Clustering

Methods Course: Gene Expression data Analysis

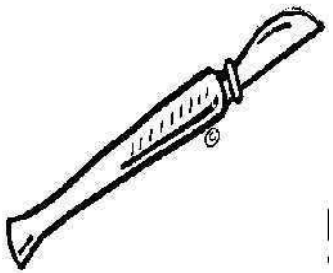
- Day Four -

Rainer Spang

Physicians are faced with treatment decisions every day

One disease:

Three alternative therapies:

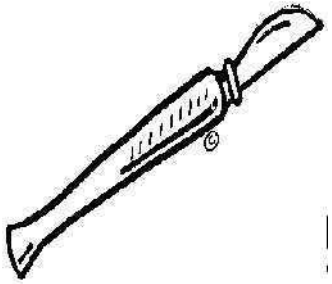


Clinical studies guide evidence based medicine

On average:

75%

55%

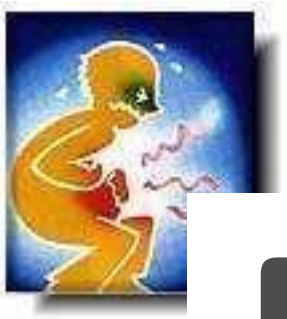


35%

therapeutic success

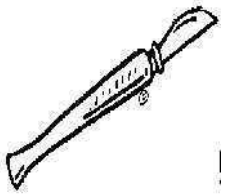
Diseases can be refined into subtypes

A

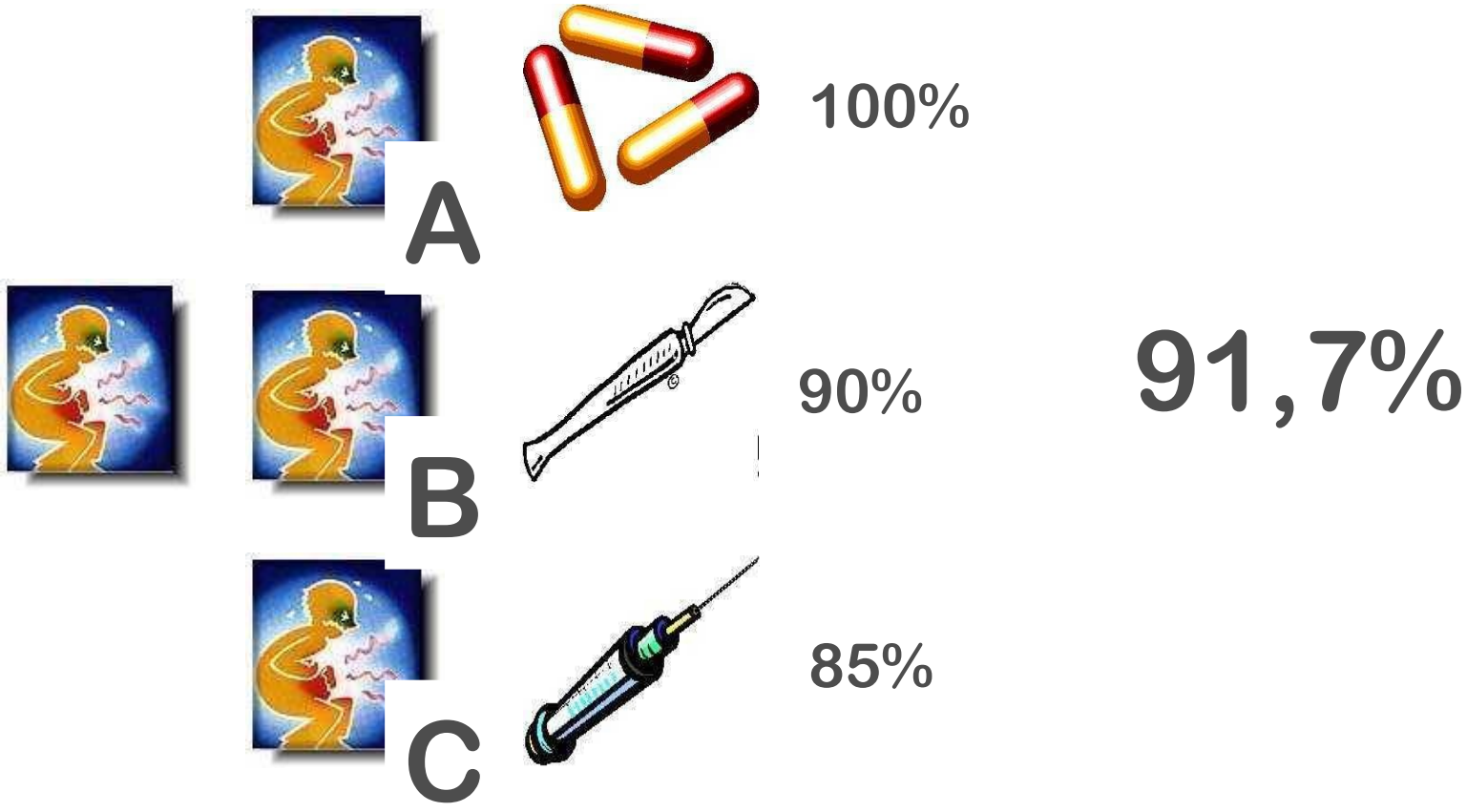


B

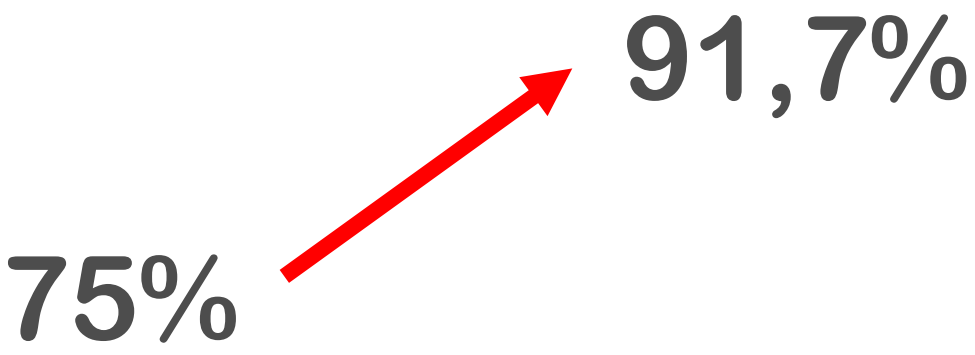
C

	 A	 B	 C
	**100%**	60%	65%
	40%	40%	**85%**
	10%	**90%**	5%

Diagnosis of subtypes leads to different treatment decisions



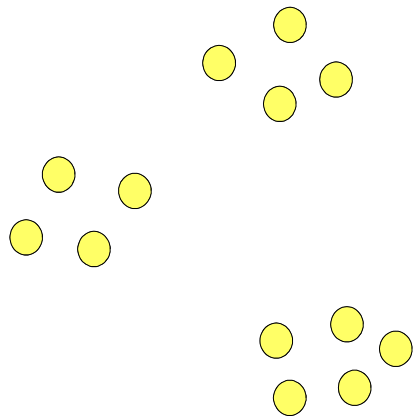
The refined diagnosis has lead to more therapeutic success



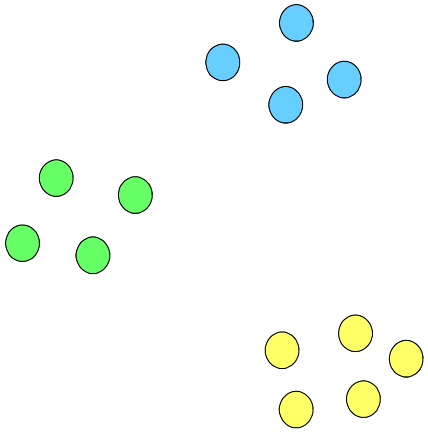
Without the development of any new therapies!

How do we define sub-entities of cancers using expression profiles?

Clustering aims at grouping similar objects (expression profiles) together



Data



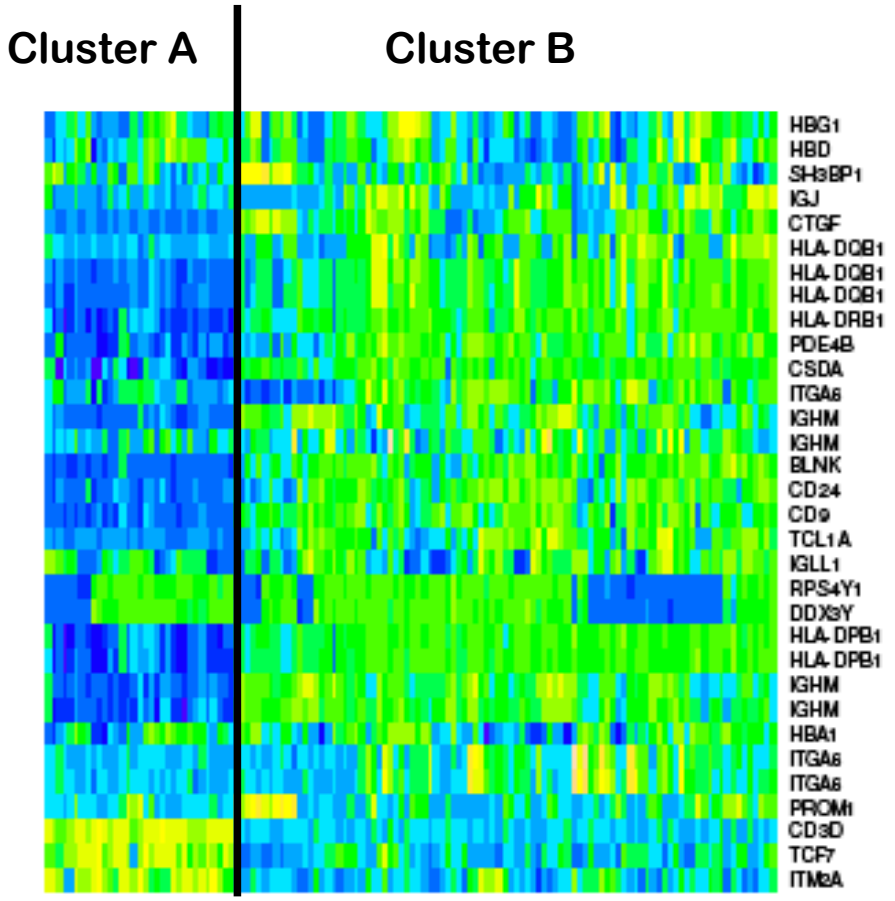
Clustering

Clustering of expression profiles can identify subtypes of tumors

Chiaretti et al. (2004)

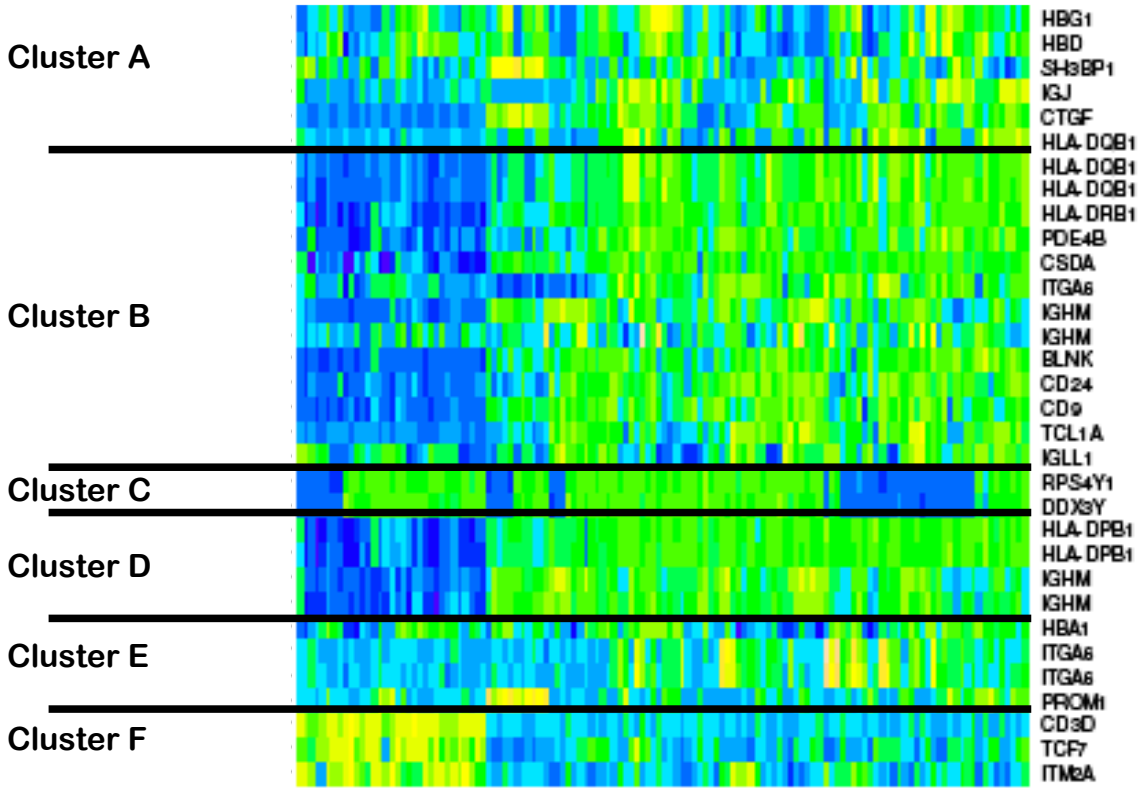
Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival.

Blood 103(7):2771-8



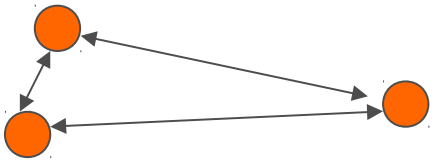
Note that the clustering does not use all genes on the array.

Clustering of genes reveals the transcriptional modules of a cell

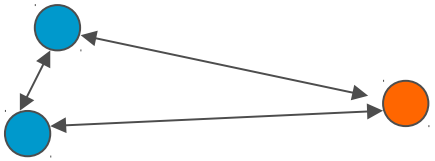


Clustering is driven by two components

Distance measure:
Quantification of the dissimilarity of objects.



Clustering-Algorithm:
Computational method to group objects based on a chosen distance measure.



The Euclidean distance is a standard distance measure

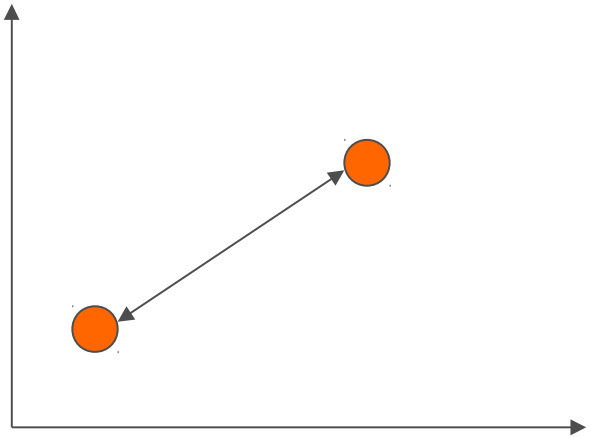
Expression profiles:

$x = (x_1, \ldots, x_n)$,

$y = (y_1, \ldots, y_n)$

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

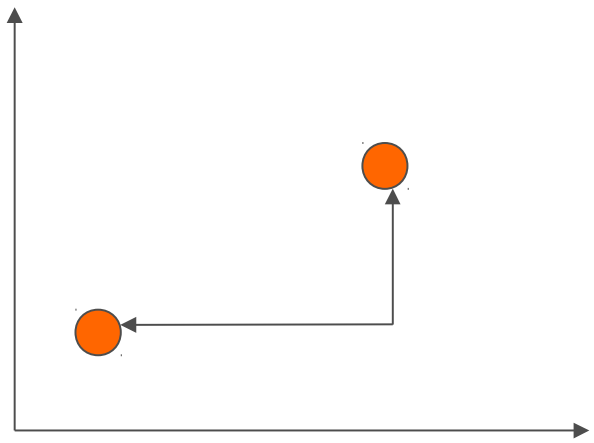
Distance as the crow flies



The square renders the distance measure sensible to outliers

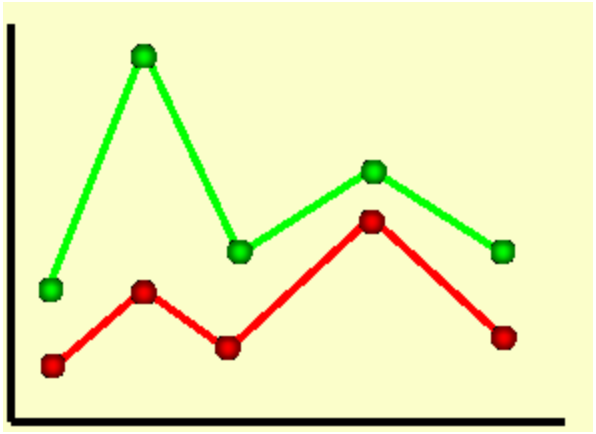
The Manhattan distance is a more robust distance measure

$$d_M(x,y) = \sum_{i=1}^{n} |x_i - y_i|.$$



The Pearson correlation is a measure that groups profiles according to their general shape

$$d_C(x,y) = 1 - \frac{\sum_{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}(x_i - \bar{x})^2 \sum_{i=1}(y_i - \bar{y})^2}}.$$

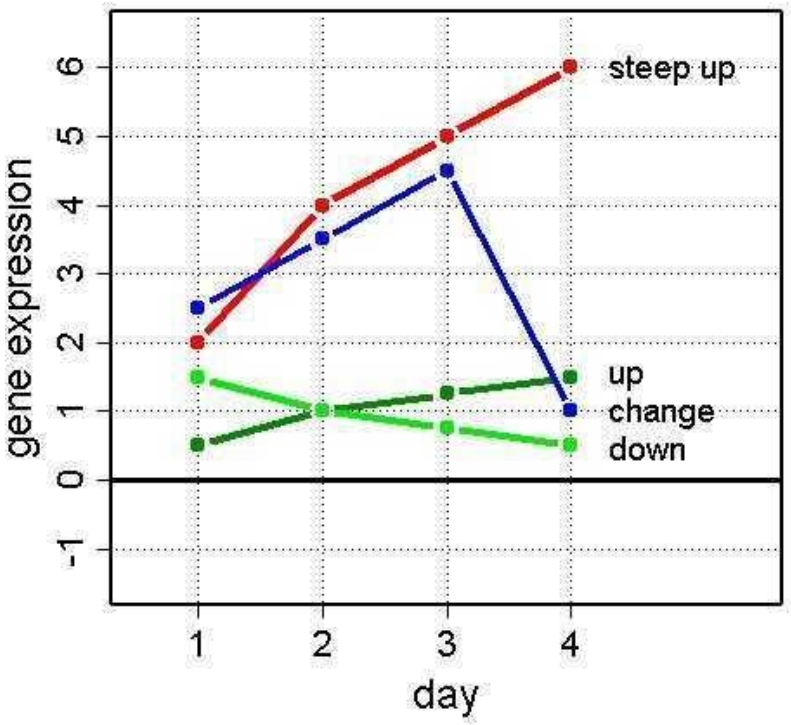


Cluster genes in time series

Pearson correlation describes the "linear dependence" of vectors:

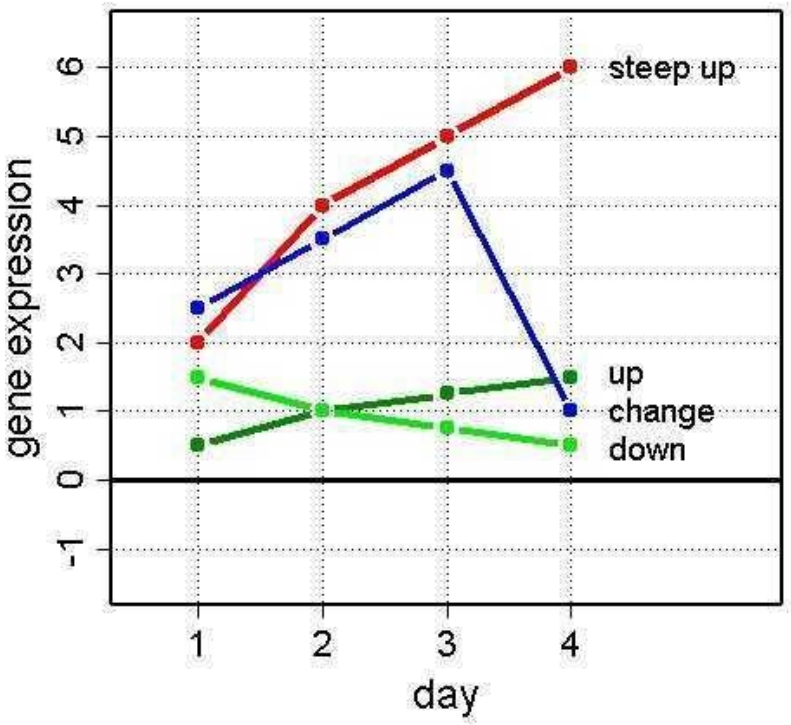
$d_c(x, y) = d_c(ax+b, y)$, $a > 0$

Example clustering of time series



steep up: x1=(2,4,5,6)
up: x2=(2/4,4/4,5/4,6/4)
down: x3=(6/4,4/4,3/4,2/4)
change: x4=(2.5,3.5,4.5,1)

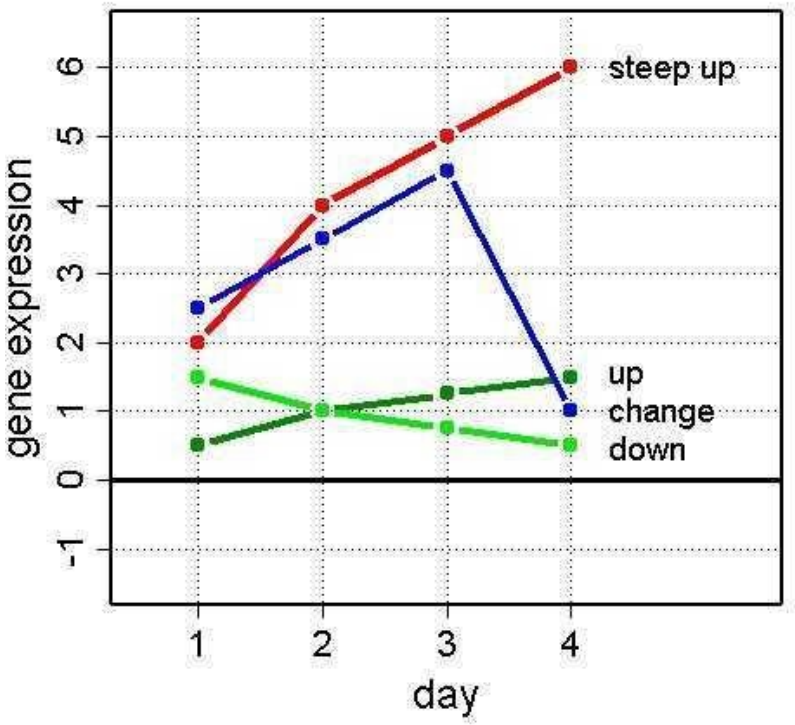
The Euclidean distance defines similarities on the time series



0	2.60	2.75	2.25
2.60	0	1.23	2.14
2.75	1.23	0	2.15
2.25	2.14	2.15	0

Matrix of pairwise distances

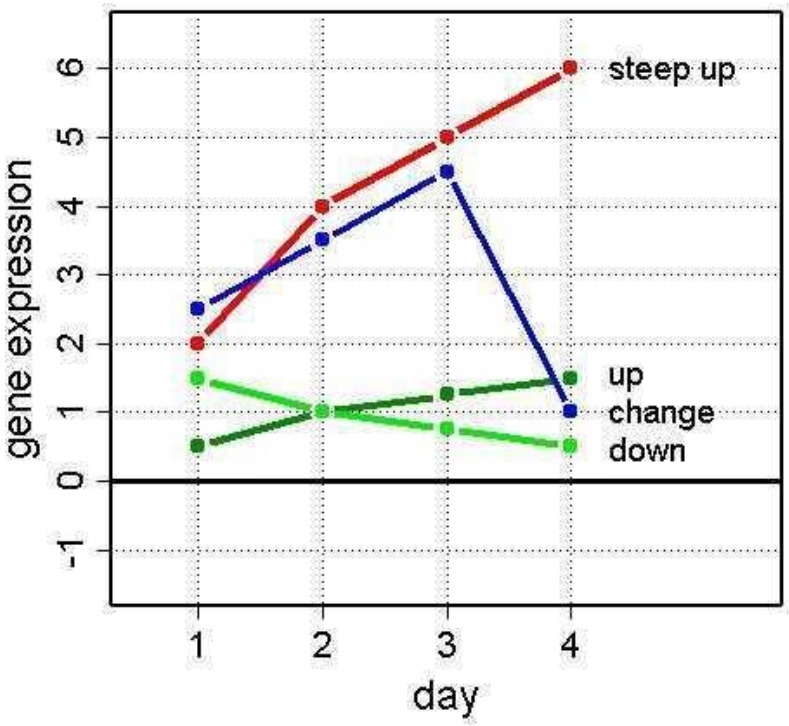
Distances generated by the Manhattan distance are different



0	12.75	13.25	6.50
12.75	0	2.50	8.25
13.25	2.50	0	7.75
6.50	8.25	7.75	0

Matrix of pairwise distances

Correlation distance is different again



0	0	2	1.18
0	0	2	1.18
2	2	0	0.82
1.18	1.18	0.82	0

Matrix of pairwise distances

d(x,y)=0 does not imply x=y

→ no metric

Clustering algorithms are methods that assign objects to groups

The algorithms build on underlying distance measures.

We will discuss:

- **Hierachichal Clustering**
- **K-Means-Clustering**
- **Partitioning around Medoids**

Hierarchical clustering builds a hierarchy of clusters

At the start every profile is a cluster of size one.

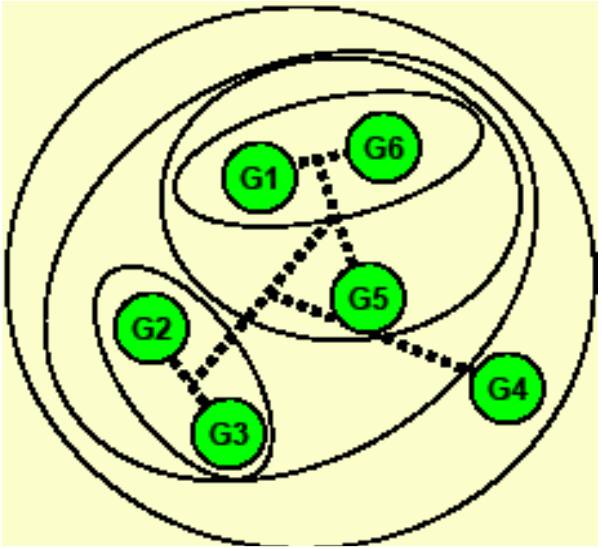
Compute distance between all profiles.

Find the pair of profiles with the smallest distance.

Join these two profiles to build a new cluster of size 2.

Compute the distance of this cluster to all other clusters.

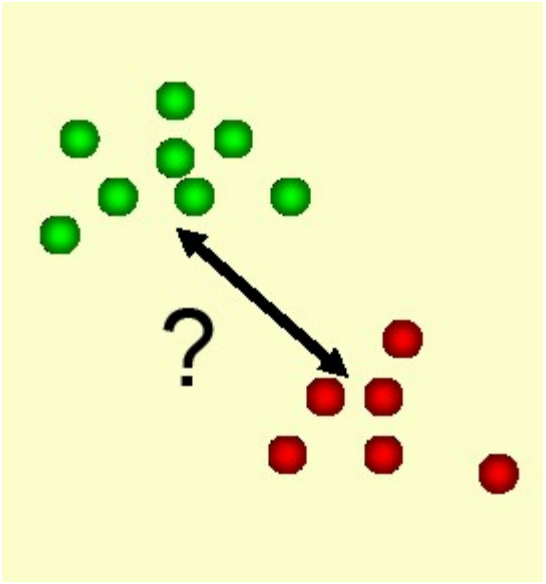
Repeat until only one cluster is left.



The algorithm asks for the distance of clusters

So far we only know about the distance of objects.

How do we compute the distance between clusters of objects?



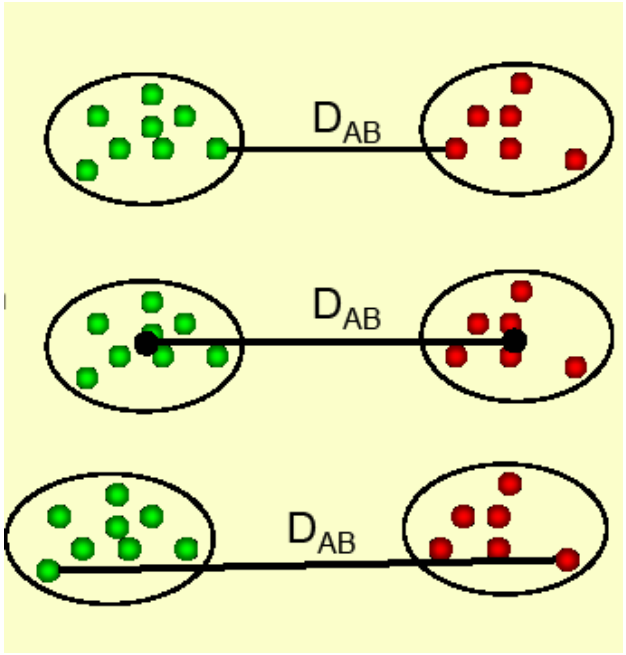
The distance of clusters can be calculated by different linkage methods

Compute all distances of every point in cluster A to every point in cluster B.
The distance of clusters can then be defined as:

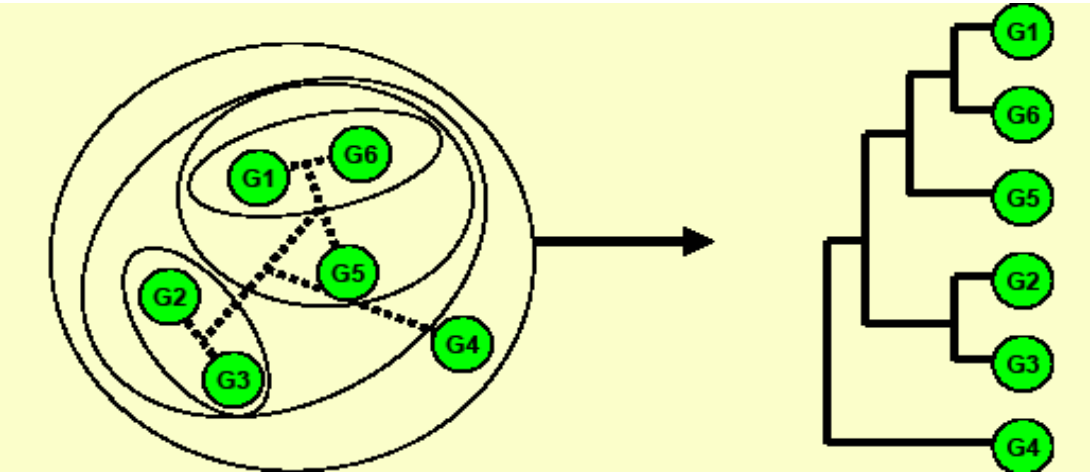
The minimum of distances (single linkage)

The average of distances (average linkage)

The maximum of distances (complete linkage)



A dendrogramm visualizes the hierarchies of clusterings



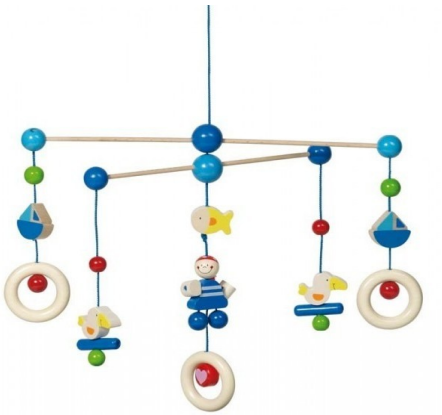
If the algorithm joins two clusters they get connected by a bracket of the dendrogram.

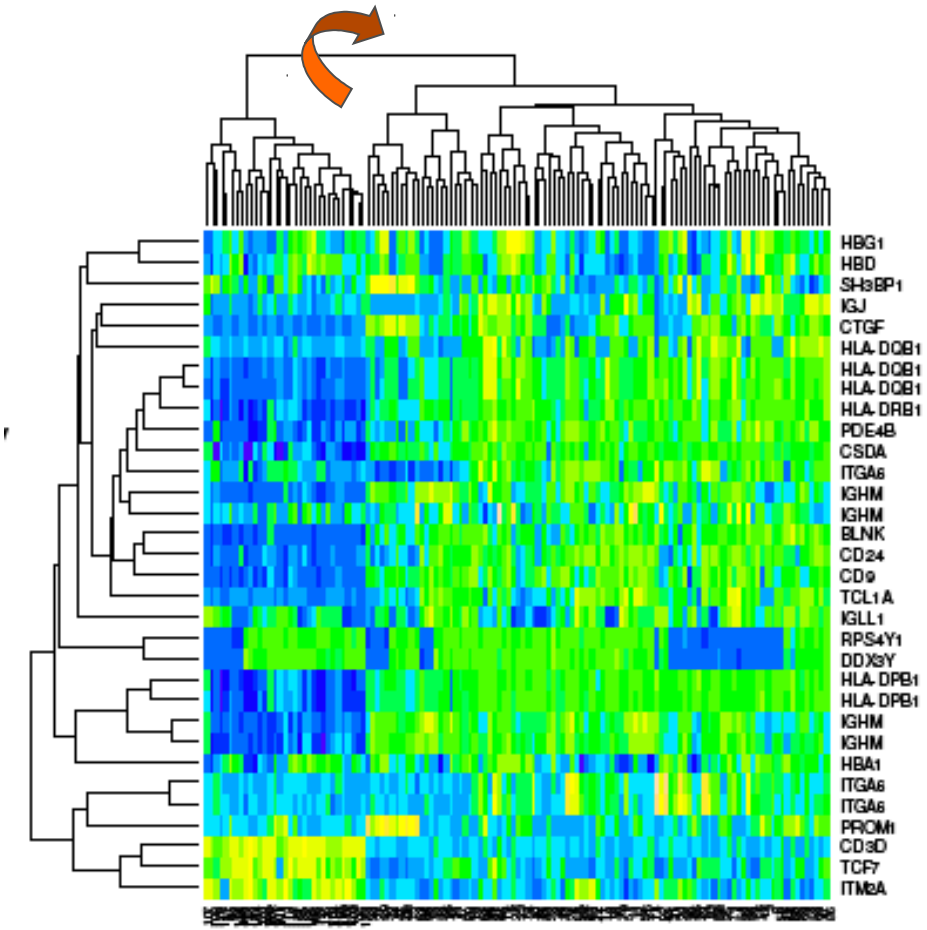
The height of the bracket reflects the similarity of the clusters.

The dendrogram is often used to order genes and samples in a heatmap

Note that the dendrogram does not uniquely define the order of samples and genes.

You can rotate clusters like in a mobile.





Hierarchies need to be cut to generate clusterings

The algorithm generates a dendrogram but no clustering.

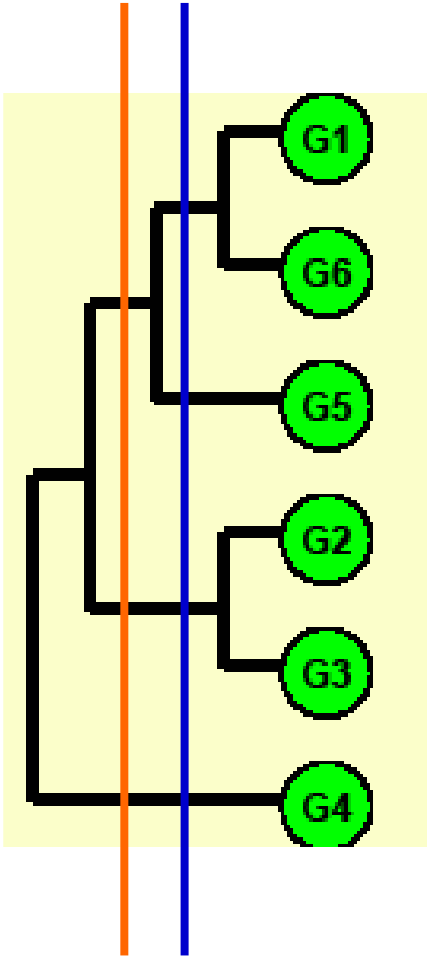
The dendrogram can be cut at different levels.

Every cut defines a different clustering.

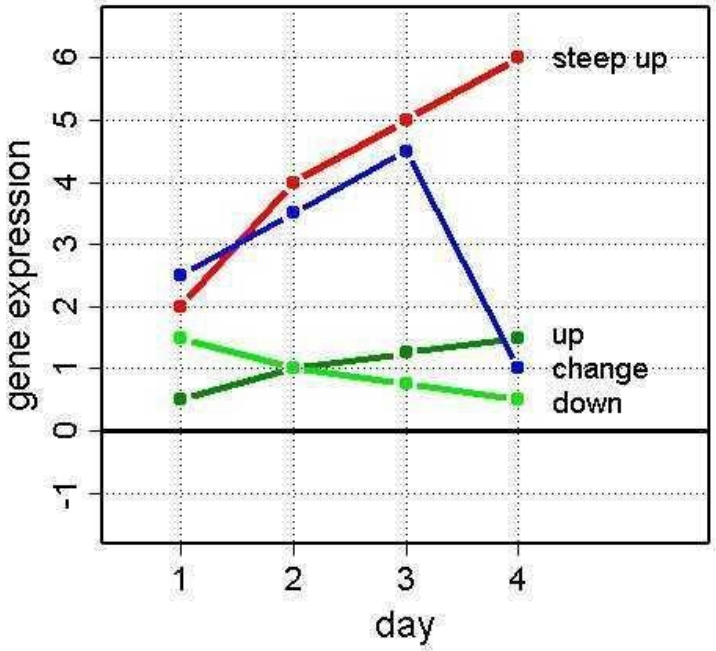
The orange cut generates 3 clusters (G1,G6,G5), (G2,G3), (G4).

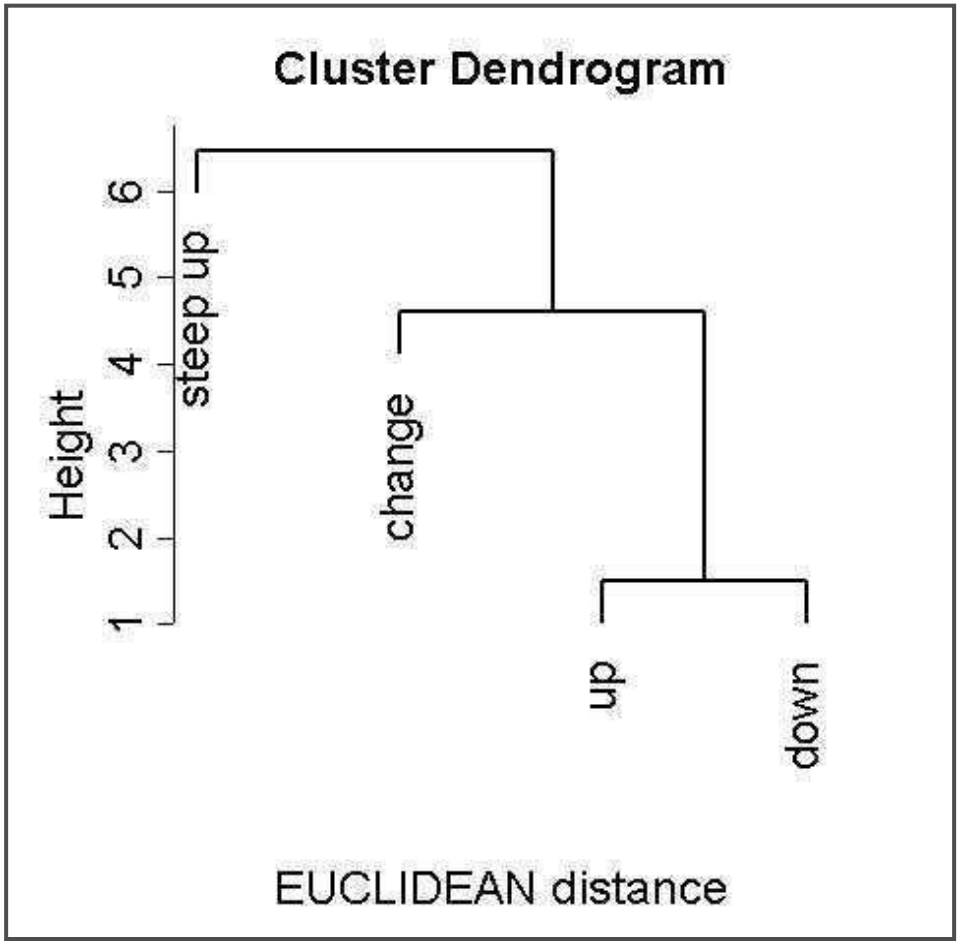
The blue cut generates 4 clusters (G1,G6), (G5), (G2,G3), (G4).

We have a hierarchy of clusterings.

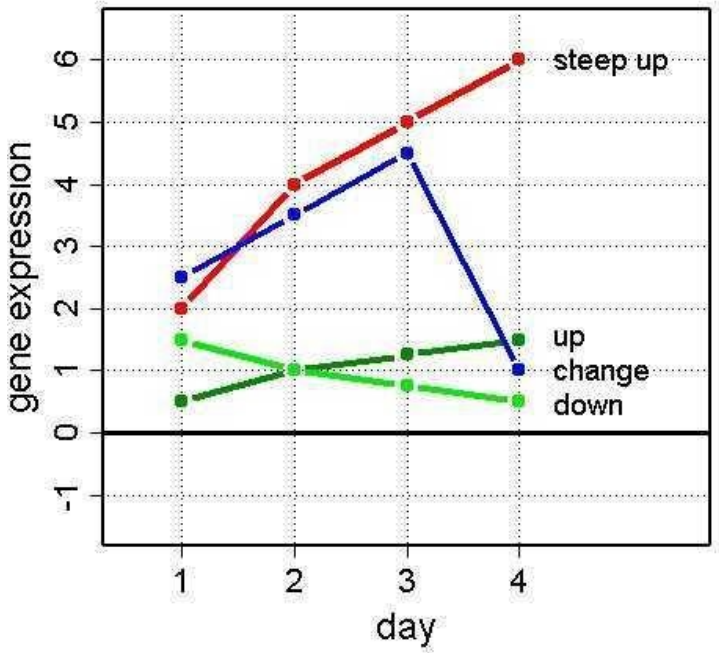


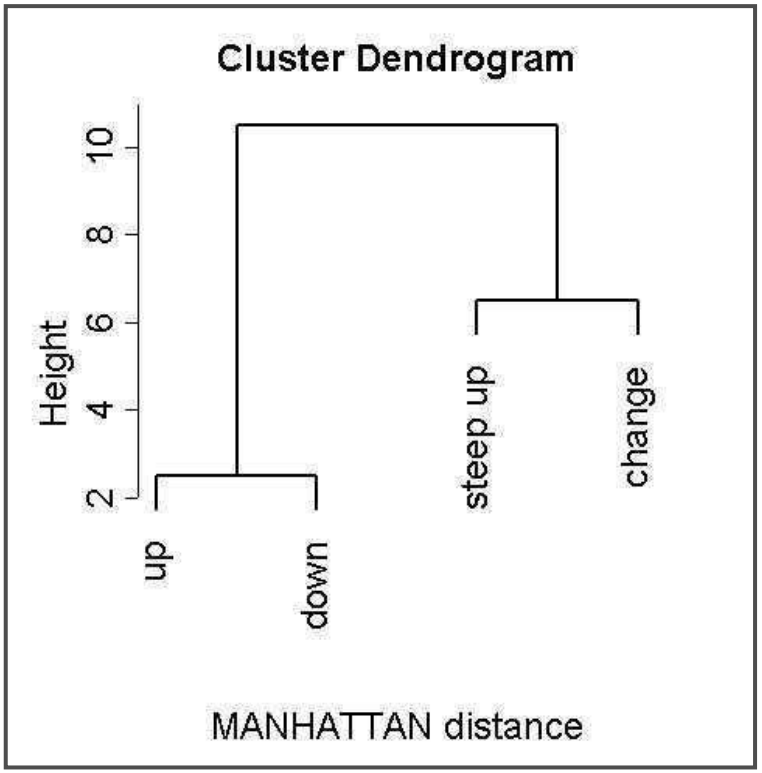
The Euclidean distance generates a dendrogram that clusters "up" and "down" together



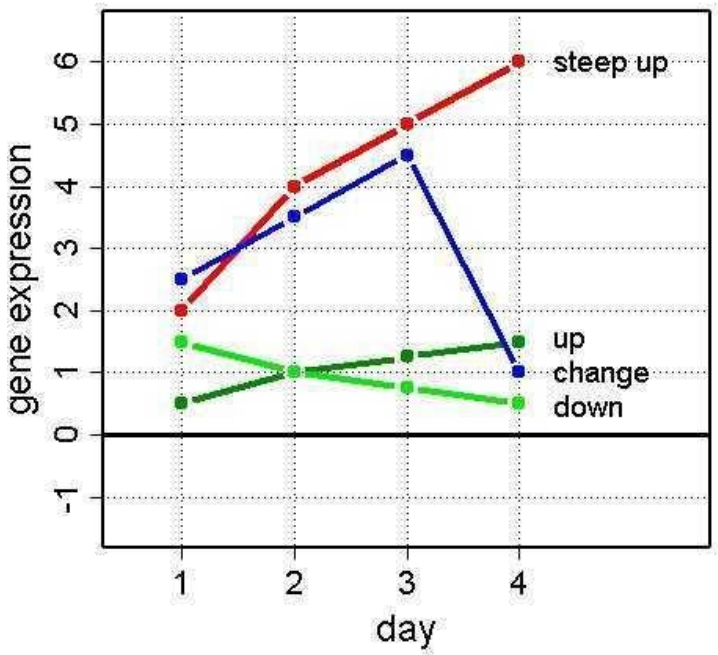


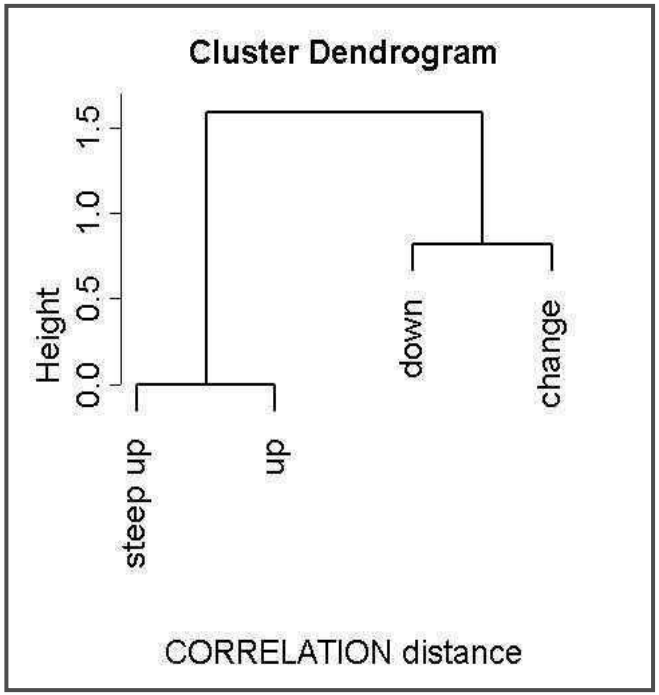
The Manhattan distance sees two close clusters



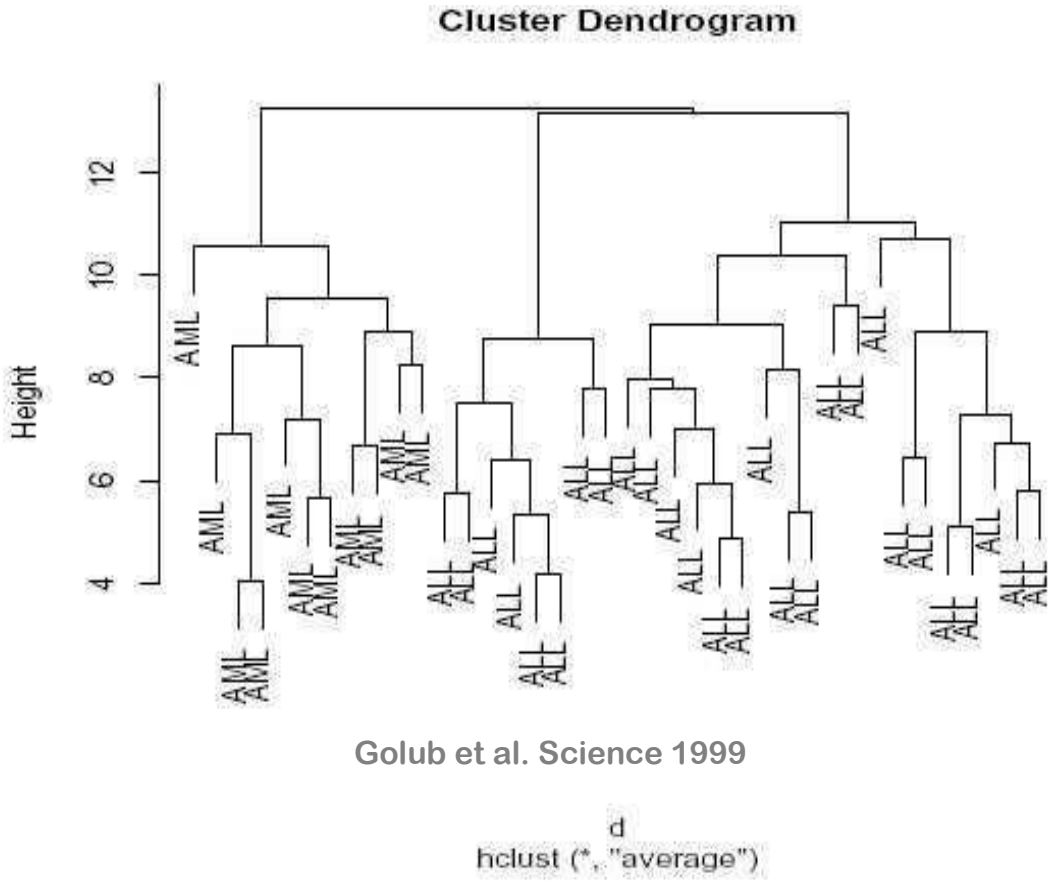


The correlation distance forms different clusters





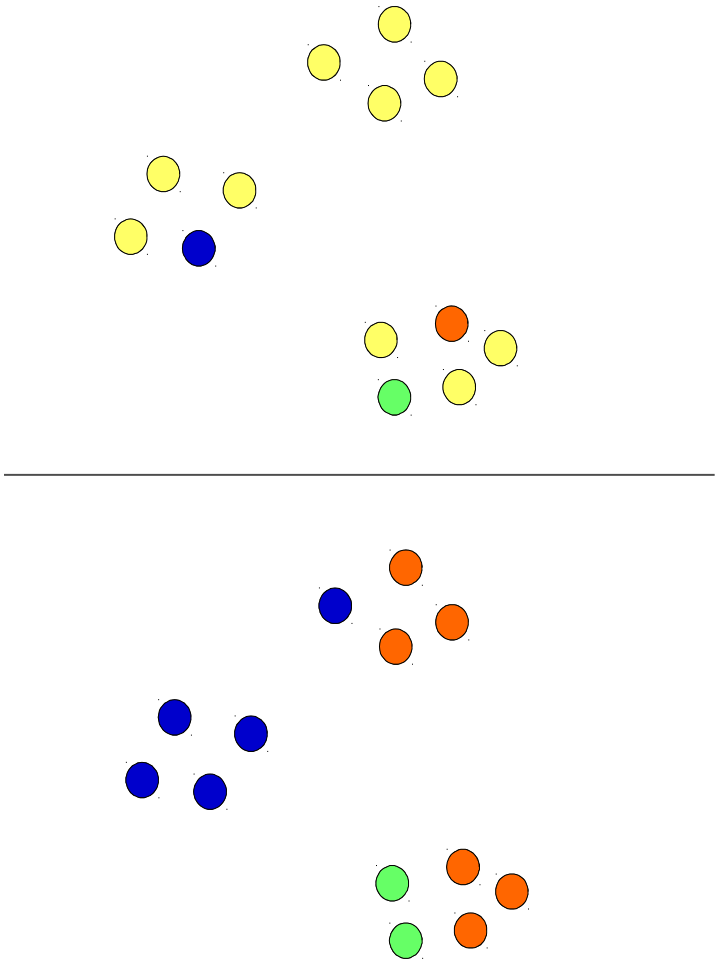
Clustering groups different types of childhood leukemia correctly



Golub et al. Science 1999

… but this is actually no clustering problem since the two groups were known a priori.

K-Means clustering generates clusters in an iterative procedure

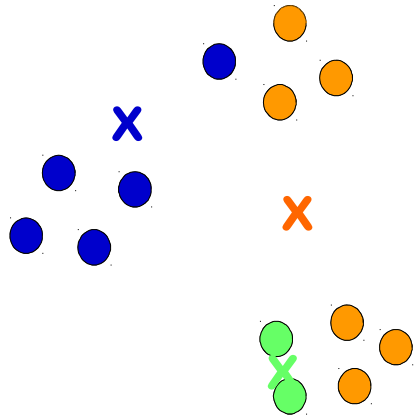


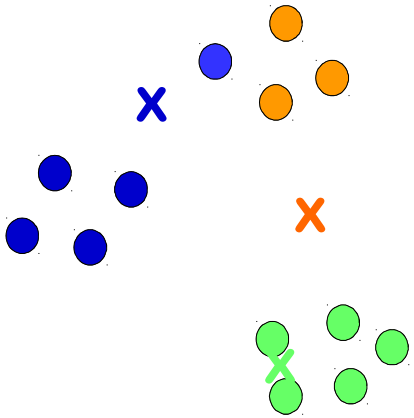
The number of clusters K is set upfront.

Pick K points at random. These are the centroids of the first iteration.

Assign each point to the cluster of the nearest centroid.

Centroids and clusters are updated iteratively





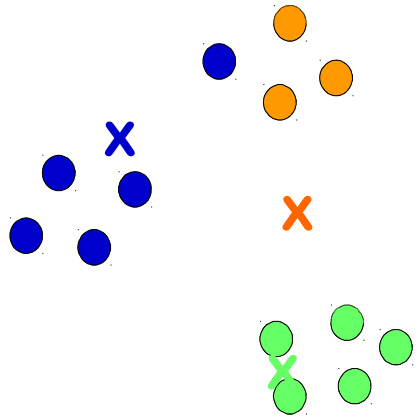
Calculate the centroid of the newly generated clusters.

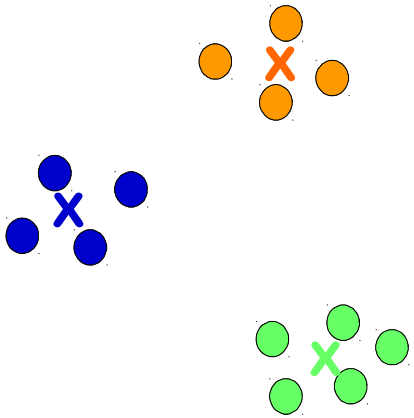
Assign points to nearest updated centroids.

Iterate until clusters do not change anymore.

The coordinates of a centroid are the averages of the corresponding coordinates of points in the cluster.

Centroids and clusters are updated iteratively





Calculate the centroid of the newly generated clusters.

Assign points to nearest updated centroids.

Iterate until clusters do not change anymore.

The coordinates of a centroid are the averages of the corresponding coordinates of points in the cluster.

K-means clustering can alternatively be described as an optimization problem

Assign points to clusters such that the following objective function is optimized:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K} \sum_{C(i)=C(j)=k} d_E(x_i, x_j)^2$$

The iterative algorithm is a heuristic to approximate the optimum of the objective function.

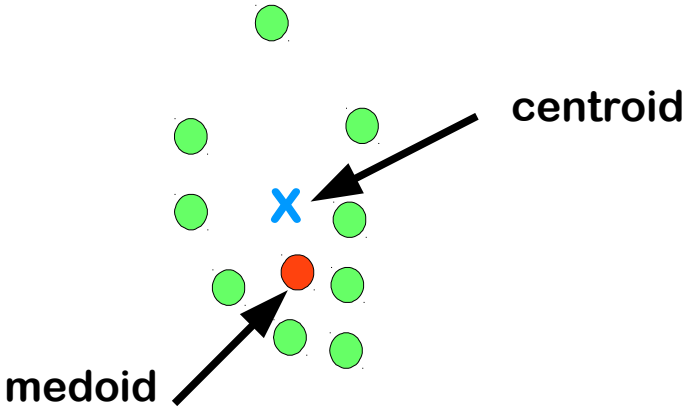
Note that the clustering is only „optimal" relative to this objective function. Different objective functions will give different "optimal" clusterings.

Partitioning around Medoids (PAM) is a clustering algorithm that uses a different objective function

Clusters are prototypical data points (medoids) instead of centroids.

The goal is to minimize the distance to the nearest medoid simultaneously for all data points.

This can be achieved by minimizing the objective function:



$$f(m_1, \ldots, m_k) = \sum_{i=1}^{n} \min_{j=1,\ldots,k} d_M(x_i, m_j)$$

The PAM algorithm also operates in iterations

Initialization: randomly choose K prototypes (medoids).

Iterate until convergence:

Swapping:

For all pairs of points (i,j) where i is a medoid and j is not:

Calculate the value of the objective function obtained by making j a medoid instead of i.

Swap medoids if the objective function improves.

The silhouette score measures whether a single data point is well clustered

For a given clustering we can calculate the silhouette s(i) of data point i:

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

b(i) := average distance of i to all points in the same cluster

a(i) := $\min_C$ d(i,C),

where d(i,C) is the average distance of i to all points in cluster C.

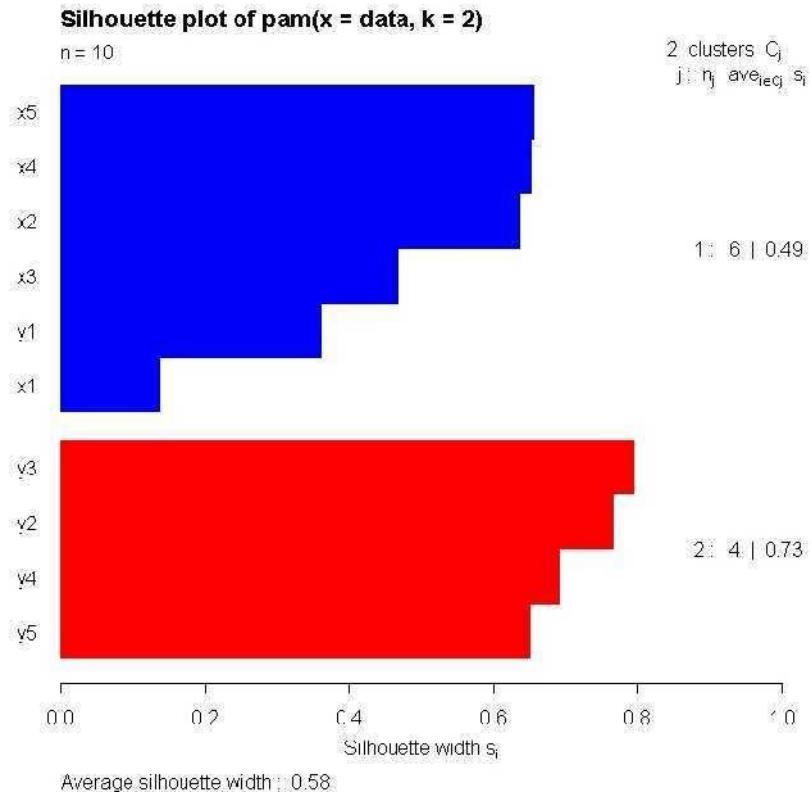
The minimum is taken across all clusters that do not contain i.

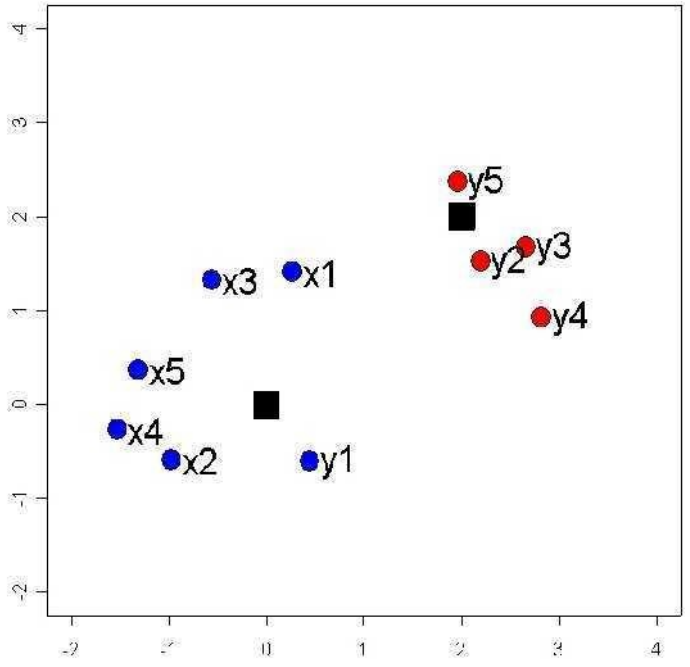
s(i) close to 1 : profile is in the „correct“ cluster

s(i) close to -1: profile is in the „wrong“ cluster

s(i) close to 0: profile can not make up its mind

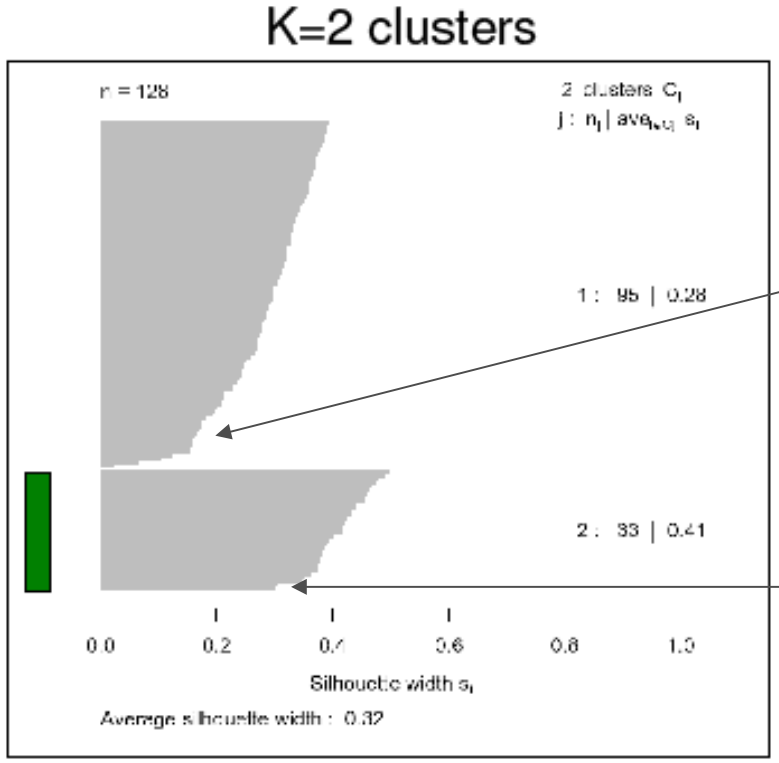
The scores combine to a silhouette that helps judging the overall quality of the clustering





The points x1 and y1 are somewhat between clusters

Are there three types of leukemia in the Chiaretti data set?



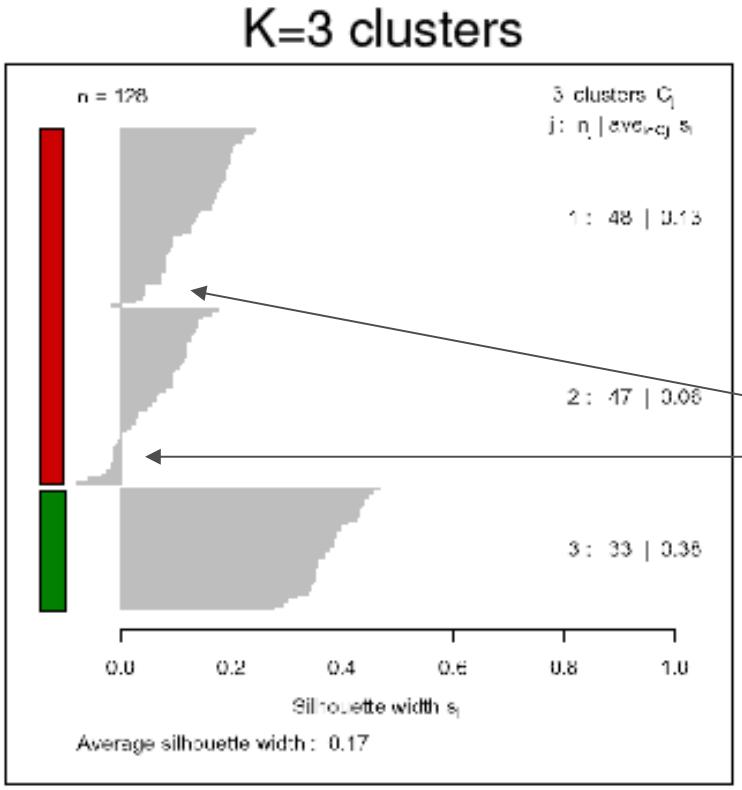
This cluster has leukemias with small silhouette scores.

All leukemias assigned to this cluster have high silhouette scores.

T-cell Leukemia

Chiaretti et al., 2004

The silhouette scores become smaller if we cluster into 3 groups



The clustering split the old cluster 1 into two parts. This has made the silhouettes worse. The data set contains only two types of leukemias.

T-cell Leukemia

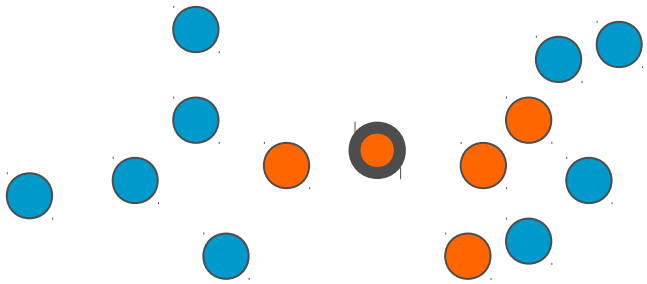
Chiaretti et al., 2004

The silhouette can be used to determine the number of clusters in a data set.

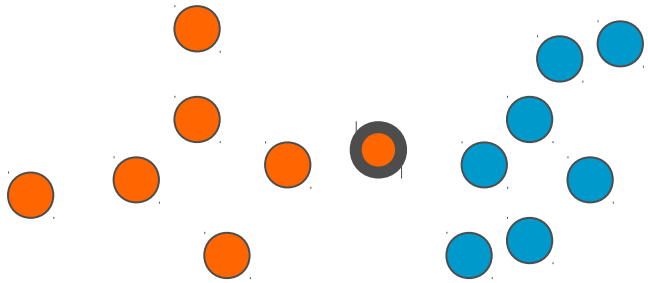
Only use clustering to solve clustering problems

You want to find all profiles that are similar to a reference profile.

Bad strategy: Cluster all profiles and chose those that are in the same cluster as the reference profile.



What you want

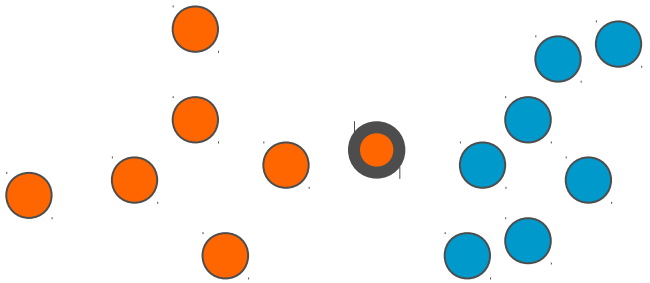


What you get

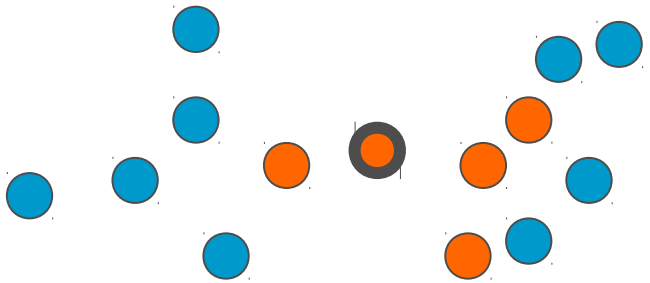
This is a screening problem and not a clustering problem

You want to find all profiles that are similar to a reference profile.

Good strategy: You calculate the distance of all profiles to the reference profile and pick the one that are closest.

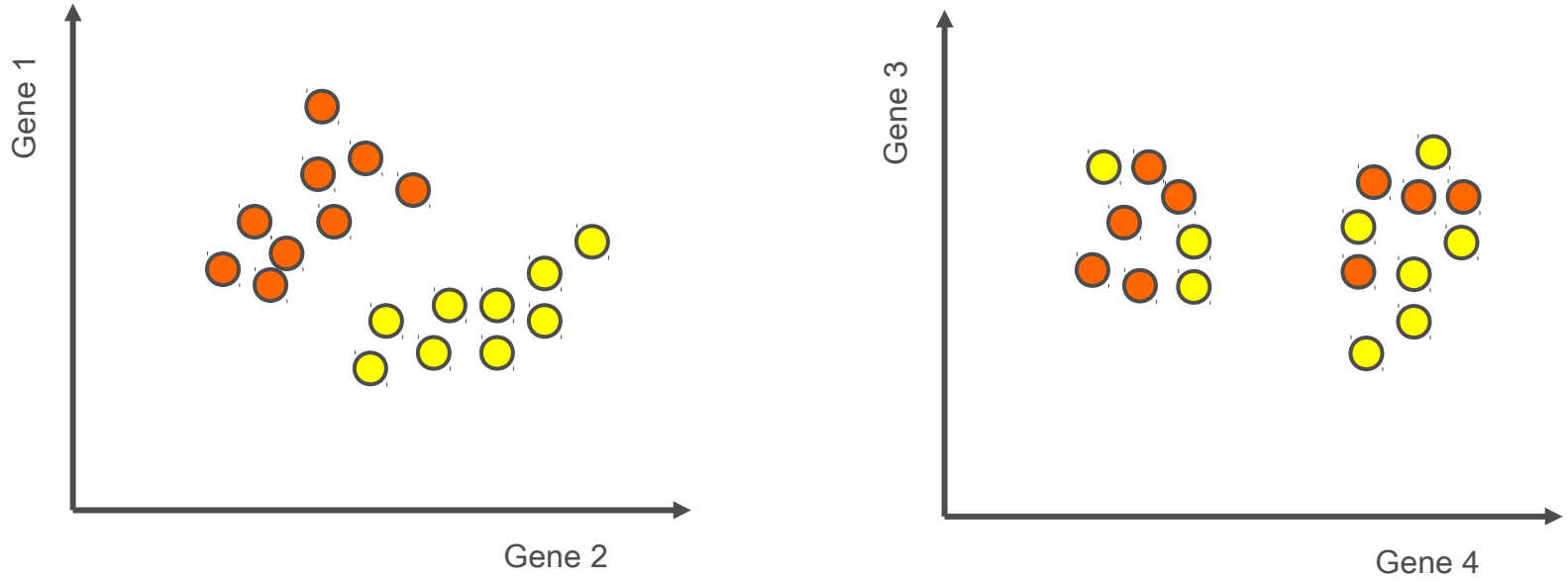


Clustering



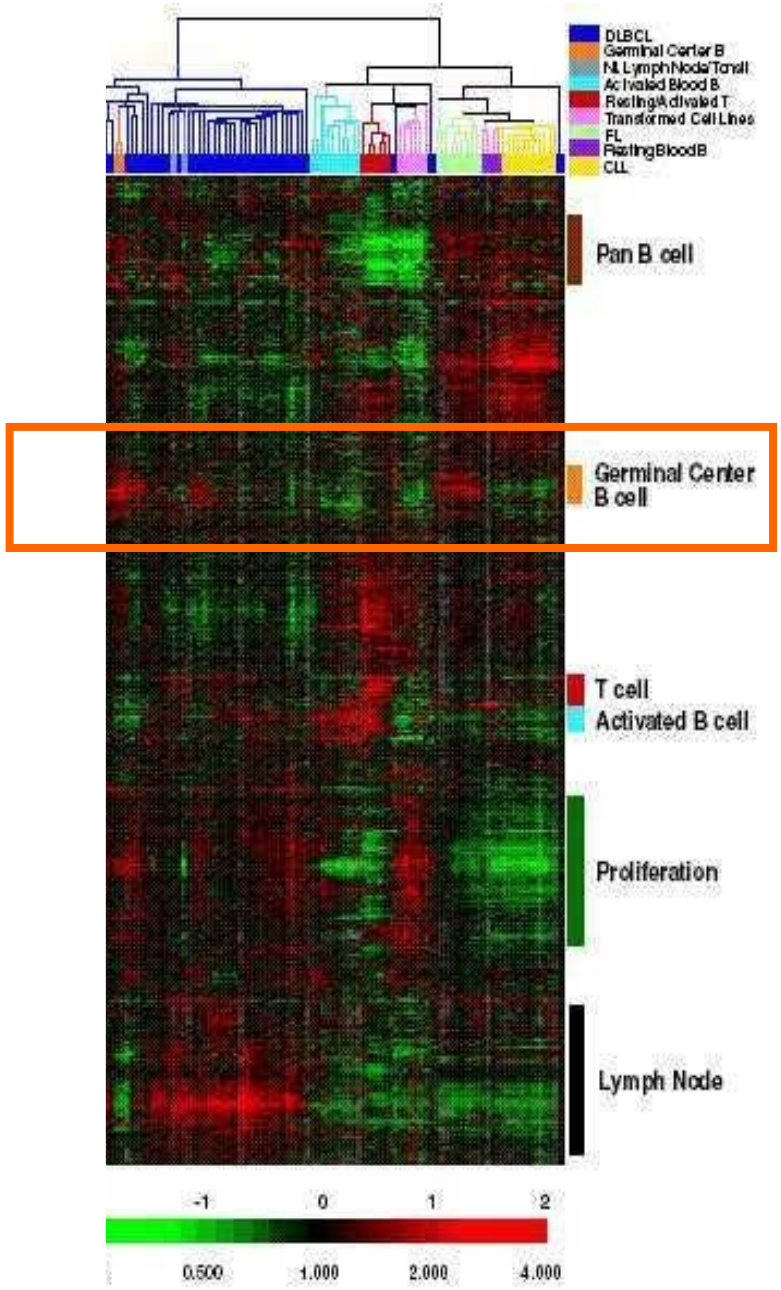
Screening

Different gene sets generate different clusterings



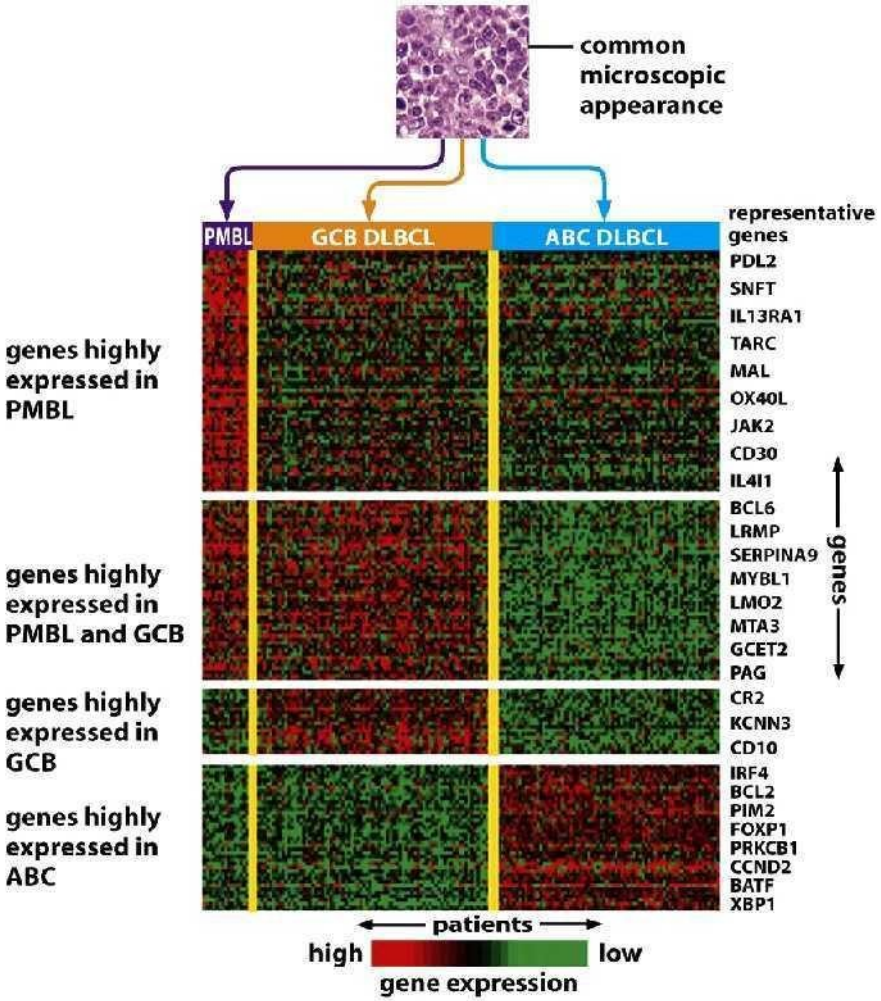
What would a clustering of tumors look like if we used only genes from the Y chromosome?

To get a meaningful clustering of tumors we need to choose genes that reflect aspects of tumor biology.



Alizadeh et al Nature 2000

Clustering lymphoma expression profiles with well chosen gene sets revealed different subtypes



The lymphoma subtypes have different prognosis

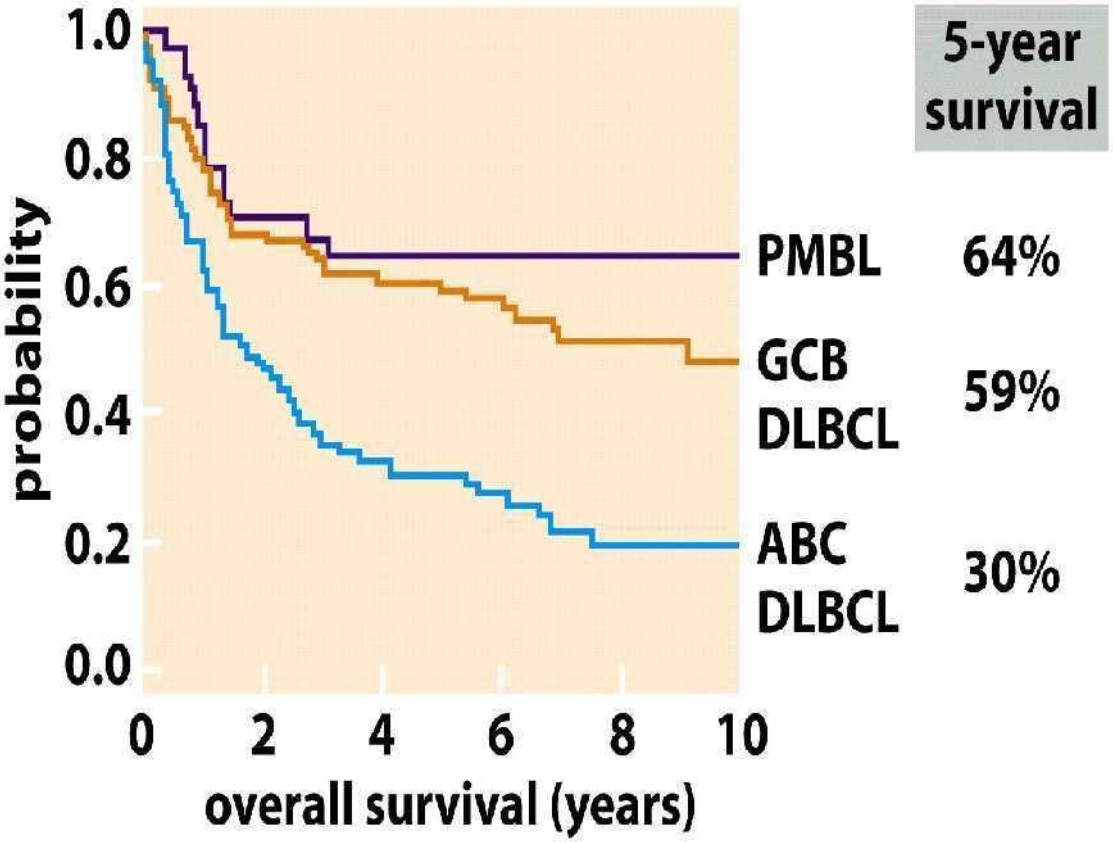


Figure 16.5b *The Biology of Cancer* (© Garland Science 2007)

The subtypes respond differently to a targeted treatment

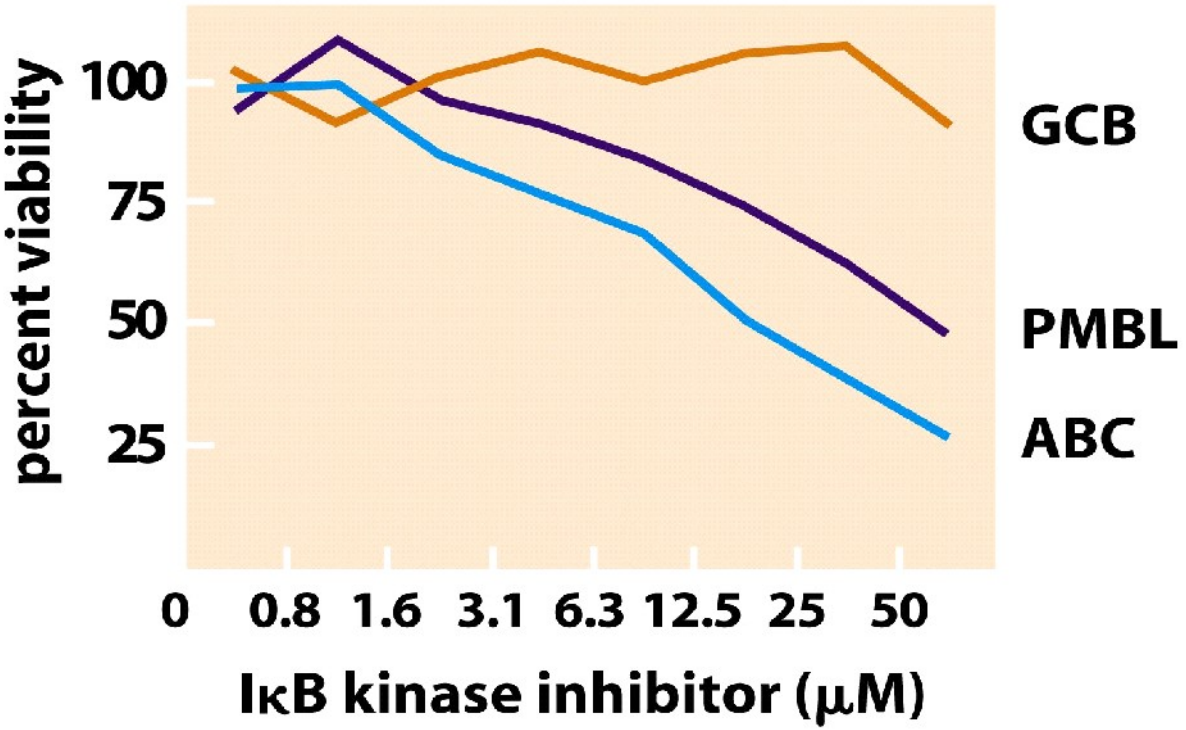


Figure 16.5d *The Biology of Cancer* (© Garland Science 2007)

Acknowledgment

Concepts, slides, and images were borrowed from:

Jörg Rahnenführer

Tobias Müller

Anja v. Heydebreck