



# Les 12 – DEG's en clustering (1)

**Emile Apol**



**Hanze University Groningen**  
APPLIED SCIENCES

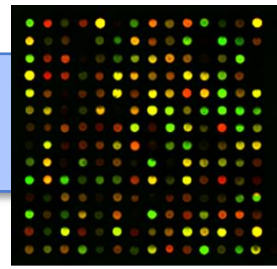
Institute for  
Life Science & Technology

# LES 12

- Clustering
- Afstand / (dis)similarity methoden
  - Euclidisch
  - Manhattan
  - Minkowski
  - Pearson
- Clustering algoritmen (linkage)
  - Minimal
  - Full
  - Average
- Cluster methoden
  - Hierarchical clustering
  - K-Means clustering



# MICROARRAY ANALYSE: STAPPENPLAN



- Background correctie
- Log transformatie
- Normalisatie (bijv. loess)
- Toetsen op DEG's:
  - $t$ -toets, 1-way ANOVA, ...
  - Wilcoxon's toets, Kruskal-Wallis toets, ...
- Aanpassen  $p$ -waarden voor multiple toetsing
- Clustering van DEG's:
  - Hiërarchisch clusteren
  - $k$ -means
  - Principale Componenten Analyse (PCA)
- Grafische weergave: heatmaps, vulcano plot, ...

## DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, 2 SAMPLES

- Data format 1:  $r$  replica's van  $M$  waarden per gen:

Gene	$M_1$	$M_2$	...	$M_i$	...	$M_r$
gene 1	0.51	0.34	...	0.55	...	0.44
gene 2	-0.14	-0.31	...	0.11	...	-0.27
...						
gene $g$	0.78	0.85	...	0.69	...	0.75
...						
gene $G$	1.15		...	0.66	...	0.91

1-sample  $t$ -toets voor  $\mu = 0$

# DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, 2 SAMPLES

- Statistische analyse per gen op  $M = \log(T)$  waarden:

- 1-sample  $t$ -toets voor  $\mu = 0$
- Wilcoxon signed-rank toets voor  $\mu = 0$

$p$ -waarde per gen

- $p$ -waarden per gen aanpassen voor multiple testing:

- Holm methode (FWER)
- Benjamini-Hochberg methode (FDR)

- Genen  $i$  met significante differentiële expressie:

$$p_{\text{adjust}, i} < \alpha$$

- Doe vervolg analyse met deze significante DEGs:

- clustering
- ...

## DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, >2 SAMPLES

- **Data format 1:**  $k$  samples,  $r$  replica's  $j$  van  $M_{ij} = \log(T_{ij})$  waarden per gen per sample  $i$  :

Gene	sample 1				sample $k$		
	$M_{11}$	...	$M_{1r}$	...	$M_{k1}$	...	$M_{kr}$
gene 1	0.51	...	0.34	...	0.55	...	0.44
gene 2	-0.14	...	-0.31	...	0.11	...	-0.27
...							
gene $g$	0.78	...	0.85	...	0.69	...	0.75
...							
gene $G$	0.45	...	0.45	...	0.66	...	0.91

1-way ANOVA

# DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, >2 SAMPLES

- Statistische analyse per gen op  $M = \log(T)$  waarden:
  - 1-way ANOVA
  - Kruskal-Wallis rank-sum toets
- $p$ -waarden per gen aanpassen voor multiple testing:
  - Holm methode (FWER)
  - Benjamini-Hochberg methode (FDR)
- Genen  $i$  met significante differentiële expressie:

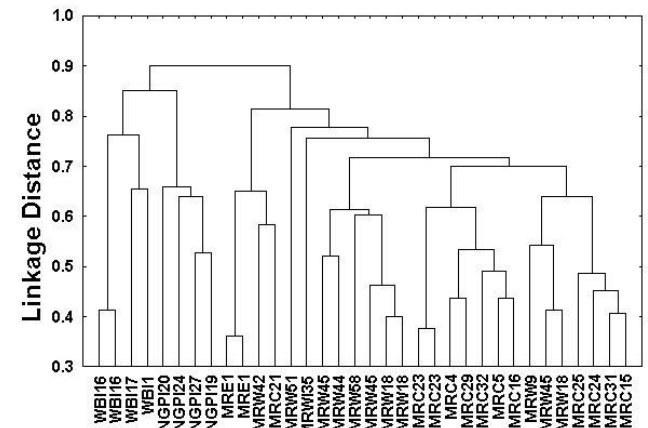
$$p_{\text{adjust}, i} < \alpha$$

- Doe vervolg analyse met deze significante DEGs:
  - clustering
  - ...

$p$ -waarde per gen

# CLUSTERING

- Groeperen van
  - samples (tumor stage I, II, III, controle)
  - tijdreeks ( $t = 0, 4, 8, 12, 16, 20, 24$  h)
  - genenop “gelijksoortig gedrag”
- In veel dimensies...
- Wat is “gelijksoortig”? → distance / (dis)similarity
- Hoe vorm ik clusters? → linkage
- Wat is biologisch relevant “gedrag”?





## CLUSTERING: GENEN

- Voorbeeld:  $n = 7$  samples (tijdpunten)
- Data matrix:  $M = \log(R/G)$  waarden per sample (norm.)

Gene	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$	$M_5$	$M_7$
gene 1	12.4	12.3	12.1	11.8	11.5	11.1	10.5
gene 2	9.8	10.5	11.5	11.8	11.2	10.6	9.9
gene 3	2.5	2.8	2.3	2.4	2.6	2.5	2.6
...							
gene $i$	6.7	6.8	6.5	6.1	5.9	5.7	5.6
...							
gene $j$	8.7	8.5	8.3	8.2	8.6	8.7	8.9
...							
gene $G$	3.8	3.9	4.9	4.2	3.8	3.6	4.2

## CLUSTERING: GENEN

- Voorbeeld:  $n = 7$  samples (tijdpunten)
- Data matrix:  $M = \log(R/G)$  waarden per sample

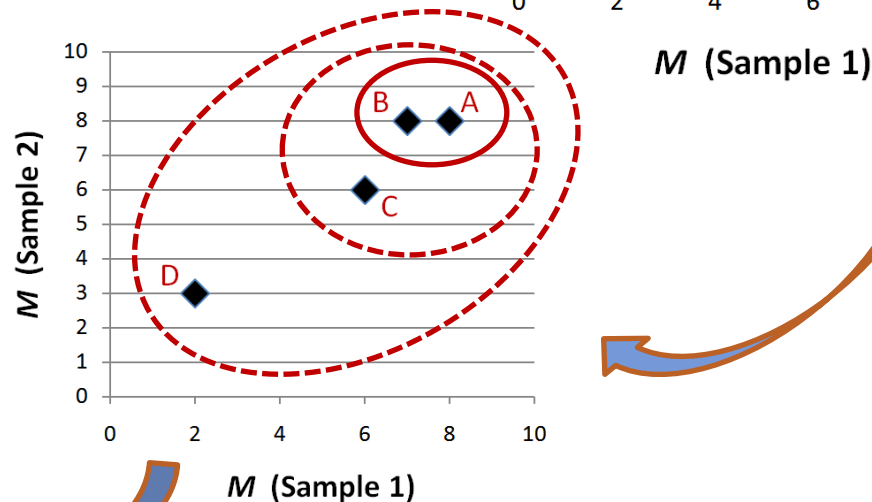
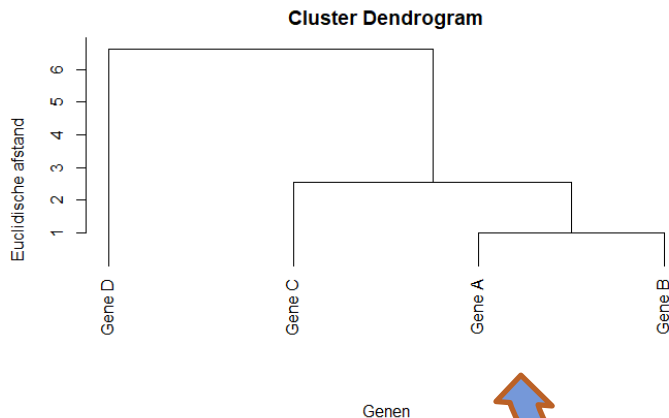
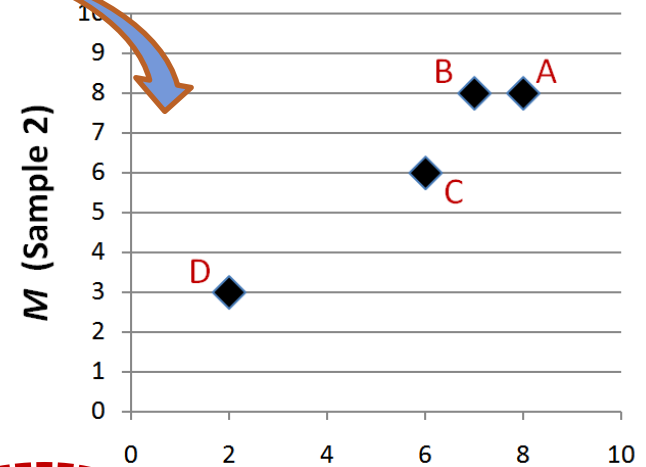
Gene	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$	$M_5$	$M_7$
gene 1	12.4	12.3	12.1	11.8	11.5	11.1	10.5
gene 2	9.8	10.5	11.5	11.8	11.2	10.6	9.9
gene 3	2.5	2.8	2.3	2.4	2.6	2.5	2.6
...							
gene $i$	6.7	6.8	6.5	6.1	5.9	5.7	5.6
...							
gene $j$	8.7	8.5	8.3	8.2	8.6	8.7	8.9
...							
						3.6	4.2

Hoe sterk “lijken” gene  $i$  en gene  $j$  op elkaar?

# CLUSTERING: AFSTAND (DISTANCE)

- Stel:  $n = 2$  samples, per sample 1  $M$ -waarde per gen

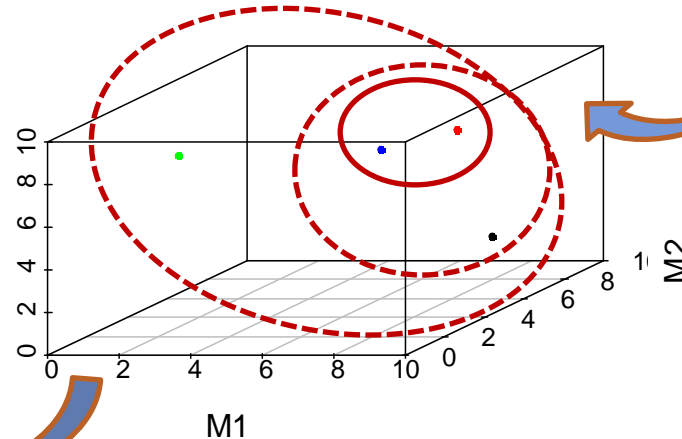
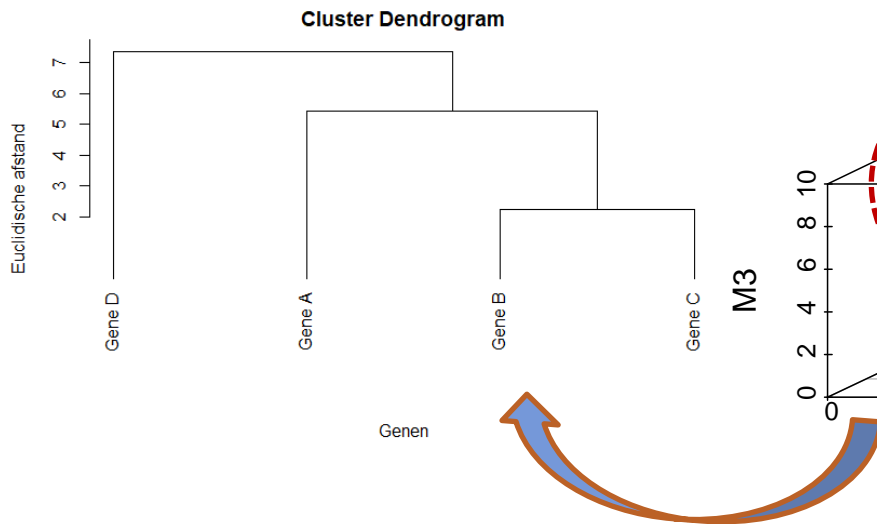
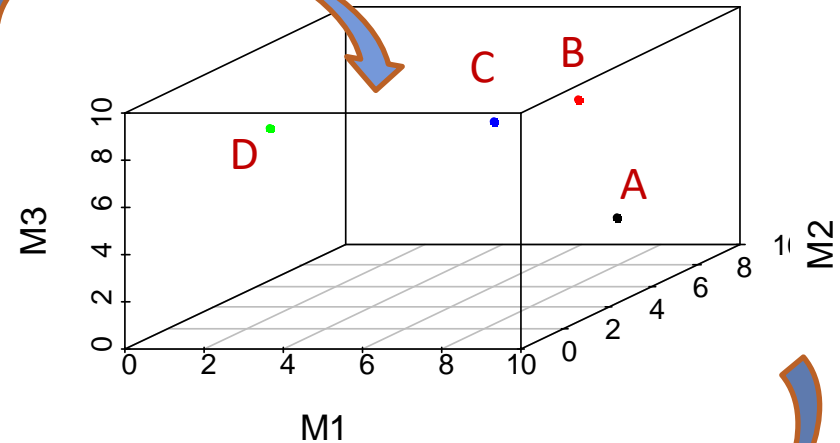
Gene	$M_1$	$M_2$
Gene A	8.0	8.0
Gene B	7.0	8.0
Gene C	6.0	6.0
Gene D	2.0	3.0



# CLUSTERING: AFSTAND (DISTANCE)

- Stel:  $n = 3$  samples, per sample 1  $M$ -waarde per gen

Gene	$M_1$	$M_2$	$M_3$
Gene A	8.0	8.0	2.0
Gene B	7.0	8.0	7.0
Gene C	6.0	6.0	7.0
Gene D	2.0	3.0	8.0



## CLUSTERING: GENEN

- Voorbeeld:  $n = 7$  samples (tijdpunten)
- Data matrix:  $M = \log(R/G)$  waarden per sample

Gene	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$	$M_5$	$M_7$
gene 1	12.4	12.3	12.1	11.8	11.5	11.1	10.5
gene 2	9.8	10.5	11.5	11.8	11.2	10.6	9.9
gene 3	2.5	2.8	2.3	2.4	2.6	2.5	2.6
...							
gene $i$	6.7	6.8	6.5	6.1	5.9	5.7	5.6
...							
gene $j$	8.7	8.5	8.3	8.2	8.6	8.7	8.9
...							
						3.6	4.2

Hoe sterk “lijken” gene  $i$  en gene  $j$  op elkaar?

## CLUSTERING: GENEN

- Op elkaar lijken = ruimtelijk dichtbij = kleine afstand
  - De log expressieratio's  $M$  per gen vormen een **punt** in een **7-dimensionale ruimte**...
  - Genen (= punten in 7D) die **sterk op elkaar lijken** hebben in 7D een **kleine afstand** (kleine **distance**  $d_{ij}$ )
- Op elkaar lijken = gecorreleerd gedrag
  - De 7 expressieratio's  $M$  per gen vormen een **tijdreeks**...
  - Genen (= tijdreeksen) die **sterk op elkaar lijken qua gedrag** zijn sterk **gecorreleerd** (hoge **correlatie coefficient**  $r_{ij}$ ), dus kleine **distance**  $1 - r_{ij}$

# CLUSTERING: DISTANCE / (DIS)SIMILARITY

○ Afstand tussen genen  $i$  en  $j$ , op basis van  $n$  samples

- Euclidische afstand:

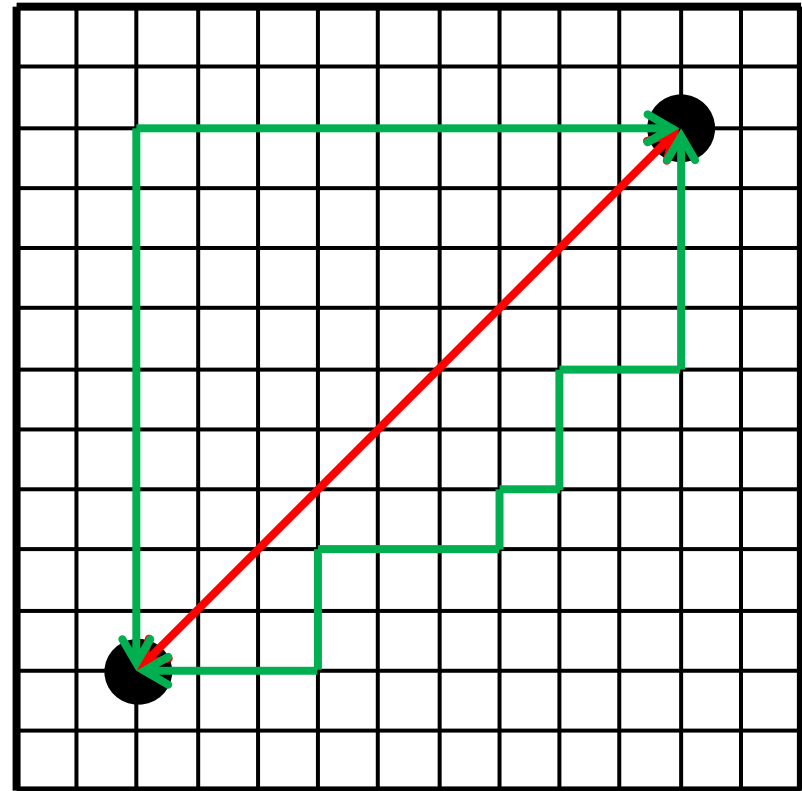
$$d_{ij} = \sqrt{\sum_{k=1}^n (M_{i,k} - M_{j,k})^2}$$

- Manhattan afstand:

$$d_{ij} = \sum_{k=1}^n |M_{i,k} - M_{j,k}|$$

- Minkowski afstand:

$$d_{ij} = \sqrt[p]{\sum_{k=1}^n |M_{i,k} - M_{j,k}|^p}$$



# CLUSTERING: DISTANCES IN R

- Dataframe **M** met  $G$  rijen (genes) en  $n$  kolommen (samples/tijdpunten), 1  $M$ -waarde per gen/sample

Gene	$M_1$	$M_2$	$M_3$	...	$M_7$
Gene 1	12.4	12.3	12.1	...	10.5
...					
Gene $i$	6.7	6.8	6.5	...	5.6
...					
Gene $G$	3.8	3.9	4.9	...	4.2

- Berekenen van distance matrix **dMat**:

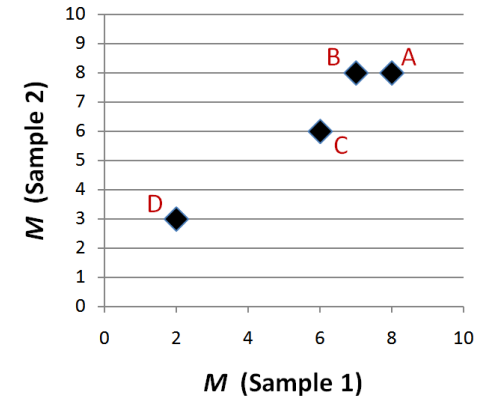
- `dMat <- dist(M, method="euclidean")`
- `dMat <- dist(M, method="manhattan")`
- `dMat <- dist(M, method="minkowski",  
p=3)`



# CLUSTERING: DISTANCES IN R

○ Voorbeeld: **M** =

Gene	$M_1$	$M_2$
Gene A	8.0	8.0
Gene B	7.0	8.0
Gene C	6.0	6.0
Gene D	2.0	3.0



• `dMat <- dist(M, method="euclidean")`

○ Resultaat:

```
> dMat
```

```

      Gene A  Gene B  Gene C
Gene B 1.000000
Gene C 2.828427 2.236068
Gene D 7.810250 7.071068 5.000000

```

$$\sqrt{(8-7)^2 + (8-8)^2}$$

$$\sqrt{(7-6)^2 + (8-6)^2}$$

$$\sqrt{(8-2)^2 + (8-3)^2}$$

## CLUSTERING: ALGORITME (HIËRARCHISCH, BOTTOM-UP)

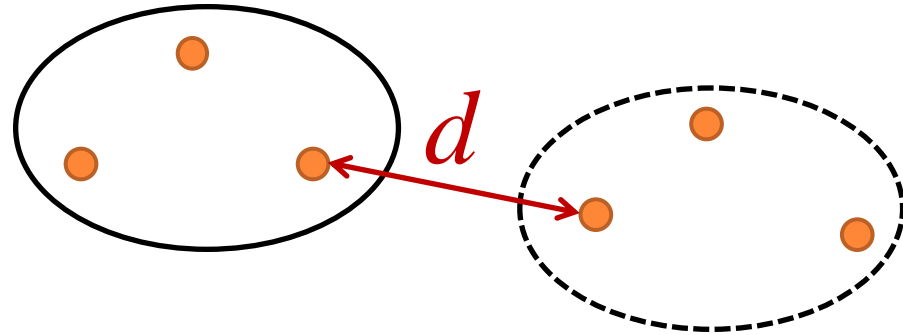
- Stap 1: bereken tussen alle  $R$  genen de distances
- Stap 2: zoek de 2 genen met de kleinste distances, dit wordt een nieuwe cluster
- Stap 3: bereken tussen alle  $R - 2$  genen en de 1<sup>e</sup> cluster de distances
- Stap 4: zoek de kleinste distance, dit wordt een nieuwe cluster
- etc.

*Linkage* = afstand tussen punt en cluster  
of tussen clusters

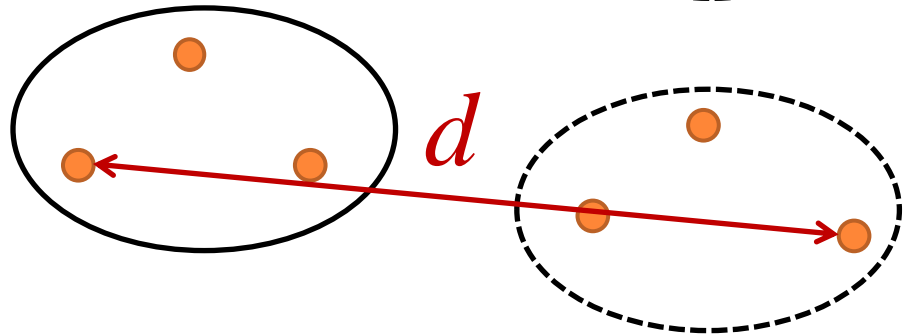
```
> dMat
      Gene A   Gene B   Gene C
Gene B 1.000000
Gene C 2.828427 2.236068
Gene D 7.810250 7.071068 5.000000
```

## CLUSTERING: LINKAGE = DISTANCE TUSSEN CLUSTERS

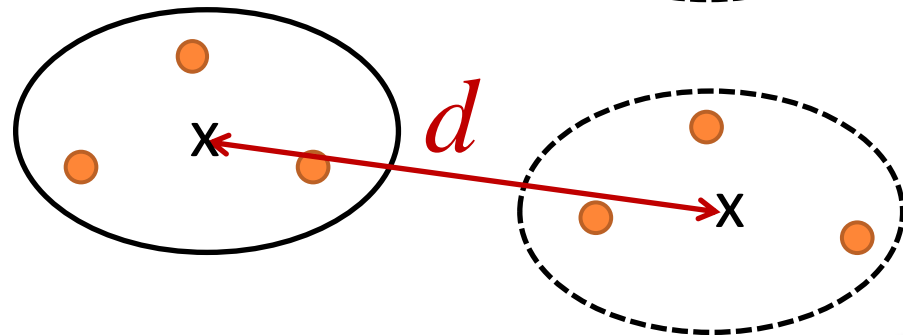
○ Single linkage:



○ Complete linkage:



○ Average linkage:



# CLUSTERING: VOORBEELD SINGLE LINKAGE (1)

○ Data:

Gene	$M_1$	$M_2$
Gene A	8.0	8.0
Gene B	7.0	8.0
Gene C	6.0	6.0
Gene D	2.0	3.0

○ Stap 1: (Euclidean) distances:

	Gene A	Gene B	Gene C
Gene B	1.000		
Gene C	2.828	2.236	
Gene D	7.810	7.071	5.000

○ Stap 2: Kleinste afstand A-B = 1.000 → cluster  $C_1$

## CLUSTERING: VOORBEELD SINGLE LINKAGE (2)

- Stap 1: (Euclidean) distances:

	Gene A	Gene B	Gene C
Gene B	1.000		
Gene C	2.828	2.236	
Gene D	7.810	7.071	5.000

- Stap 2: Kleinste afstand  $A-B = 1.000 \rightarrow$  cluster  $C_1$
- Stap 3: Bereken nieuwe distances (linkage = single)

	$C_1$	Gene C
Gene C	2.236	
Gene D	7.071	5.000

## CLUSTERING: VOORBEELD SINGLE LINKAGE (3)

- Stap 3: (Euclidean) distances:

	$C_1$	Gene C
Gene C	2.236	
Gene D	7.071	5.000

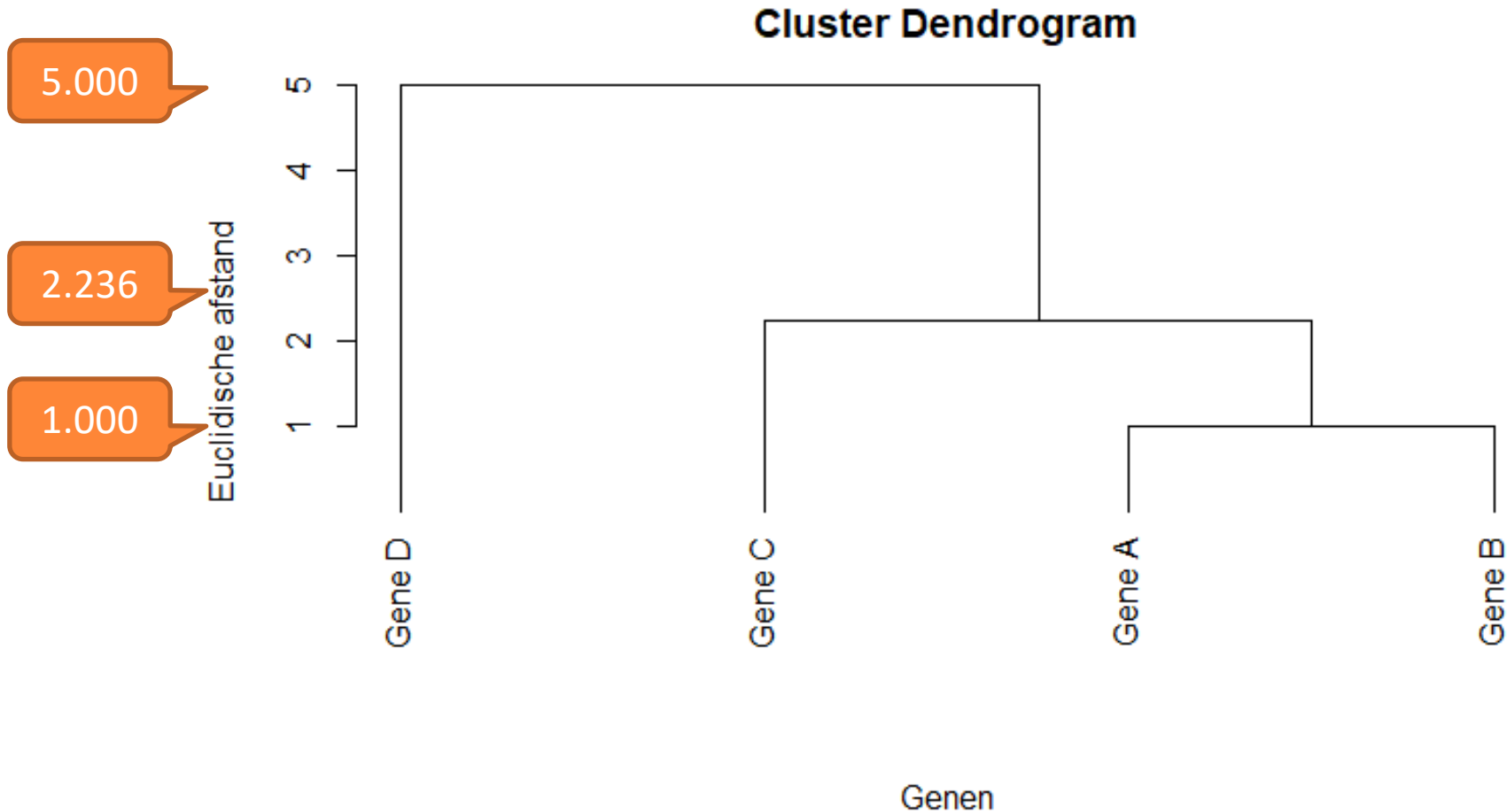
- Stap 4: Kleinste afstand  $C-C_1 = 2.236 \rightarrow$  cluster  $C_2$
- Stap 5: Bereken nieuwe distances (linkage = single)

	$C_2$
Gene D	5.000

- Stap 6: Kleinste afstand  $D-C_2 = 5.000 \rightarrow$  cluster  $C_3$

# CLUSTERING: VOORBEELD SINGLE LINKAGE (4)

○ Output:



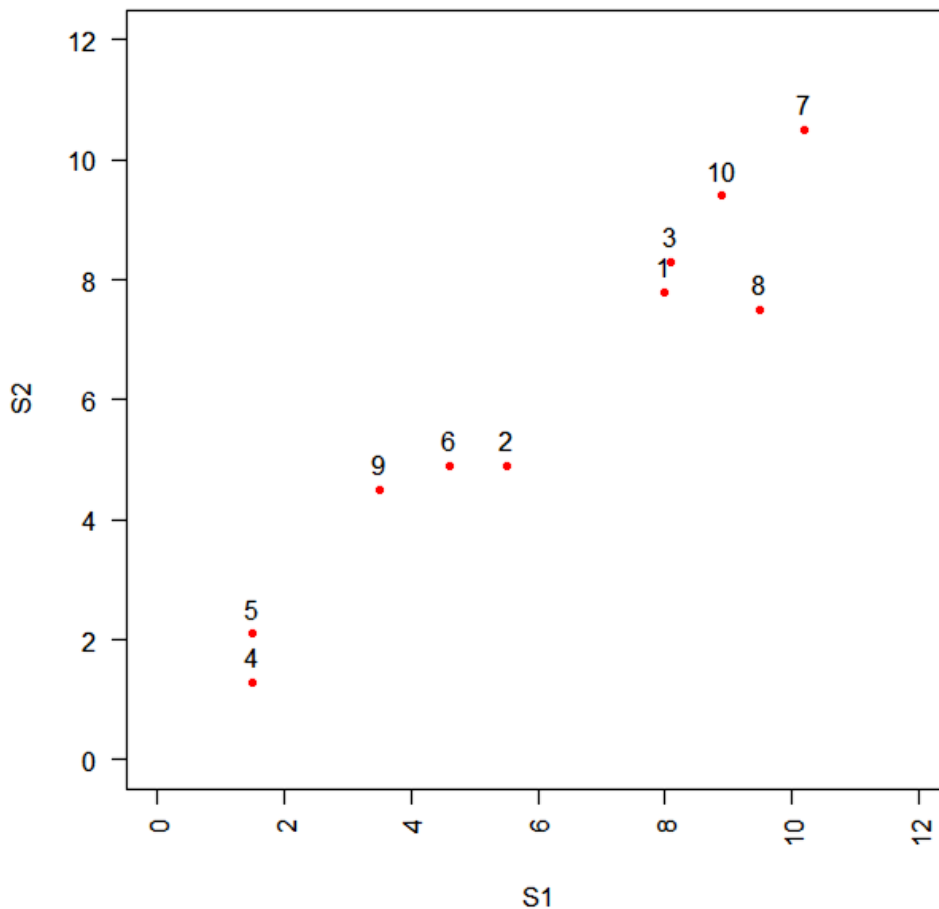
## CLUSTERING: HIËRARCHISCH CLUSTEREN IN R

- Distance matrix **dMat** bevat alle afstanden tussen alle genen
- Berekenen van dendrogram **clust**:
  - `clust <- hclust(dMat, method="single")`
  - `clust <- hclust(dMat, method="average")`
  - `clust <- hclust(dMat, method="complete")`
- Plotten dendrogram:
  - `plot(clust, xlab= , ylab= , ...)`



# CLUSTERING DEG's

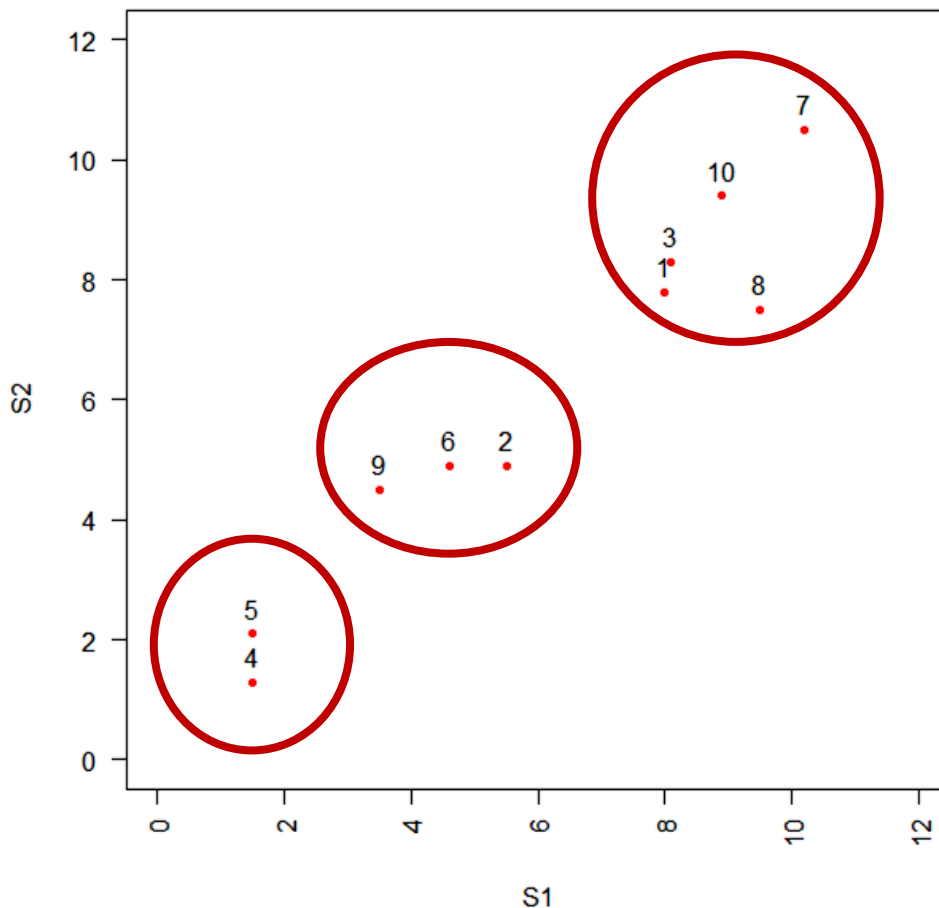
- Voorbeeld: 10  $M$ -waarden van DEG's in 2 situaties S1 en S2 (bijv. tijdstippen, gehalten trigger molecuul X, etc.)



	row.names	S1	S2
1	gene 1	8	7.8
2	gene 2	5.5	4.9
3	gene 3	8.1	8.3
4	gene 4	1.5	1.3
5	gene 5	1.5	2.1
6	gene 6	4.6	4.9
7	gene 7	10.2	10.5
8	gene 8	9.5	7.5
9	gene 9	3.5	4.5
10	gene 10	8.9	9.4

# CLUSTERING DEG's

- Voorbeeld: 10 *M*-waarden van DEG's in 2 situaties S1 en S2 (bijv. tijdstippen, gehalten trigger molecuul X, etc.)



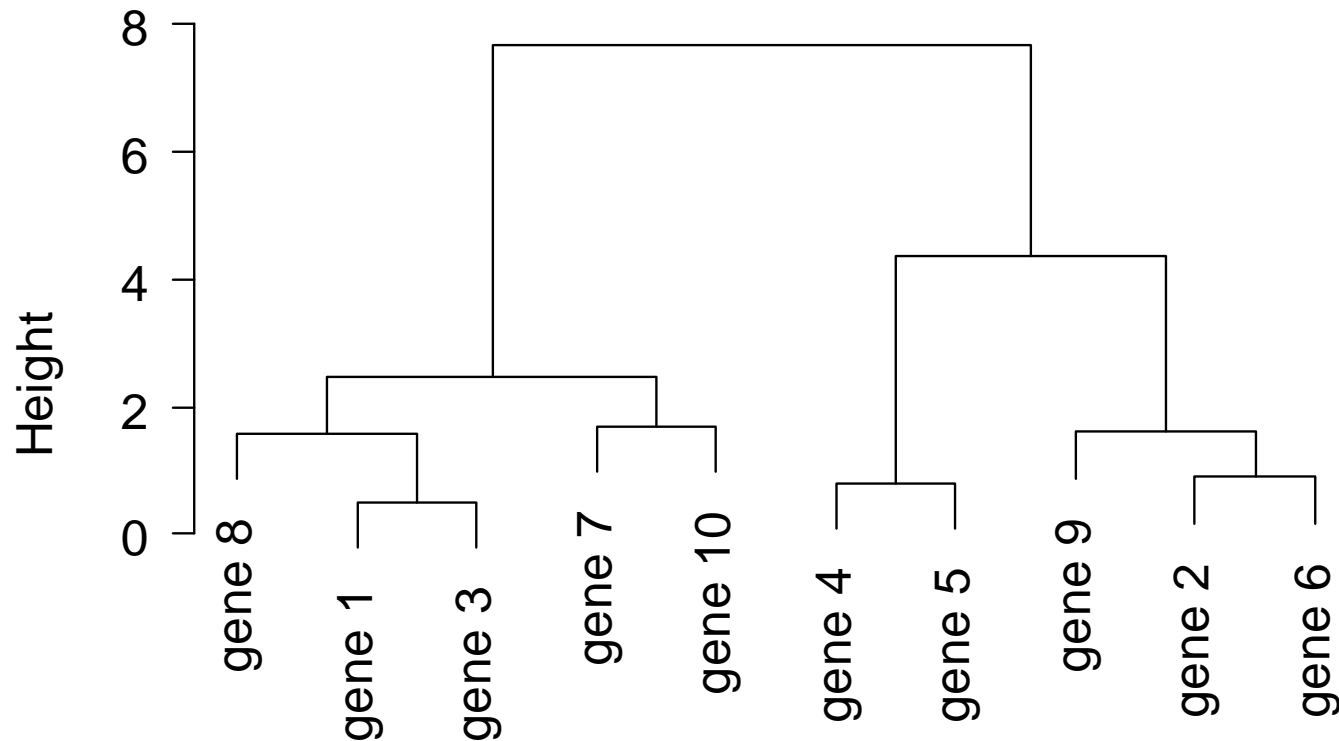
	row.names	S1	S2
1	gene 1	8	7.8
2	gene 2	5.5	4.9
3	gene 3	8.1	8.3
4	gene 4	1.5	1.3
5	gene 5	1.5	2.1
6	gene 6	4.6	4.9
7	gene 7	10.2	10.5
8	gene 8	9.5	7.5
9	gene 9	3.5	4.5
10	gene 10	8.9	9.4

## CLUSTERING DEG's

- Clustering van dataframe/matrix **X** met 2 samples per row, en 10 genen (=rows):
- Berekenen van afstandsmatrix:
  - **dMat <- dist(X, method="euclidean")**
- Bereken van clustering:
  - **clust <- hclust(dMat, method="average")**
- Plotten van dendrogram:
  - **plot(clust)**

# CLUSTERING DEG's

○ Resultaat: **Cluster Dendrogram**



d.E  
hclust (\*, "average")

## CLUSTERING: PEARSON DISTANCE

- Euclidische/Manhattan/Minkowski distance:
  - gelijkenis in (absolute) waarden
- Pearson distance:
  - gelijkenis in variatie (“gedrag/patroon”)

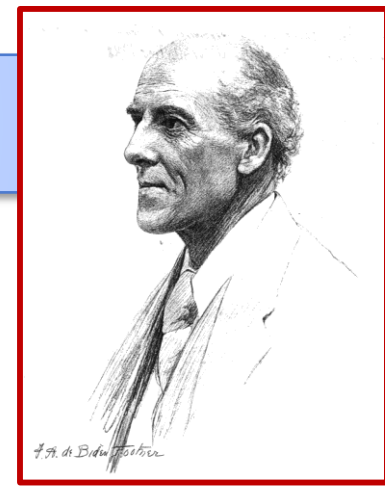
$$d_{ij} = 1 - r_{ij}$$

met

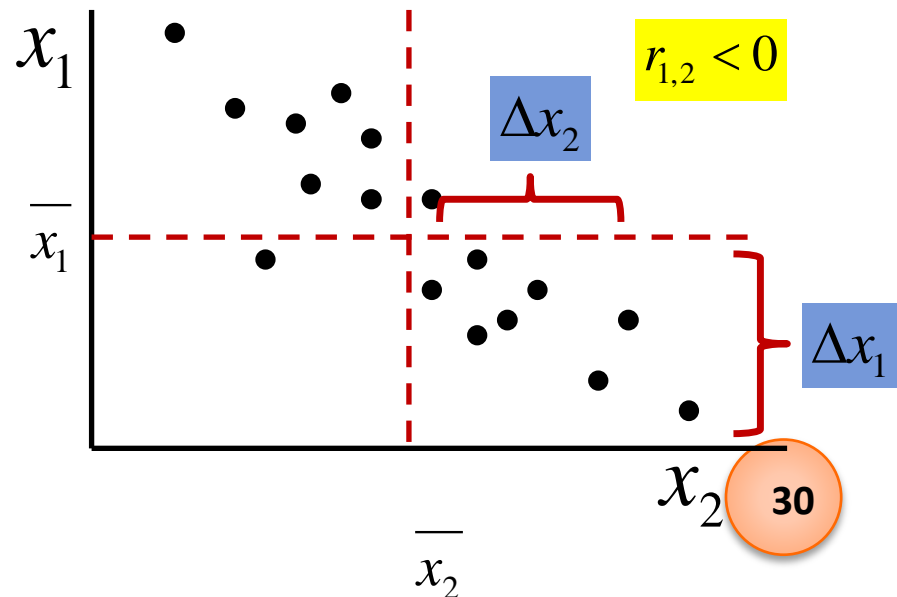
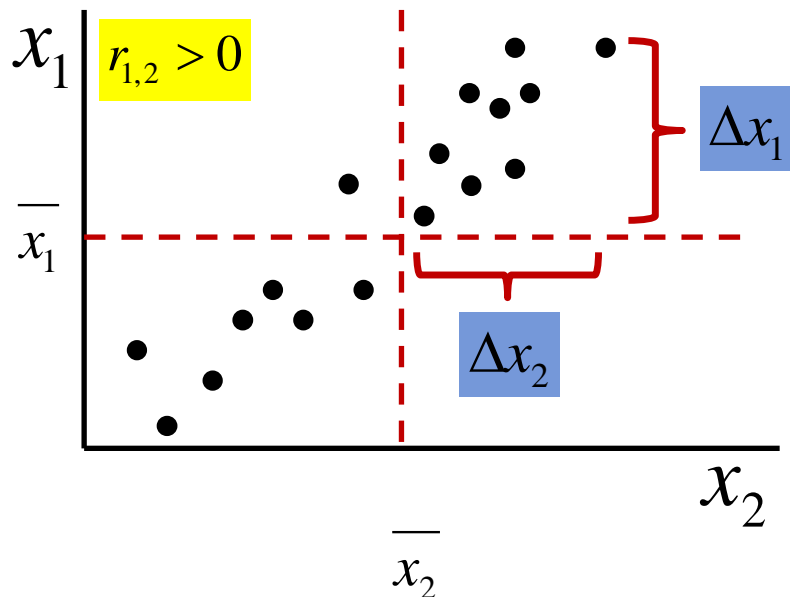
Pearson correlatie  
coëfficiënt

$$r_{ij} = \frac{\sum_{k=1}^n (M_{i,k} - \overline{M_{i\bullet}}) \cdot (M_{j,k} - \overline{M_{j\bullet}})}{\sqrt{\sum_{k=1}^n (M_{i,k} - \overline{M_{i\bullet}})^2} \cdot \sqrt{\sum_{k=1}^n (M_{j,k} - \overline{M_{j\bullet}})^2}}$$

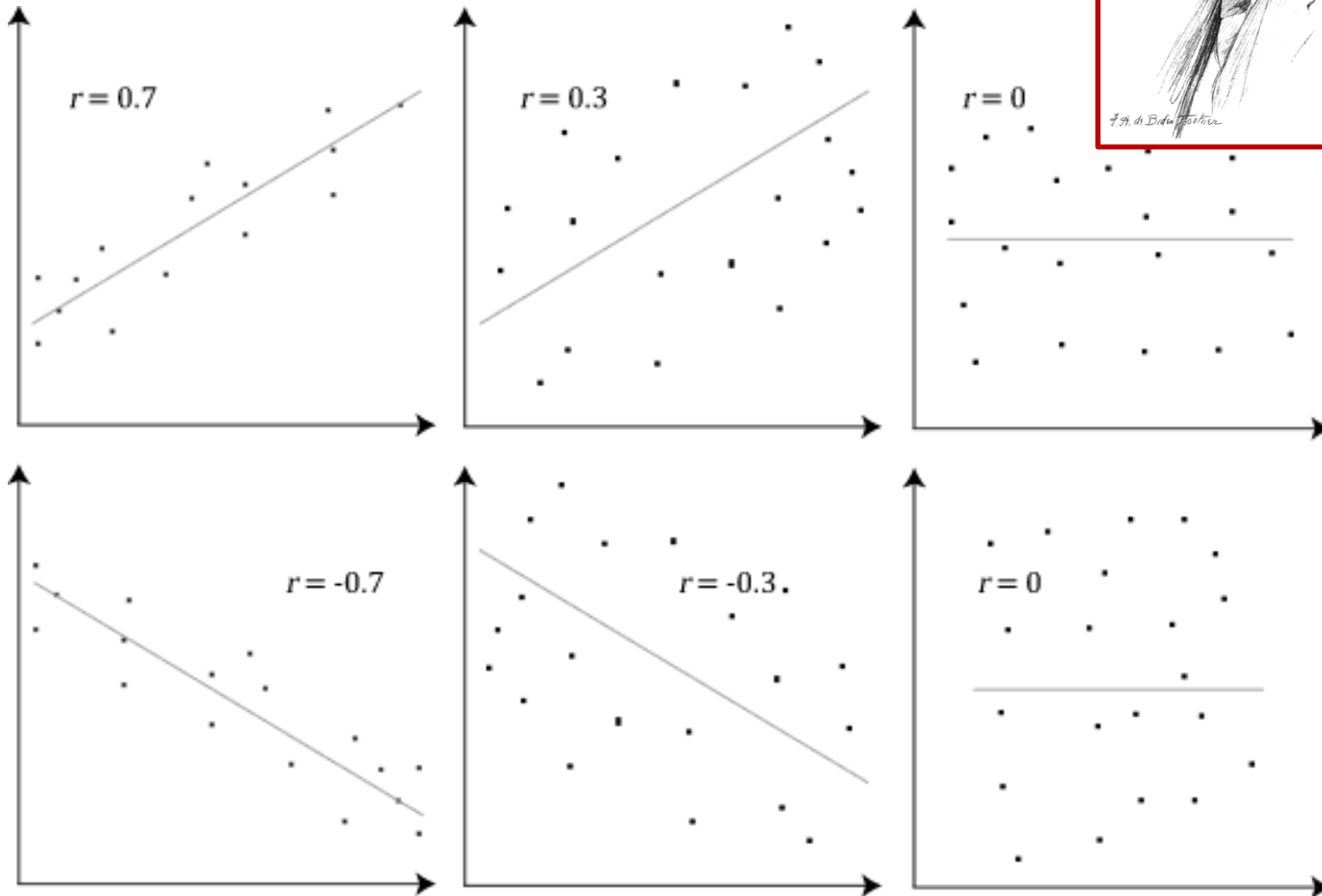
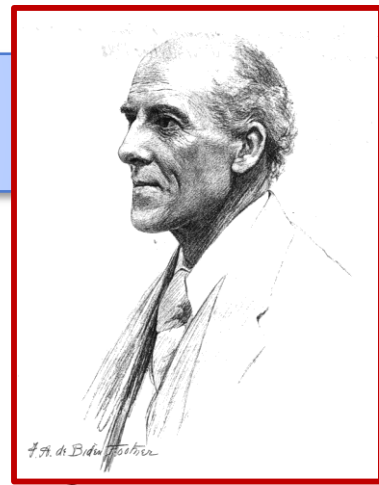
# PEARSON CORRELATIE COËFFICIËNT $r_{ij}$



$$r_{1,2} = \frac{\sum_{t=1}^N (x_{1,t} - \bar{x}_1) \cdot (x_{2,t} - \bar{x}_2)}{\sqrt{\sum_{t=1}^N (x_{1,t} - \bar{x}_1)^2} \cdot \sqrt{\sum_{t=1}^N (x_{2,t} - \bar{x}_2)^2}} = \frac{\sigma_{1,2}^2}{\sigma_1 \cdot \sigma_2} = \frac{\text{cov}(x_1, x_2)}{\sqrt{\text{var}(x_1)} \cdot \sqrt{\text{var}(x_2)}}$$



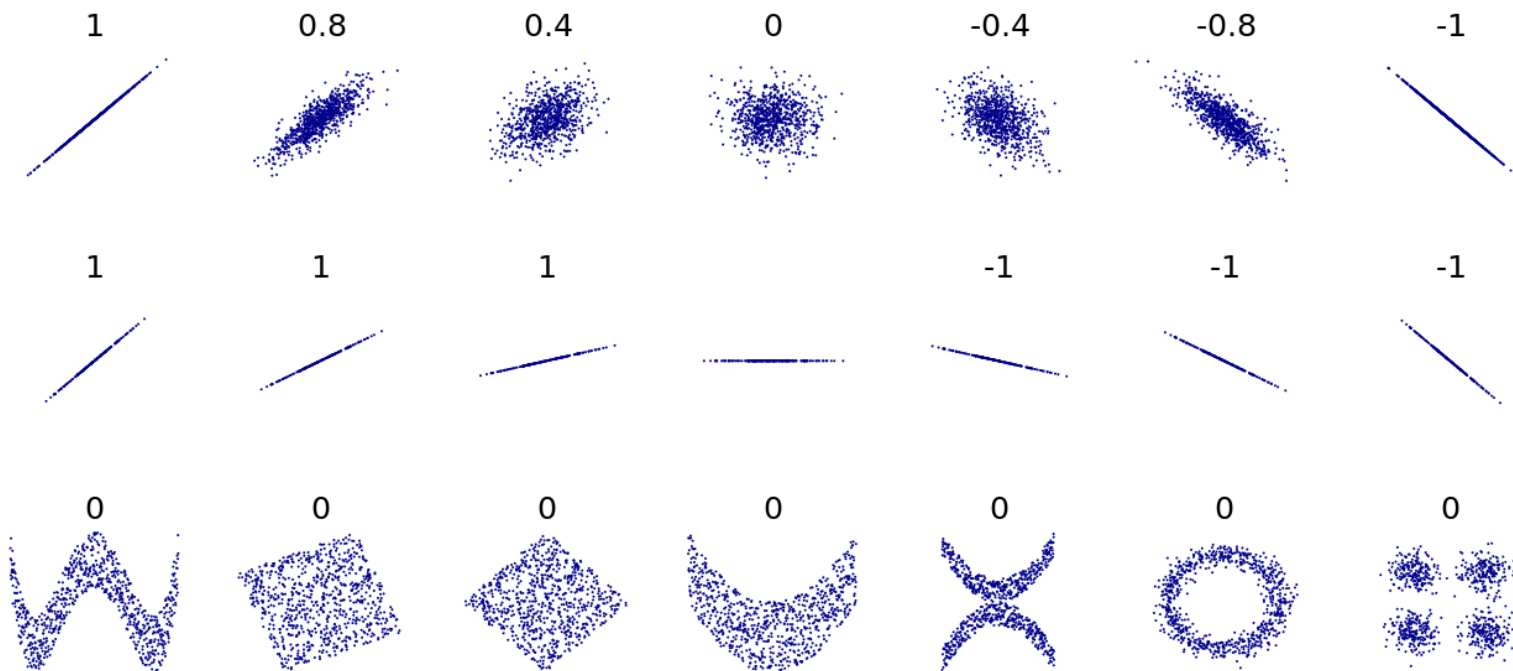
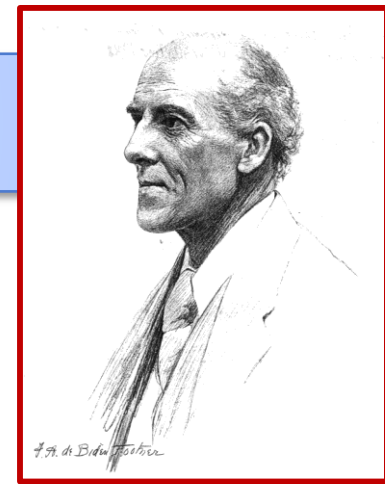
# PEARSON CORRELATIE COËFFICIËNT $r_{ij}$



# PEARSON CORRELATIE COËFFICIËNT $r_{ij}$

Let op!

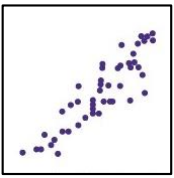
- Correlatie gaat over lineaire relaties
- Correlatie is niet oorzakelijk verband



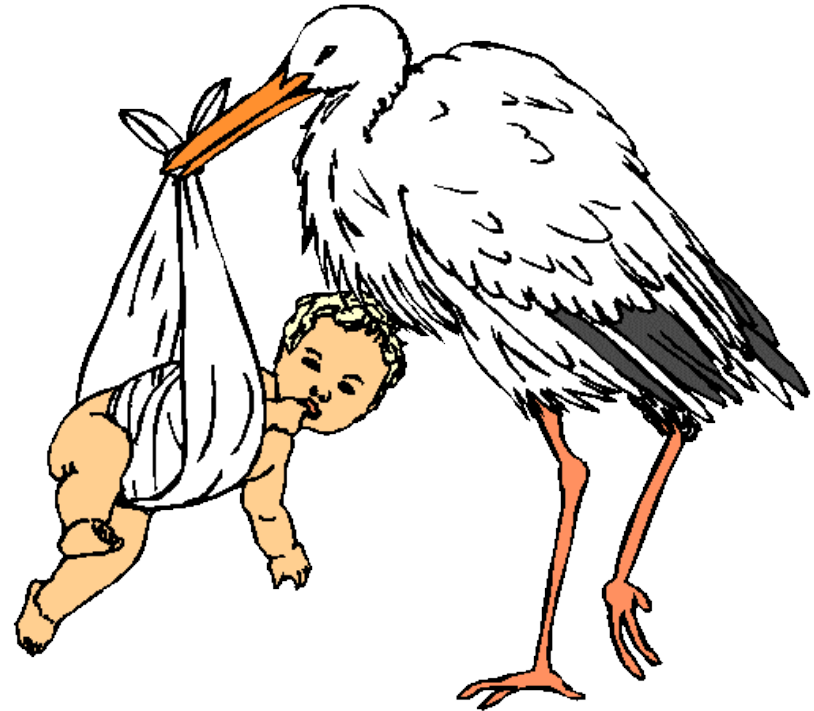
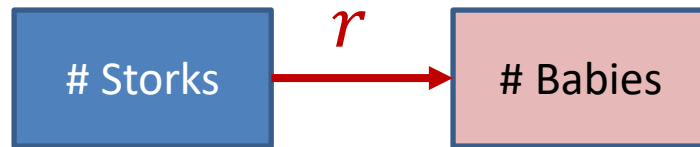




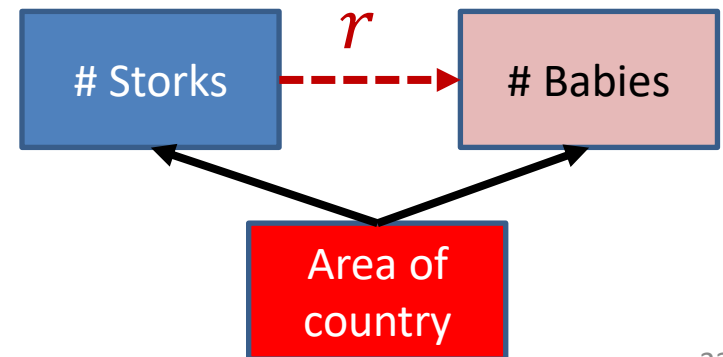
# Correlation: interpretation



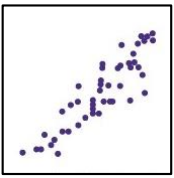
There is a significant correlation between number of storks in a country and the number of babies born! Do storks therefore deliver babies?



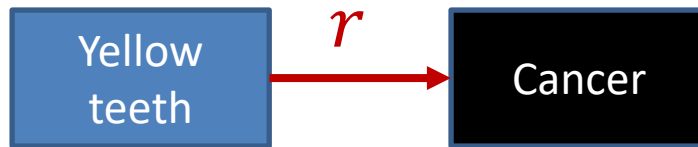
- What could be a 3<sup>rd</sup> factor?
- Size of the country!



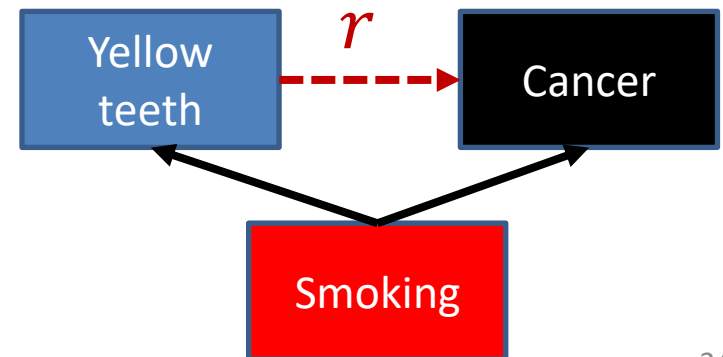
# Correlation: interpretation



There is a significant correlation between having yellow teeth and dying from lung cancer! Do yellow teeth therefore cause cancer?



- What could be a 3<sup>rd</sup> factor?
- Smoking!

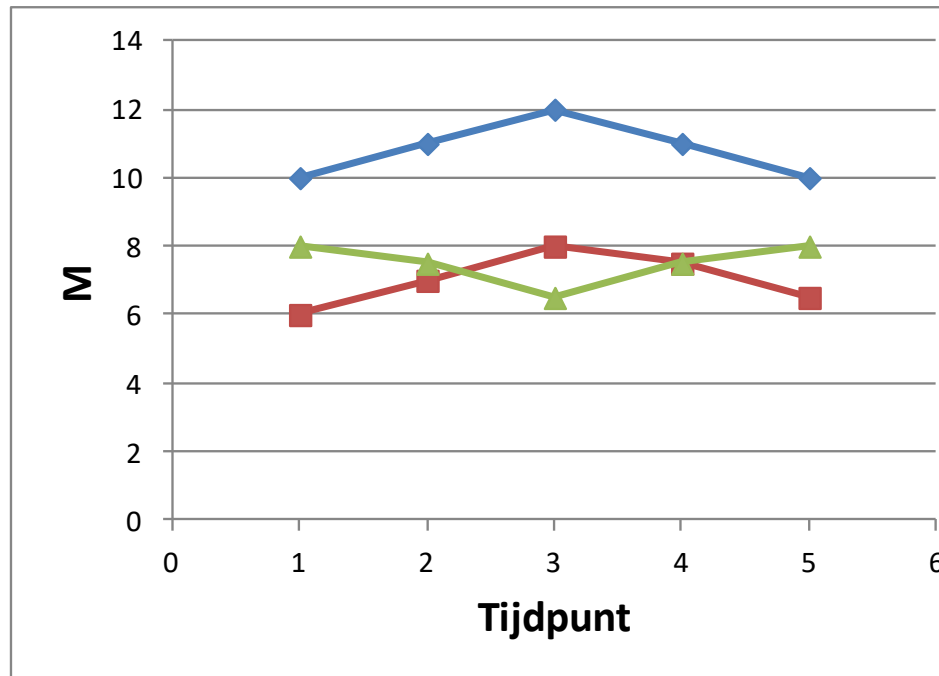


# PEARSON CORRELATIE COËFFICIËNT

## ○ Betekenis:

$r_{ij}$	Betekenis
1	Perfekte overeenkomst (correlatie) tussen $i$ en $j$
0	Geen overeenkomst (correlatie) tussen $i$ en $j$
-1	Perfekte anticorrelatie tussen $i$ en $j$

## ○ Voorbeelden:



$r_{ij}$	M1	M2	M3
M1	1		
M2	0,945	1	
M3	-0,976	-0,904	1

◆ M1  
 ■ M2  
 ▲ M3

$$d_{ij} = 1 - r_{ij}$$

$d_{ij}$	M1	M2	M3
M1	0		
M2	0,055089	0	
M3	1,976	1,904	0

## PEARSON DISTANCE IN R

- Pearson distance  $d = 1 - r$  ligt tussen 0 (nl.  $r = 1$ , perfecte correlatie) en 2 (nl.  $r = -1$ , antigeccorreleerd)
- Berekenen distance matrix van frame **M**:
  - `dMat <- as.dist( 1 - cor( t(M) ) )`

`hclust( )` wil een  
“distance” object

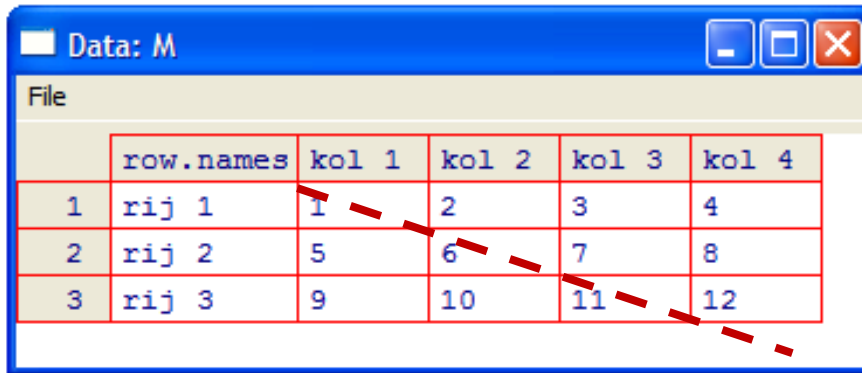
`cor( )` berekent  $r$   
tussen de  
KOLOMMEN van  
matrix

“getransponeerde”  
van M

correlatie tussen rijen van M

# TRANSPONEREN VAN MATRIX (DATAFRAME)

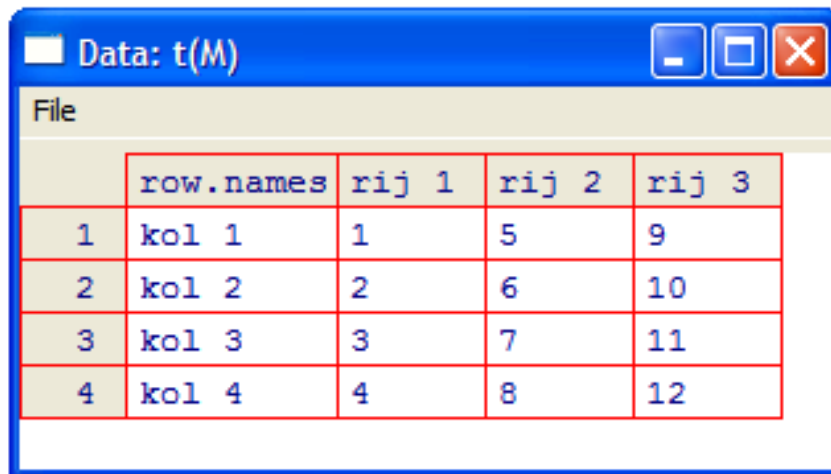
- Transponeren (functie  $\mathbf{t}(\ )$  ) betekent omklappen in de diagonaal, oftewel: kolom  $\rightarrow$  rij, en rij  $\rightarrow$  kolom:



	row.names	kol 1	kol 2	kol 3	kol 4
1	rij 1	1	2	3	4
2	rij 2	5	6	7	8
3	rij 3	9	10	11	12

$$M = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}$$

- $\mathbf{t}(M)$



	row.names	rij 1	rij 2	rij 3
1	kol 1	1	5	9
2	kol 2	2	6	10
3	kol 3	3	7	11
4	kol 4	4	8	12

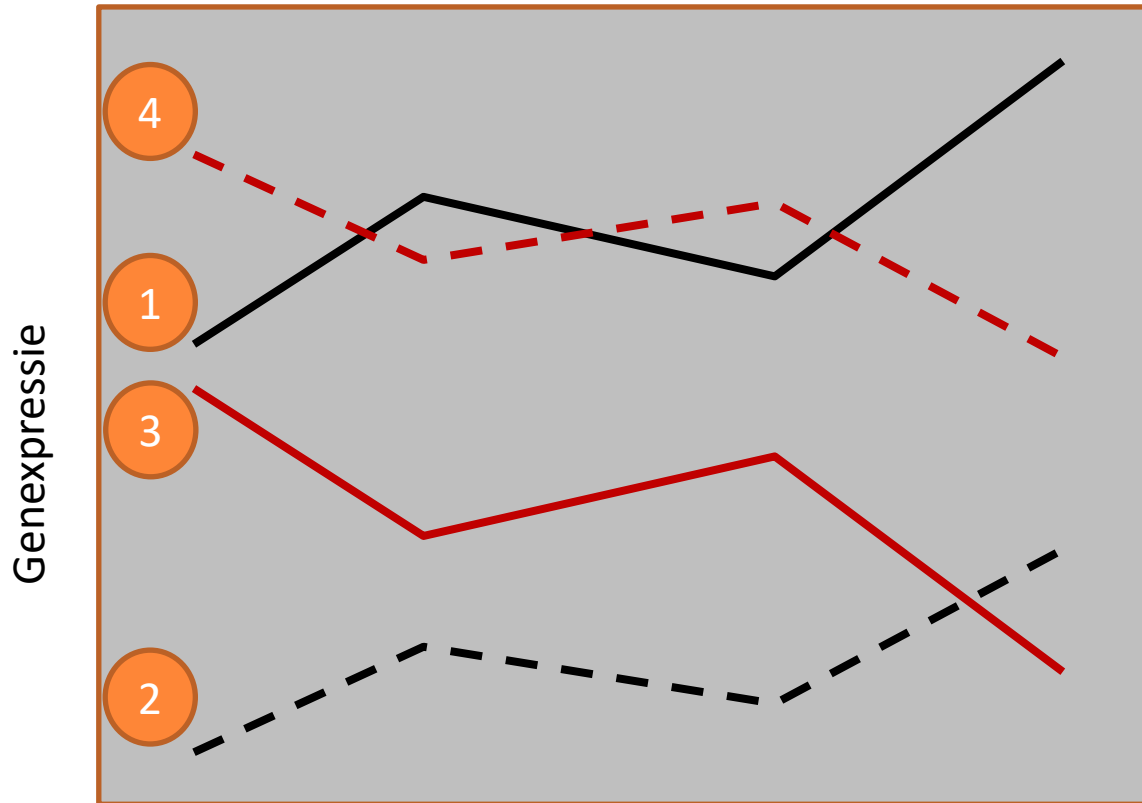
$$M^T = \begin{pmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{pmatrix}$$

## VARIATIES PEARSON DISTANCE

- “Gewone” Pearson distance:  $d_{ij} = 1 - r_{ij}$ 
  - Eigenschap: groot verschil tussen positieve en negatief gecorreleerde genen
  - Bereik:  $0 < d_{ij} < 2$
- “Absolute” Pearson distance:  $d_{ij} = 1 - |r_{ij}|$
- of “Squared” Pearson distance:  $d_{ij} = 1 - r_{ij}^2$ 
  - Bereik:  $0 < d_{ij} < 1$
  - Eigenschap: verschil tussen wel (positief/negatief) en niet gecorreleerde genen

# VARIATIES PEARSON DISTANCE

- Gewone vs. Absolute/squared Pearson distance:



Euclidisch:

$$d_{14} < d_{12}$$

Correlatie:

$$\begin{aligned} r_{12} &\approx r_{34} \approx 1 \\ r_{13} &\approx r_{24} \approx -1 \\ r_{14} &\approx r_{23} \approx -1 \end{aligned}$$

Pearson:

$$d_{14} \gg d_{12}$$

Absolute/squared Pearson:

$$d_{14} \approx d_{12}$$

## “WIJZE” ADVIEZEN

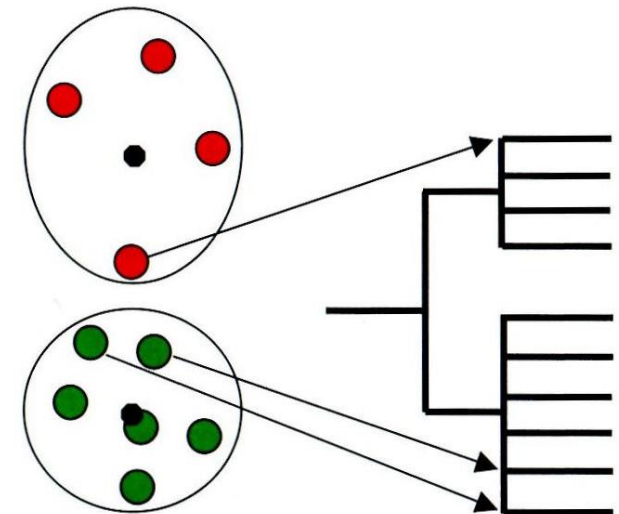
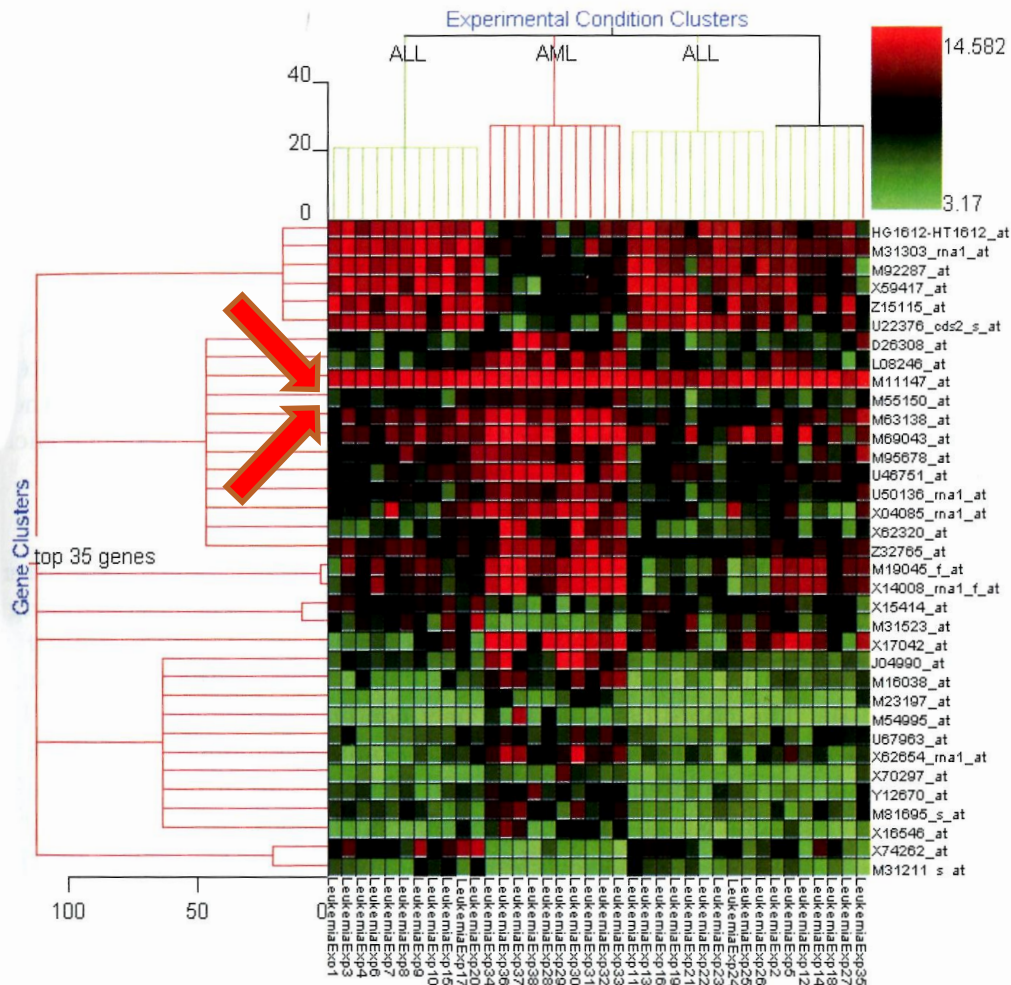


- Alles kun je clusteren
- Clustering is geen doel op zich!
- Clustering hangt sterk af van details:
  - welke distance maat?
  - welke linkage?
  - welke methode?
- Volgorde van (genen) in een cluster zegt niets!
- Cluster alleen een selectie van genen:
  - DEG's
  - Biologisch relevante set, bijv. centraal metabolisme



# VOLGORDE VAN GENEN IN CLUSTER

- Volgorde zegt niets!



## DOEL VAN CLUSTERING

- Vinden van “groep” genen (samples) die biologisch samenhangen
- Link genen binnen cluster aan biologische functie
  - embryonale ontwikkeling
  - centraal metabolisme
  - signaling pathway
  - ...



# WELKE MANIER IS “BESTE”/MEEST GEBRUIKT ?

## ○ Distance maat $d_{ij}$ :

- Clustering genen:
  1. Pearson
  2. Euclidisch
- Clustering samples/tijdpunten:
  1. Euclidisch
  2. Pearson

Geen verschil  
tussen clustering  
van absolute en  
relatieve  
expressies  
(Draghici, 2012)

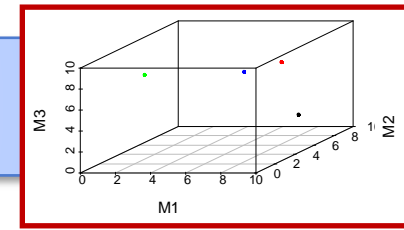
## ○ Linkage:

- average linkage

## ○ Methode:

- Hiërarchisch clusteren, bottum-up (langzaam maar goed)

# 3D PLOTS IN R



- Installeren van package **scatterplot3d**
  - `install.packages("scatterplot3d")`
- Laden package:
  - `library("scatterplot3d")`
- 3D-plot van dataframe **X** (rij = gen, kolom = sample)
- `scatterplot3d(X, xlab="M1", ylab="M2", zlab="M3", pch=20, color=c("black", "red", "blue"), xlim=c(0,10), ylim=c(0,10), zlim=c(0,10))`

# Jullie kunnen nu de opdrachten van les 12 maken



Hanze University Groningen  
APPLIED SCIENCES

Institute for  
Life Science & Technology