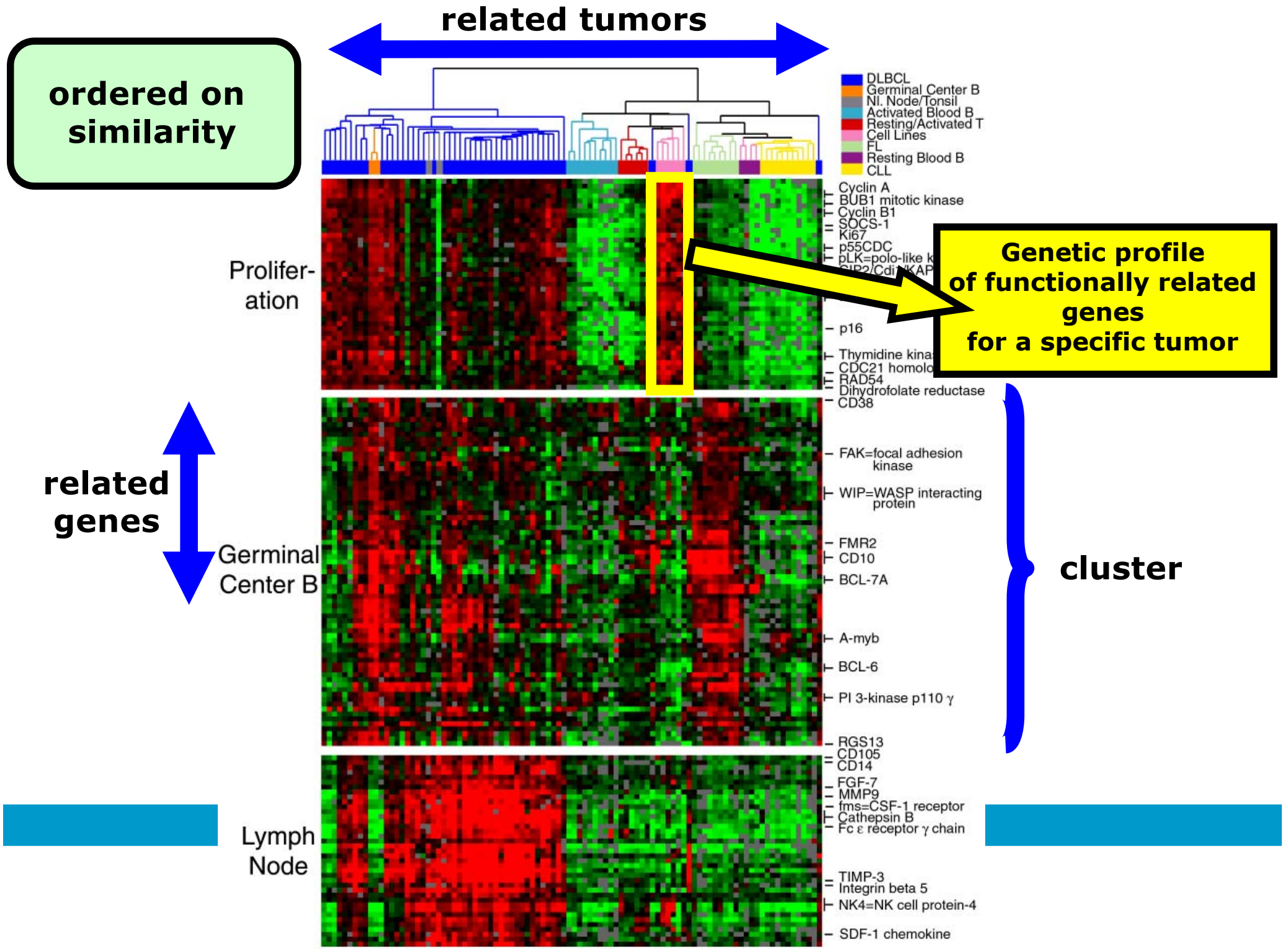


complete linkage

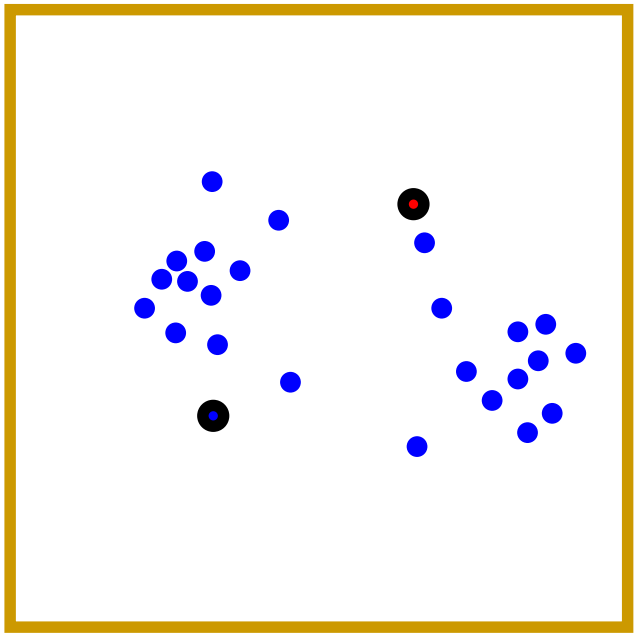
single linkage

SL vs CL: Outlier influences



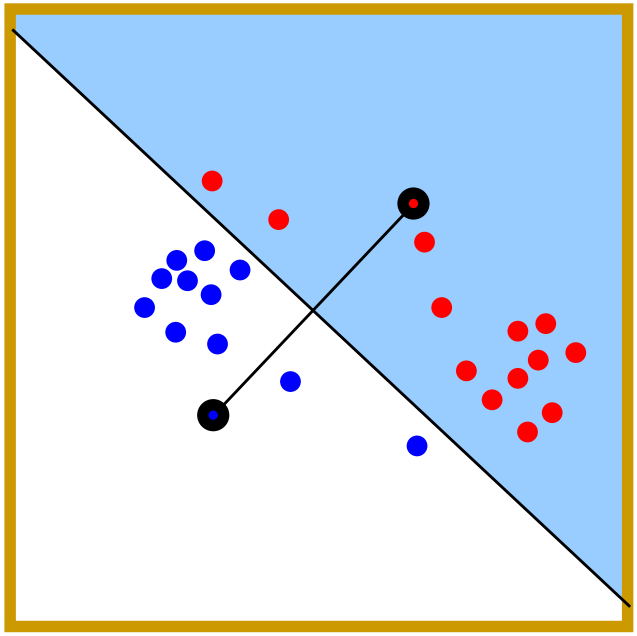

related tumors
ordered on similarity
DLBCL
Germinal Center B
NI. Node/Tonsil
Activated Blood B
Resting/Activated T
Cell Lines
FL
Resting Blood B
CLL
Cyclin A
BUB1 mitotic kinase
Cyclin B1
SOCS-1
Ki67
p55CDC
pLK=polo-like k
Prolifer-ation
p16
Thymidine kina
CDC21 homolo
RAD54
Dihydrofolate reductase
CD38
Genetic profile of functionally related genes for a specific tumor
related genes
FAK=focal adhesion kinase
WIP=WASP interacting protein
Germinal Center B
FMR2
CD10
BCL-7A
A-myb
BCL-6
PI 3-kinase p110 γ
cluster
RGS13
CD105
CD14
FGF-7
MMP9
fms=CSF-1 receptor
Cathepsin B
Fc ε receptor γ chain
Lymph Node
TIMP-3
Integrin beta 5
NK4=NK cell protein-4
SDF-1 chemokine

K-means clustering: Explanation by example



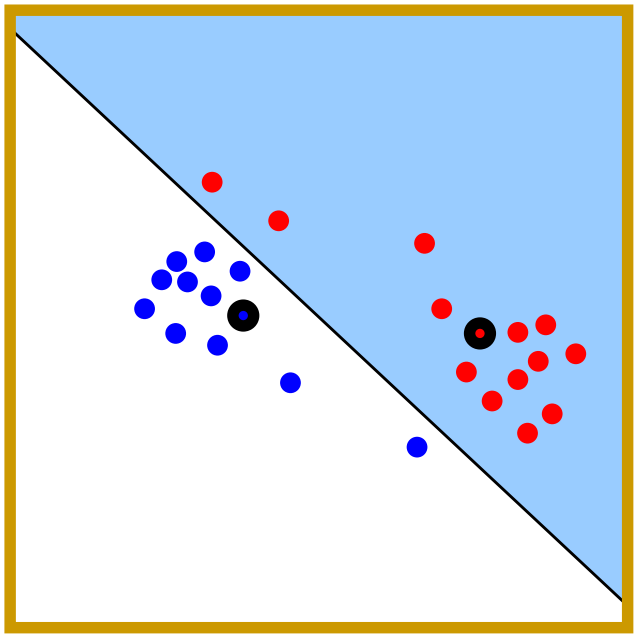
Choose randomly 2 prototypes

K-means clustering: Explanation by example



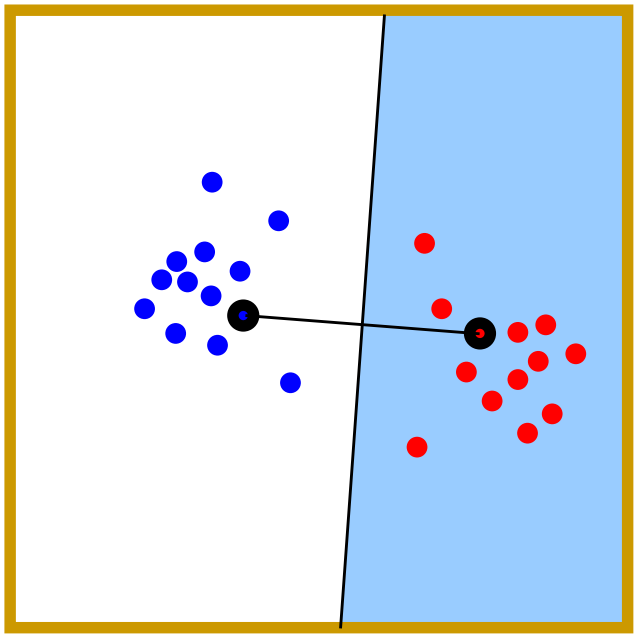
Assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2

K-means clustering: Explanation by example



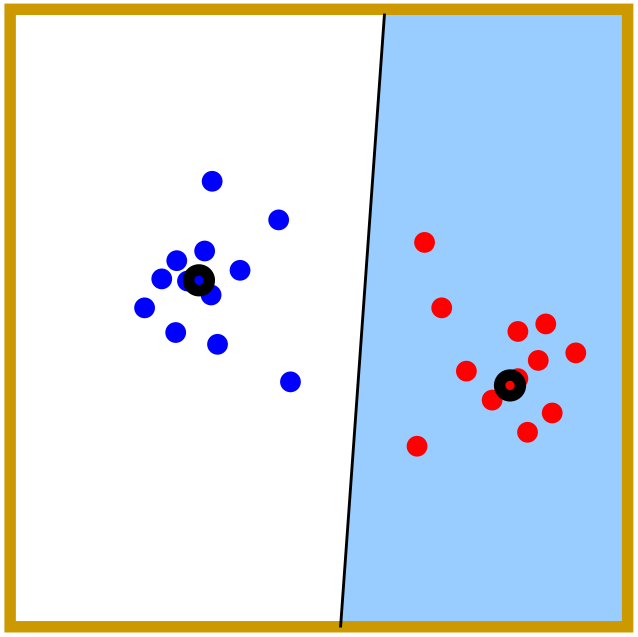
Calculate new cluster prototypes
By averaging objects

K-means clustering: Explanation by example



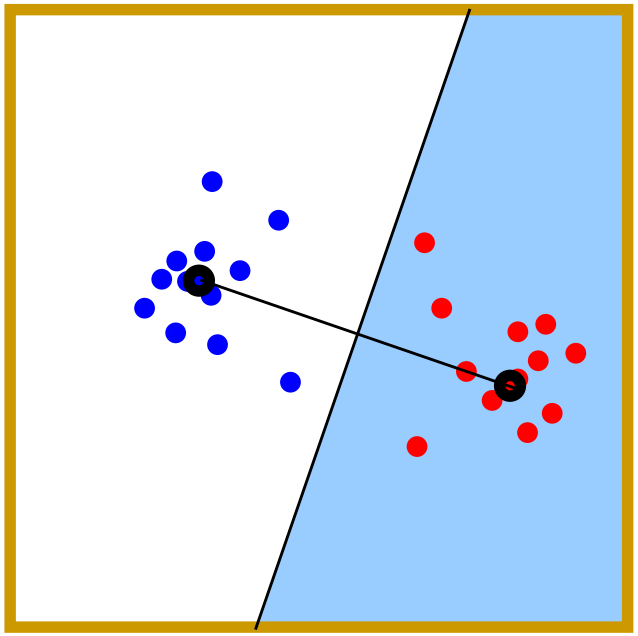
Re-assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2

K-means clustering: Explanation by example



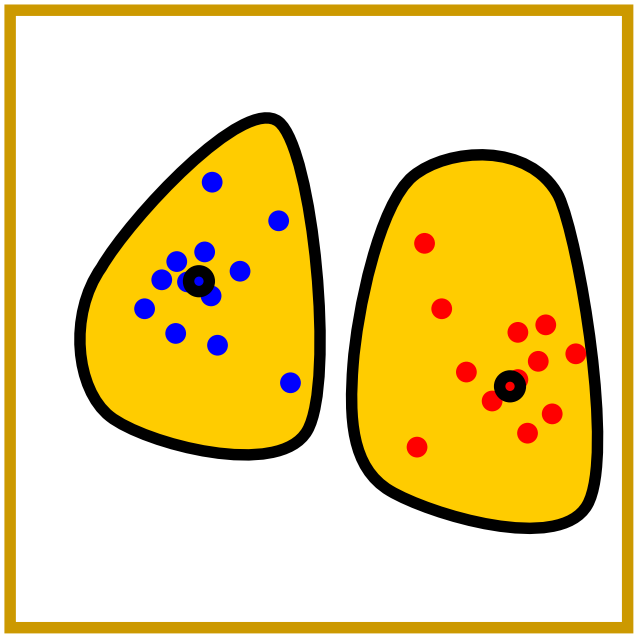
Re-calculate new cluster prototypes

K-means clustering: Explanation by example



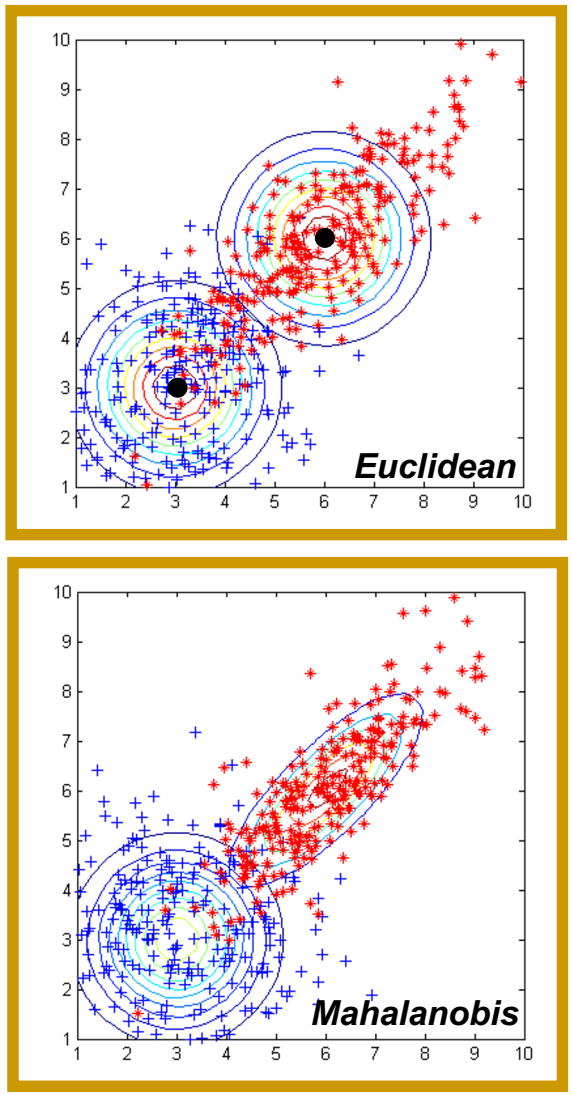
Re-assign objects to closest prototype
If no objects change cluster then finished

K-means clustering: Explanation by example



Establish clusters

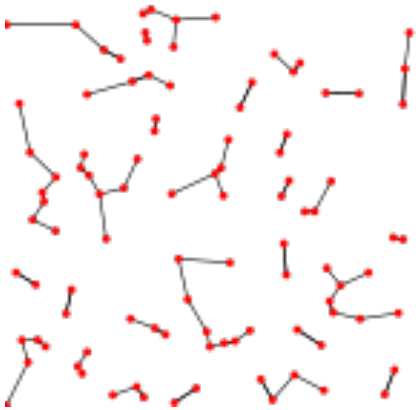
K-means clustering: Parameters



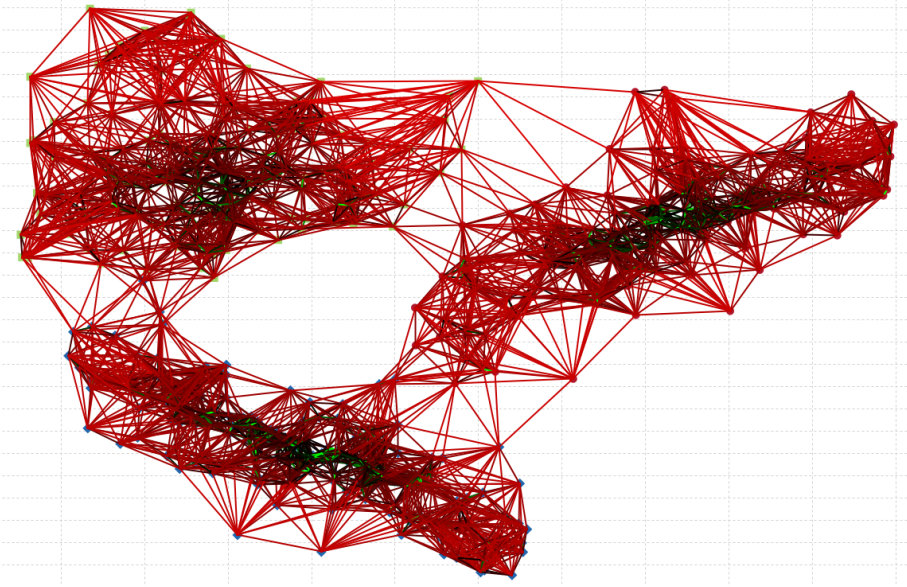
- **K-means**
  - Fixed number of clusters (need to know a priori)
  - Choice of distance measure
  - Prototype choice
- **Distance measure**
  - Euclidean: Round clusters
  - Mahalanobis: Elongated clusters
- **Prototype choice**
  - Point
  - Line etc.
- **Number of clusters**
  - Validate clustering!

Graph-based Clustering

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



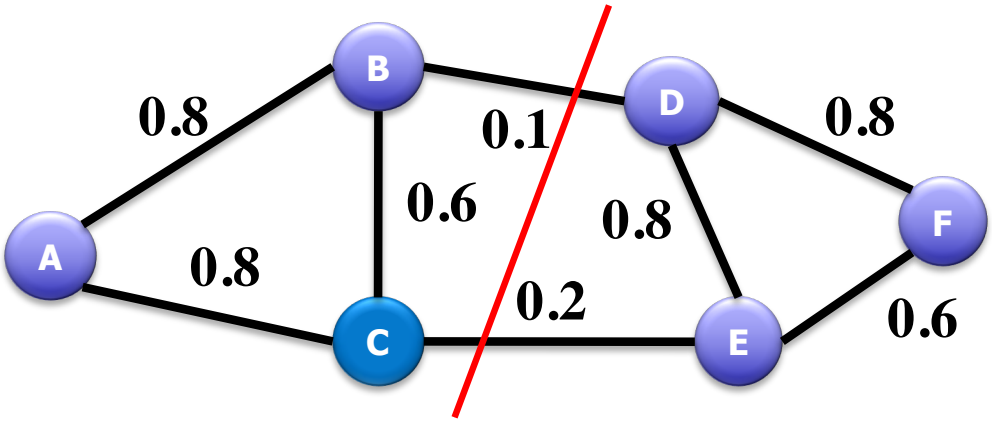
$k=1$



$k=20$

Spectral clustering (1)

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

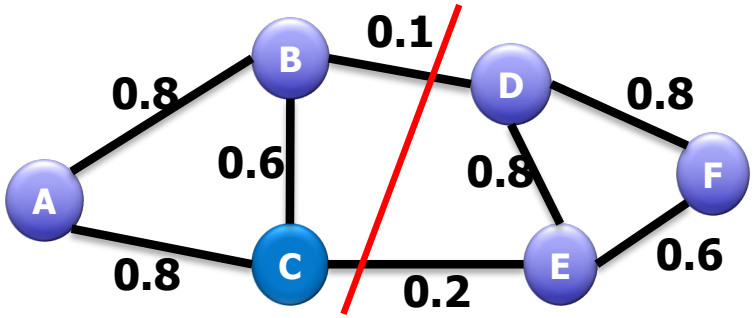


$$NC(C_1,C_2) = \frac{\text{cut}(C_1,C_2)}{assoc(C_1,V)} + \frac{\text{cut}(C_2,C_1)}{assoc(C_2,V)} = 2 - \left(\frac{assoc(C_1,C_1)}{assoc(C_1,V)} + \frac{assoc(C_2,C_2)}{assoc(C_2,V)}\right)$$

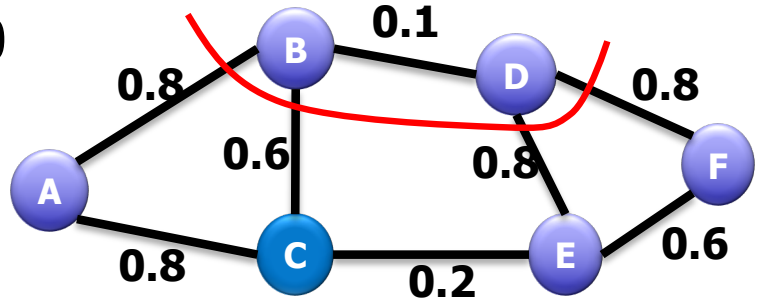
- cut(C_1, C_2) = weight of links between C_1 and C_2
- cut(C_2, C_1) = same
- assoc(C_1, V) = total weight of links from nodes in C_1 to entire graph
- assoc(C_2, V) = total weight of links from nodes in C_2 to entire graph

Normalized cut (example)

- cut(S,T) = 0.1 + 0.2 = 0.3
- vol(S)= 0.3 + 0.6 + 0.8 + 0.8 = 2.5
- vol(T)= 0.3 + 0.8 + 0.8 + 0.6 = 2.5
- Ncut(S,T) = 0.3/2.5 + 0.3/2.5 = 0.24



- cut(S,T) = 0.8 + 0.6 + 0.8 + 0.8 = 3.0
- vol(S) = 3.0 + 0.1 = 3.1
- vol(T) = 3.0 + 0.8 + 0.2 + 0.6 = 4.6
- Ncut(S,T) = 3.0/3.1 + 3.0/4.6 = 1.62



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

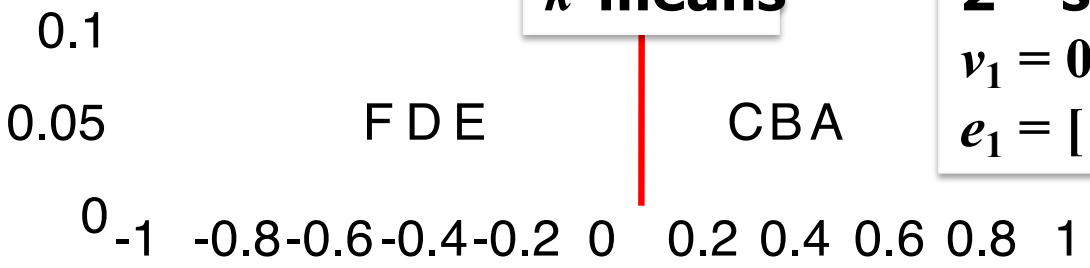
$$W = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{pmatrix} \end{array} \xrightarrow[L=\mathrm{diag}(1^T W)-W]{\text{Laplacian}} L = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 1.6 & -0.8 & -0.8 & 0 & 0 & 0 \\ -0.8 & 1.5 & -0.6 & -0.1 & 0 & 0 \\ -0.8 & -0.6 & 1.6 & 0 & -0.2 & 0 \\ 0 & -0.1 & 0 & 1.7 & -0.8 & -0.8 \\ 0 & 0 & -0.2 & -0.8 & 1.6 & -0.6 \\ 0 & 0 & 0 & -0.8 & -0.6 & 1.4 \end{pmatrix} \end{array}$$

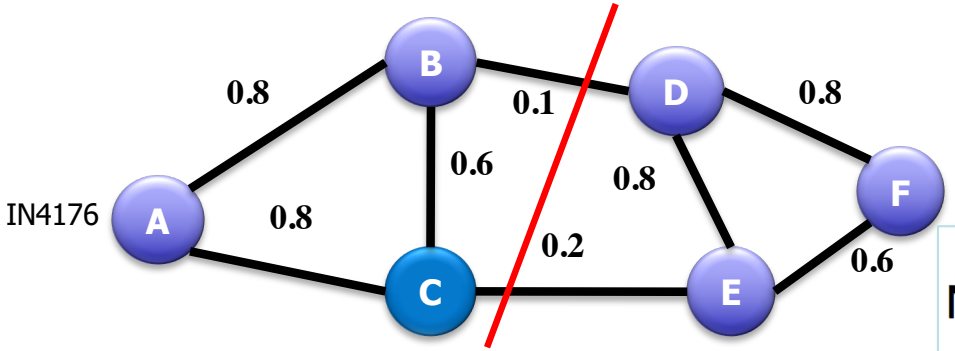
2nd smallest eigenvalue:

$v_1 = 0.19$

$e_1 = [\ 0.44\ 0.41\ 0.37\ -0.40\ -0.37\ -0.45\ ]^T$

***k*-means**





IN4176

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

Normalized Cuts and Image Segmentation

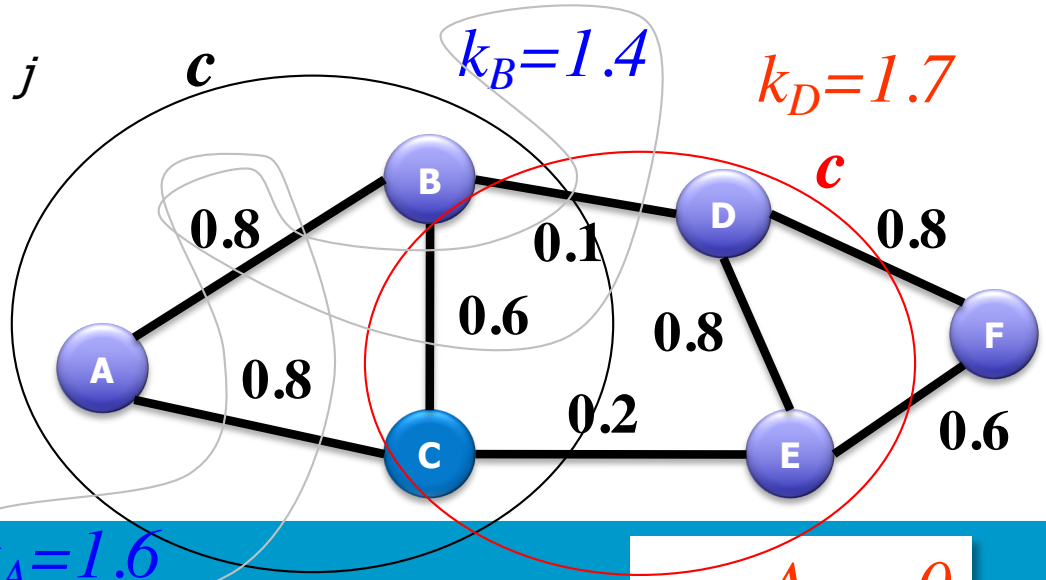
Jianbo Shi and Jitendra Malik, *Member, IEEE*

Louvain method for community detection

Maximize modularity

$$Q = \frac{1}{2m} \sum_{ij} \underbrace{\left[A_{ij} - \frac{k_i k_j}{2m} \right]}_{Q_{ij}} \delta(c_i, c_j) \qquad (Q^*=2mQ)$$

i , j	nodes
c_i , c_j	community of node i and j
A_{ij}	weight edge node i and j
k_i , k_j	total weight of edges of nodes i and j
$\delta(c_i,c_j)$	1 if $c_i=c_j$, zero otherwise
m	total weight of edges



$A_{AB}=0.8$

$\delta(c_A,c_B)=1$

$Q_{AB}=0.8-1.6*1.4/2m$

$Q_{CD}=0-1.6*1.7/2m$

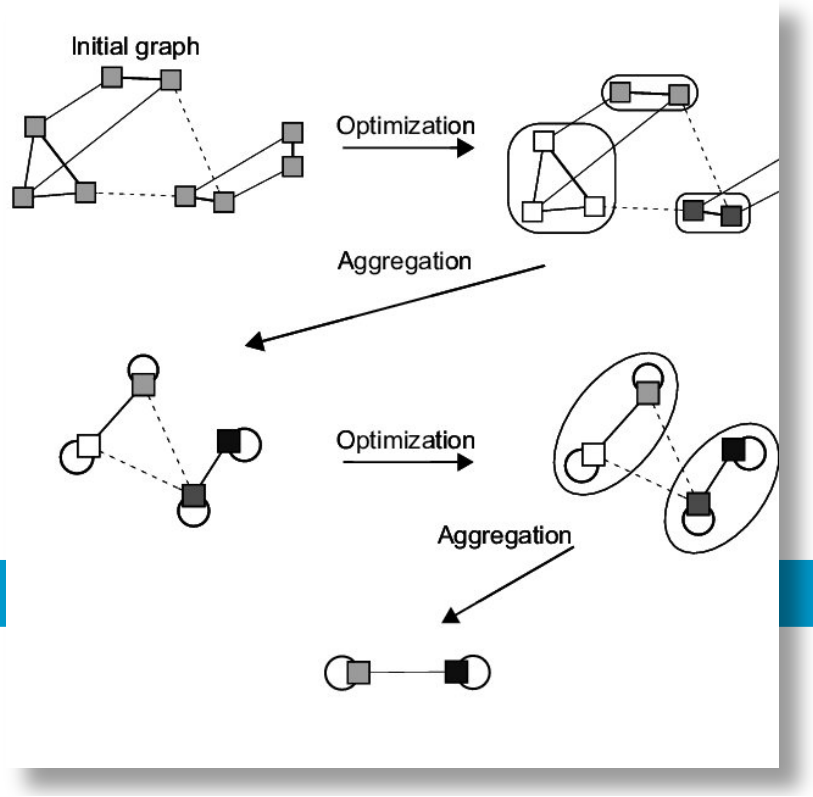
The Delft Bioinformatics Lab

Louvain method for community detection

Modularity optimized in two phases, applied iteratively

Phase 1: Optimization (communities); moving nodes from own community to community of neighboring node (*improve Q*)

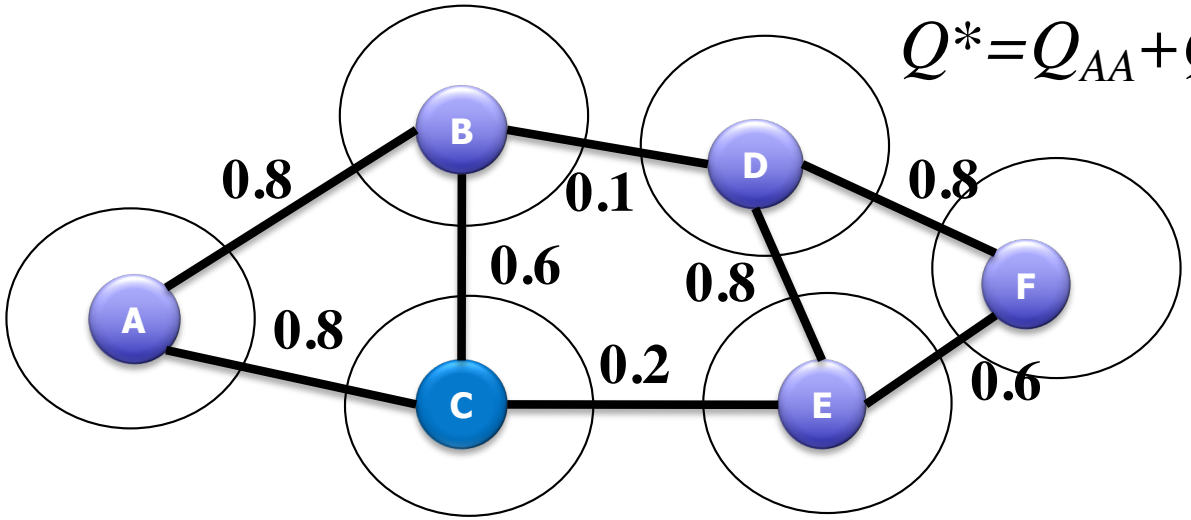
Phase 2: Aggregation; building new network by aggregating nodes in detected communities



Louvain method, running example
(start: every node in separate community)

$M=0.8+0.8+0.6+0.1+0.2+0.8+0.8+0.6=4.7$

$Q^*=Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}$



$Q_{AA} = 0 - 1.6*1.6/9.4$
$Q_{BB} = 0 - 1.5*1.5/9.4$
$Q_{CC} = 0 - 1.6*1.6/9.4$
$Q_{DD} = 0 - 1.7*1.7/9.4$
$Q_{EE} = 0 - 1.6*1.6/9.4$
$Q_{FF} = 0 - 1.4*1.4/9.4$

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

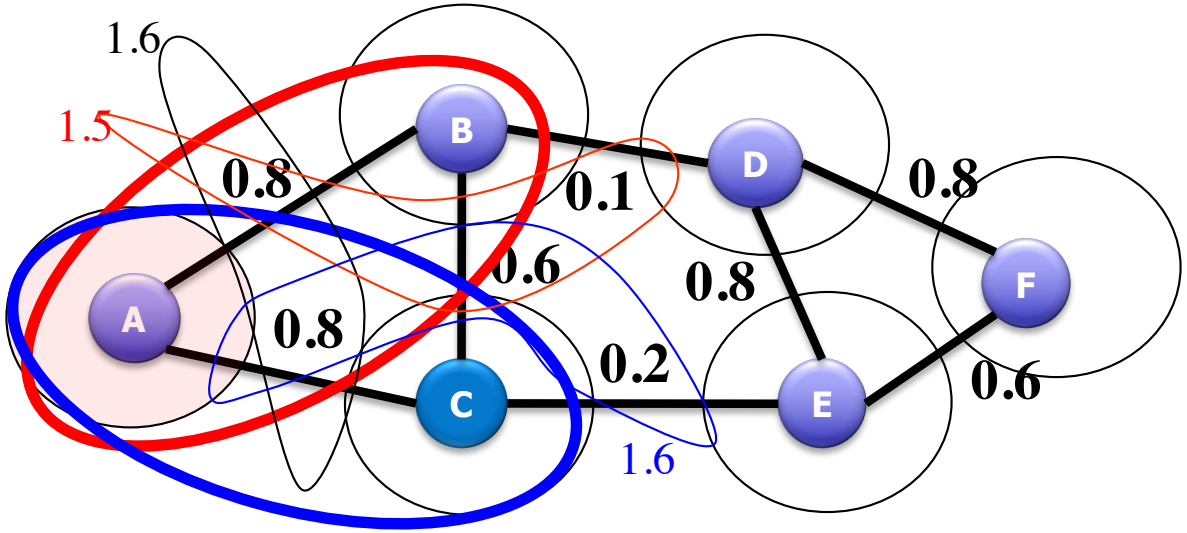
Only 1 for AA,BB,CC,DD,EE,FF

The Delft Bioinformatics Lab

TU Delft

1: Optimize (node A)

$$Q^*=Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{AB}$$
$$Q^*=Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{AC}$$



Q_{AB}=0.8-1.6*1.5/9.4=0.54

Q_{AC}=0.8-1.6*1.6/9.4=0.53

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

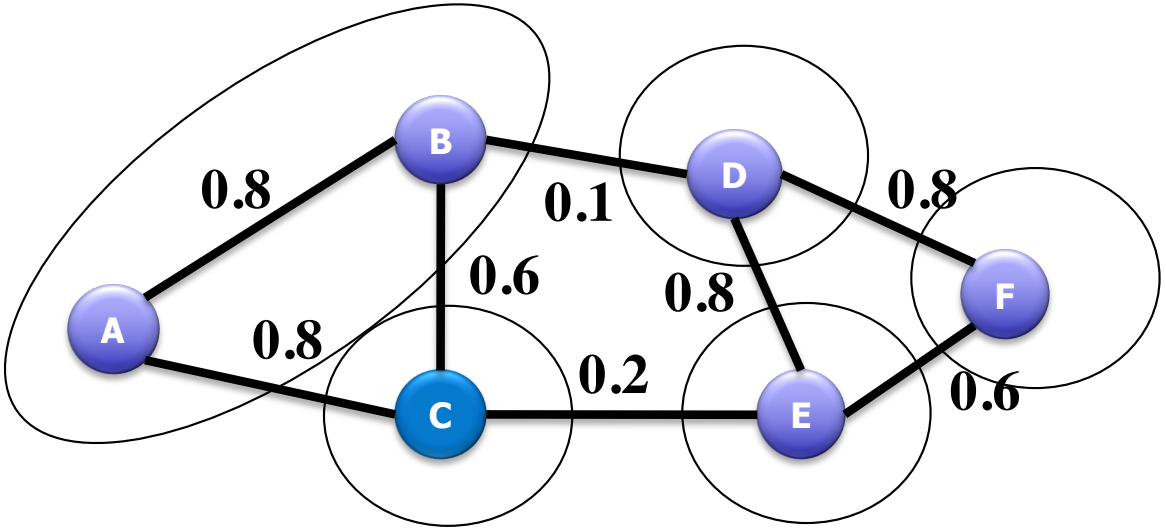
Now 1 for AA,BB,CC,DD,EE,FF AND AB or AC

The Delft Bioinformatics Lab

TU Delft

1: Optimize (node A)

$$Q^*=Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{AB}$$
$$Q^*= Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{AC}$$



Q_{AB}=0.8-1.6*1.5/9.4=0.54

Q_{AC}=0.8-1.6*1.6/9.4=0.53

$Q_{AB}>Q_{AC}$ and positive

Q_{AB} improves Q*

Move A with B

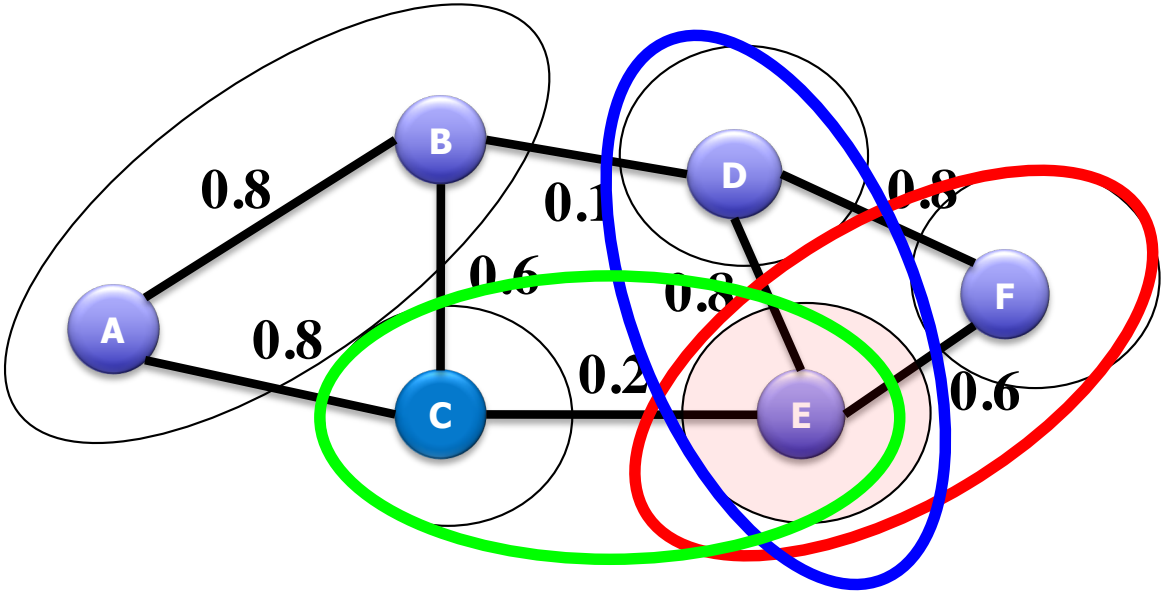
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

1: Optimize (node E)

$$Q^* = Q_{AA} + Q_{BB} + Q_{CC} + Q_{DD} + Q_{EE} + Q_{FF} + Q_{AB} + Q_{EC}$$
$$Q^* = Q_{AA} + Q_{BB} + Q_{CC} + Q_{DD} + Q_{EE} + Q_{FF} + Q_{AB} + Q_{ED}$$
$$Q^* = Q_{AA} + Q_{BB} + Q_{CC} + Q_{DD} + Q_{EE} + Q_{FF} + Q_{AB} + Q_{EF}$$



Q_{EC}=0.2-1.6*1.6/9.4=-0.07

Q_{ED}=0.8-1.6*1.7/9.4=0.51

Q_{EF}=0.6-1.6*1.4/9.4=0.36

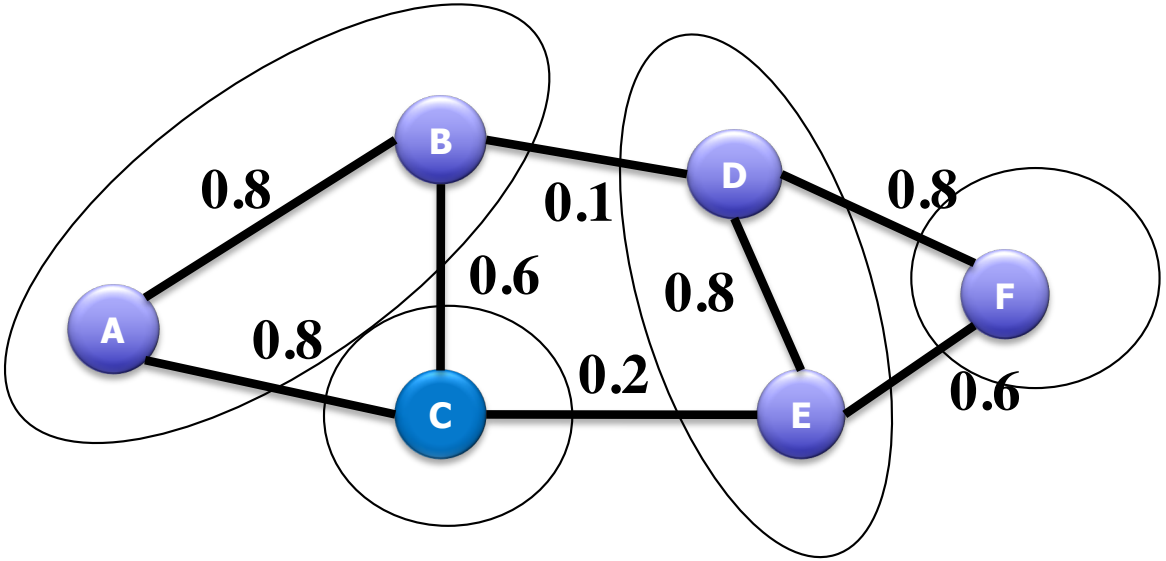
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

1: Optimize (node E)

$$Q^* = Q_{AA} + Q_{BB} + Q_{CC} + Q_{DD} + Q_{EE} + Q_{FF} + Q_{AB} + Q_{EC}$$
$$Q^* = Q_{AA} + Q_{BB} + Q_{CC} + Q_{DD} + Q_{EE} + Q_{FF} + Q_{AB} + Q_{ED}$$
$$Q^* = Q_{AA} + Q_{BB} + Q_{CC} + Q_{DD} + Q_{EE} + Q_{FF} + Q_{AB} + Q_{EF}$$



Q_{EC}=0.2-1.6*1.6/9.4=-0.07

Q_{ED}=0.8-1.6*1.7/9.4=0.51

Q_{EF}=0.6-1.6*1.4/9.4=0.36

Q_{ED} largest & positive

Move E with D

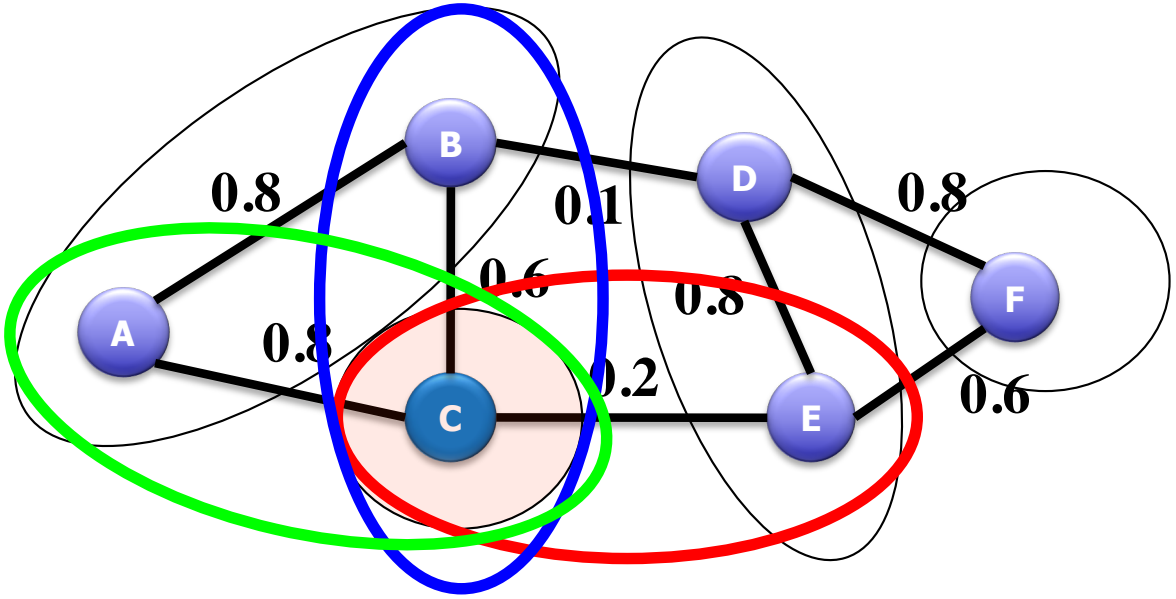
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

1: Optimize (node C)

$$Q^* = Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{ED}+Q_{CA}$$
$$Q^* = Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{ED}+Q_{CB}$$
$$Q^* = Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{ED}+Q_{CE}$$



Q_{CA}=0.8-1.6*1.6/9.4=0.53
Q_{CB}=0.6-1.6*1.5/9.4=0.34
Q_{CE}=0.2-1.6*1.6/9.4=-0.07

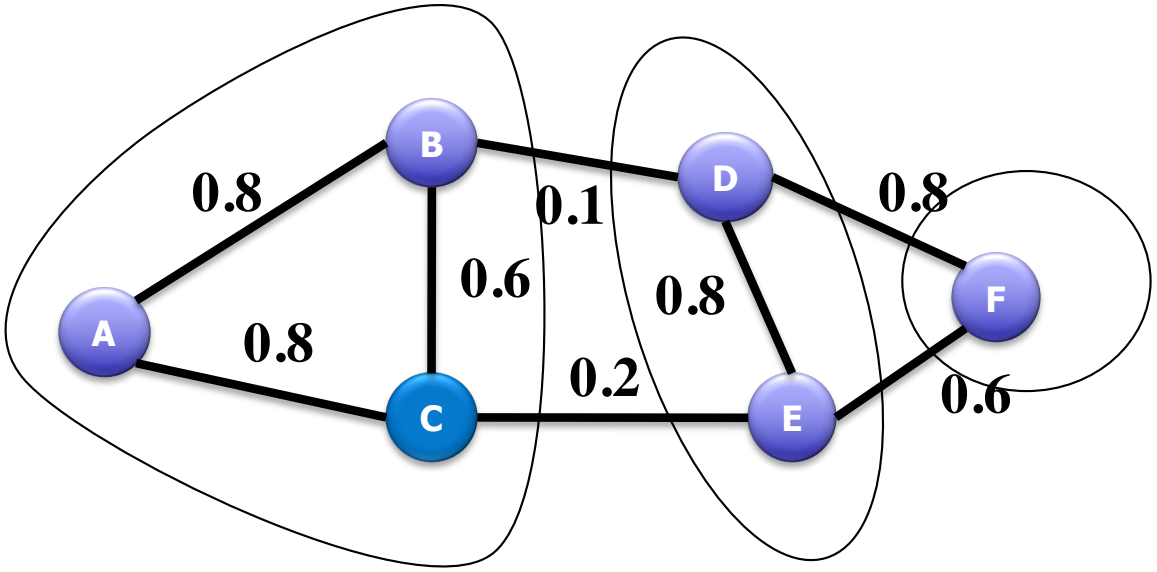
$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

1: Optimize (node C)

$$Q^* = Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{ED}+Q_{CA}$$
$$Q^* = Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{ED}+Q_{CB}$$
$$Q^* = Q_{AA}+Q_{BB}+Q_{CC}+Q_{DD}+Q_{EE}+Q_{FF}+Q_{ED}+Q_{CE}$$



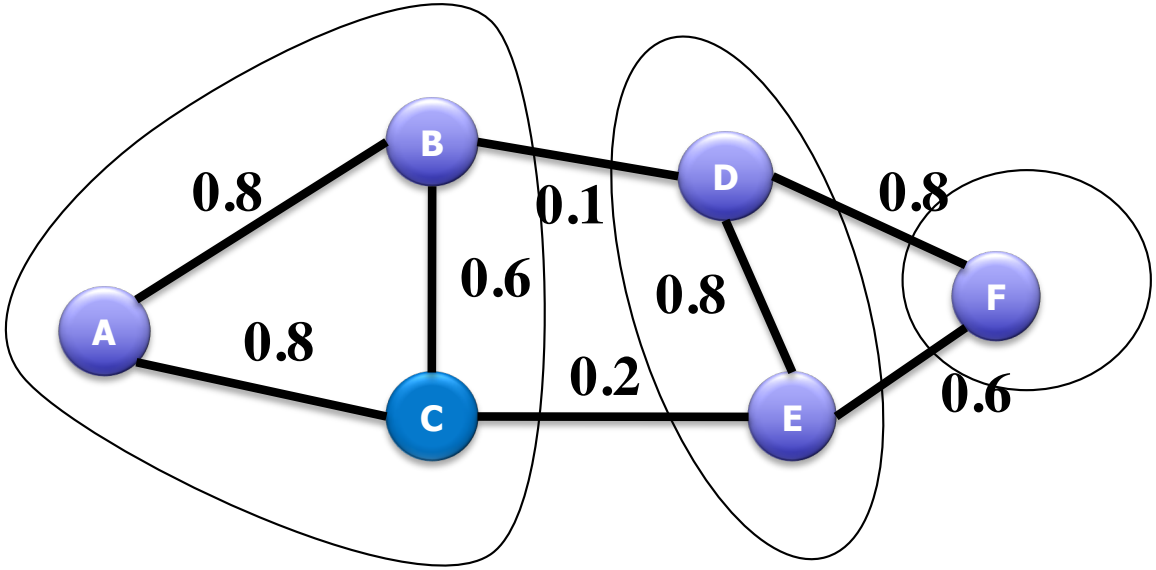
Q_{CA}=0.8-1.6*1.6/9.4=0.53
Q_{CB}=0.6-1.6*1.5/9.4=0.34
Q_{CE}=0.2-1.6*1.6/9.4=-0.07
Q_{CA} largest
Q_{CA} still postive
Move C with A (and B)

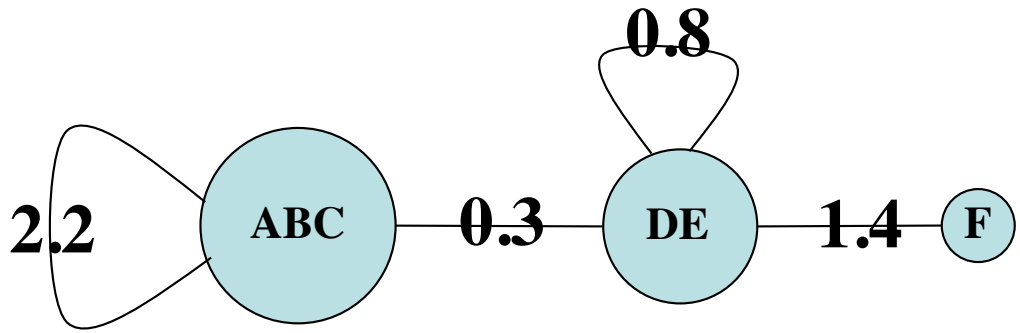
$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

1: Aggregate



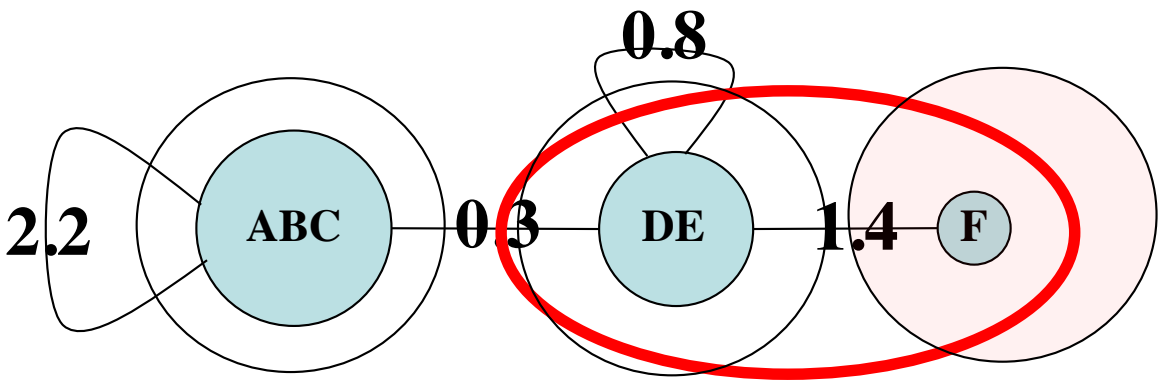


The Delft Bioinformatics Lab

TU Delft

2: Optimize (Node F)

$$Q^* = Q_{ABC} + Q_{DE} + Q_F + Q_{F\text{-}DE}$$



$$Q_{F\text{-}DE} = 1.4 - 0 * 1.7/9.4 = 1.4$$

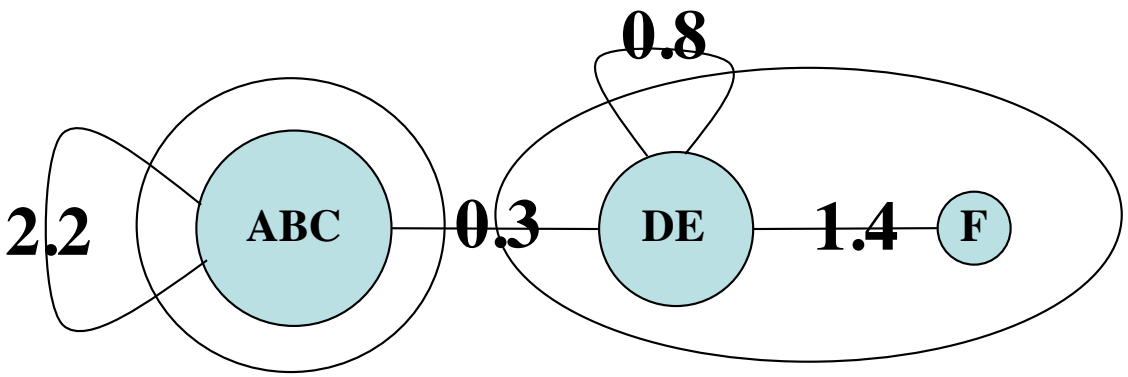
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

2: Optimize (Node F)

$$Q^* = Q_{ABC} + Q_{DE} + Q_F + Q_{F\text{-}DE}$$



$Q_{F\text{-}DE}$=1.4-0*1.7/9.4=1.4

$Q_{F\text{-}DE}$ postive

Move F with DE

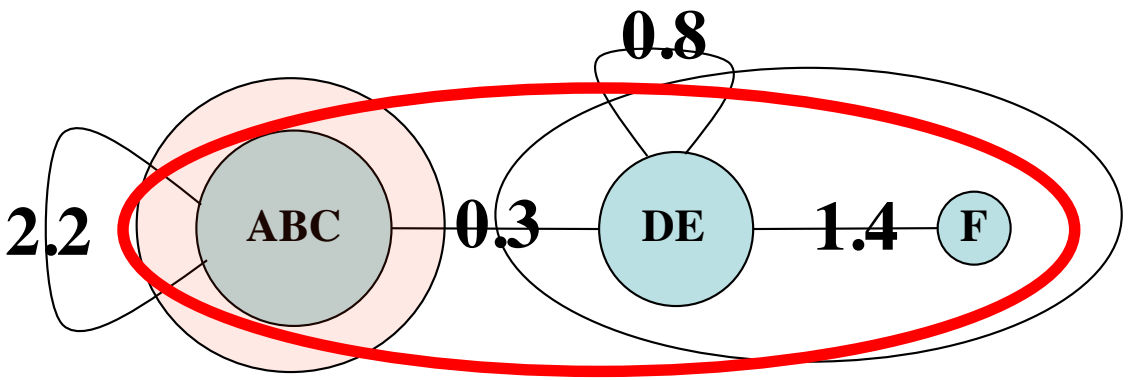
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

2: Optimize (Node ABC)

$$Q^* = Q_{ABC} + Q_{DE} + Q_F + Q_{F\text{-}DE} + Q_{ABC\text{-}DE}$$



$Q_{ABC\text{-}DE} = 0.3 - 2.2 * 1.8 / 9.4 = -0.12$

$Q_{ABC\text{-}DE} < 0$

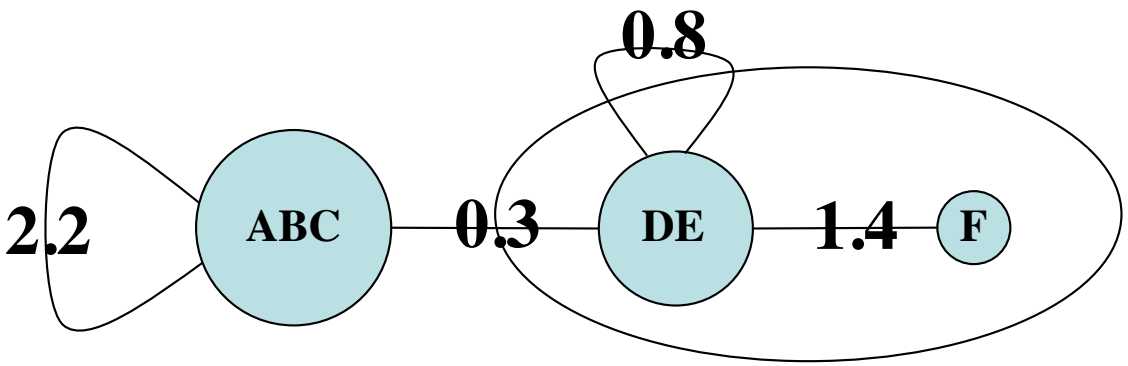
STOP (no improvement of Q)

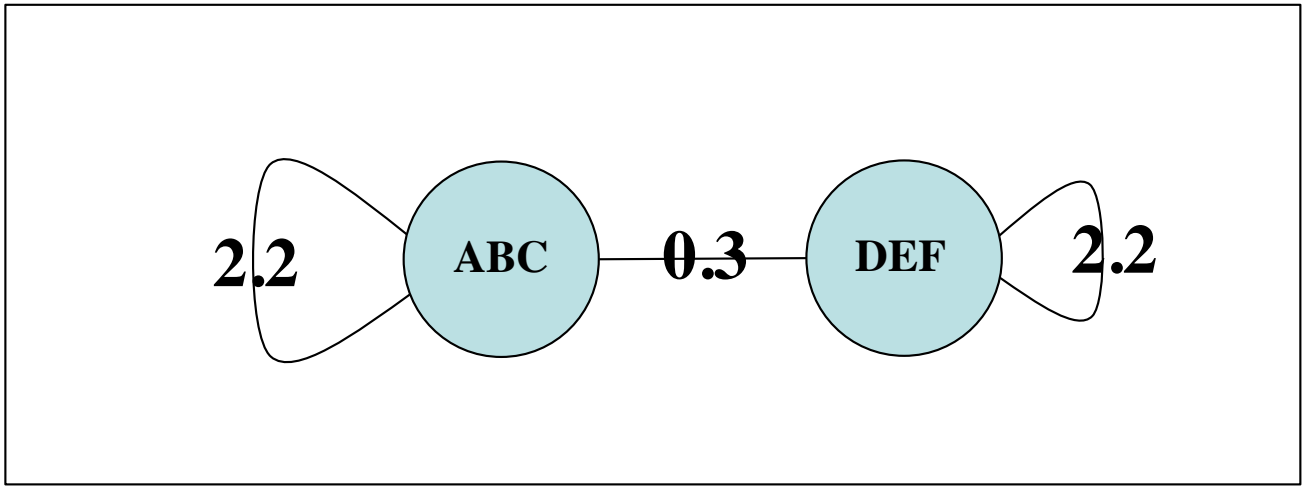
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

The Delft Bioinformatics Lab

TU Delft

2: Agregate, finish





The Delft Bioinformatics Lab

TU Delft

Markov clustering (1)

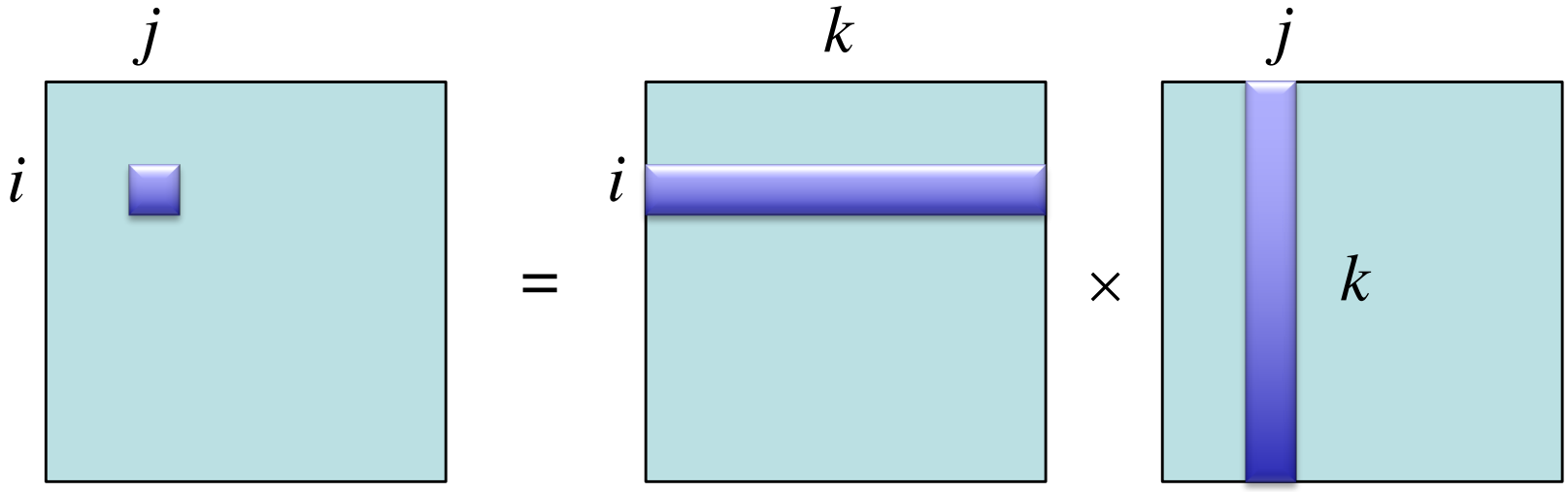
- **Markov clustering**
  - Random walks are Markov chains
  - Start with Markov matrix: $M_{ij} = p(j \rightarrow i)$
  - (optionally add loops (diagonal))

$$W = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{pmatrix} \end{array} \xrightarrow[\mathbf{M} = \mathbf{W}\mathrm{diag}(\mathbf{1}^T \mathbf{W})^{-1}]{\text{Markov}} M = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 0.53 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.38 & 0.06 & 0 & 0 \\ 0.5 & 0.4 & 0 & 0 & 0.12 & 0 \\ 0 & 0.07 & 0 & 0 & 0.5 & 0.57 \\ 0 & 0 & 0.12 & 0.47 & 0 & 0.43 \\ 0 & 0 & 0 & 0.47 & 0.38 & 0 \end{pmatrix} \end{array}$$

$$(sum\ columns = 1)$$

Markov clustering (2)

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

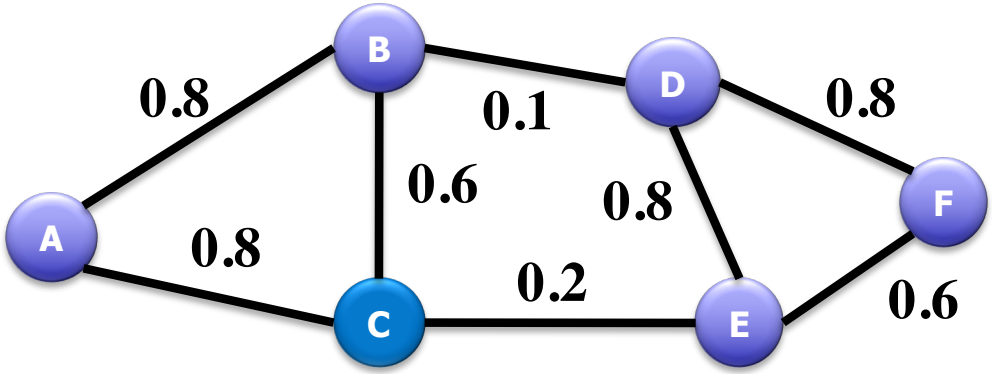


$$\left(M^2\right)_{ij} = \sum_k p(j \rightarrow k)p(k \rightarrow i)$$

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

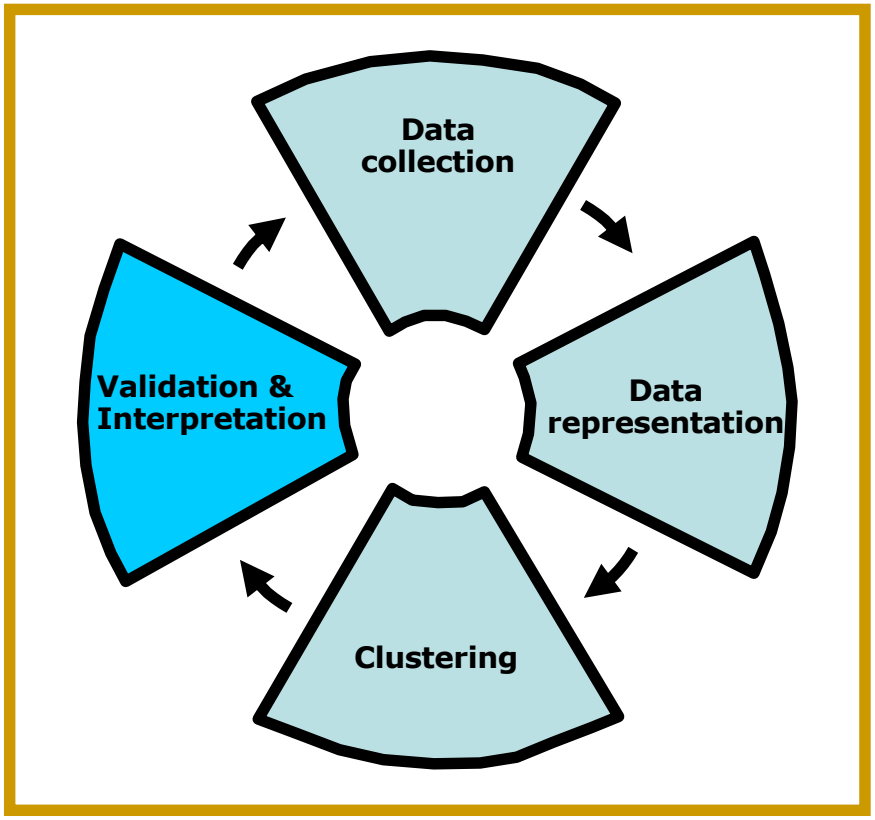
Markov clustering (3)

- **Markov clustering:**
  - Iterate
    1. $M = M^{\alpha}$
    2. $M_{ij} = M_{ij}^{\beta} \Big/ \sum_i M_{ij}^{\beta}$
- 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 $\alpha = \beta = 2$



$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0.\blacksquare & 0.\blacksquare & 0.\blacksquare & 0.00 & 0.0\blacksquare & 0.00 \\ 0.\blacksquare & 0.\blacksquare & 0.\blacksquare & 0.0\blacksquare & 0.0\blacksquare & 0.00 \\ 0.\blacksquare & 0.\blacksquare & 0.\blacksquare & 0.0\blacksquare & 0.0\blacksquare & 0.0\blacksquare \\ 0.00 & 0.0\blacksquare & 0.0\blacksquare & 0.\blacksquare & 0.\blacksquare & 0.\blacksquare \\ 0.0\blacksquare & 0.0\blacksquare & 0.0\blacksquare & 0.\blacksquare & 0.\blacksquare & 0.\blacksquare \\ 0.00 & 0.00 & 0.0\blacksquare & 0.\blacksquare & 0.\blacksquare & 0.\blacksquare \end{pmatrix} \end{array}$$

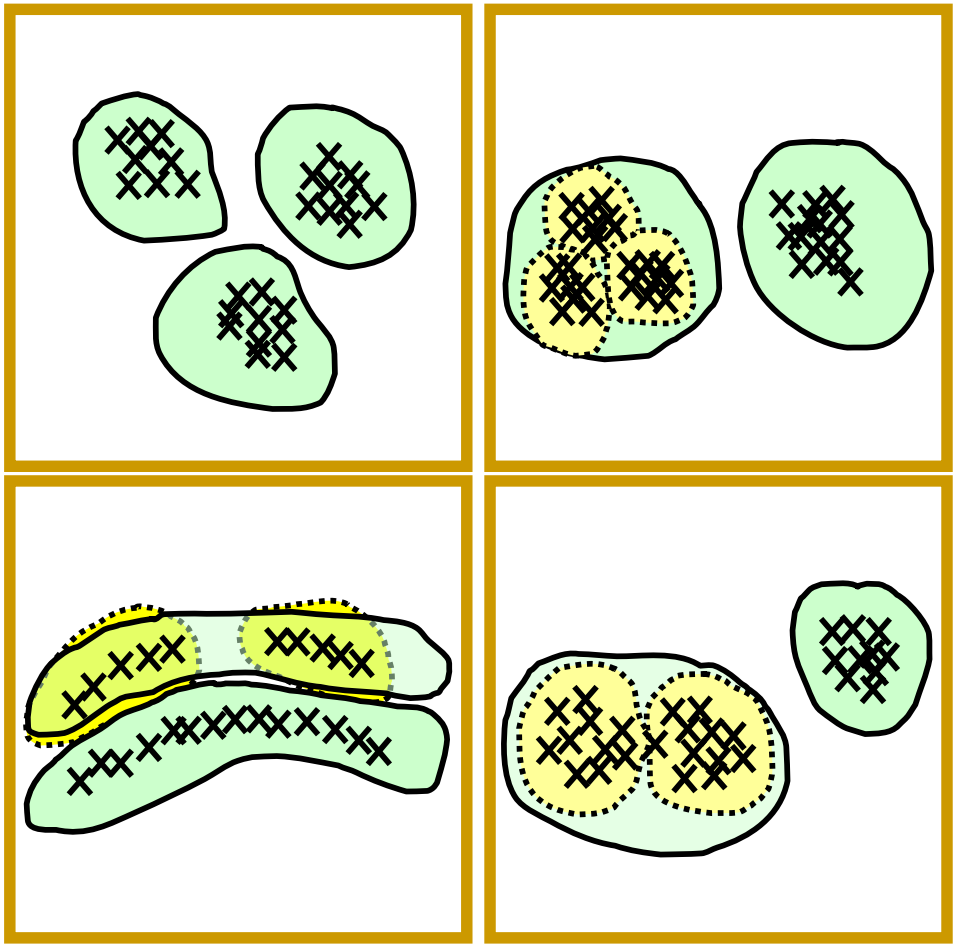
Validation



TOPICS

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

Subjectivity

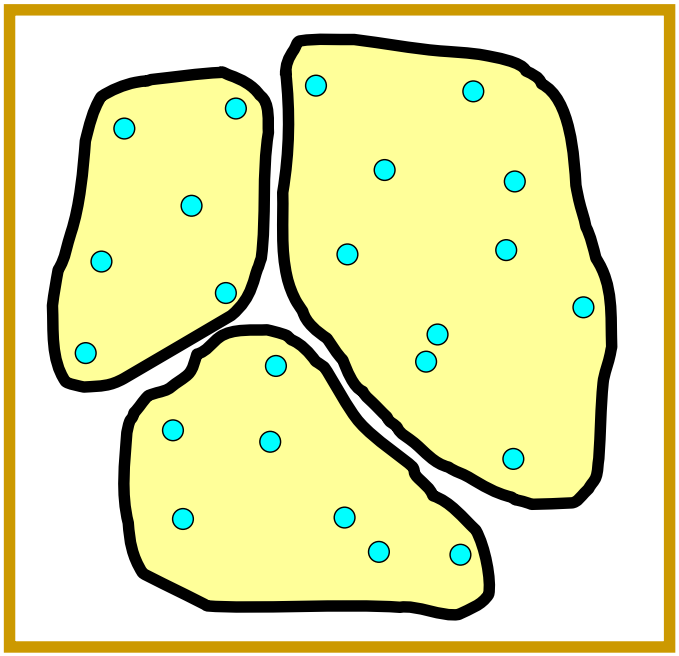


- **Principle choices:**
  - **similarity**
  - **algorithm**
- **Different choice leads to different results**

  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

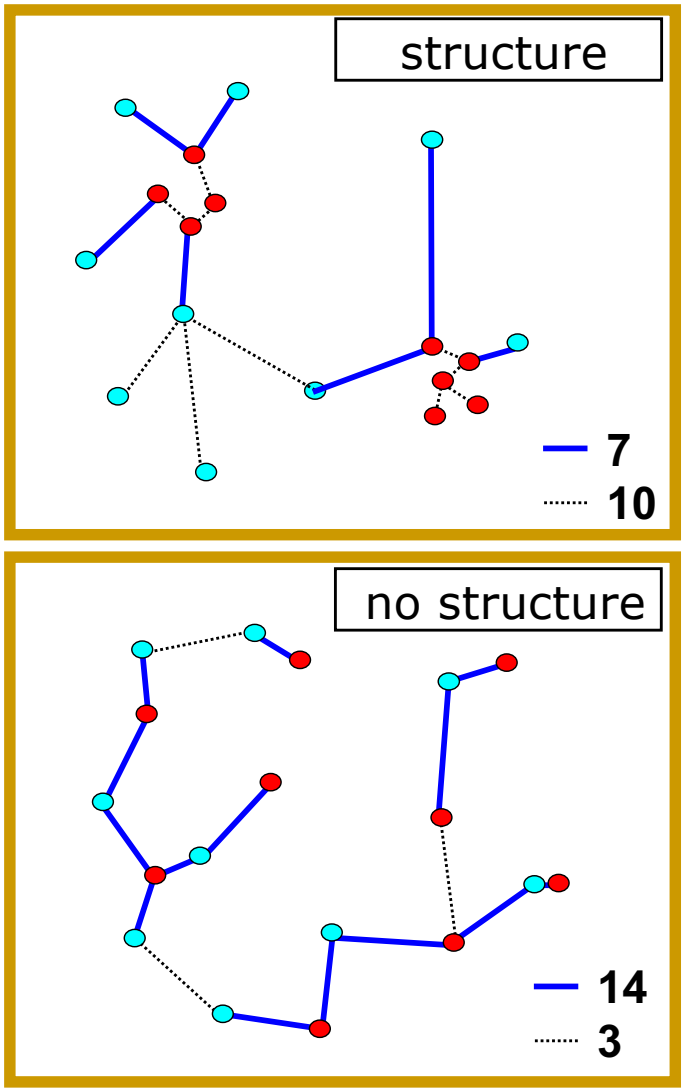
  Aim: Test whether data possesses structure

- **Cluster validity**

  Choices impose restrictions on for example shape

  Aim: Quantitative evaluation of the clustering results

Test for spatial randomness



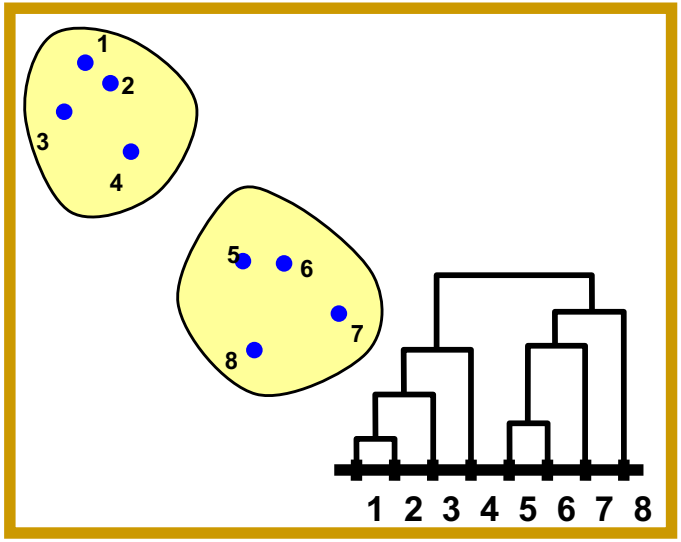
- **Test**

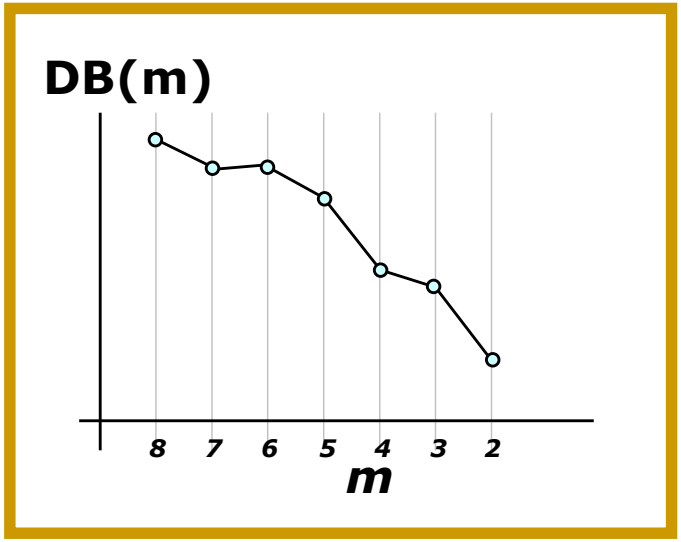
  If data (●) clusters frequently with random data (●) then data structureless

- **Approach**
  - Generate random vectors (**Y**) uniformly over observed region of data (**X**)
  - Find MST (single linkage HC) of **X** ∨ **Y**
  - Determine number of edges *q* that connect vectors of **X** with **Y**
  - If **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**

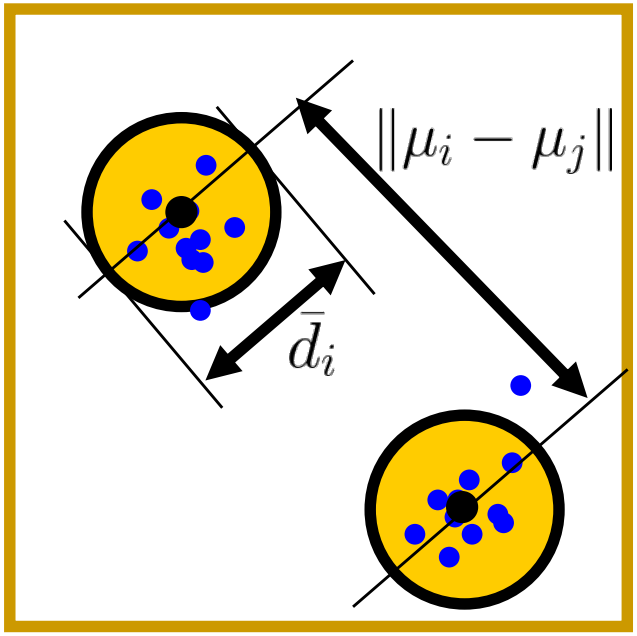
  For a specific clustering m, **DB(m)** : Average similarity of a cluster with its most similar cluster

- **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**

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

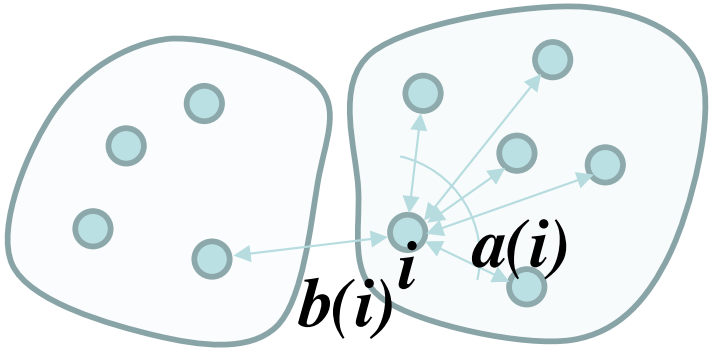
- $\bar{d}_i$:average distance within cluster *i*, μ_i:centroid of cluster *i*
- **Most similar cluster to C_i**

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

- **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:

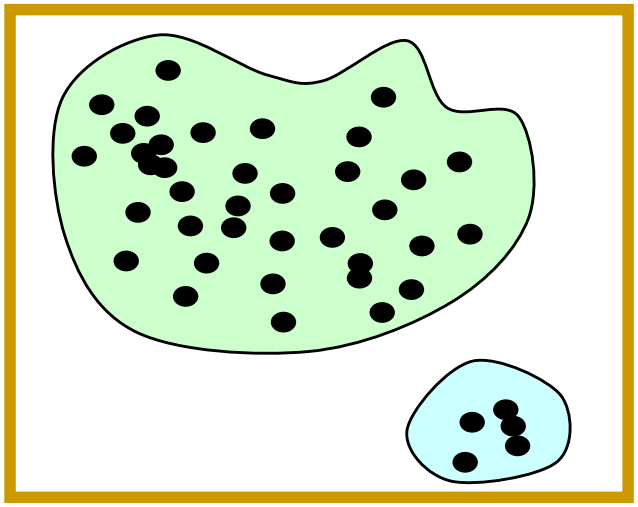
$$a(i) = \frac{1}{|C_i|} \sum_{\forall j} d(x_i, x_j) \qquad b(i) = min_{\forall j, j \notin C_i} d(x_i, x_j)$$

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

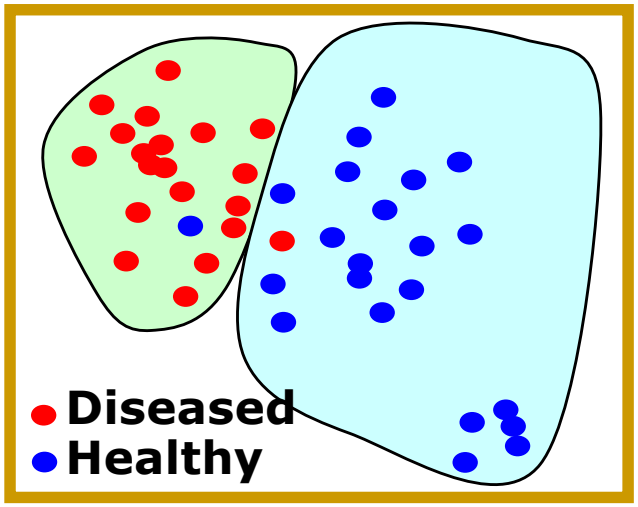
- **Silhouette score is average of all these similarities**

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

Clustering vs classification

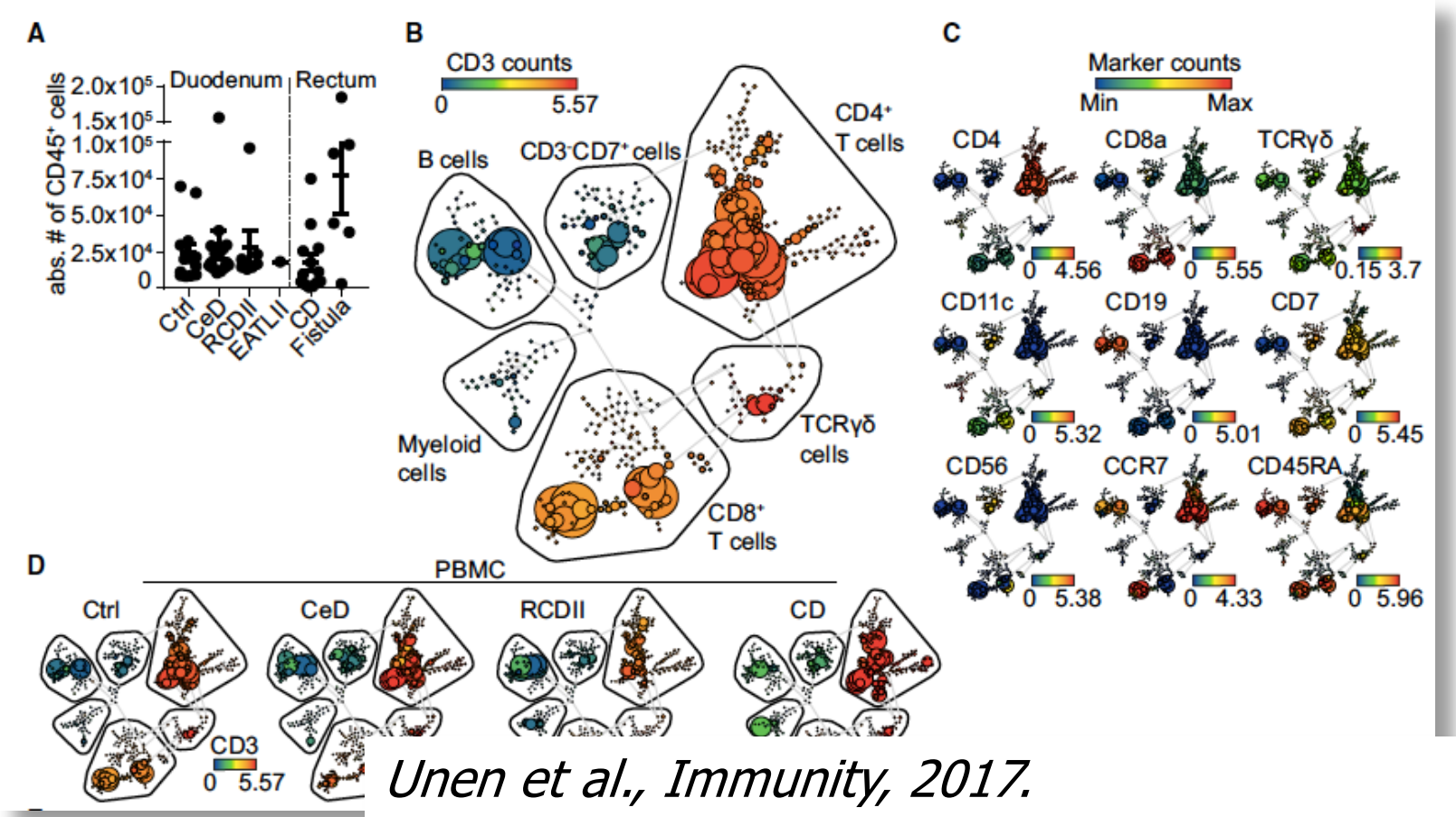


- **Machine learning**
- **Clustering**
  - **unsupervised** learning
  - discovering structure/relations



- **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….

The Delft Bioinformatics Lab

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!

TU Delft

Delft University of Technology

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, 6, 303 - 304 (2008)
https://www.nature.com/nbt/journal/v26/n3/full/nbt0308-303.html