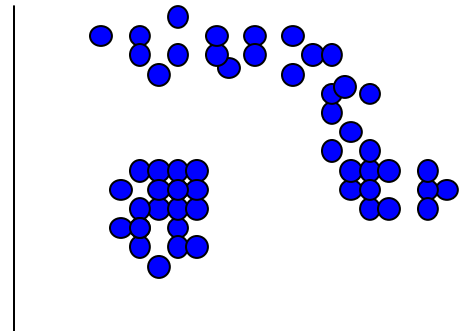
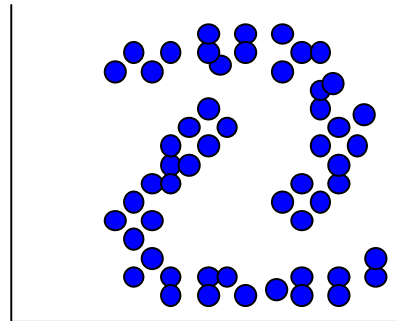
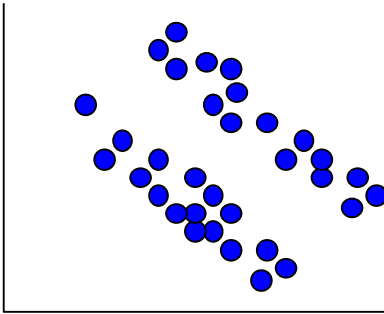
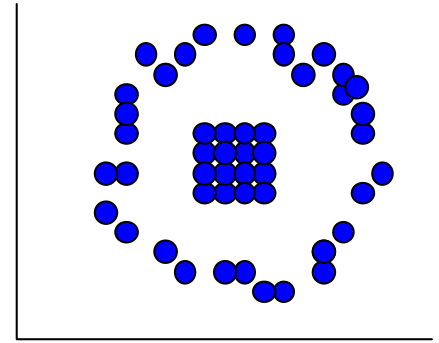
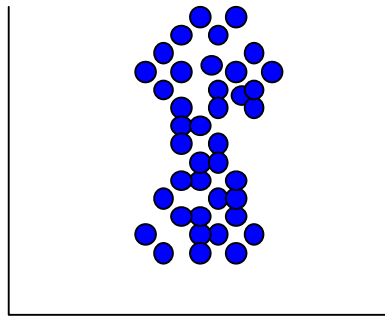
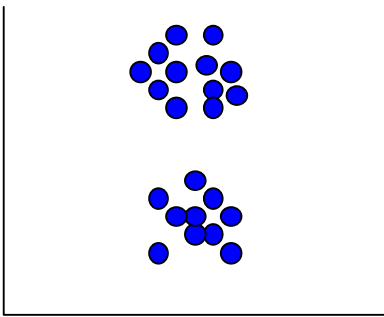


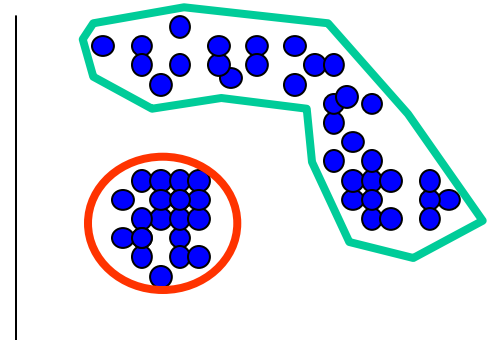
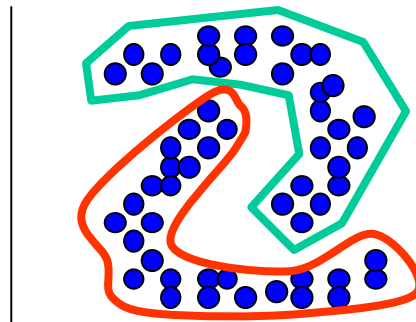
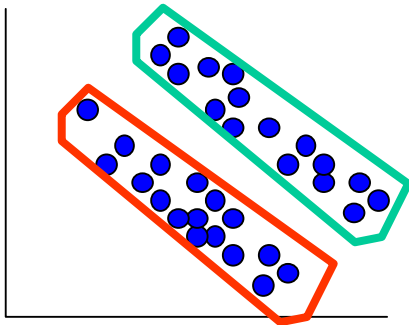
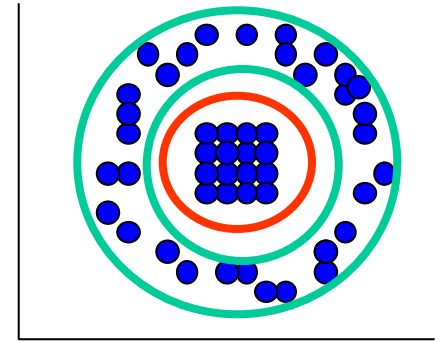
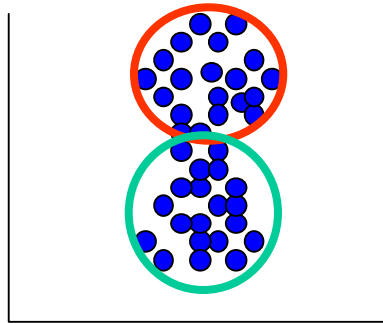
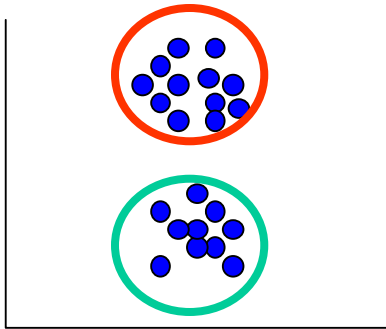
# Clustering

Antoine van Kampen  
Bioinformatics Lab, AMC

# Clusters of Two-Dimensional Data



# Clusters of Two-Dimensional Data



# DNA Microarray

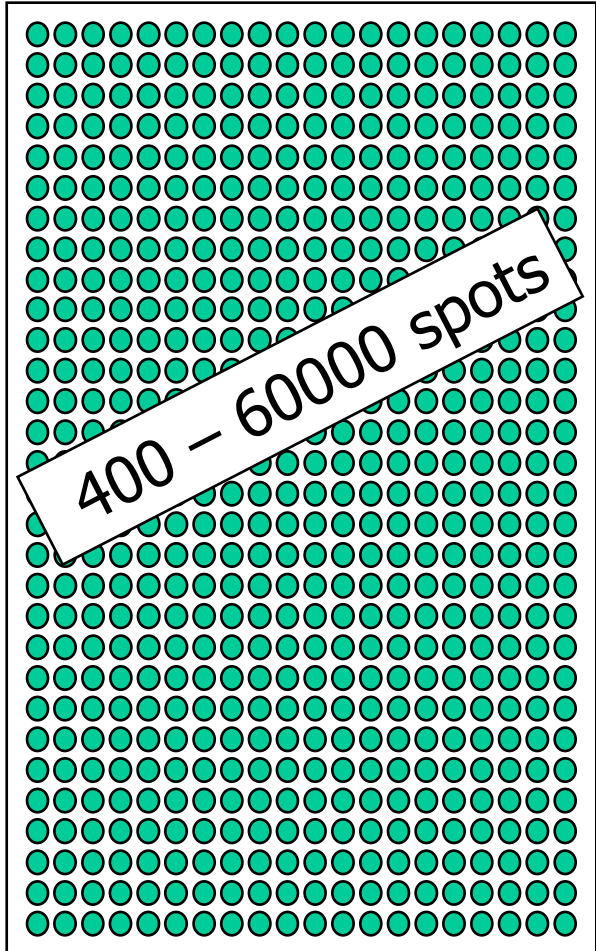
Genome

PCR-amplified

Spotted on slide

GenBank	Name
K00558	tubulin, alpha, ubiquitous
M11886	major histocompatibility complex, class I, C
M26880	ubiquitin C
M86400	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
U14971	ribosomal protein S9
V00530	hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome)
X00351	actin, beta
X01677	glyceraldehyde-3-phosphate dehydrogenase
X56932	ribosomal protein L13a
NM_000476	adenylate kinase 1
NM_003666	basic leucine zipper nuclear factor 1 (JEM-1)
NM_001399	ectodermal dysplasia 1, anhidrotic
NM_005294	G protein-coupled receptor 21
NM_001509	glutathione peroxidase 5 (epididymal sperm-head-related protein)
NM_004145	myosin IXB
NM_002531	neurotensin receptor 1 (high affinity)
NM_006213	phosphorylase kinase
NM_003223	transcription factor 10 (transcription factor 4)
NM_003352	ubiquitin-like protein
NM_005000	...
NM_001000	... (gene)
NM_001000	... repeat domains)
NM_001000	... factor 2 binding protein
NM_001000	... cell signalling 17
NM_001000	... protein
NM_001000	... (nucleoside diphosphate linked moiety X)-type motif 4
NM_001000	prostate tumor over expressed gene 1
U15300	nuclear transcription factor, X-box binding 1
M67454	tumor necrosis factor receptor superfamily, member 6
J03143	interleukin 6 gamma receptor 1
NM_000014	alpha-2-macroglobulin
NM_001693	ATPase, H+ transporting, lysosomal (vacuolar proton pump), beta polypeptide, 56/58kD, isoform 2
NM_001236	carbonyl reductase 3
NM_005248	Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog
NM_002066	GPI anchored molecule like protein
NM_002336	low density lipoprotein receptor-related protein 6
NM_000306	POU domain, class 1, transcription factor 1 (Pit1, growth hormone factor 1)
NM_005012	receptor tyrosine kinase-like orphan receptor 1
NM_003092	small nuclear ribonucleoprotein polypeptide B''
NM_002687	pinin, desmosome associated protein
NM_003792	endothelial differentiation-related factor 1
NM_004294	mitochondrial translational release factor 1
NM_005831	nuclear domain 10 protein
NM_006696	thyroid hormone receptor coactivating protein

Yeast: 6000 genes  
Man: 30000 genes



spots = probes  
(single-stranded cDNA)

# DNA Microarray

Gene Bank

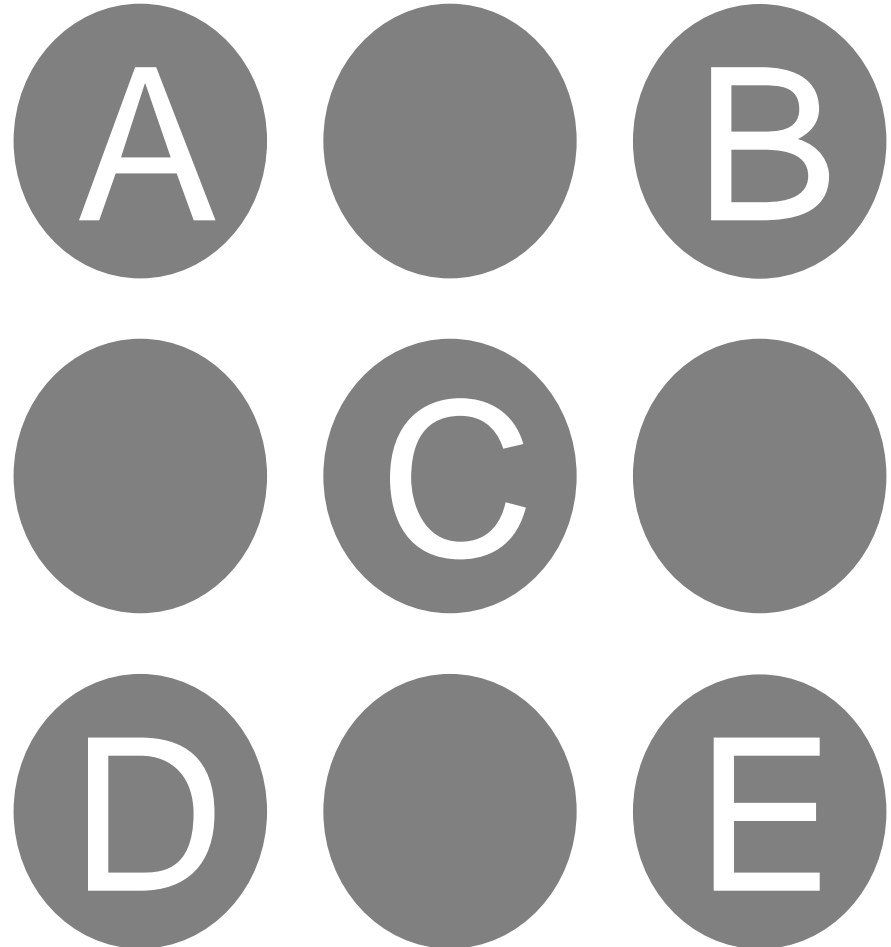
gene A

gene B

gene C

gene D

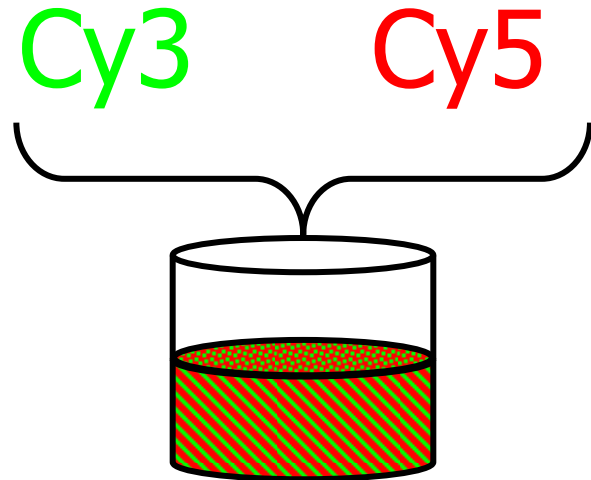
gene E



# mRNA Target Preparation

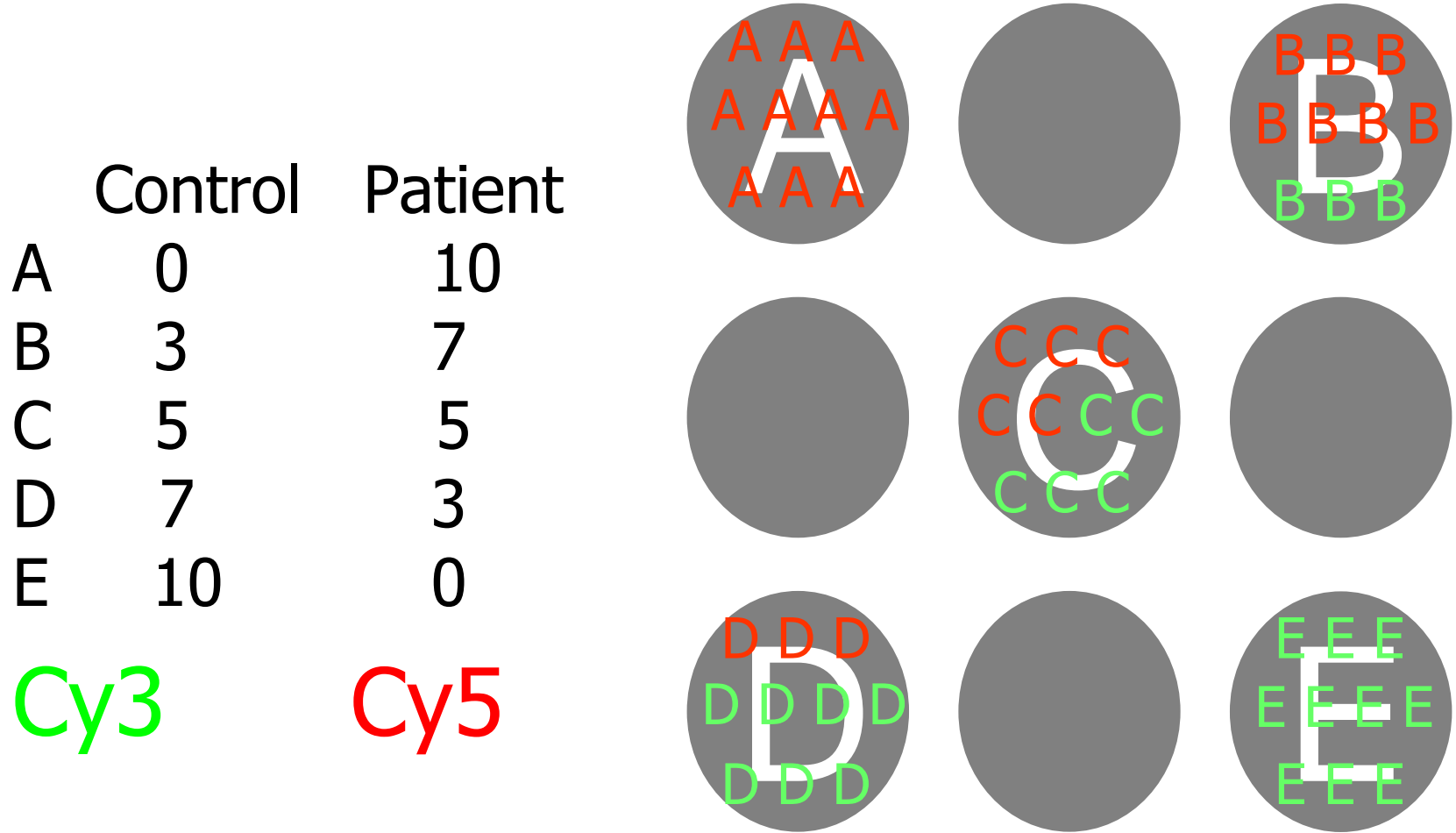
Genes	Control	Patient
gene A	0	10
gene B	3	7
gene C	5	5
gene D	7	3
gene E	10	0

mRNA isolation  
cDNA synthesis  
fluorescent labeling  
mix both targets  
spread over microarray

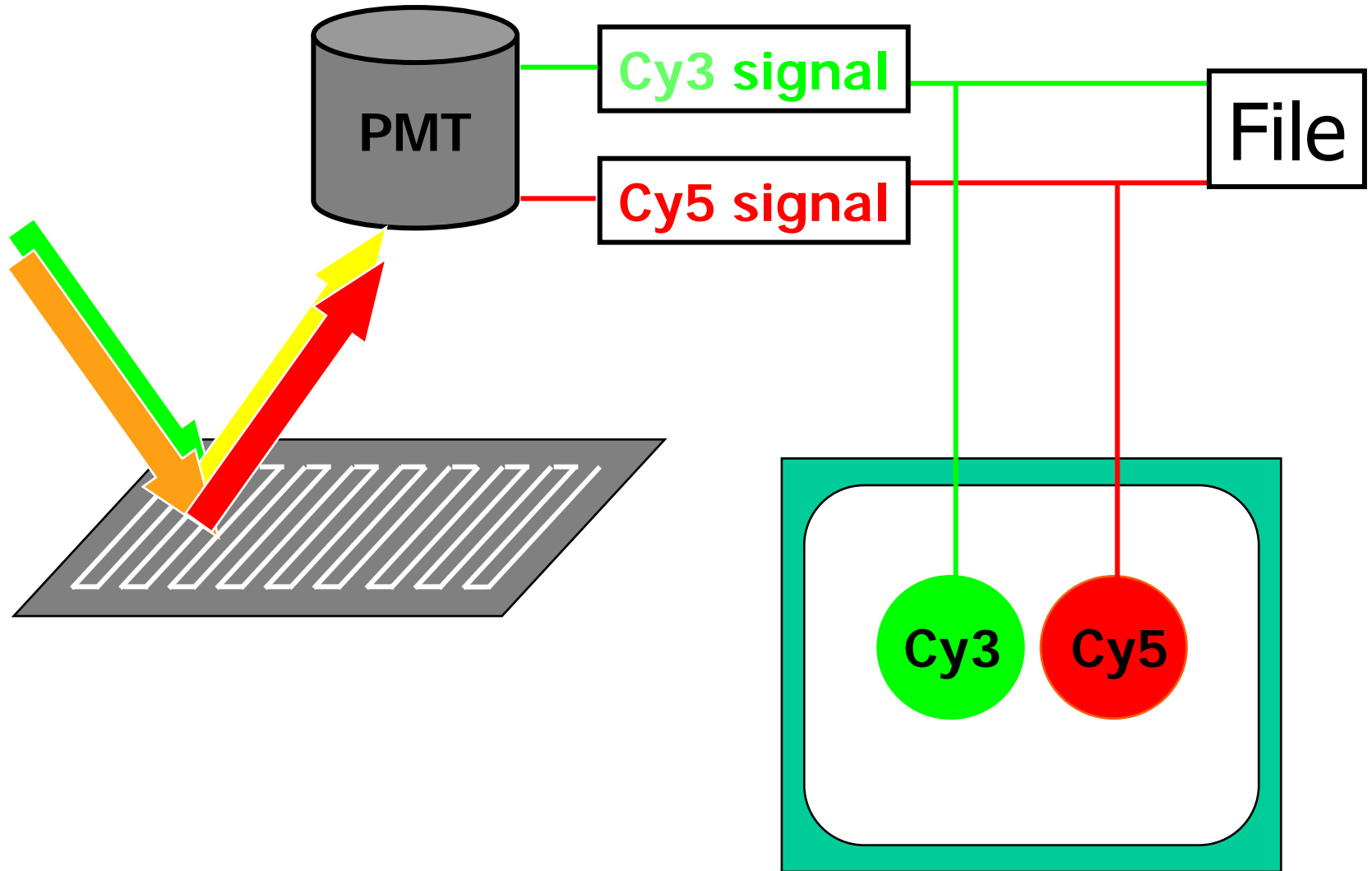


# Array Hybridization

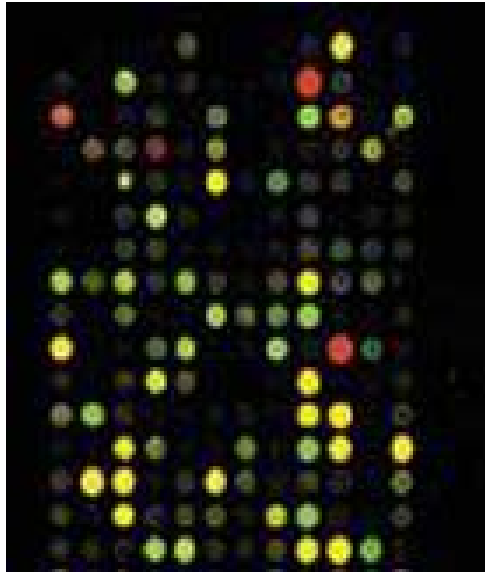
Base pairing between probe & target



# Array Scanning



# Array Measurement

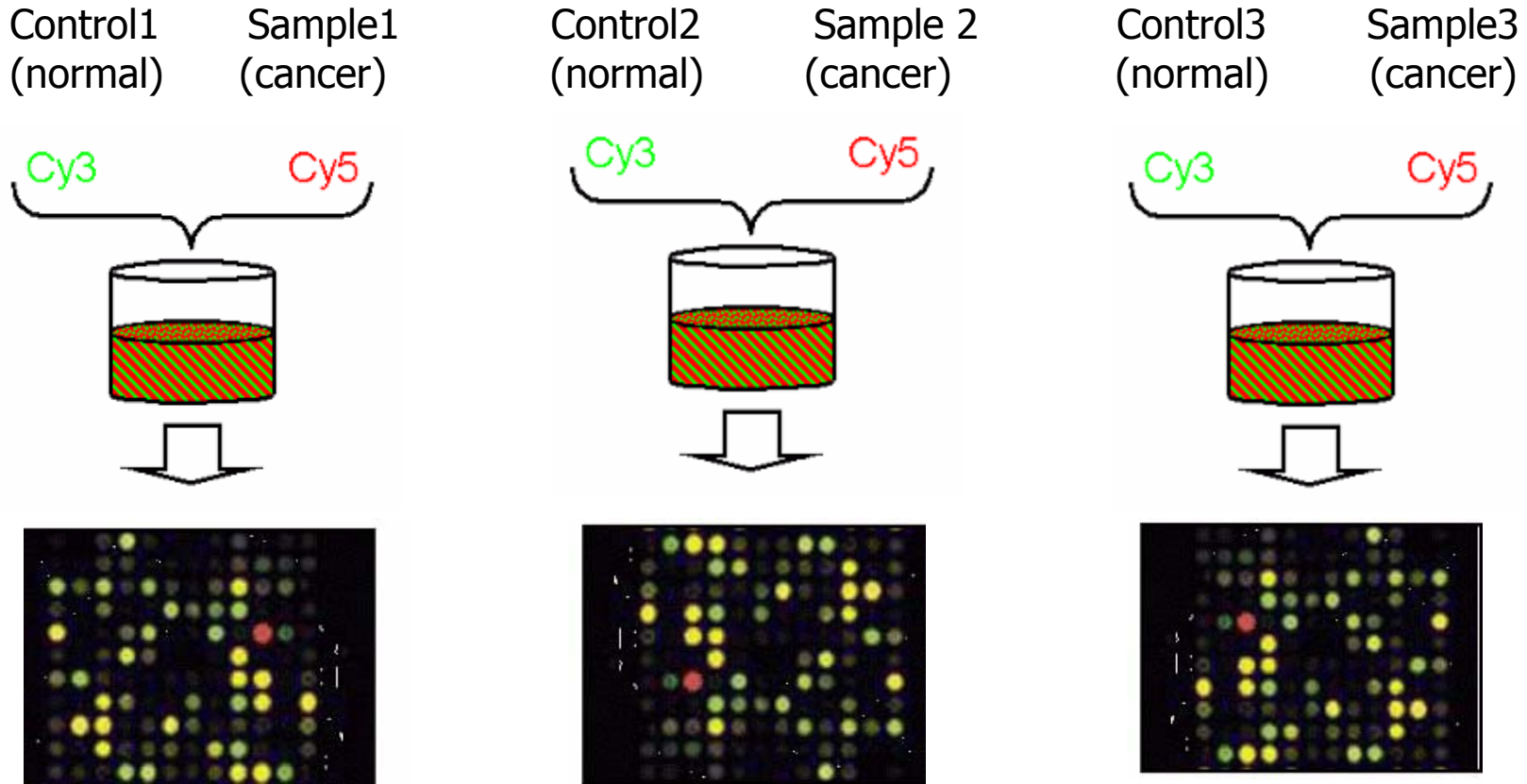


row	col	Cy3	Cy5	description
1	1	3	1416	Unsequenced
1	2	22	1315	Unsequenced
1	3	4	7205	Unsequenced
1	4	3	7780	Unsequenced
1	5	2	14214	Unsequenced
1	6	2	15237	Unsequenced
1	7	25966	2144	Unsequenced
1	8	16	4461	Unsequenced
1	9	11	6909	Unsequenced
1	10	30	14007	M mus. helix loop helix protein
2	1	16	13363	M mus. helix loop helix protein
2	2	11	8783	Unsequenced
2	3	21	8657	Unsequenced
2	4	4	48066	R nor TM4 fibroblast tropomyosin 4
2	5	5	50831	R nor TM4 fibroblast tropomyosin 4
2	6	1344	2342	Unsequenced
2	7	33	5473	Unsequenced
2	8	15	3092	Unsequenced
2	9	1358	791	Unsequenced
2	10	135	2450	Unsequenced
3	1	25	4726	R nov. Ribosomal protein S11
3	2	4	4469	R nov. Ribosomal protein S11
3	3	177	2251	Unsequenced
3	4	2544	4503	M mus. follistatin-like (Fst1)
3	5	3	2057	R nor zinc finger protein
3	6	2	1918	R nor zinc finger protein
3	7	6535	5206	Unsequenced
3	8	12	133	Ins/EV
3	9	5	560	Ins/EV
3	10	112	6664	H sap.mRNA for protein phosphatase
4	1	45	3384	Rat ASM15 gene
4	2	34	3141	Rat ASM15 gene
4	3	6	2276	R nor. mRNA (pJG116)
4	4	27	2921	R nor. mRNA (pJG116)
4	5	1500	3826	Small piece no EST
4	6	18	6463	R nor. mRNA for RNA polymerase II
4	7	29	9873	R nor. mRNA for RNA polymerase II

## image analysis

- grid finding / spot finding
- spot fitting / segmentation
- spot measurement

# Microarray Experiment



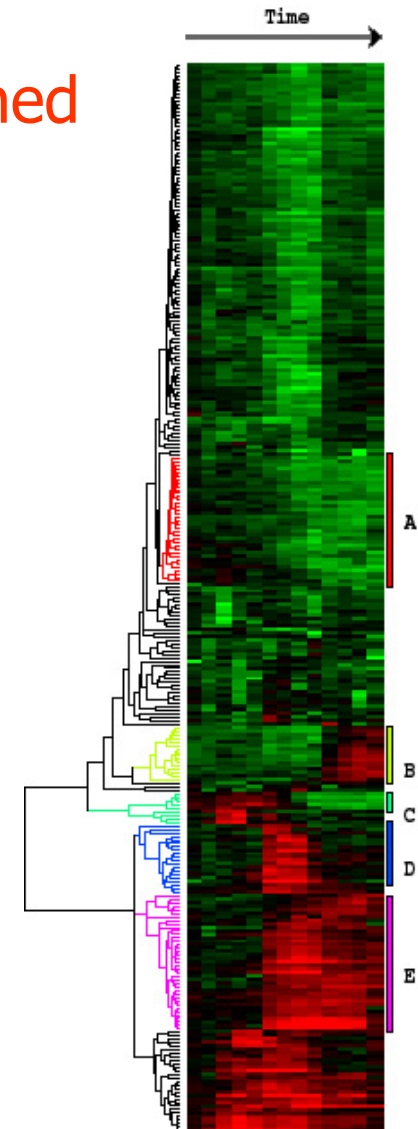
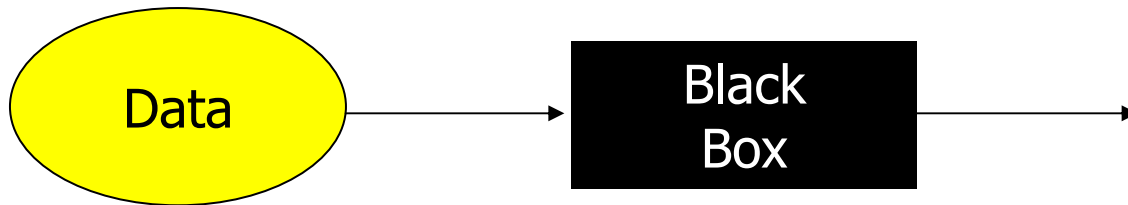
- Differential expression: which genes are different in **normal** tissue versus **cancer**?
- Clustering: which genes have similar profiles (co-expression)?
- Classification: can we predict whether an unknown sample is **normal** or **cancer**

# Outline: Unsupervised Learning

Unsupervised: classes are not pre-defined

Goal: grouping and visualization

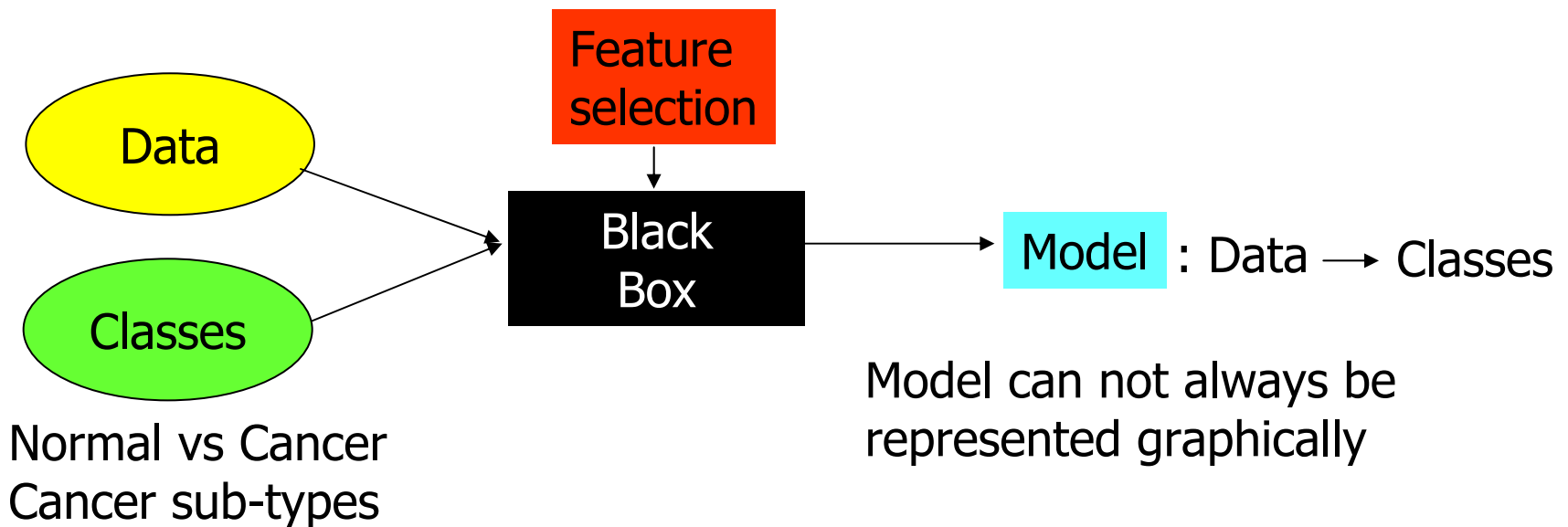
- Hierarchical clustering
- K-means
- Principal component analysis



# Outline: Supervised Learning

Supervised: classes are defined

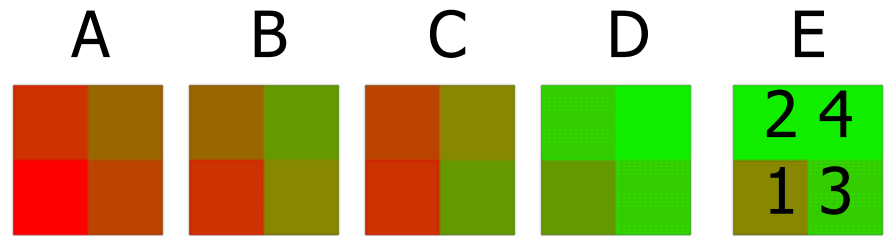
Goal: building models for future classifications



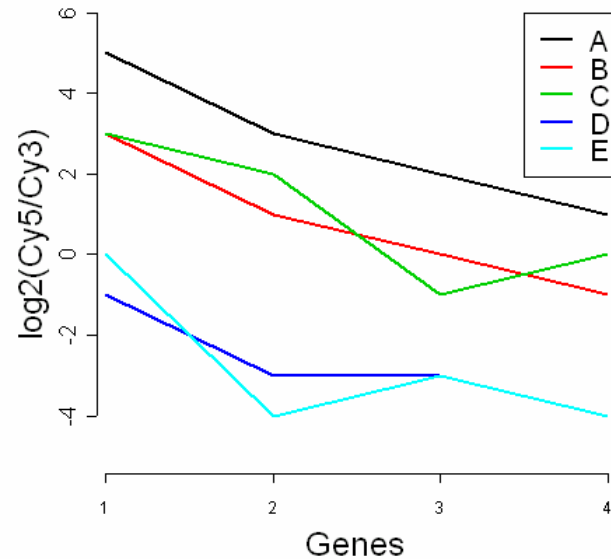
# Visualisation of Array Data

array \ gene	A	B	C	D	E
gene1	5	3	3	-1	0
gene2	3	1	2	-3	-4
gene3	2	0	-1	-3	-3
gene4	1	-1	0	-4	-4

$\log(\text{Cy5}/\text{Cy3})$



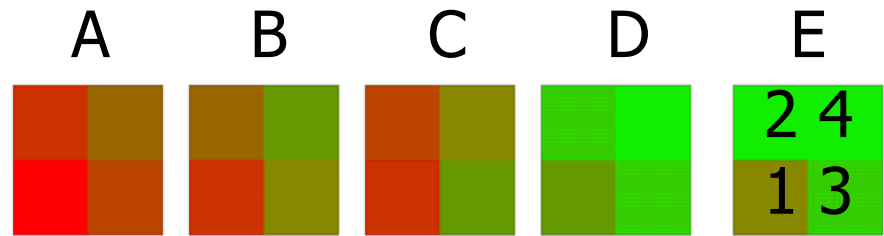
Profiles (arrays):



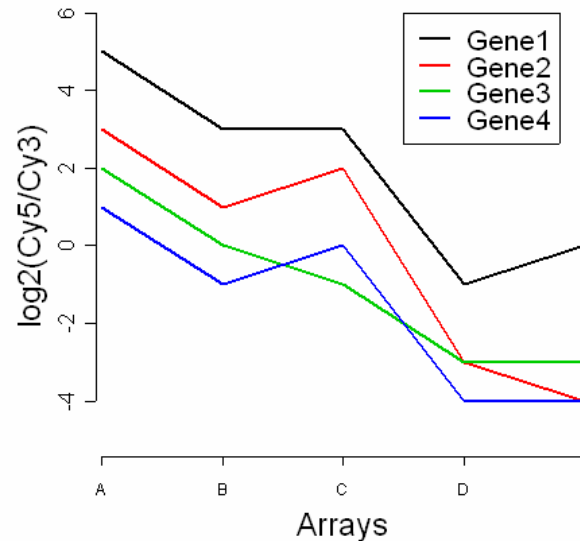
# Visualisation of Array Data

gene \ array	A	B	C	D	E
gene1	5	3	3	-1	0
gene2	3	1	2	-3	-4
gene3	2	0	-1	-3	-3
gene4	1	-1	0	-4	-4

$\log(\text{Cy5}/\text{Cy3})$



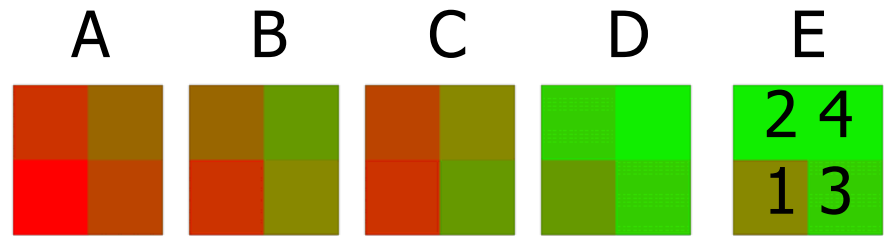
Profiles (genes):



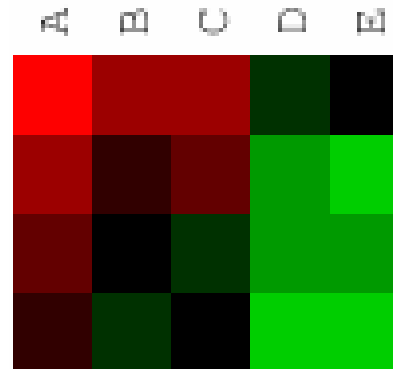
# Visualisation of Array Data

gene \ array	A	B	C	D	E
gene1	5	3	3	-1	0
gene2	3	1	2	-3	-4
gene3	2	0	-1	-3	-3
gene4	1	-1	0	-4	-4

$\log(\text{Cy5}/\text{Cy3})$



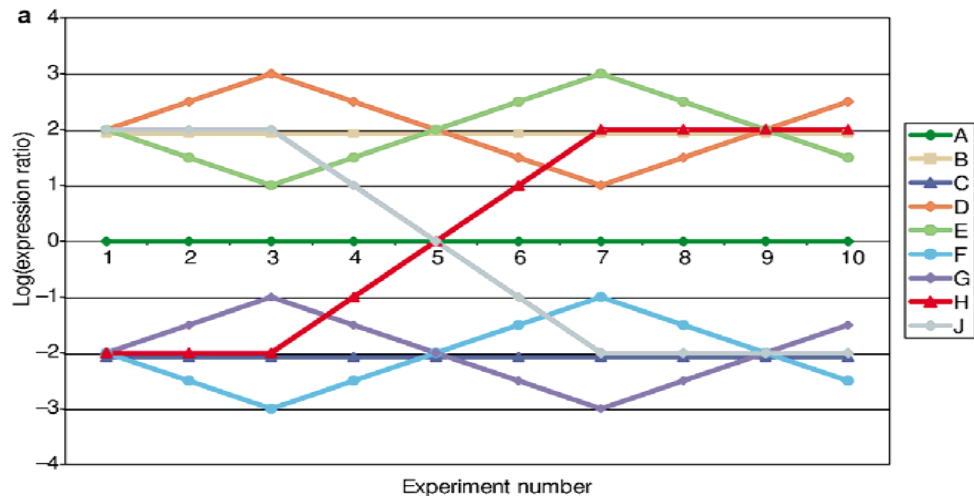
Matrix representation:



gene1  
gene2  
gene3  
gene4

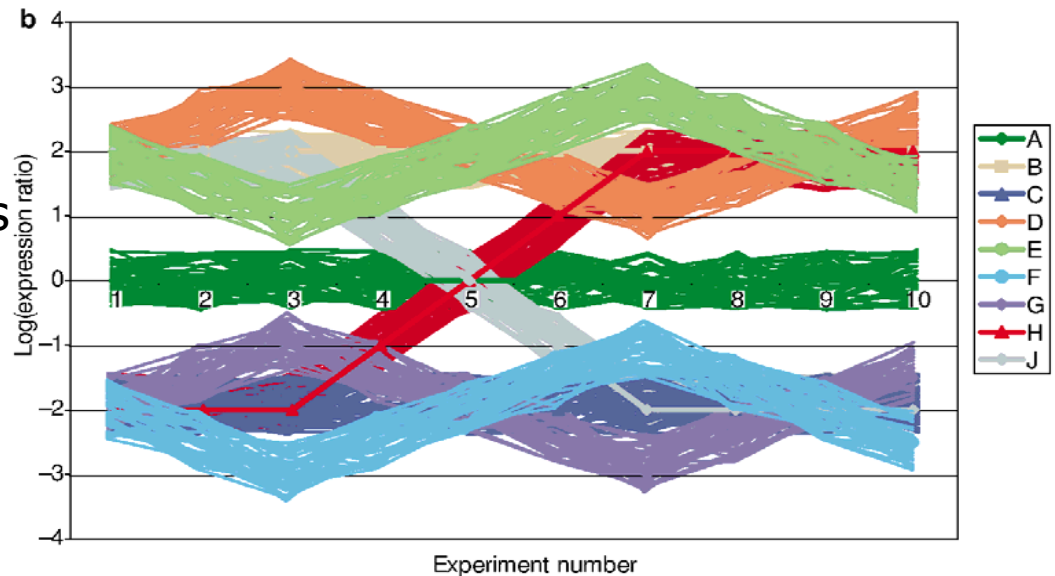
# Visualisation of Array Data

- 10 arrays (x-axis)
- 9 different profiles (A-J)
- log-ratio (y-axis)



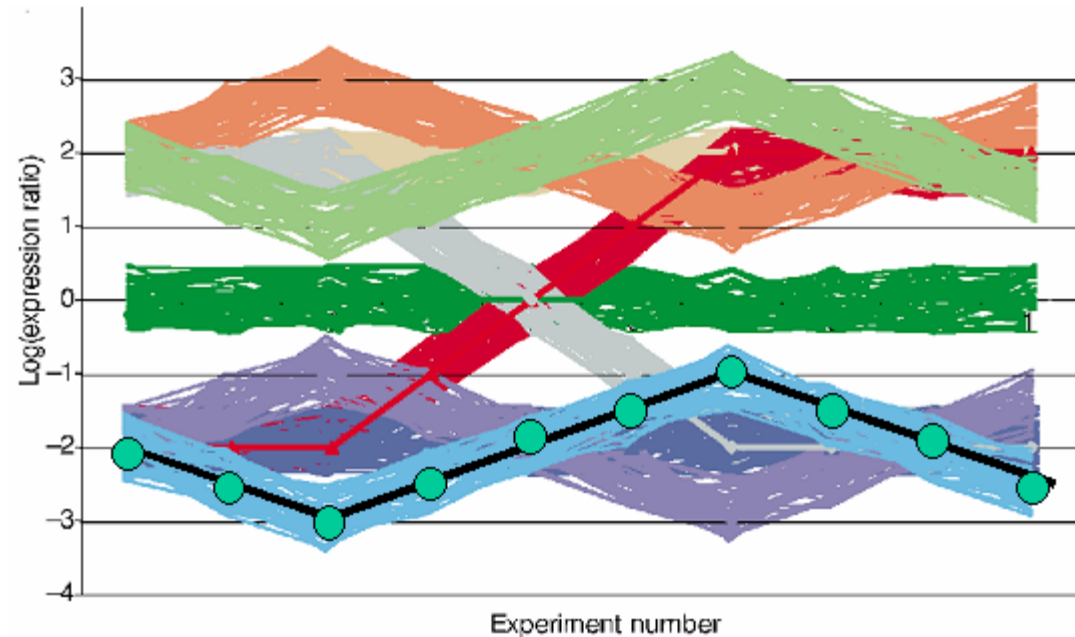
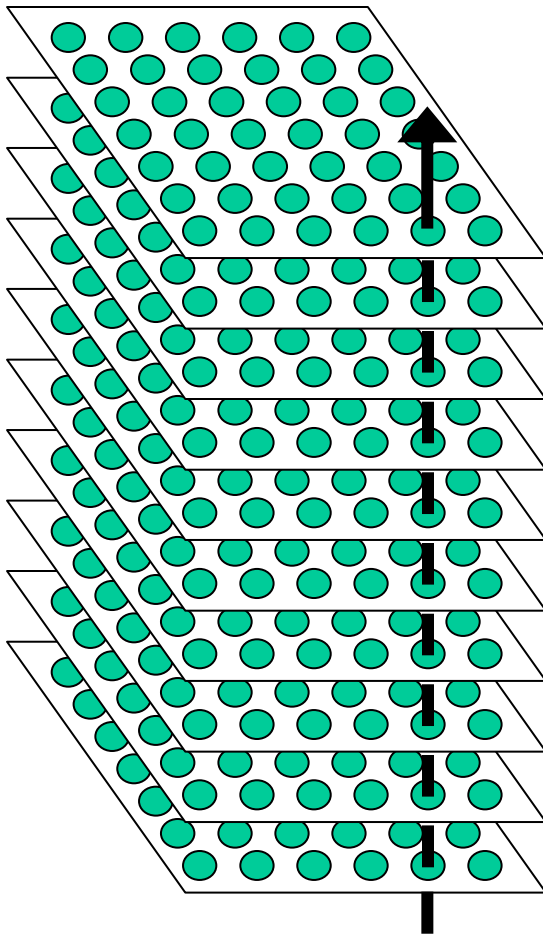
- round each profile 50  
"noisy profiles"

Each simulated array contains  
 $9 \times (1 + 50) = 459$  spots



Quackenbush, Nature Reviews  
Genetics, (2), 418-427 (2001)

# Gene profiles

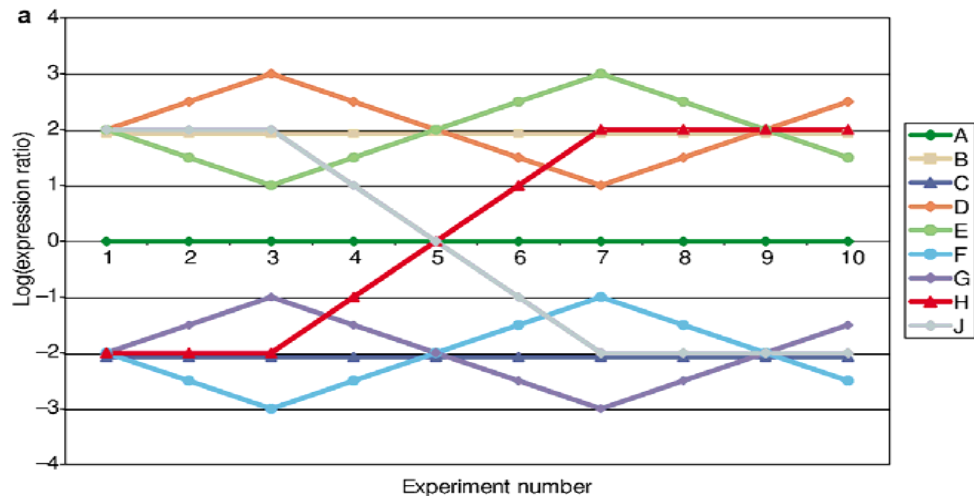


Each experiment represents e.g. a time-point

gene = (h1,h2,h3,h4,h5,h6,h7,h8,h9,h10)  
10 dimensional vector (cannot be visualized)

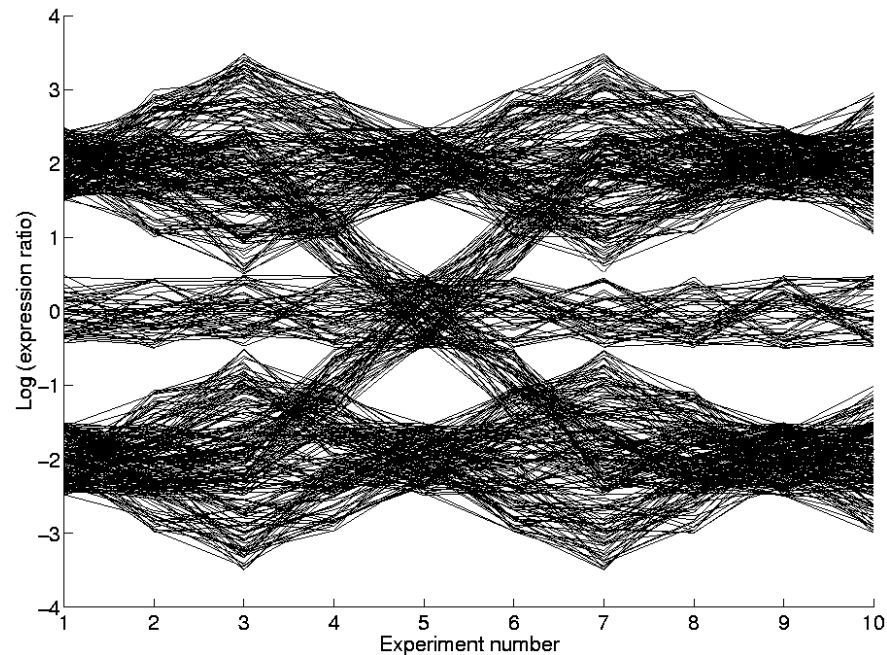
# Visualisation of Array Data

- 10 arrays (x-axis)
- 9 different profiles (A-J)
- log-ratio (y-axis)



- round each profile 50  
"noisy profiles"

Each simulated array contains  
 $9 \times (1 + 50) = 459$  spots

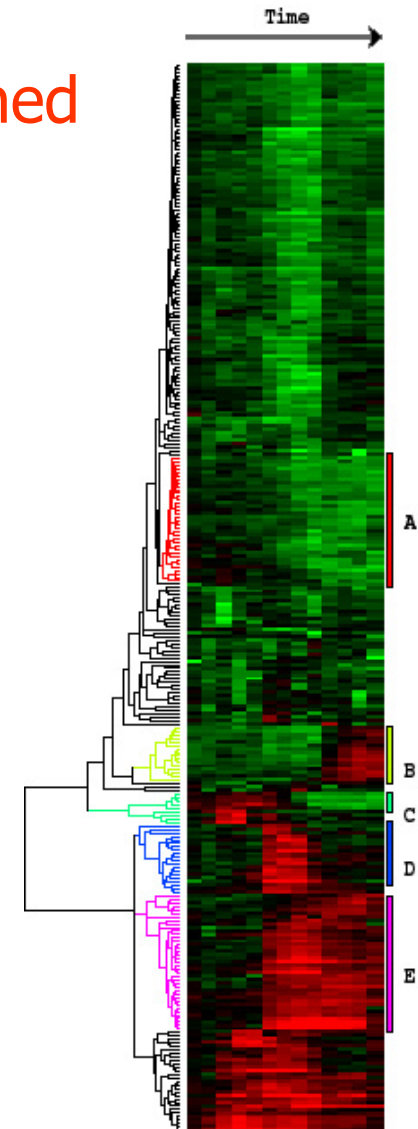
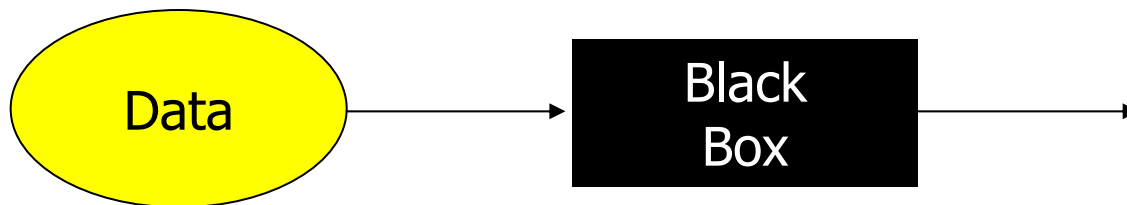


# Outline: Unsupervised Learning

$\log(\text{Cy5}/\text{Cy3})$   $\begin{cases} \rightarrow < 0: \text{green} \\ \rightarrow \geq 0: \text{red} \end{cases}$

Unsupervised: classes are not pre-defined  
Goal: grouping and visualization

- Hierarchical clustering
- K-means
- Principal component analysis

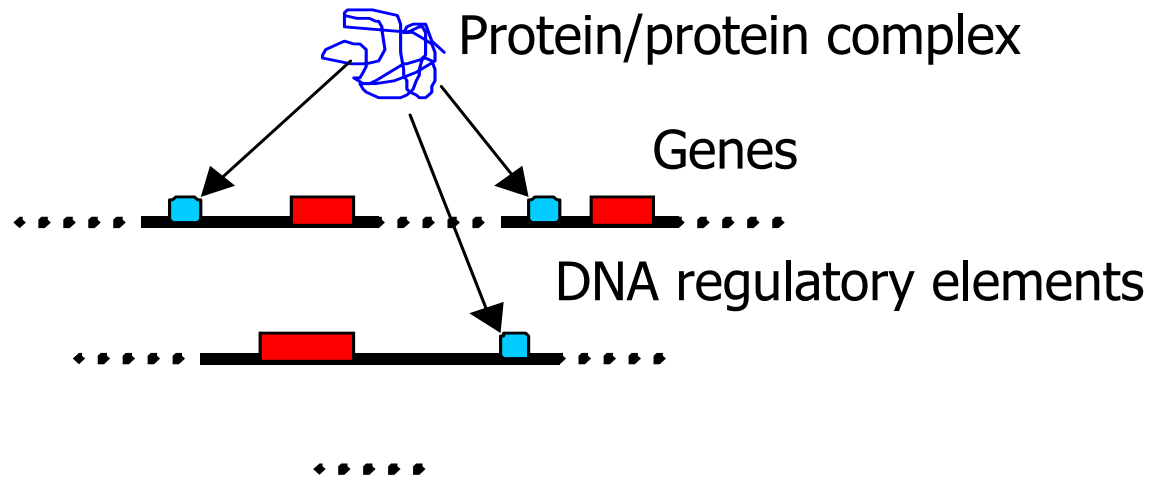


# Unsupervised Learning: Clustering

Clustering: grouping together similar objects.

Microarray data

**Genes:** similar  $\sim$  co-expression  $\sim$  co-regulation  $\sim$   
same pathway / same function



**Samples:** similar  $\sim$  same type of tissue

Used for discovery of new subclasses in a form of cancer

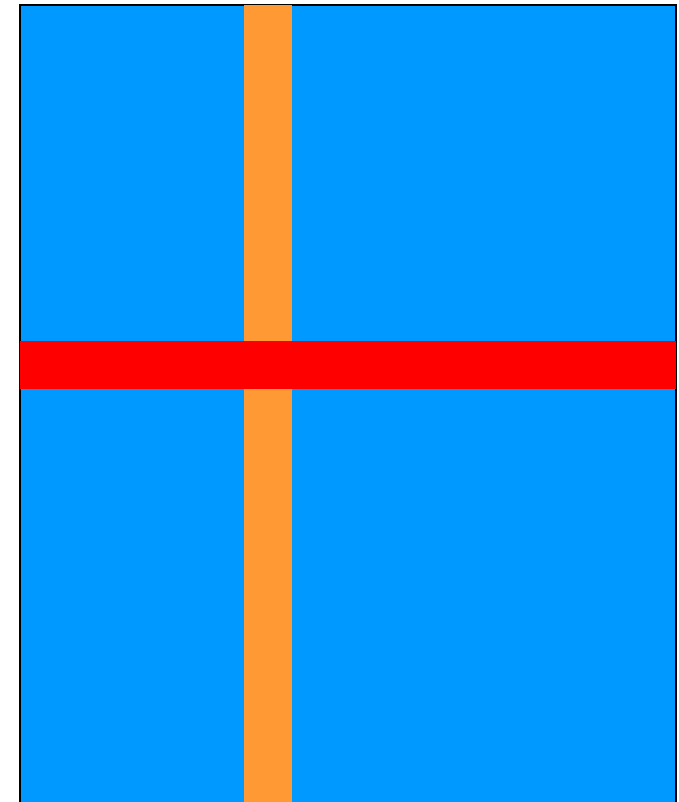
# Unsupervised Learning: Clustering

array \ gene	A	B	C	D	E
gene1	5	3	3	-1	0
gene2	3	1	2	-3	-4
gene3	2	0	-1	-3	-3
gene4	1	-1	0	-4	-4

gene expression data matrix

$n$  experiments

$p$  genes

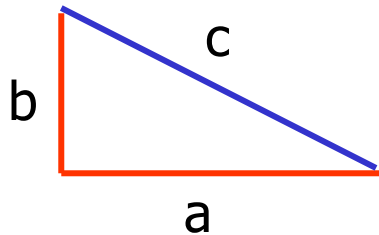


genes: vectors in  $\mathbf{R}^n$

arrays: vectors in  $\mathbf{R}^p$

Often  $p \gg n$

# Distance Measures: Euclidean (2D)

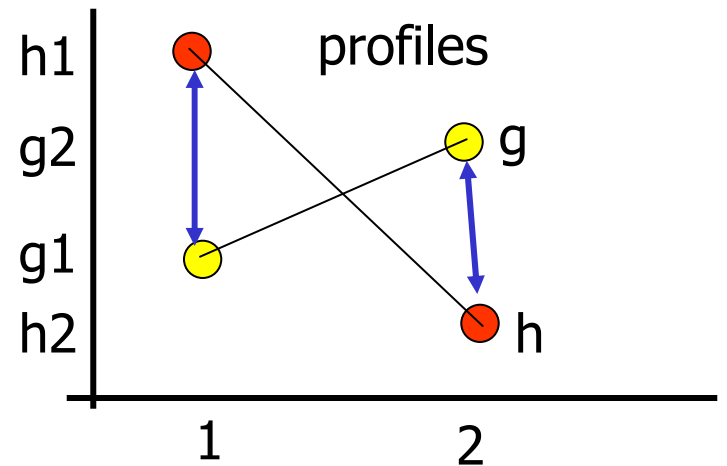
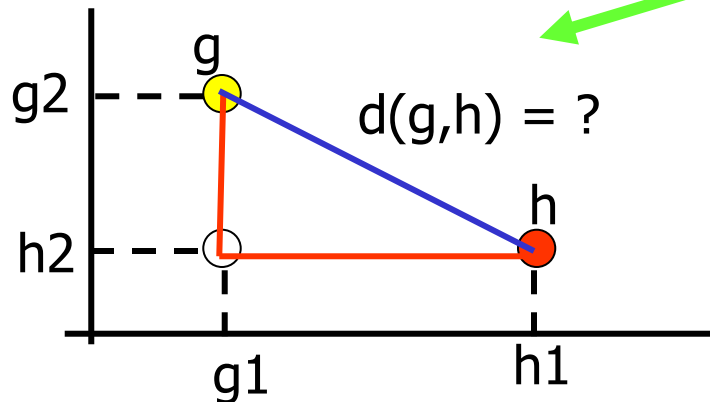


Pythagoras

$$c^2 = a^2 + b^2$$

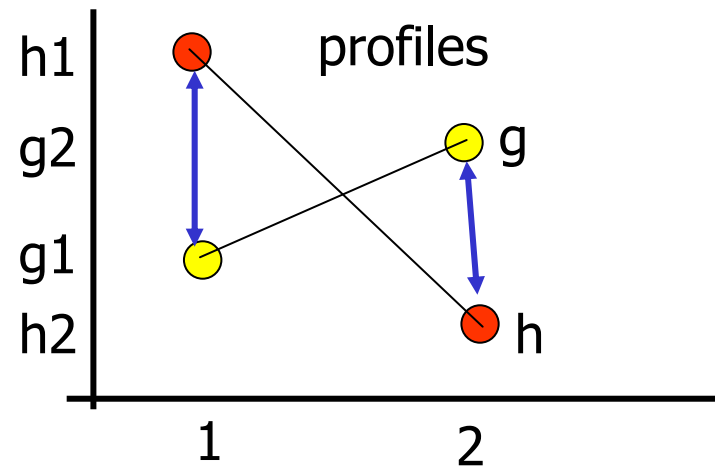
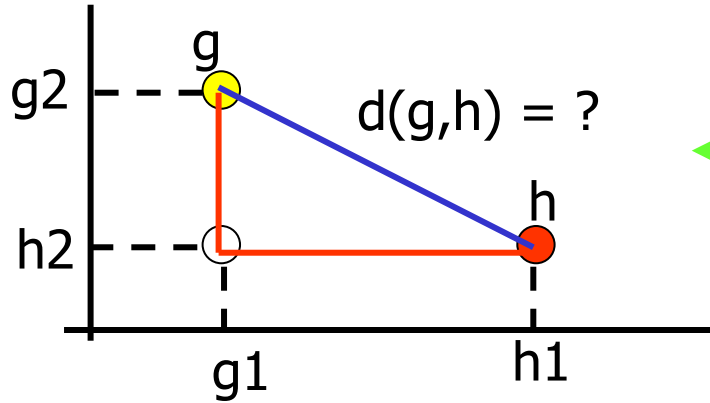
gene  $g = (g_1, g_2)$

gene  $h = (h_1, h_2)$



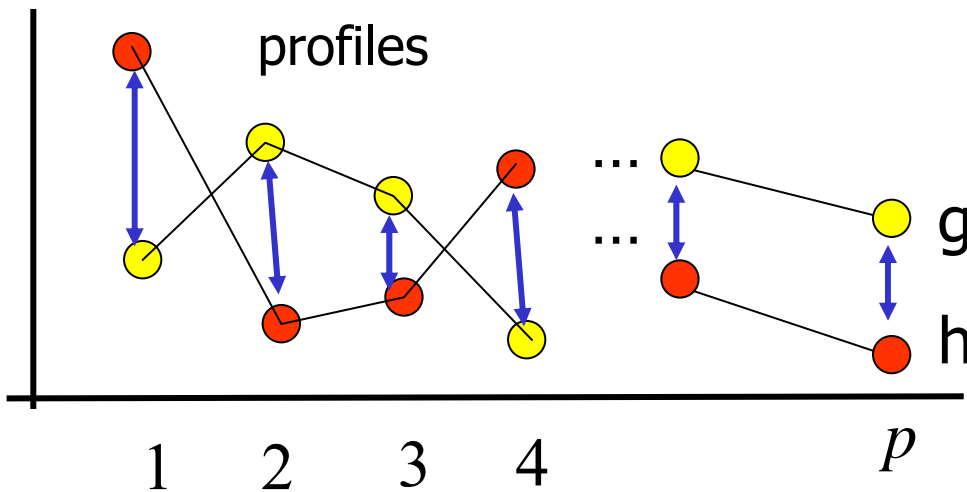
$$d(g, h) = \sqrt{(g_1 - h_1)^2 + (g_2 - h_2)^2}$$

# Euclidean (general D)



$$d(g, h) = \sqrt{(g_1 - h_1)^2 + (g_2 - h_2)^2}$$

Generalization to  $p$  dimensions:

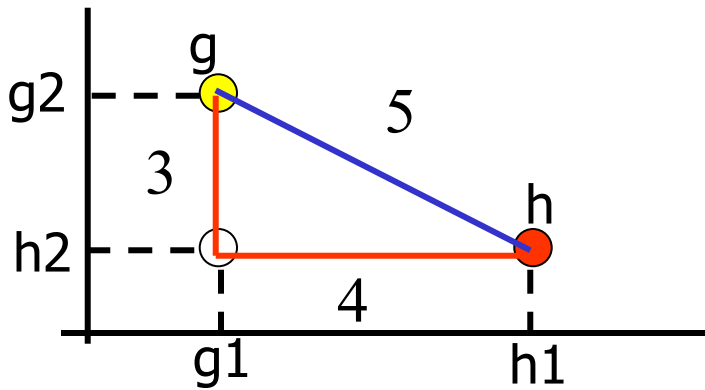


$$d(g, h) = \sqrt{\sum_{i=1}^p (g_i - h_i)^2}$$

# Distance Measures: Manhattan

Distance along path parallel to the axes

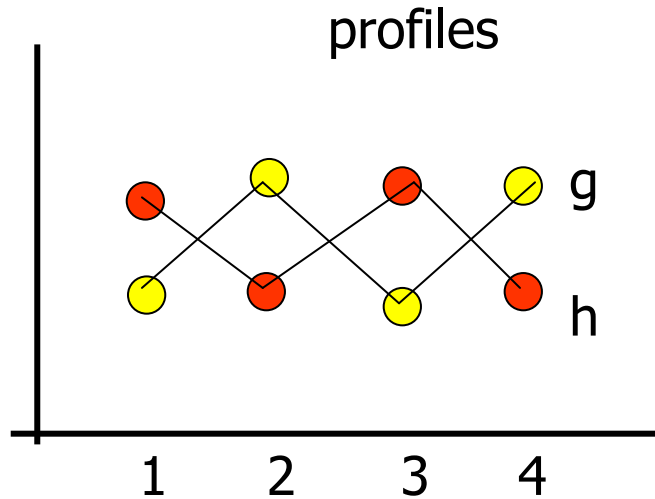
Manhattan: 
$$d_1 = \sum_{i=1}^p |x_i - y_i|$$



$$d_2 = \sqrt{4^2 + 3^2} = 5$$

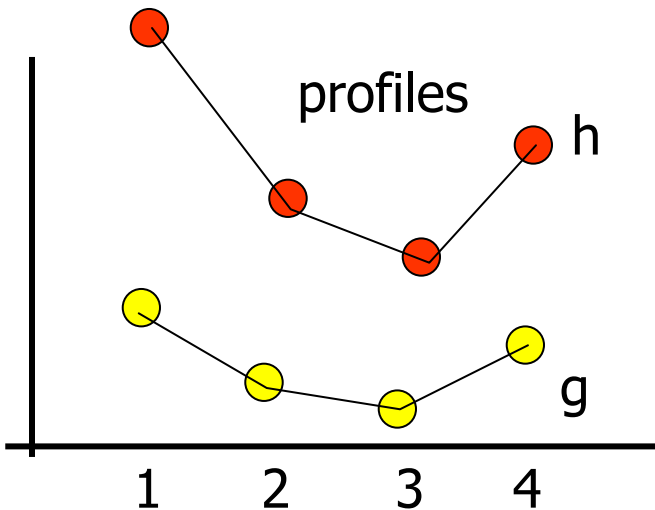
$$d_1 = 4 + 3 = 7$$

# Choice of distance measure



small Euclidean distance  
large correlation distance  
( $d=1-r$ )

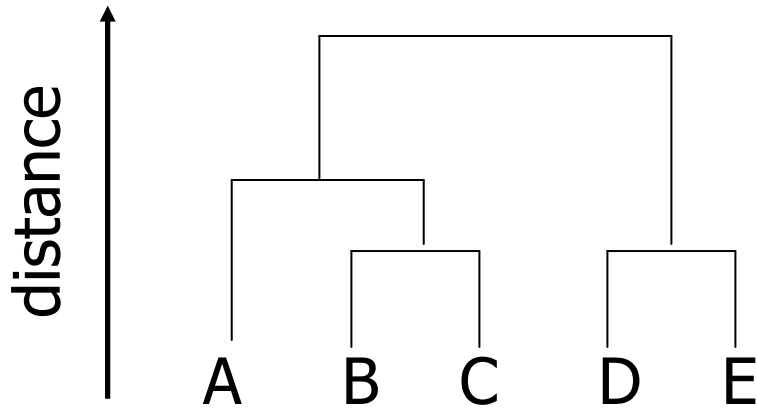
Correlation distance: same pattern



large Euclidian distance  
small correlation distance  
( $d=1-r$ )

# Hierarchical Clustering

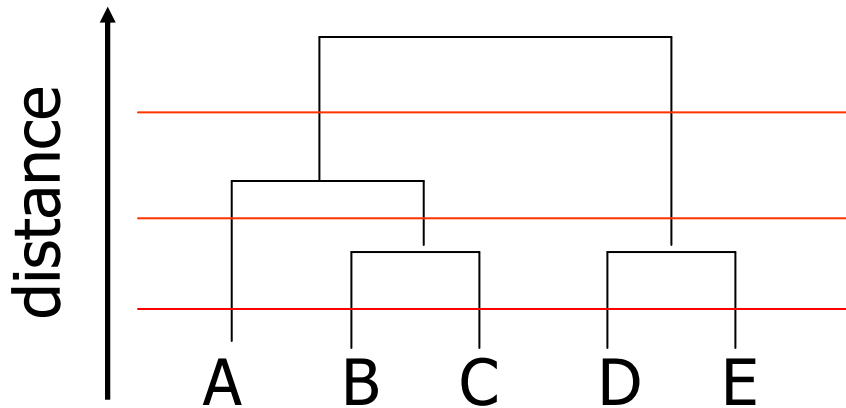
Group elements in a tree-like structure



The more similar objects are, the shorter the path

# Hierarchical Clustering

Group elements in a tree-like structure:



2 clusters:  $\{A, B, C\}\{D, E\}$

3 clusters:  $\{A\}\{BC\}\{DE\}$

5 clusters:  $\{A\}\{B\}\{C\}\{D\}\{E\}$

The more similar objects are, the shorter the path



# Hierarchical Clustering

## Algorithm (agglomerative clustering)

- Start: all objects in a separate cluster
- Clustering: combine the 2 clusters with the shortest distance

## Distance between clusters is:

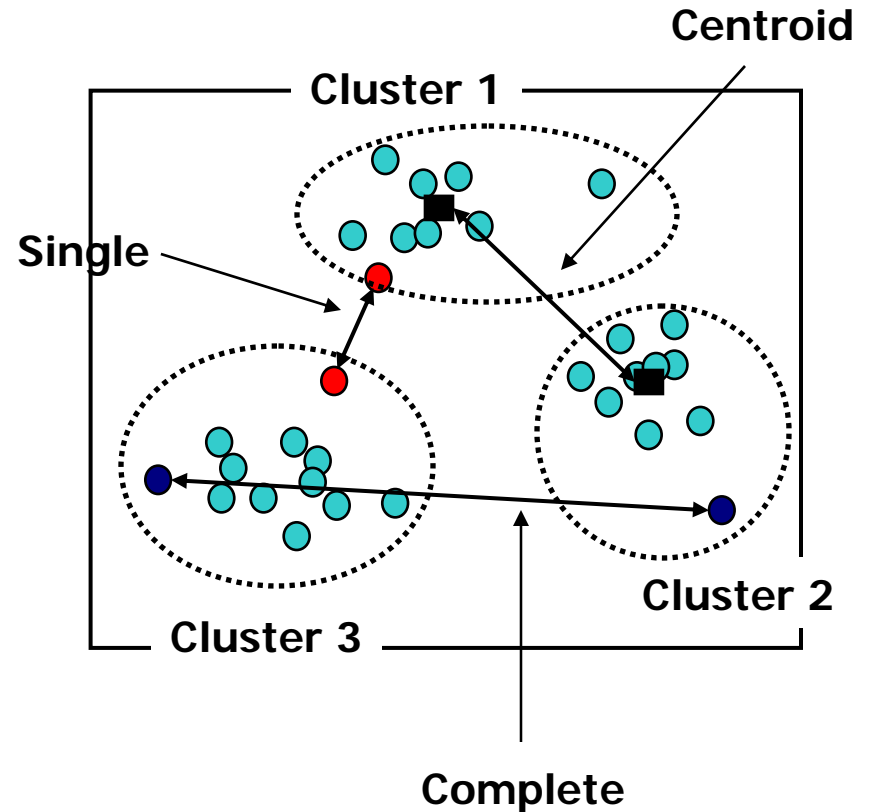
single, complete, centroid, average, median, Ward, ...

- Repeat till only 1 cluster is left

# Combining Clusters

Distance between two clusters is:

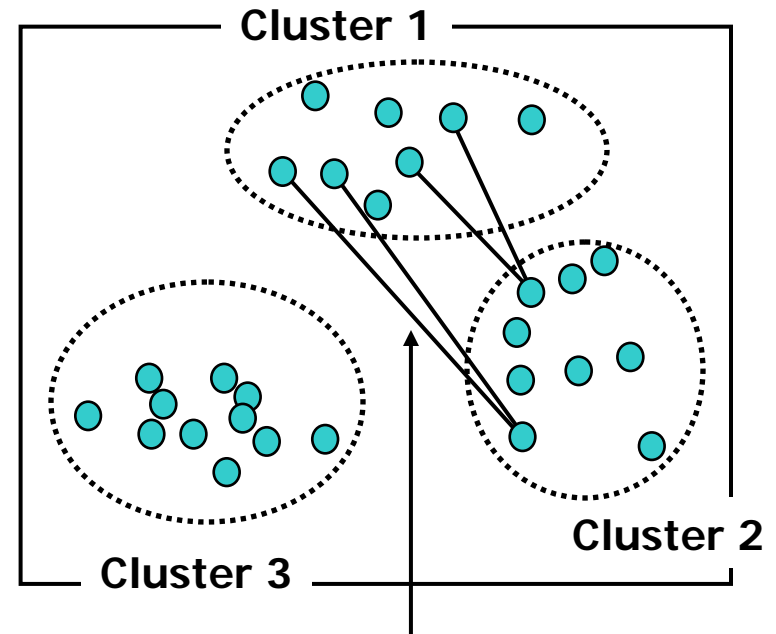
- **Single**  
distance between two closest cluster members
- **Complete**  
distance between two most distant cluster members
- **Centroid**  
distance between means of each cluster



# Combining Clusters

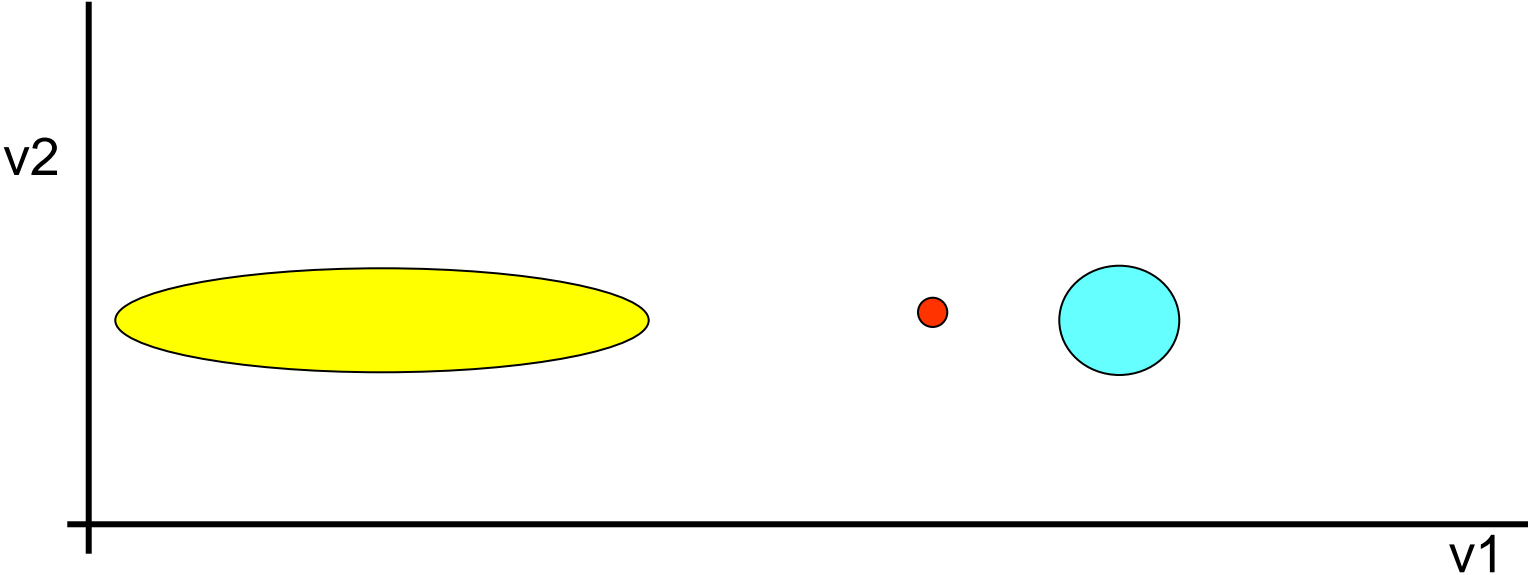
Distance between two clusters is:

- **Average**  
average distance between all members of the two clusters
- **Median**  
median distance between all members of the two clusters
- **Ward**  
average distance between all members of the two clusters with adjustment for covariance

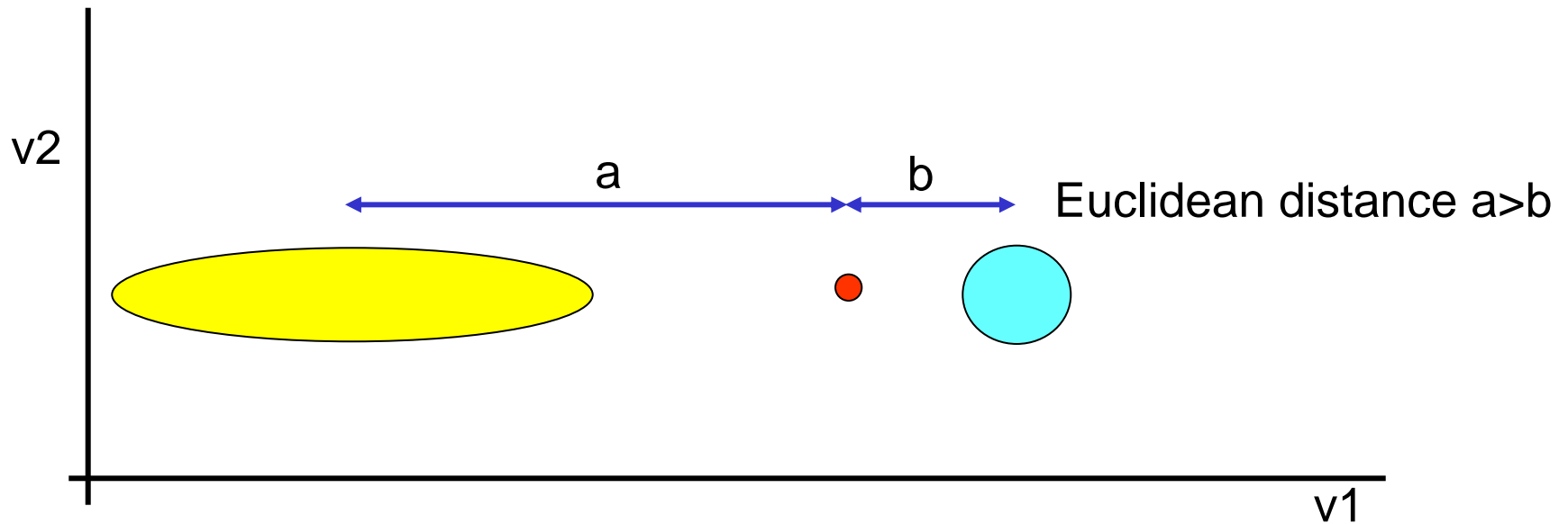


**Mean/median/adjusted mean  
of all pairwise distances**

# Mahalanobis distance



# Mahalanobis distance



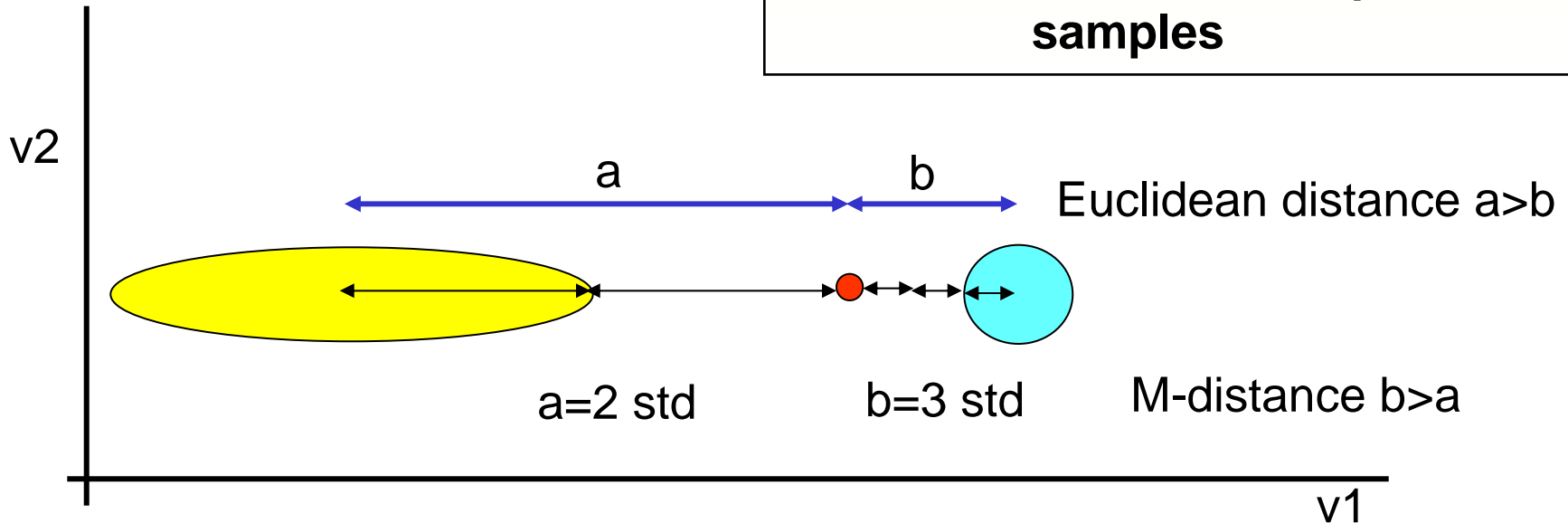
# Mahalanobis distance

$$D_{ij}^2 = \sum_{r=1}^p \sum_{s=1}^p (\mu_{ri} - \mu_{rj}) c_{rs}^{-1} (\mu_{si} - \mu_{sj})$$

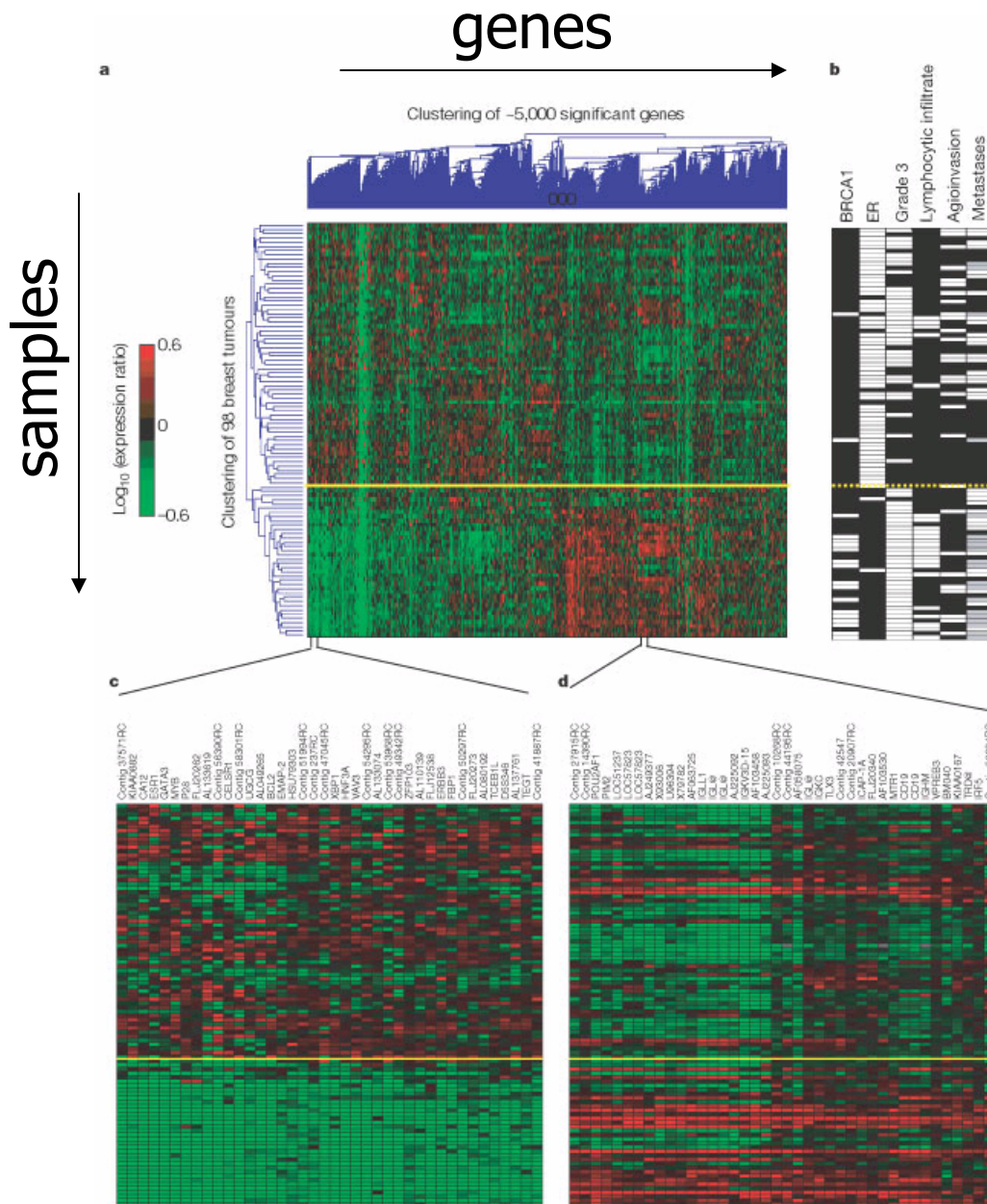
$$= (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \mathbf{C}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

$\mu_{ri}, \mu_{sj}$  Means of variable  $r$  and  $s$  in samples  $i$  and  $j$

$c_{rs}^{-1}$  Inverse of covariance between variables  $r$  and  $s$ , assumed equal for all samples



# Example



■ negative  
□ positive

histopathological data

# Hierarchical Clustering: Example

Distance matrix:

array \ gene	A	B	C	D	E		array	A	B	C	D
gene1	5	3	3	-1	0		B	8			
gene2	3	1	2	-3	-4	→	C				
gene3	2	0	-1	-3	-3		D				
gene4	1	-1	0	-4	-4		E				

Manhattan distance:

$$\begin{aligned} \text{dist}(A,B) &= |5-3| + |3-1| + |2-0| + |1+1| \\ &= 8 \end{aligned}$$

# Hierarchical Clustering: Example

Distance matrix:

array \ gene	A	B	C	D	E		array	A	B	C	D
gene1	5	3	3	-1	0		B	8			
gene2	3	1	2	-3	-4	→	C	7	3		
gene3	2	0	-1	-3	-3		D	22	14	15	
gene4	1	-1	0	-4	-4		E	22	14	15	2

Manhattan distance:

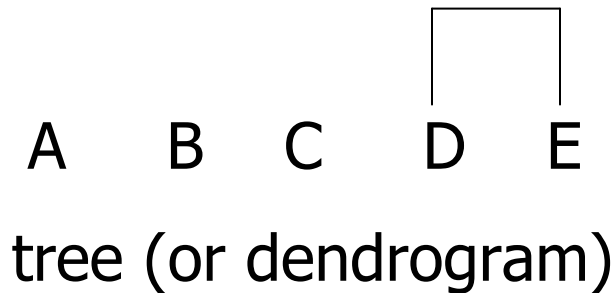
$$\begin{aligned} \text{dist}(A,B) &= |5-3| + |3-1| + |2-0| + |1+1| \\ &= 8 \end{aligned}$$

# Hierarchical Clustering: Example

Distance matrix:

array \ gene	A	B	C	D	E		array	A	B	C	D
gene1	5	3	3	-1	0		B	8			
gene2	3	1	2	-3	-4	→	C	7	3		
gene3	2	0	-1	-3	-3		D	22	14	15	
gene4	1	-1	0	-4	-4		E	22	14	15	2

↑  
minimum



# Hierarchical Clustering: Example

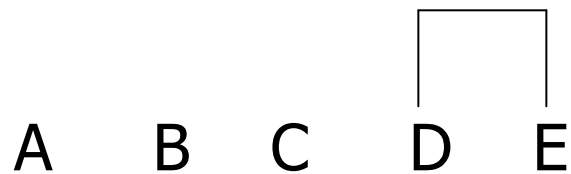
array	A	B	C	D
B	8			
C	7	3		
D	22	14	15	
E	22	14	15	2



dendrogram

# Hierarchical Clustering: Example

array	A	B	C	D		array	A	B	C
B	8				→	B			
C	7	3				C			
D	22	14	15			DE			
E	22	14	15	2					

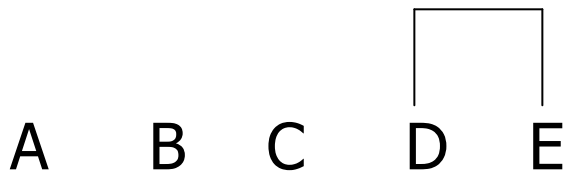


dendrogram

# Hierarchical Clustering: Example

array	A	B	C	D	array	A	B	C
B	8				B	8		
C	7	3			C	7	3	
D	22	14	15		DE	22		
E	22	14	15	2				

$$\text{dist}(A, DE) = \min(\text{dist}(A, D), \text{dist}(A, E)) = 22$$

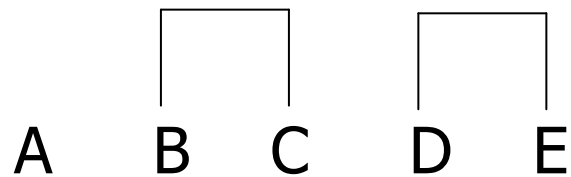


dendrogram

Single linkage: minimum distance

# Hierarchical Clustering: Example

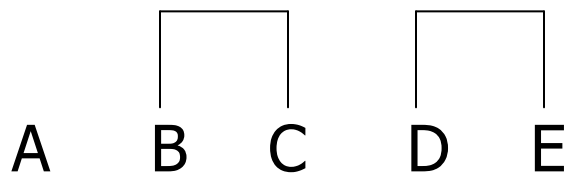
array	A	B	C	D	array	A	B	C
B	8				B	8		
C	7	3			C	7	3	
D	22	14	15		DE	22	14	15
E	22	14	15	2				



dendrogram

# Hierarchical Clustering: Example

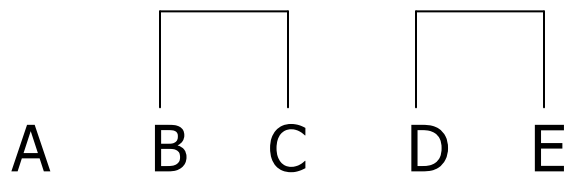
array	A	B	C
B	8		
C	7	3	
DE	22	14	15



dendrogram

# Hierarchical Clustering: Example

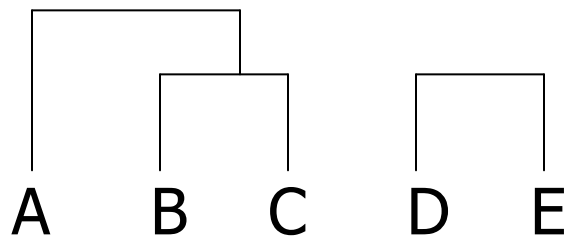
array	A	B	C		array	A	BC
B	8			→	BC	7	
C	7	3			DE	22	14
DE	22	14	15				



dendrogram

# Hierarchical Clustering: Example

array	A	B	C		array	A	BC
B	8			→	BC	7	
C	7	3			DE	22	14
DE	22	14	15				



dendrogram

# Hierarchical Clustering: Example

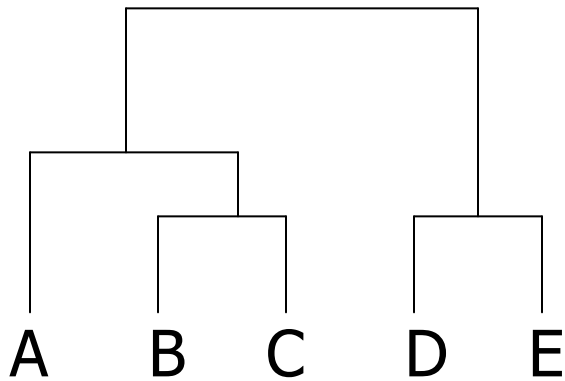
array	A	B	C
B	8		
C	7	3	
DE	22	14	15



array	A	BC
BC	7	
DE	22	14

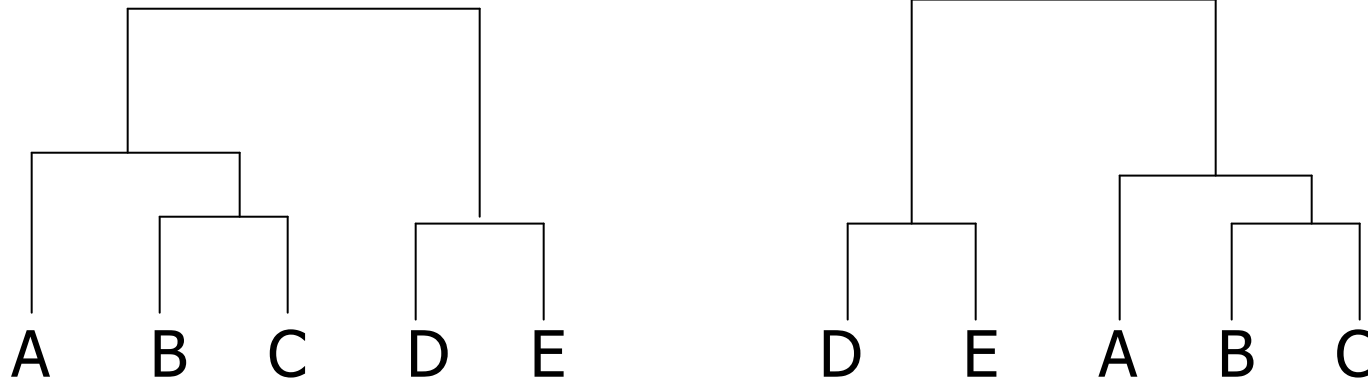


array	ABC
DE	14



dendrogram

# Tree Spotting



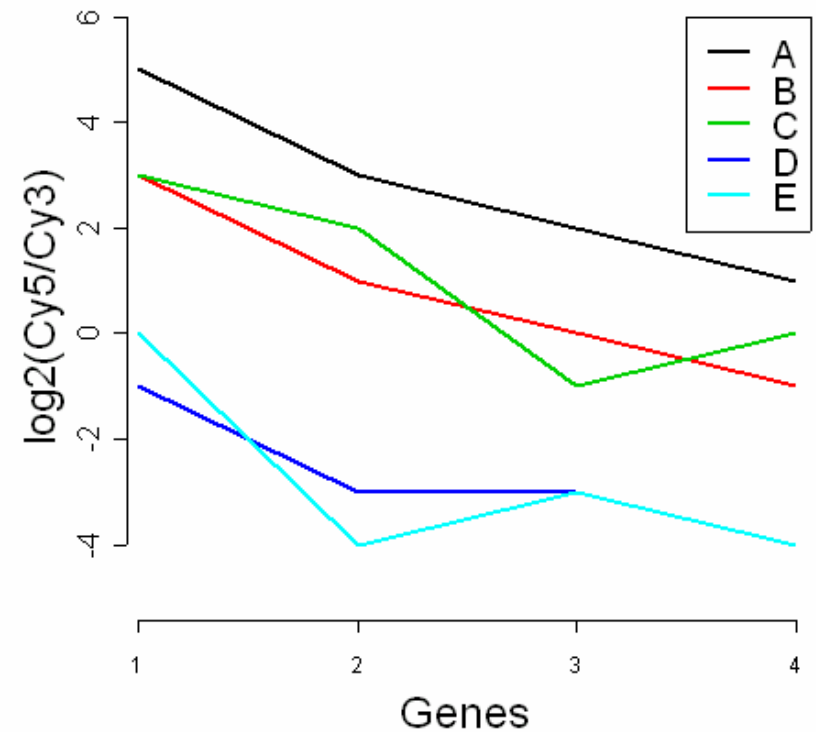
Both are valid representations of a tree found with single linkage & Manhattan distance

When clustering  $N$  elements, the tree contains  $N-1$  splits.  
Which subtree goes left/right does not matter →

$2^{(N-1)}$  equivalent trees

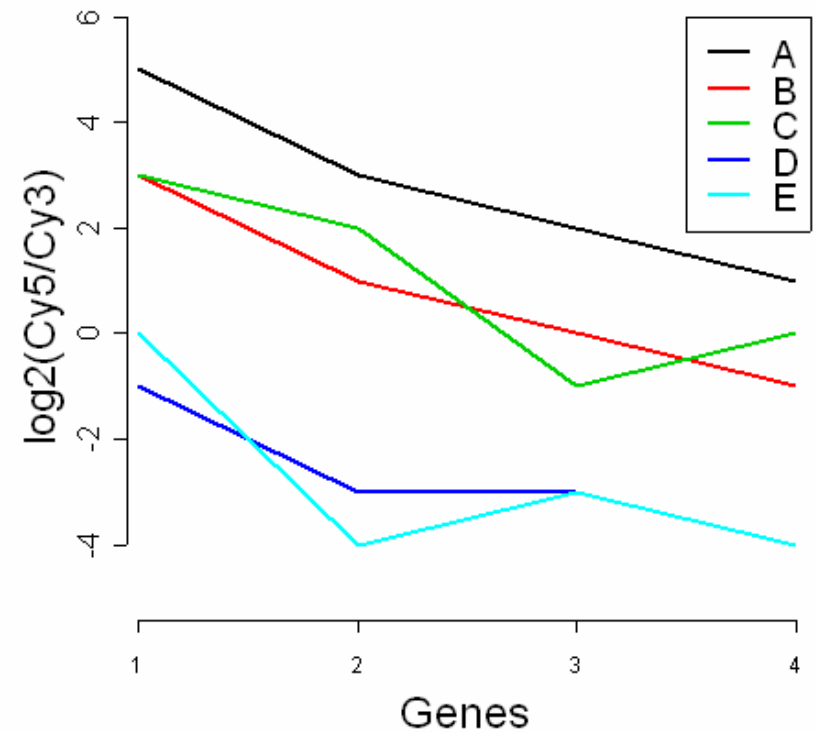
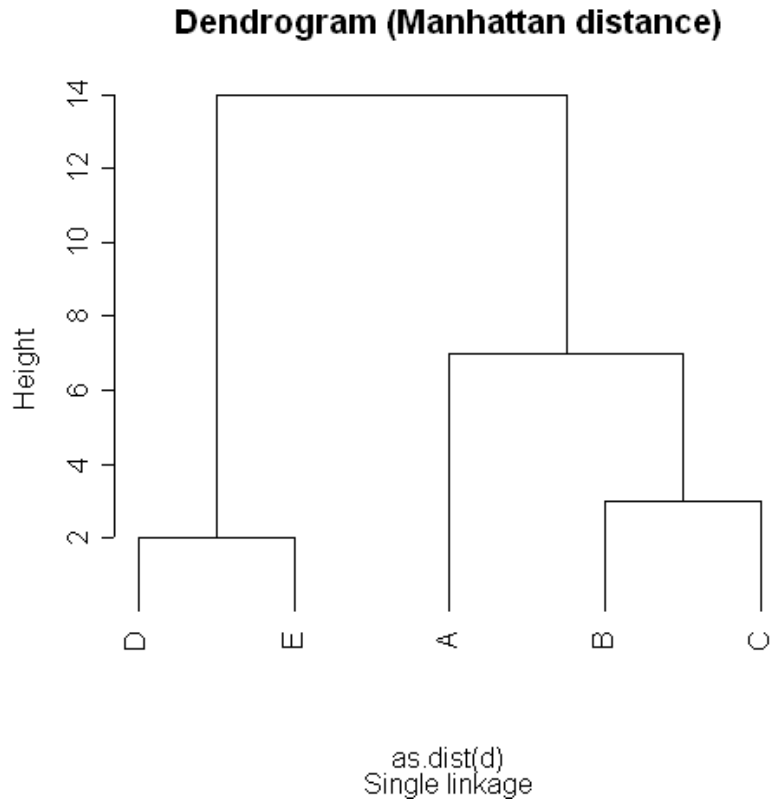
# Hierarchical Clustering: Example

Clustering found (Manhattan distance):



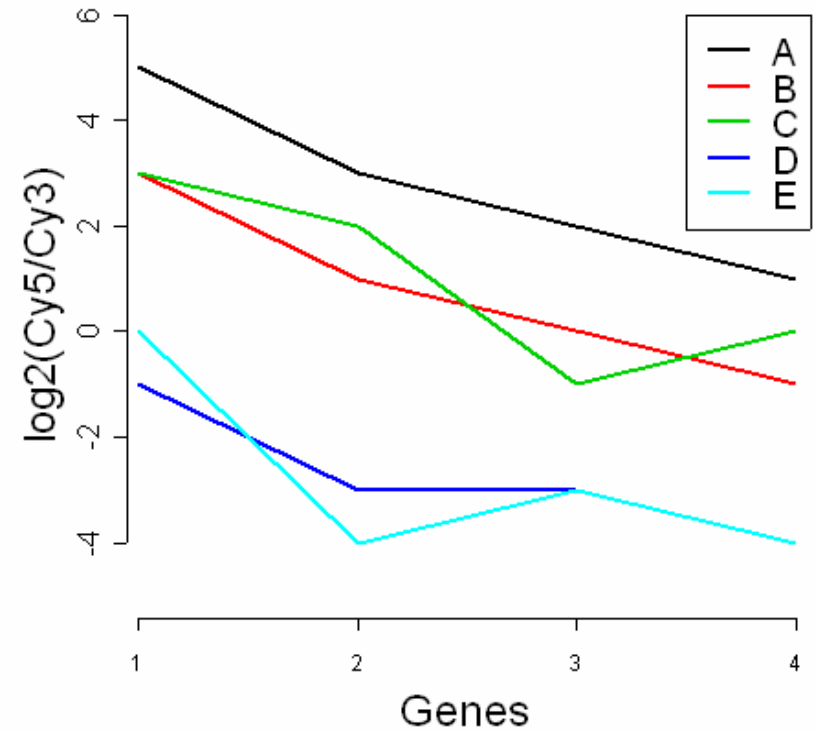
# Hierarchical Clustering: Example

Clustering found (Manhattan distance):



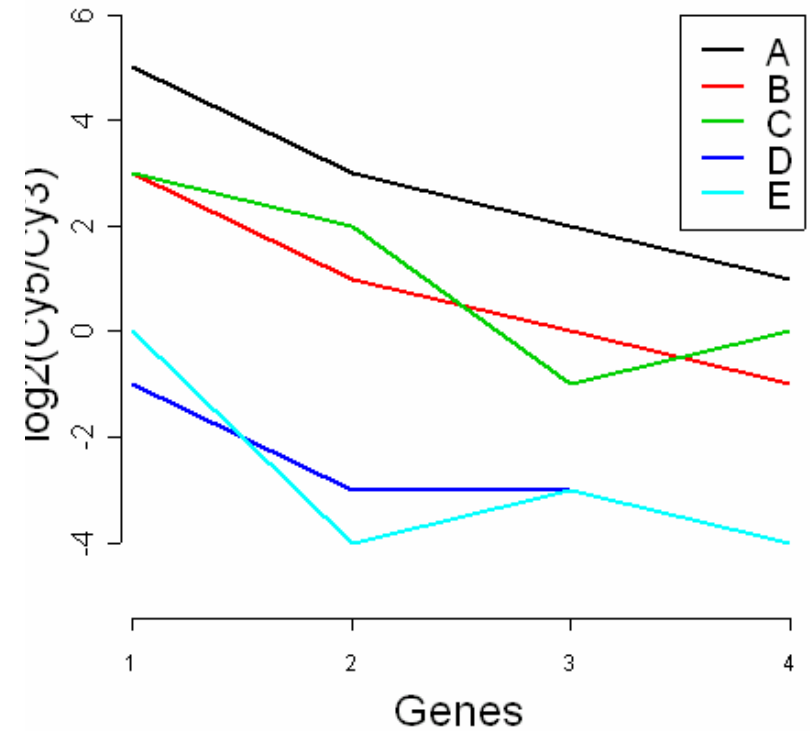
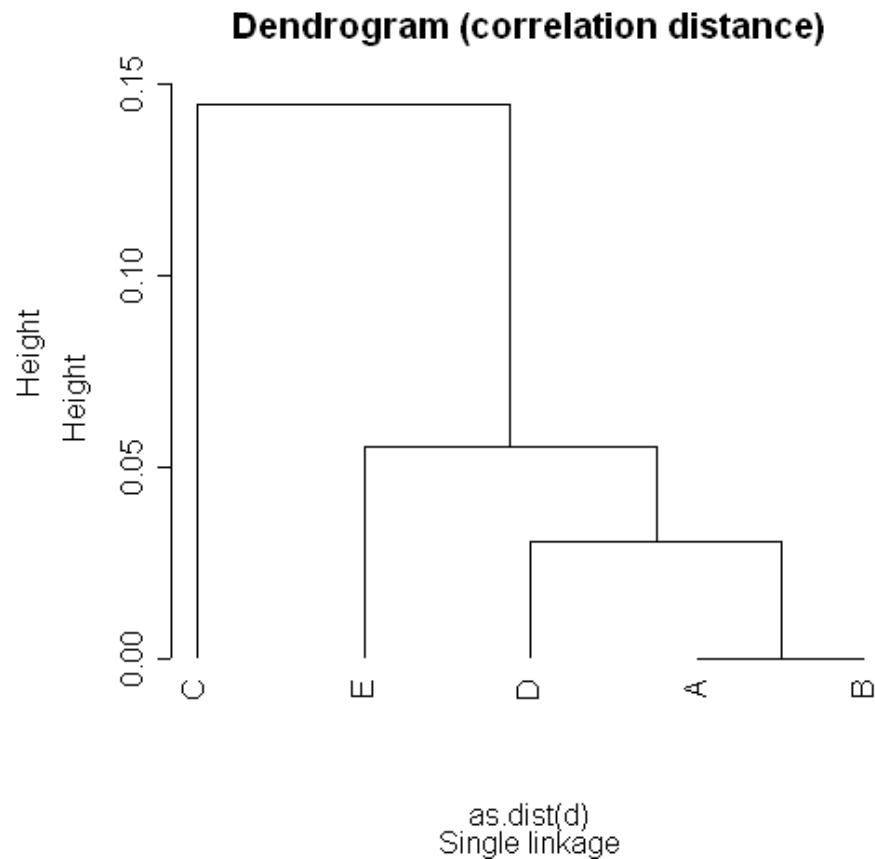
# Hierarchical Clustering: Example

Clustering found (correlation distance):



# Hierarchical Clustering: Example

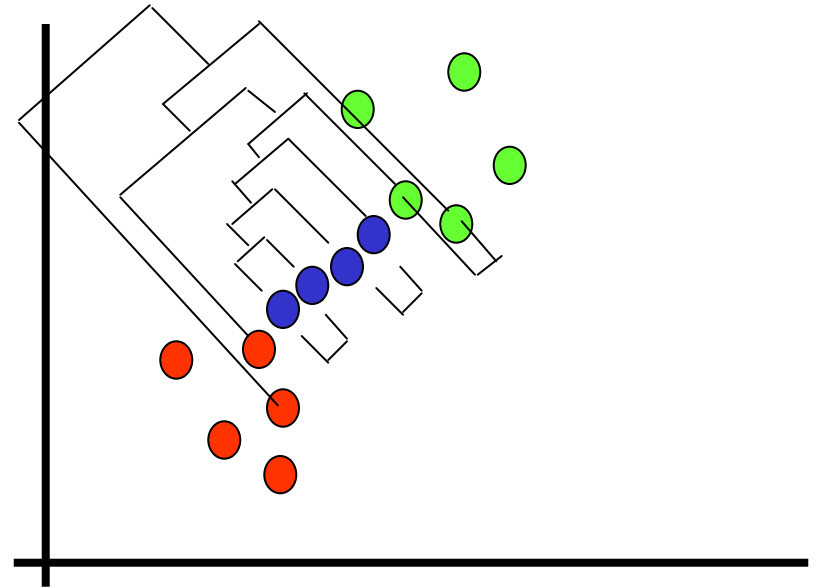
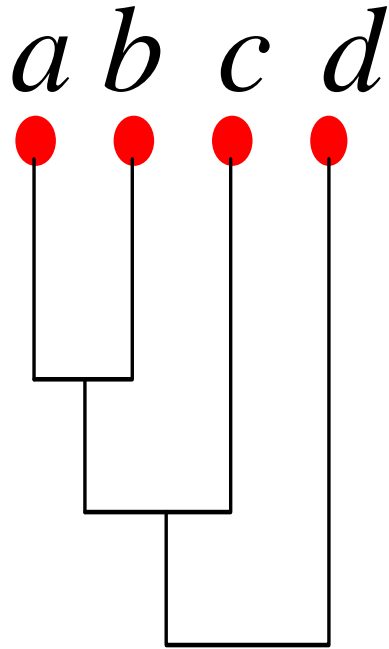
Clustering found (correlation distance):



# Clustering: Warning

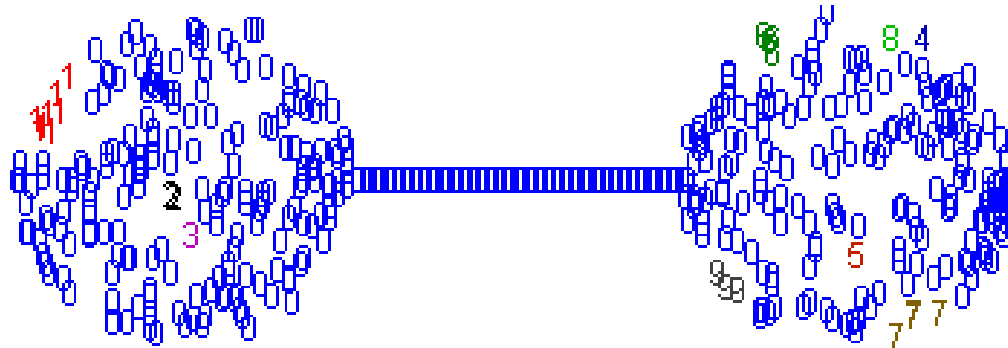
## Single linkage and chaining effect

Single-Link

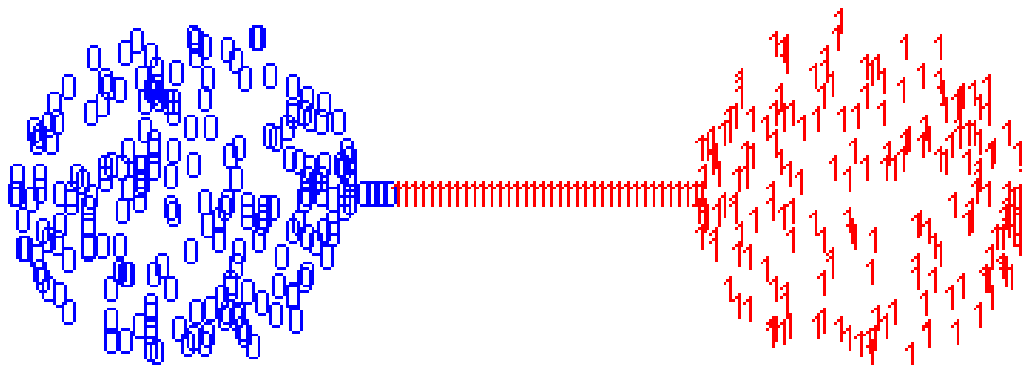


One growing cluster

# Example of chaining effect

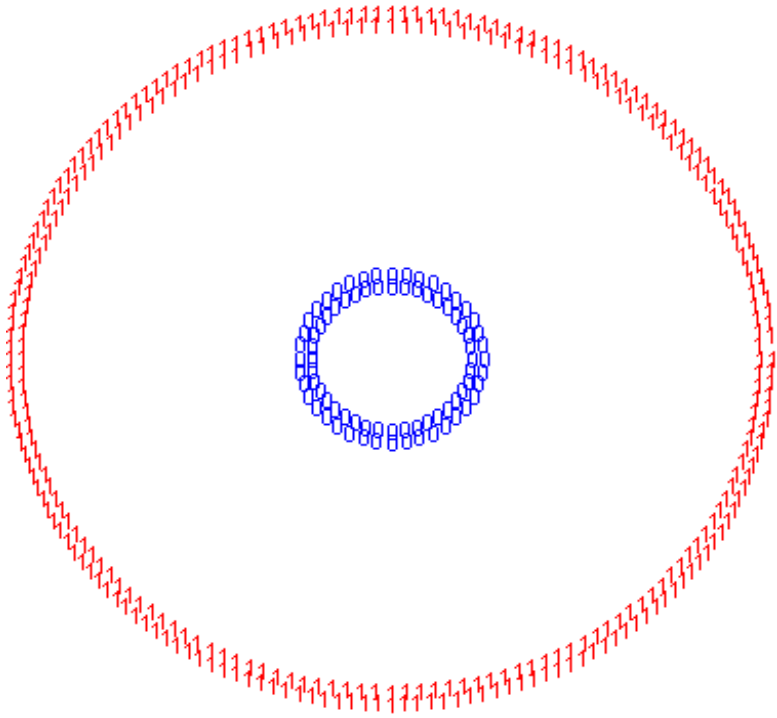


Single-link (10 clusters)

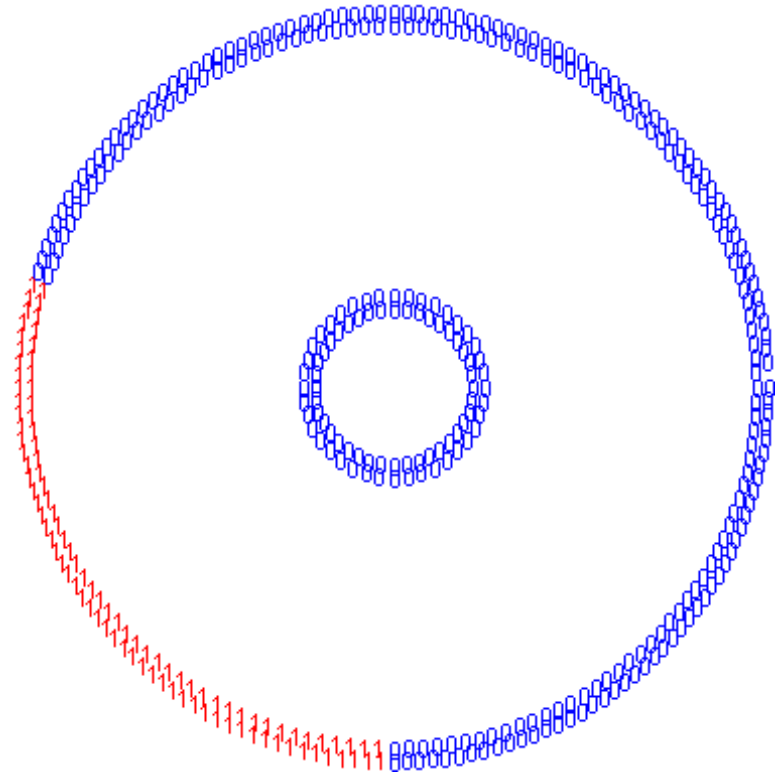


Complete-link (2 clusters)

# Single vs complete linkage

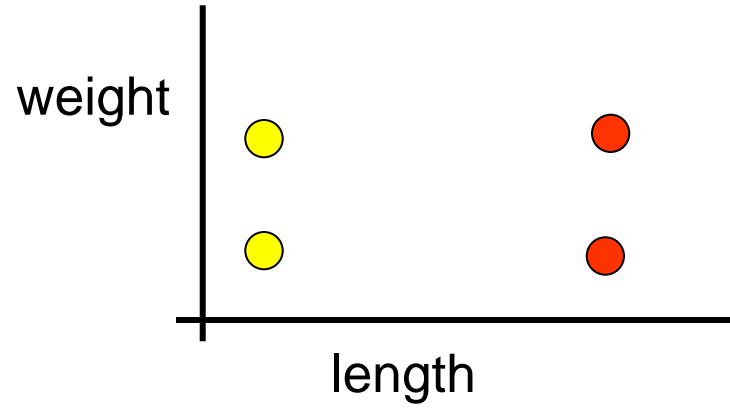


Single-link (2 clusters)

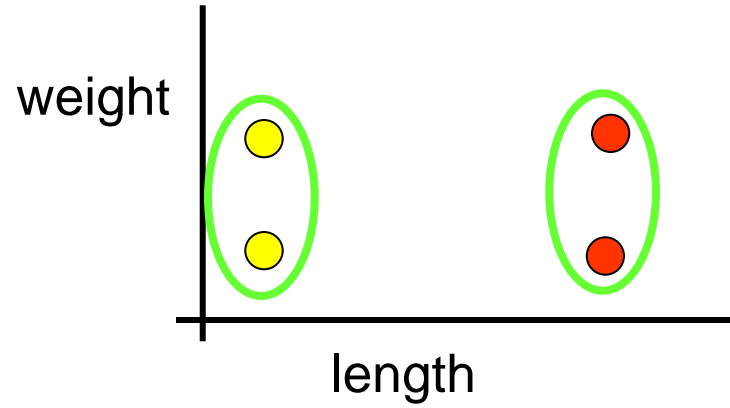


Complete-link (2 clusters)

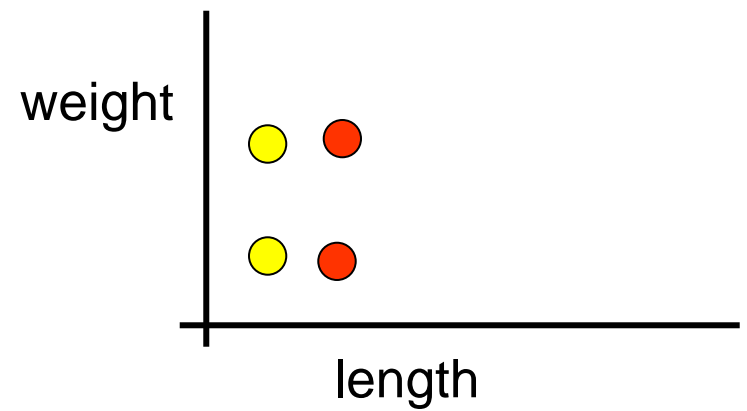
# Clustering: Warning



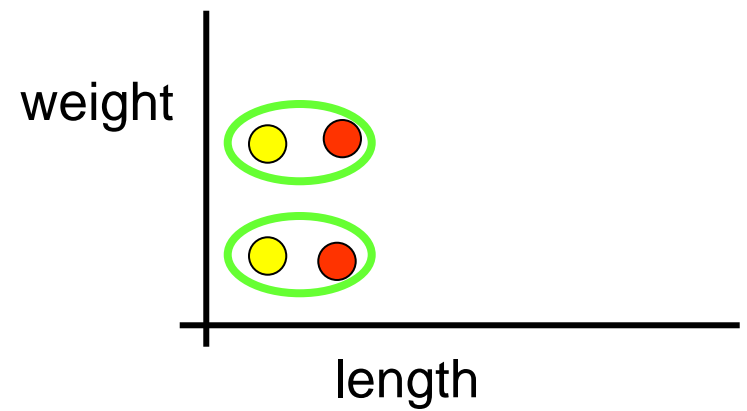
# Clustering: Warning



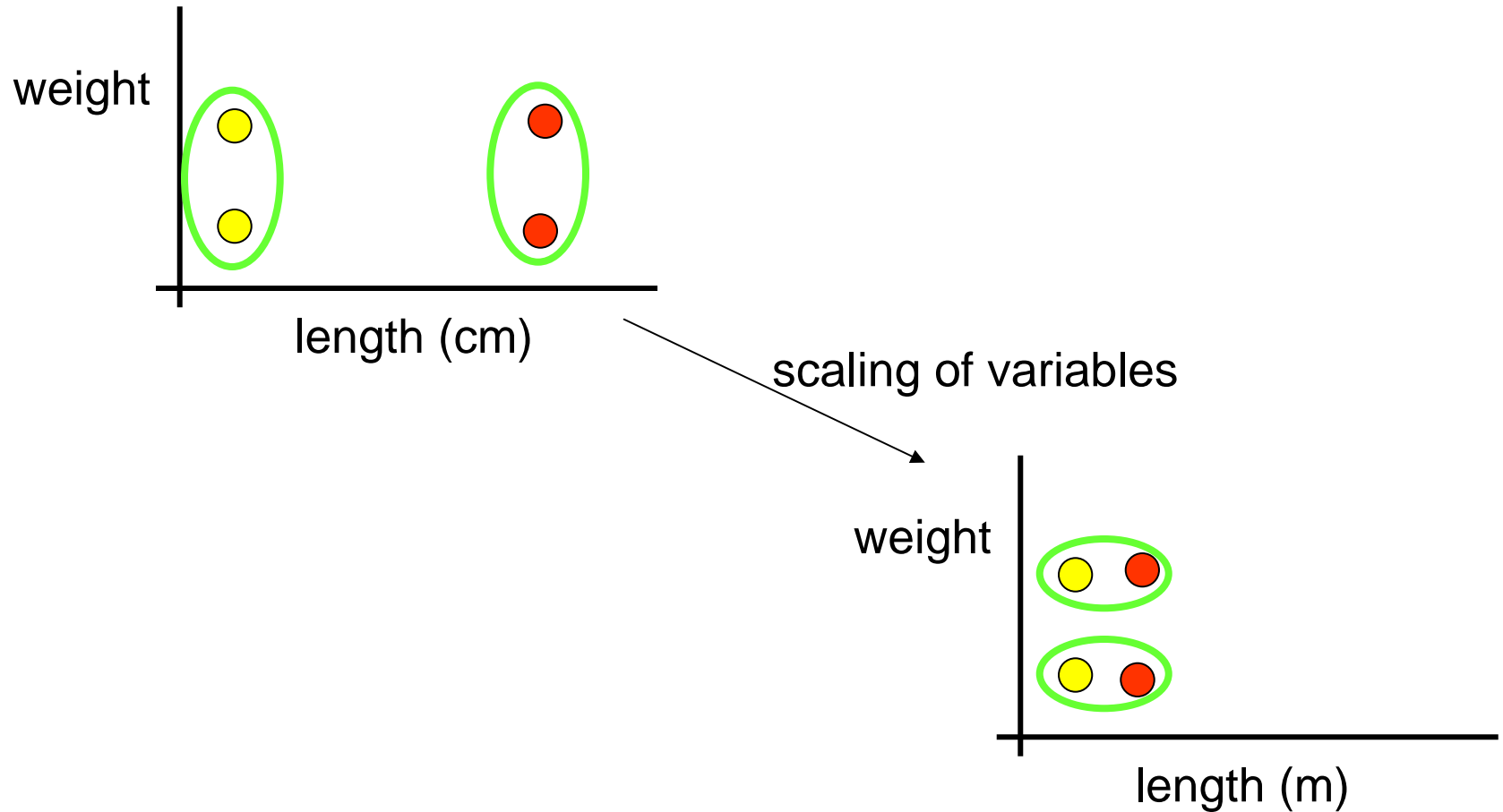
# Clustering: Warning



# Clustering: Warning



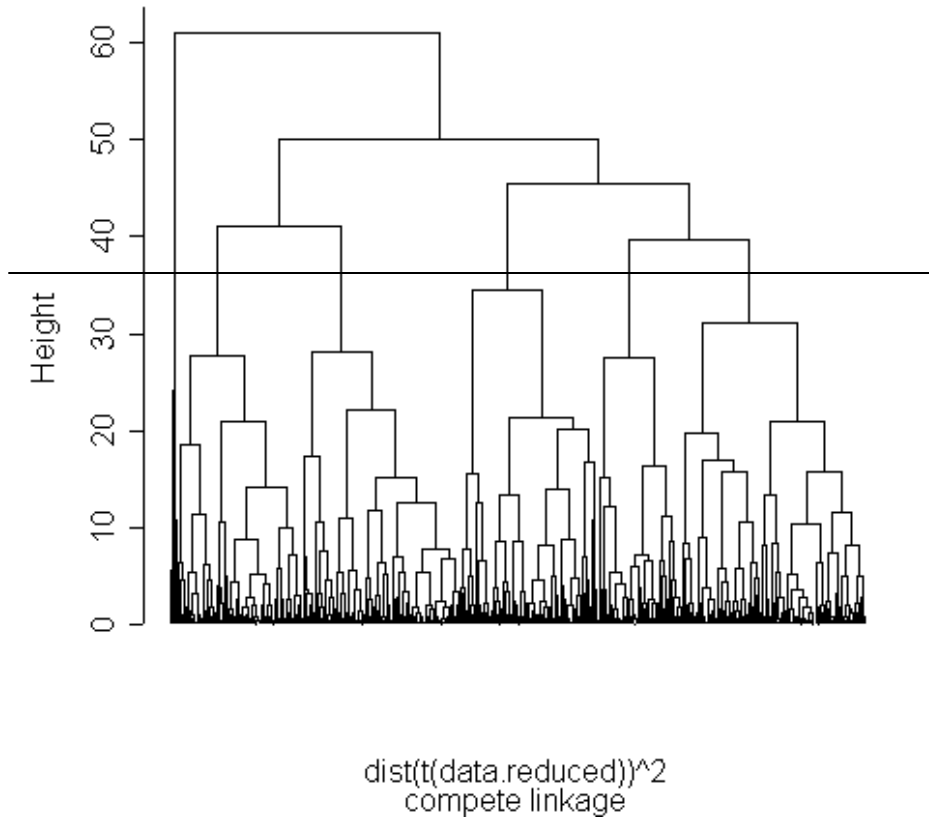
# Clustering: Warning Measurements on different scales



# Clustering: Warning

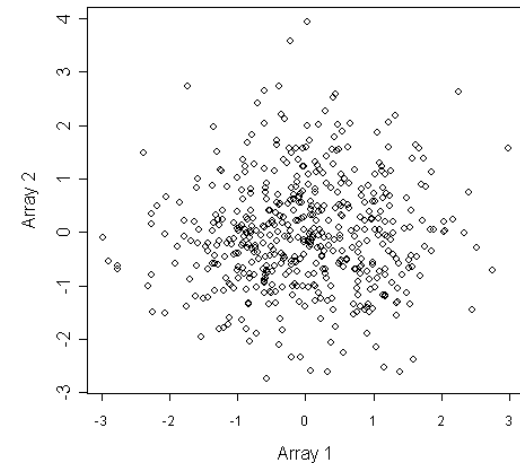
Cluster 500 genes, 5 arrays:

Dendrogram (Euclidian distance)



6 clusters

Data were random ...



# Clustering: Warning

- Select differentially expressed genes (t-test) for tumor subtype A versus tumor subtype B
- Cluster samples based on expression level of those genes

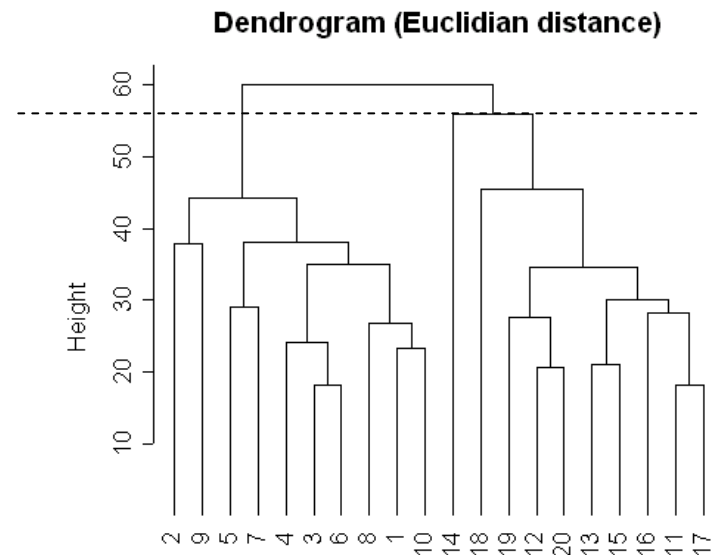
You find a cluster for subtype A and a cluster for subtype B:

subtype A and B are indeed different – publish paper

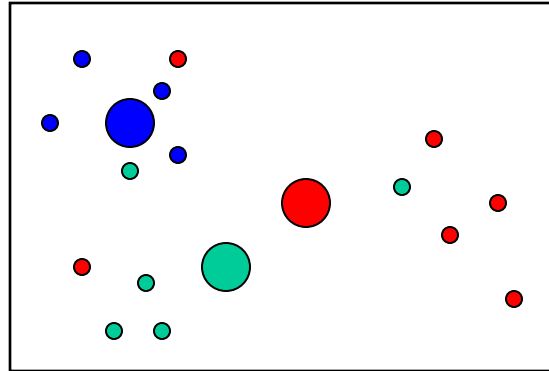
No: genes were selected for this!

Result can be obtained with

random data

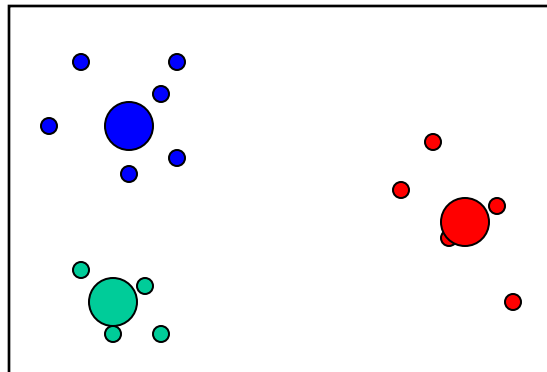
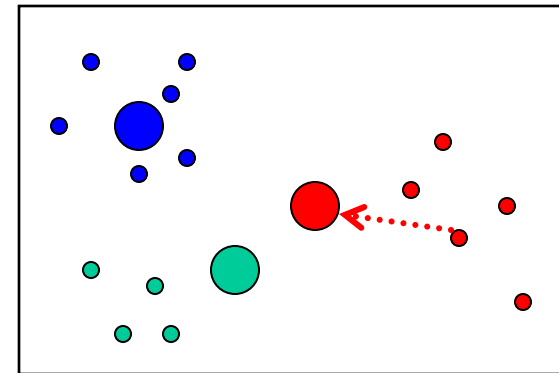


# K-means: Example, $K = 3$



**Step 1:** Make random assignments and compute centroids (big dots)

**Step 2:** Assign points to nearest centroids



**Step 3:** Re-compute centroids (in this example, solution is now stable)

# K-means: Weaknesses

- Must choose parameter  $K$  in advance, or try many values
- The algorithm works best on data which contains spherical clusters; clusters with other geometry may not be found
- The algorithm is sensitive to *outliers* --- points which do not belong to any cluster. These can distort the centroid positions and heavily bias clustering

# Visualization

Imagine 30,000 genes on 100 arrays:

30,000×100 expression matrix

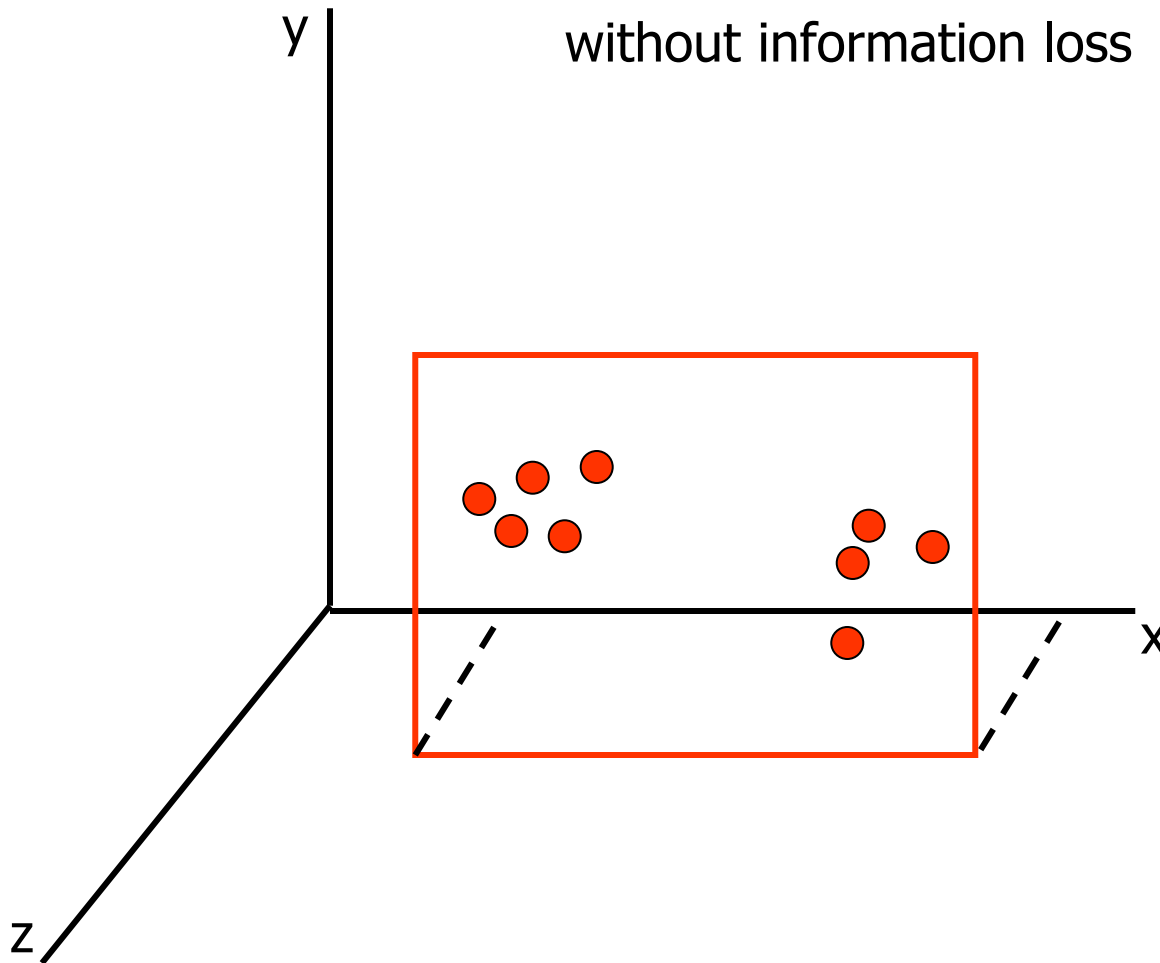
- computationally complex: reduce dimensionality
- difficult to plot: reduce to 2D or 3D

**Principal component analysis (PCA):** linear projection on selected rotated axes

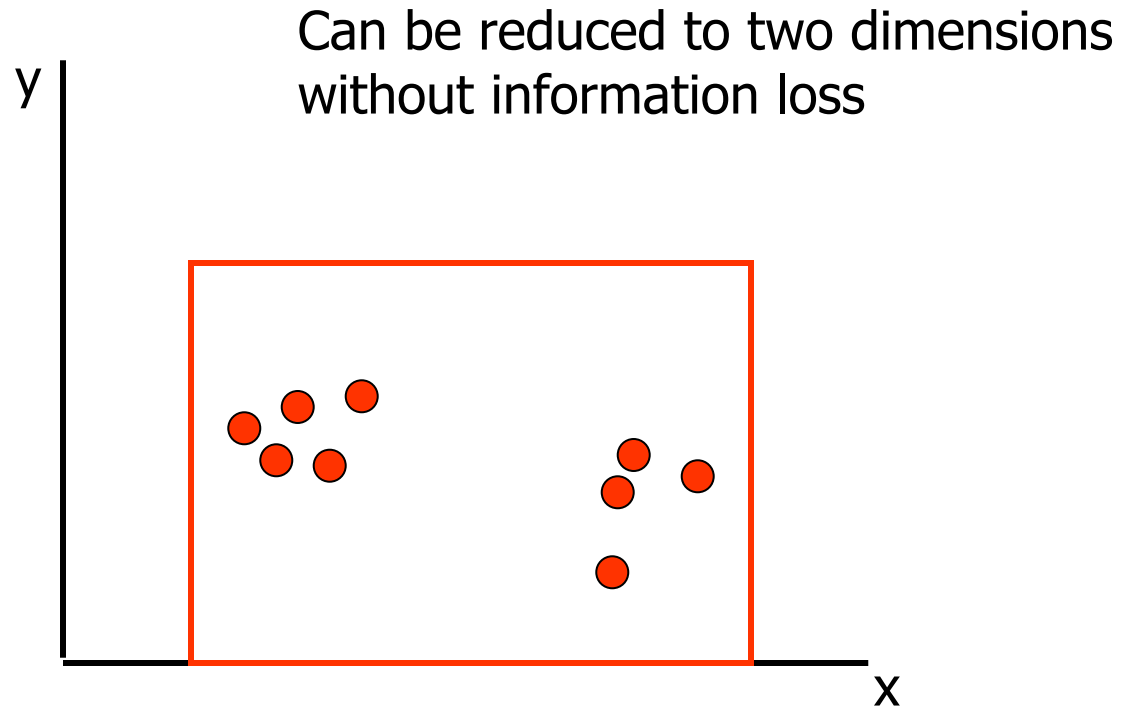
Retain as much of the structure/information as possible

# PCA and Dimensionality Reduction

Can be reduced to two dimensions without information loss



# PCA and Dimensionality Reduction

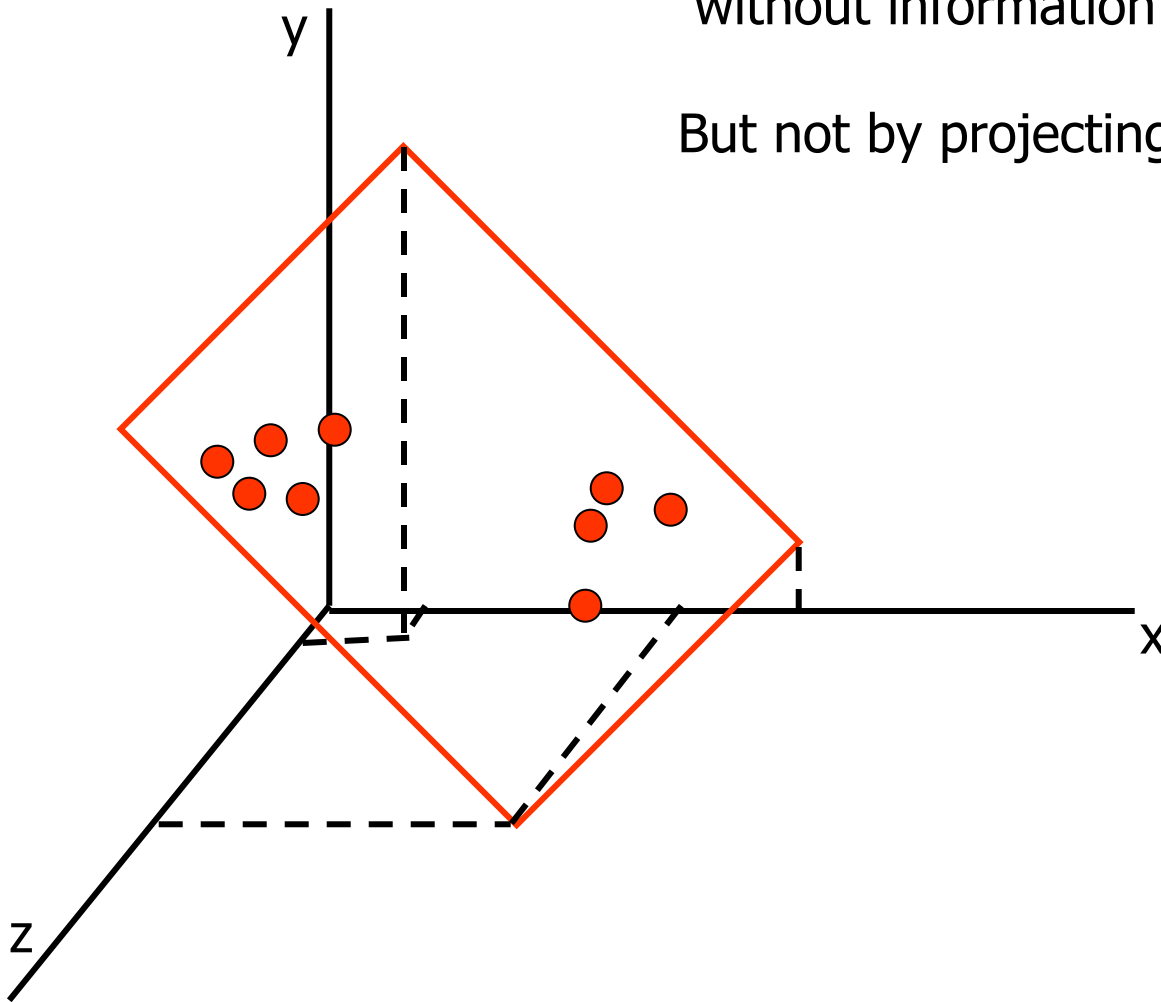


Simply project on x-y plane

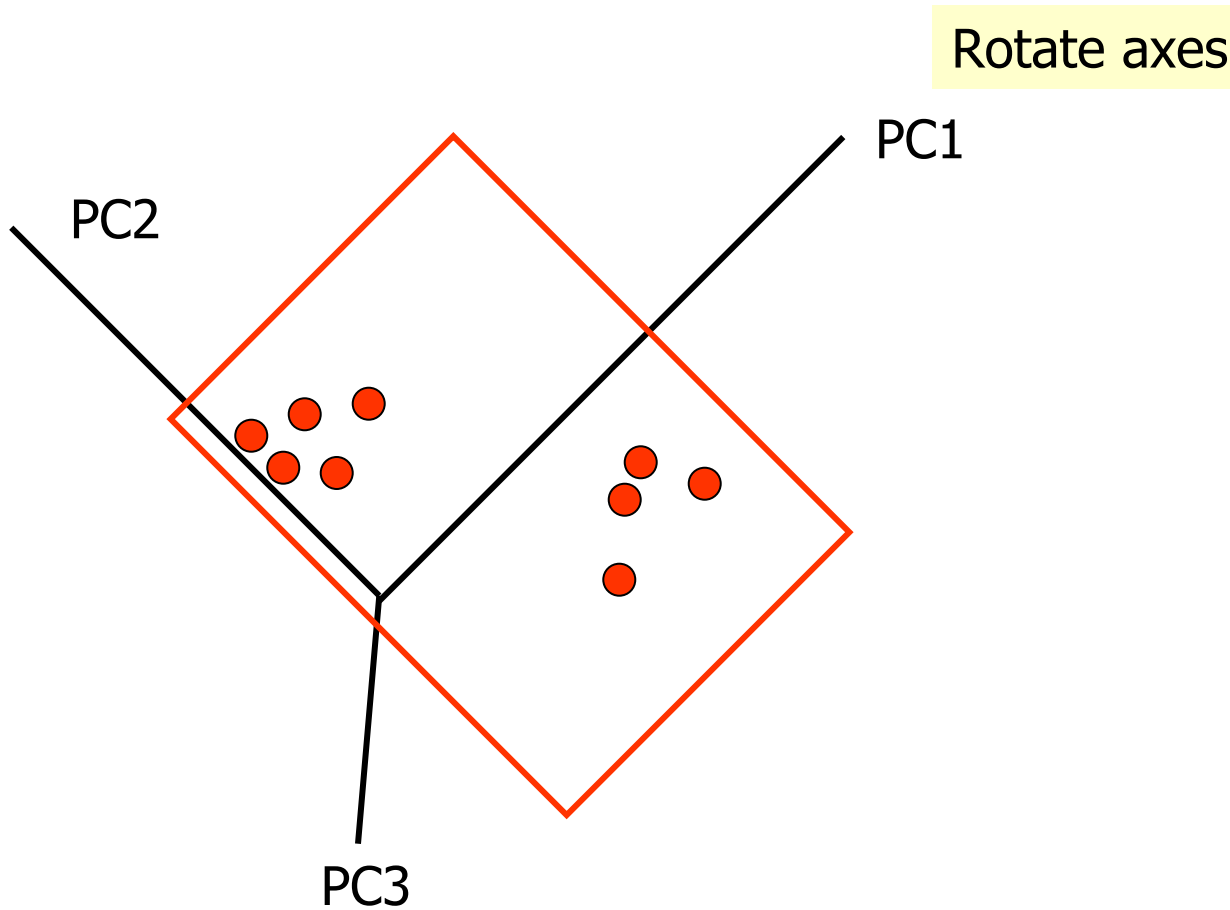
# PCA and Dimensionality Reduction

Can be reduced to two dimensions  
without information loss

But not by projecting on the axes

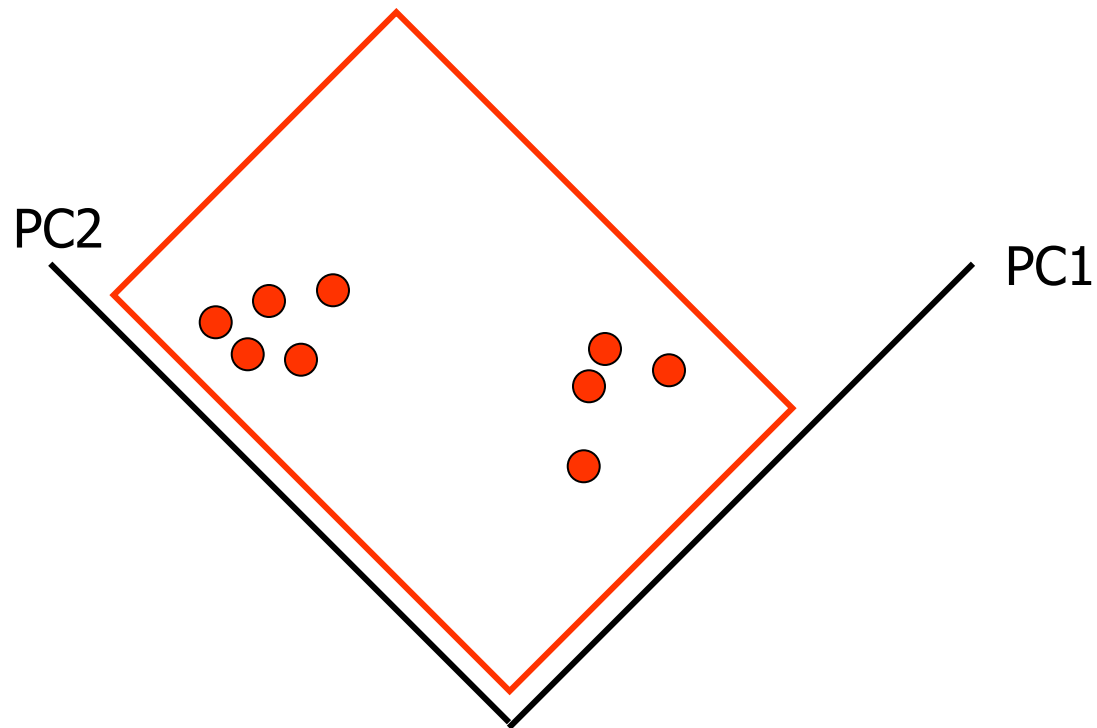


# PCA and Dimensionality Reduction

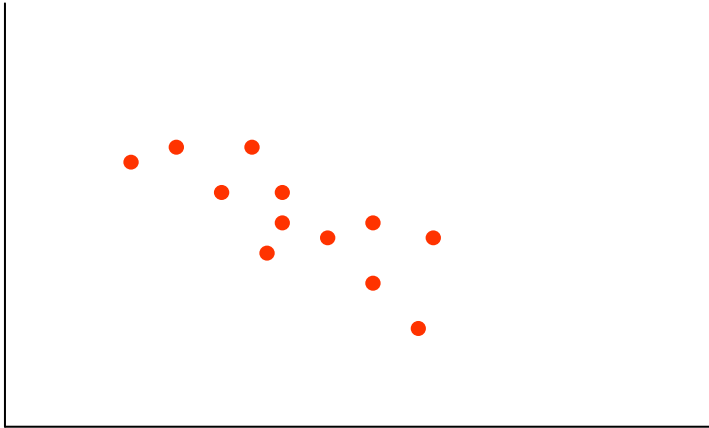


# PCA and Dimensionality Reduction

Rotate axes and project



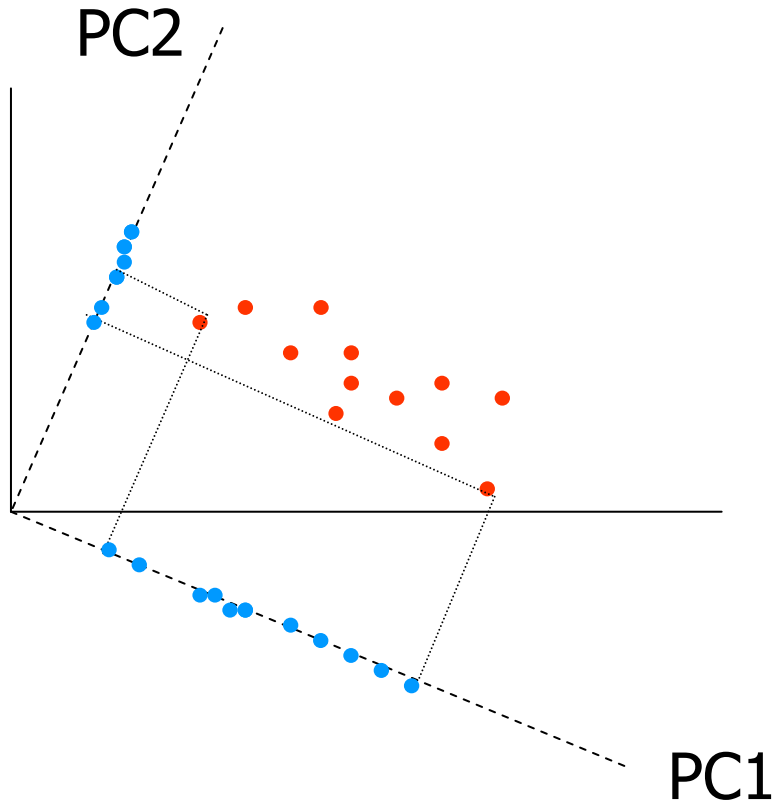
# PCA: Largest Variance



In general, one loses some information when projecting onto a smaller number of axes

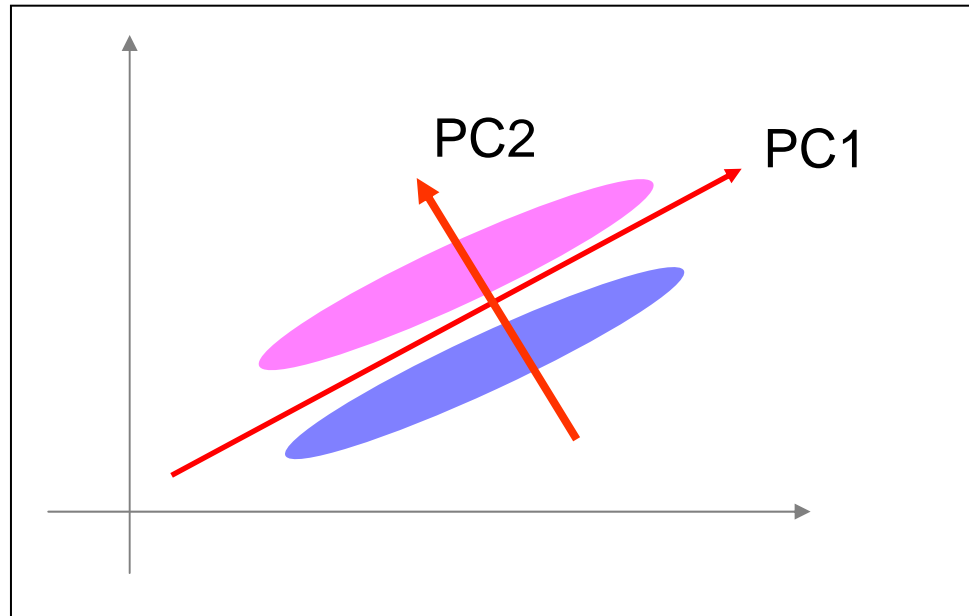
PCA: keep axes along which the spread (variance) is largest

# PCA: Largest Variance

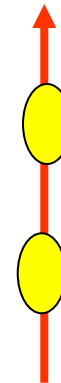


PC1 retains most of the variance

# PC in direction of largest variance: situation 1



PC2

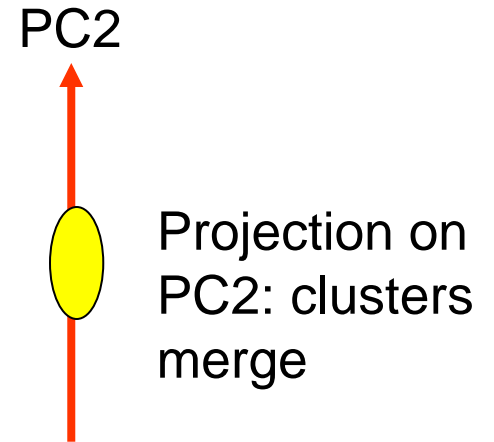
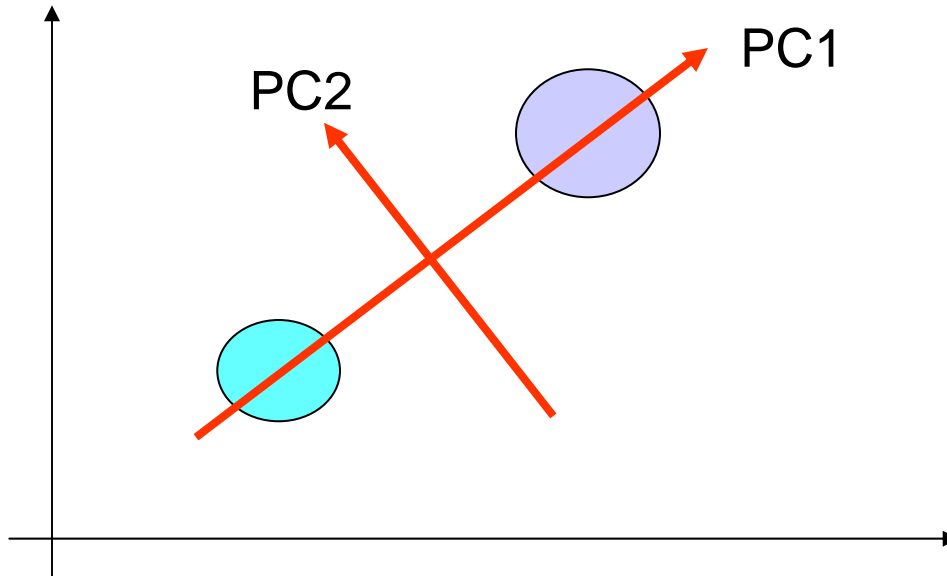


Projection on  
PC2: two  
clusters

Projection on PC1: clusters are merged



# PC in direction of largest variance: situation 2



Projection on PC1: 2 clusters



# PCA: Math

Data:  $x = (x_1, x_2, x_3, \dots, x_p)$

Projection:  $z = (z_1, z_2, z_3, \dots, z_d)$   $d < p$

Principal components:  $u_i = (u_{i1}, u_{i2}, u_{i3}, \dots, u_{ip})$  ( $1 \leq i \leq d$ )

$$z_1 = u_{11}x_1 + u_{12}x_2 + \dots + u_{1p}x_p$$

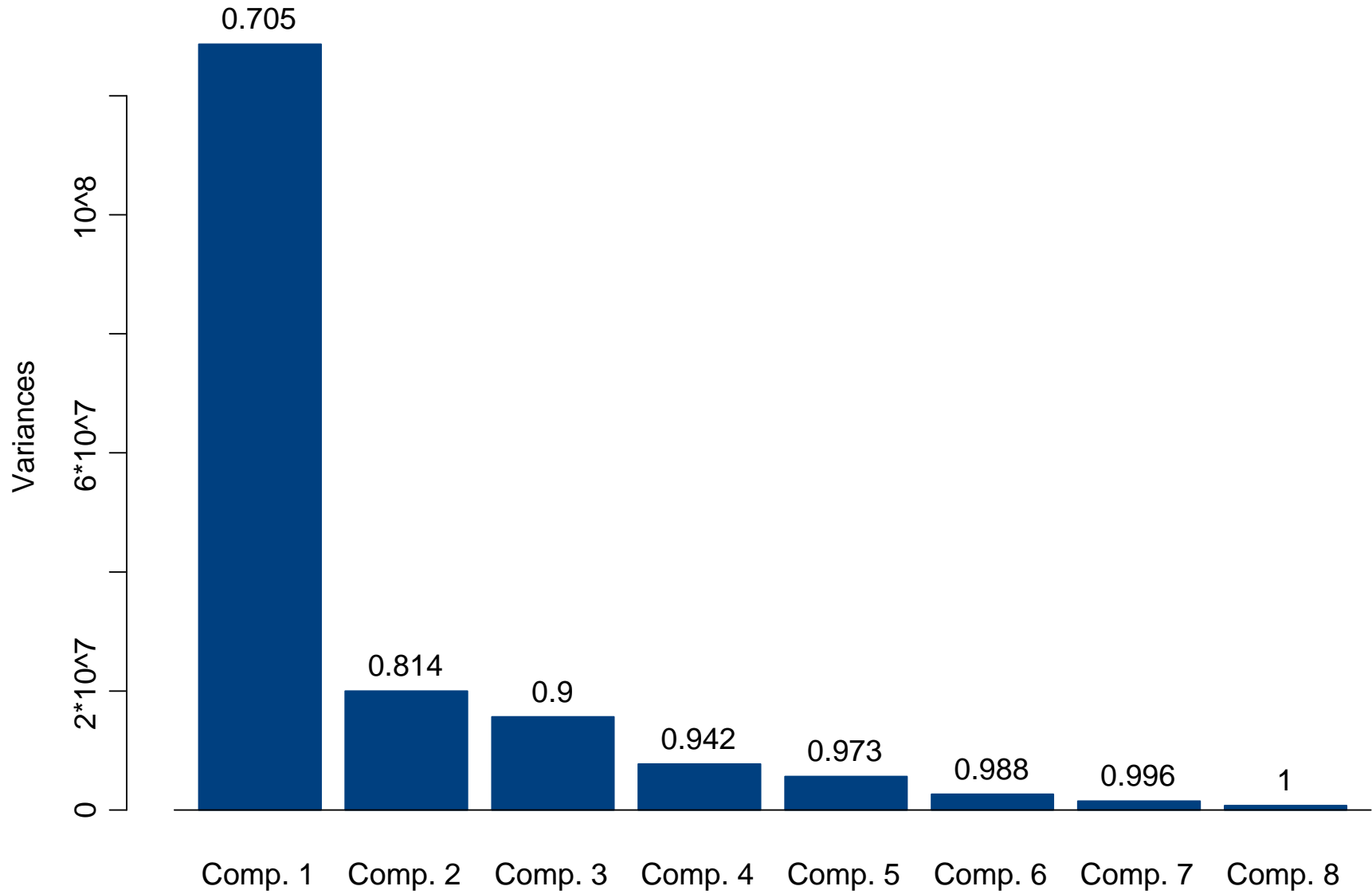
$$z_2 = u_{21}x_1 + u_{22}x_2 + \dots + u_{2p}x_p$$

...

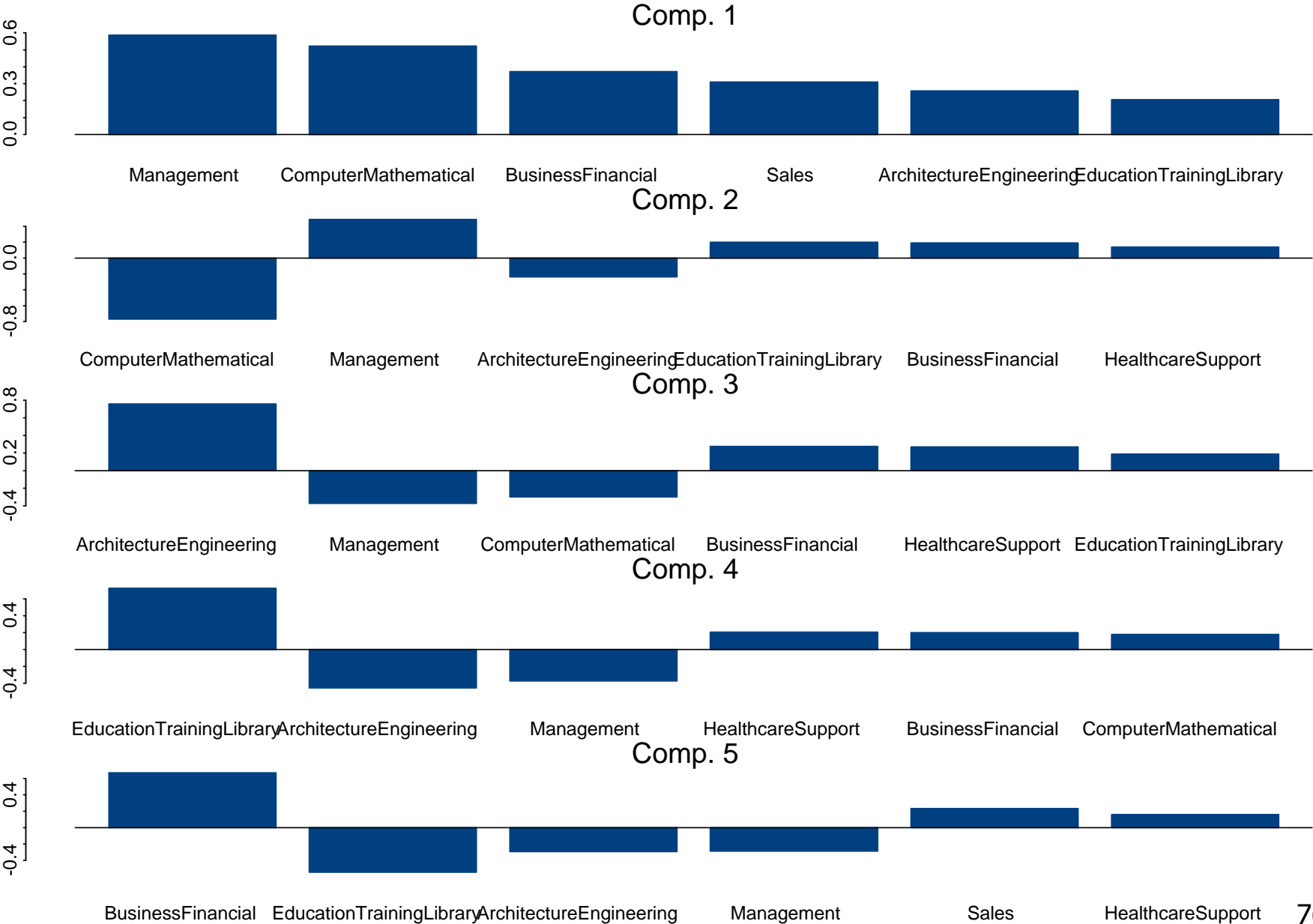
$$z_d = u_{d1}x_1 + u_{d2}x_2 + \dots + u_{dp}x_p$$

**Coefficients (“loadings”),  $u$ , represent the linear correlation between the original variables,  $x_n$ , and the PC $n$**

# Relative Importance of Principal Components



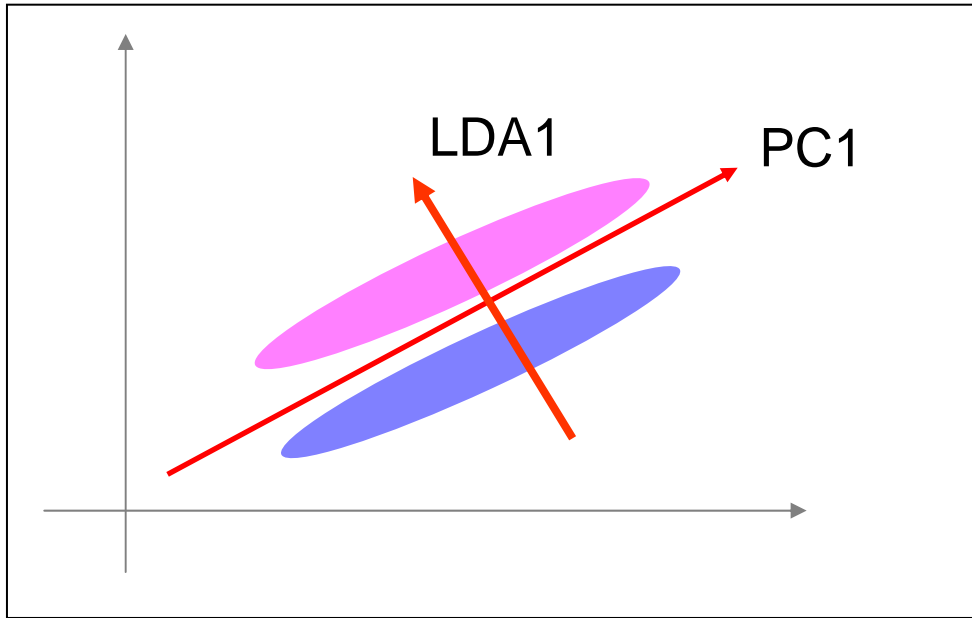
# PCA Loadings



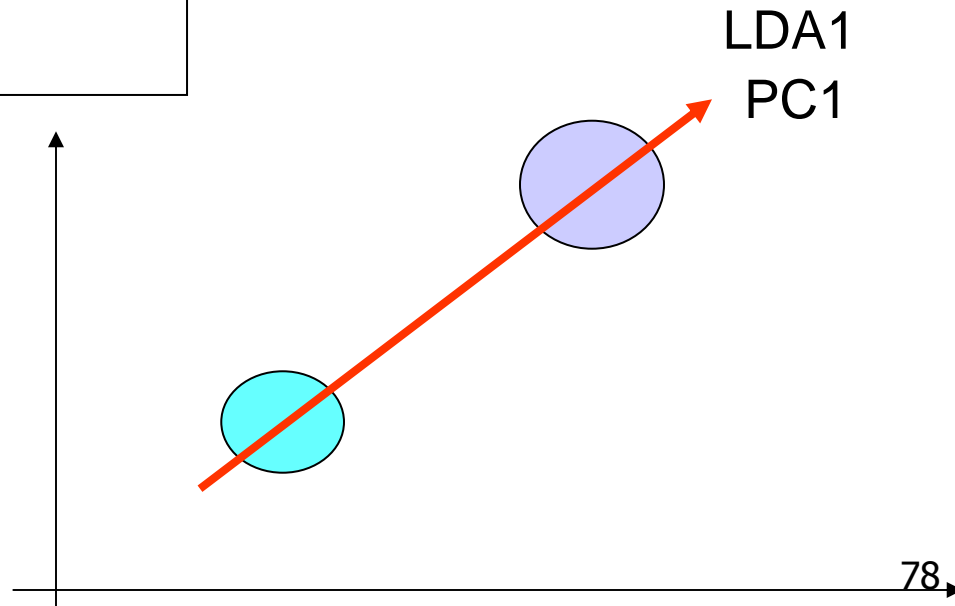
# Linear discriminant analysis (LDA)

- Supervised; used for classification of unknown samples/genes
- Need to build model first (training)
- May require feature selection
- LDA calculates hyperplanes to separate classes
- Need for cross-validation, bootstrapping, to test validity of model

# PCA vs LDA

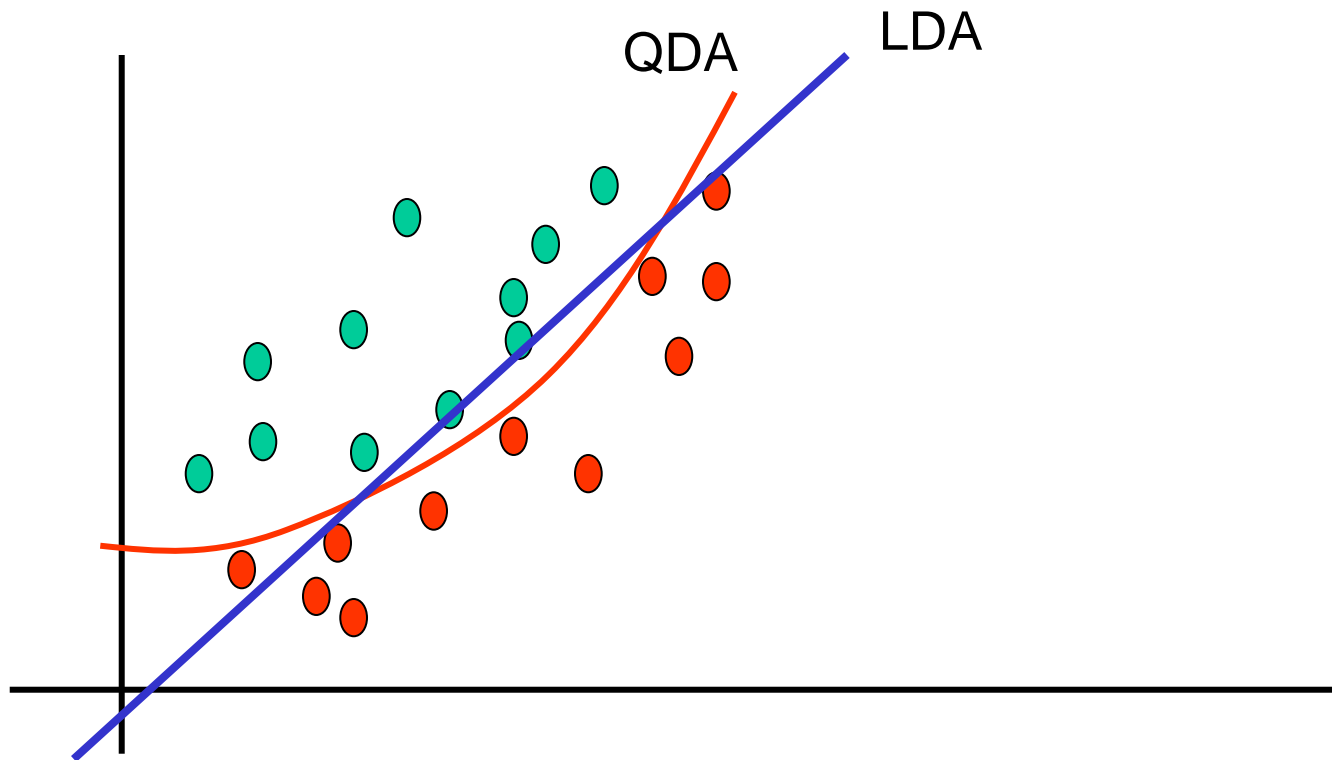


PC1: max variance  
LDA1: max separation



# Quadratic Discriminant Analysis

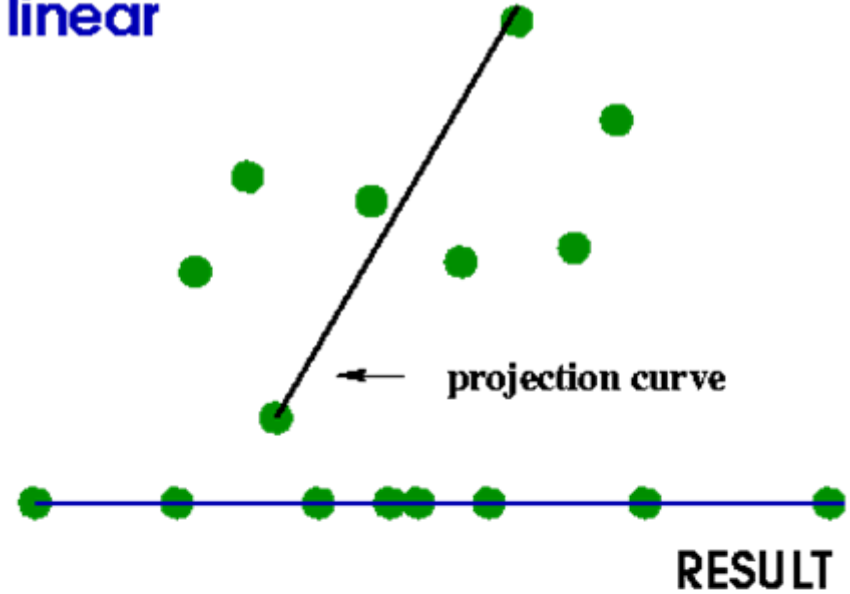
Quadratic hyperplanes to separate clusters (with different variances)



# Multi dimensional scaling

PCA

linear



MDS

nonlinear

