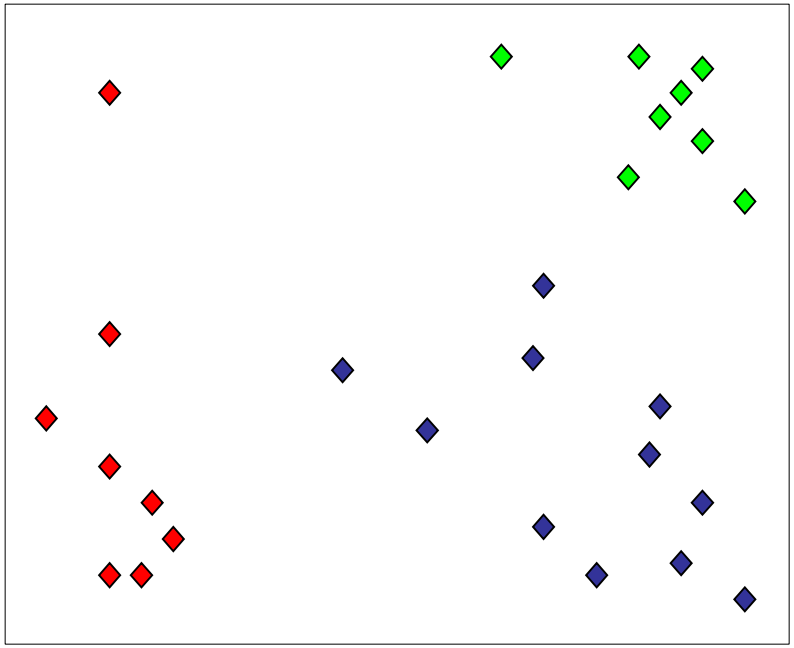
10-810 /02-710
Computational Genomics

Clustering expression data

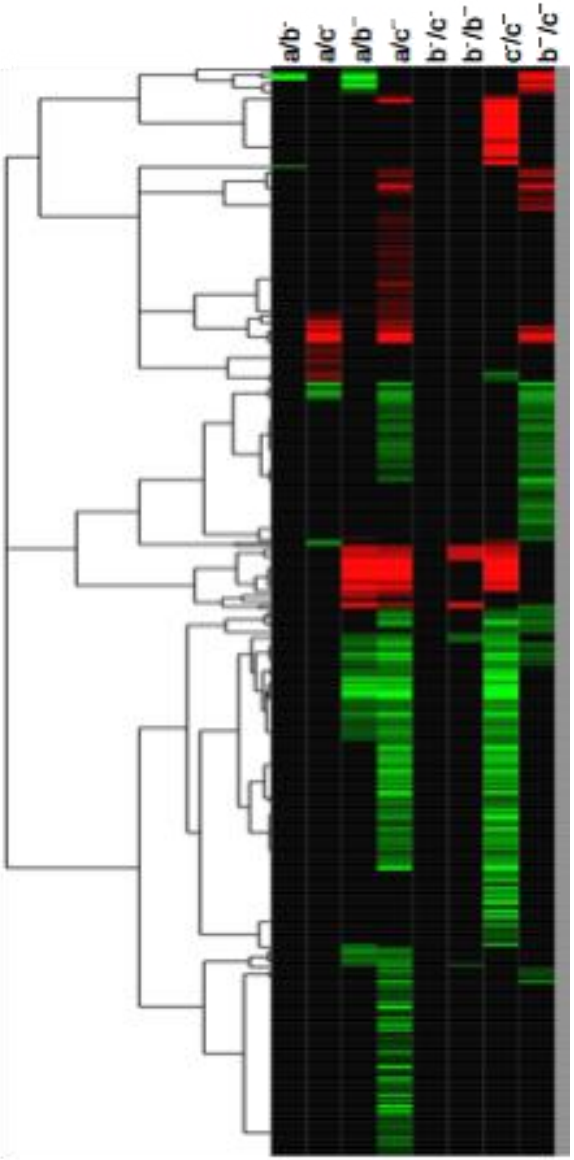
What is Clustering?

- Organizing data into *clusters* such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects.



Clustering gene expression data

- Organizing data into *clusters* such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects.



Goal

- Data organization (for further study)
- Functional assignment
- Determine different patterns

Genes

- Classification
- Relations between experimental conditions

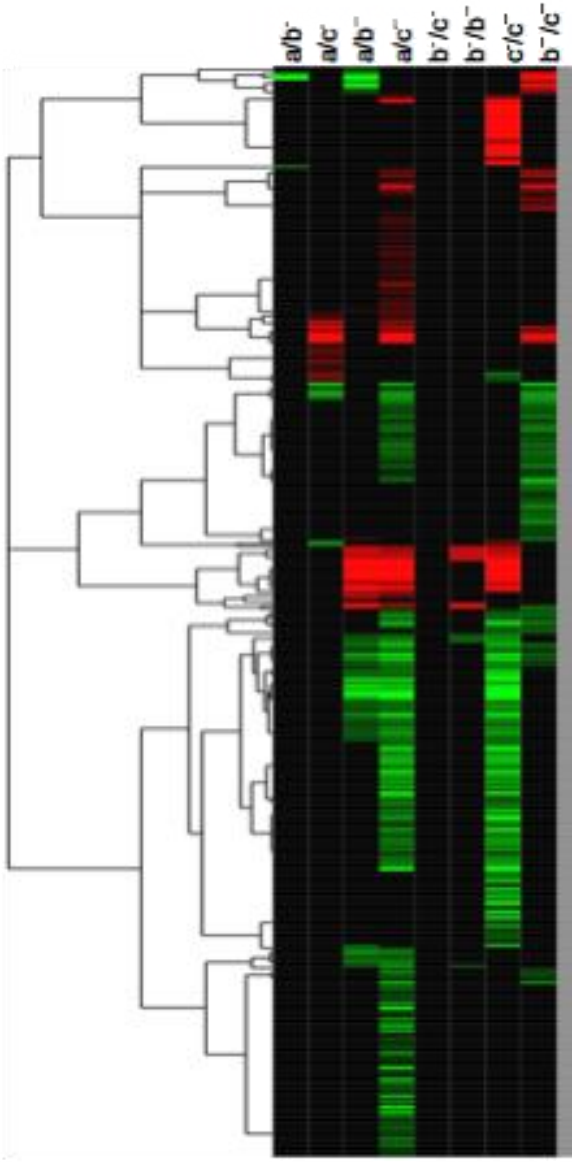
Experiments

- Subsets of genes related to subset of experiments

Both

Example: clustering genes

- Clustering genes helps determine new functions for unknown genes
- An early “killer application” in this area
  - The most cited paper in PNAS! (13,844, Google Scholar)



What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

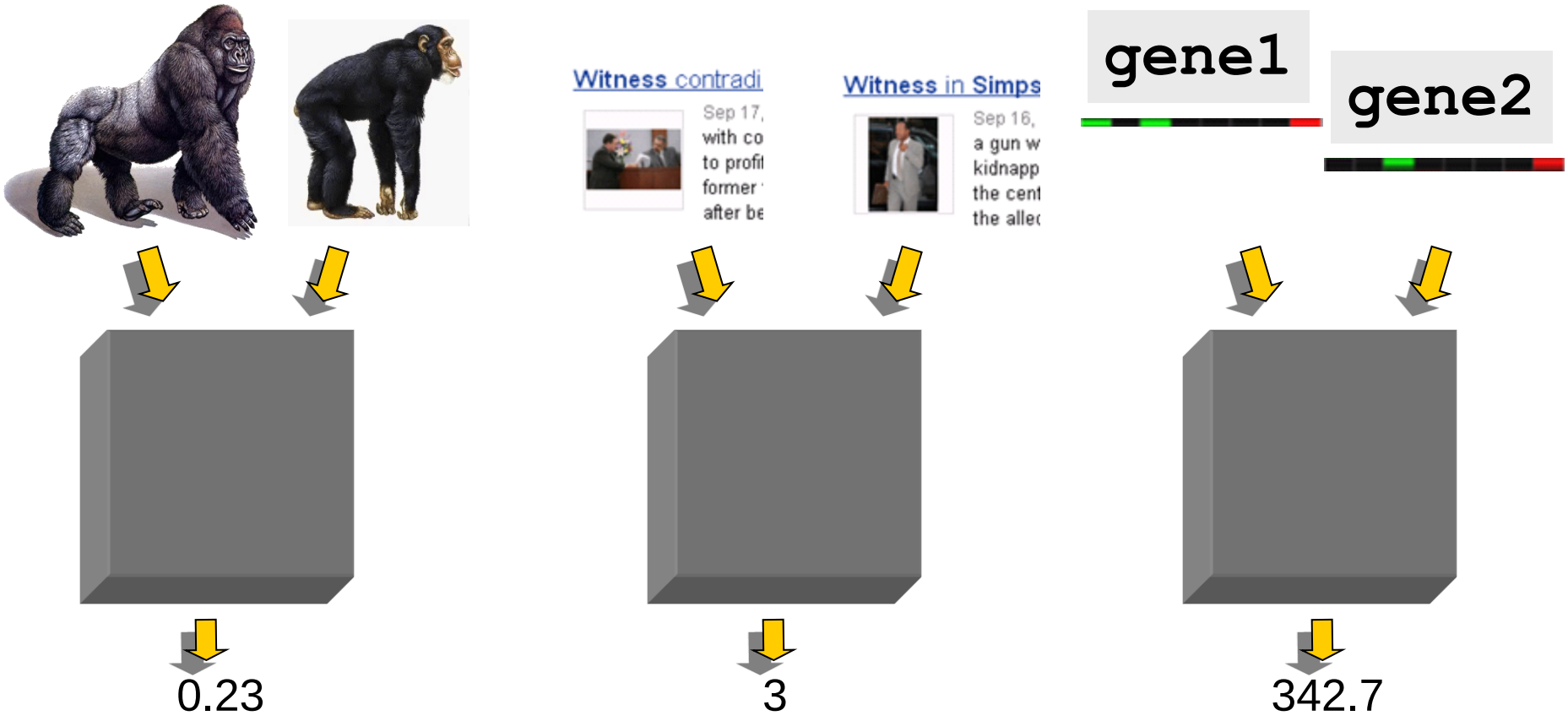
Webster's Dictionary

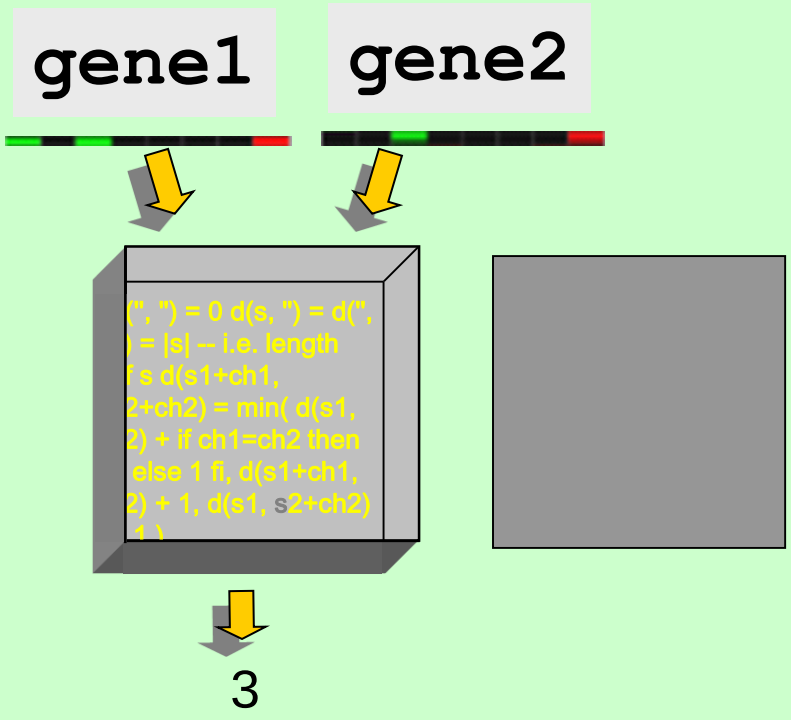
Similarity is hard to define, but…
"*We know it when we see it*"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Defining Distance Measures

Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O_1 and O_2 is a real number denoted by $D(O_1,O_2)$





Inside these black boxes: some function on two variables (might be simple or very complex)

A few examples:

- Euclidian distance

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Correlation coefficient

$$s(x, y) = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}$$

- Similarity rather than distance
- Can determine similar trends

Clustering algorithms

We can divide clustering methods into roughly three types:

1. Hierarchical agglomerative clustering
   - eg. Ward hierarchical clustering
2. Model based
   - eg. k-means, k-medioids, Gaussian mixtures, dimensionality reduction (PCA [eigengene*], ICA, ..)
3. Iterative partitioning (top down)
   - eg. graph based algorithms (spectral), local-linear embedding (LLE), isomap,..

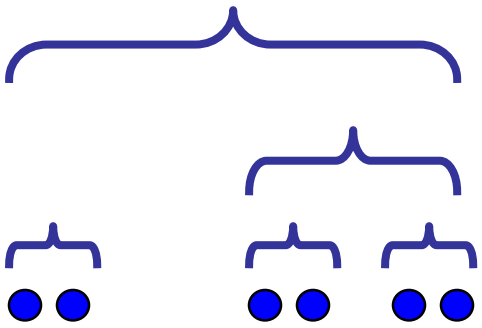
Proc Natl Acad Sci U S A. 2000 Aug 29;97(18):10101-6.

Singular value decomposition for genome-wide expression data processing and modeling. Alter O, Brown PO, Botstein D.

Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14863-8.
Cluster analysis and display of genome-wide expression patterns.

Hierarchical clustering

- Probably the most popular clustering algorithm in this area
- First presented in this context by Eisen in 1998



dendrogram

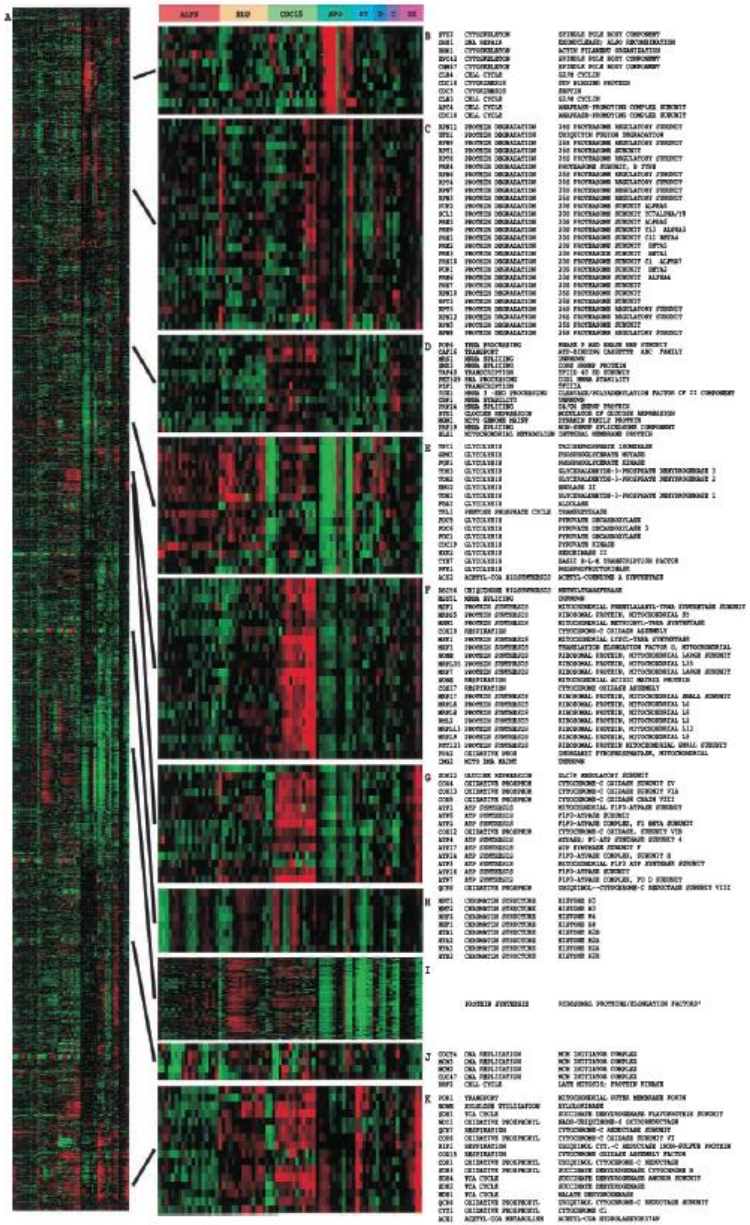
- Agglomerative **(**bottom-up)
- Algorithm:
  1. Initialize: each item a cluster
  2. Iterate:
     - select two most *similar* clusters
     - merge them
  3. Halt: when there is only one cluster left

Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14863-8.

Cluster analysis and display of genome-wide expression patterns.

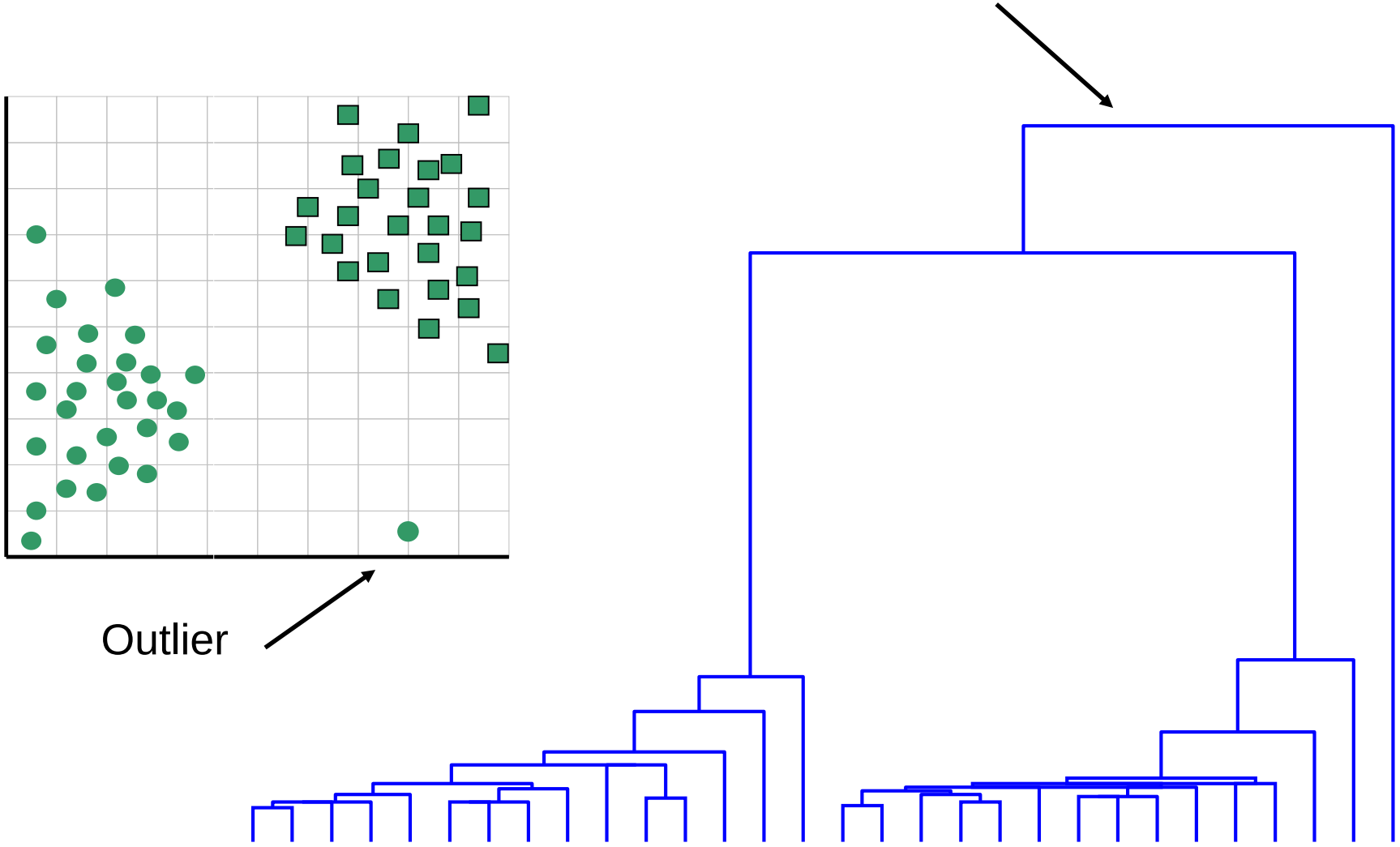
Cluster results

Combining several time series yeast datasets



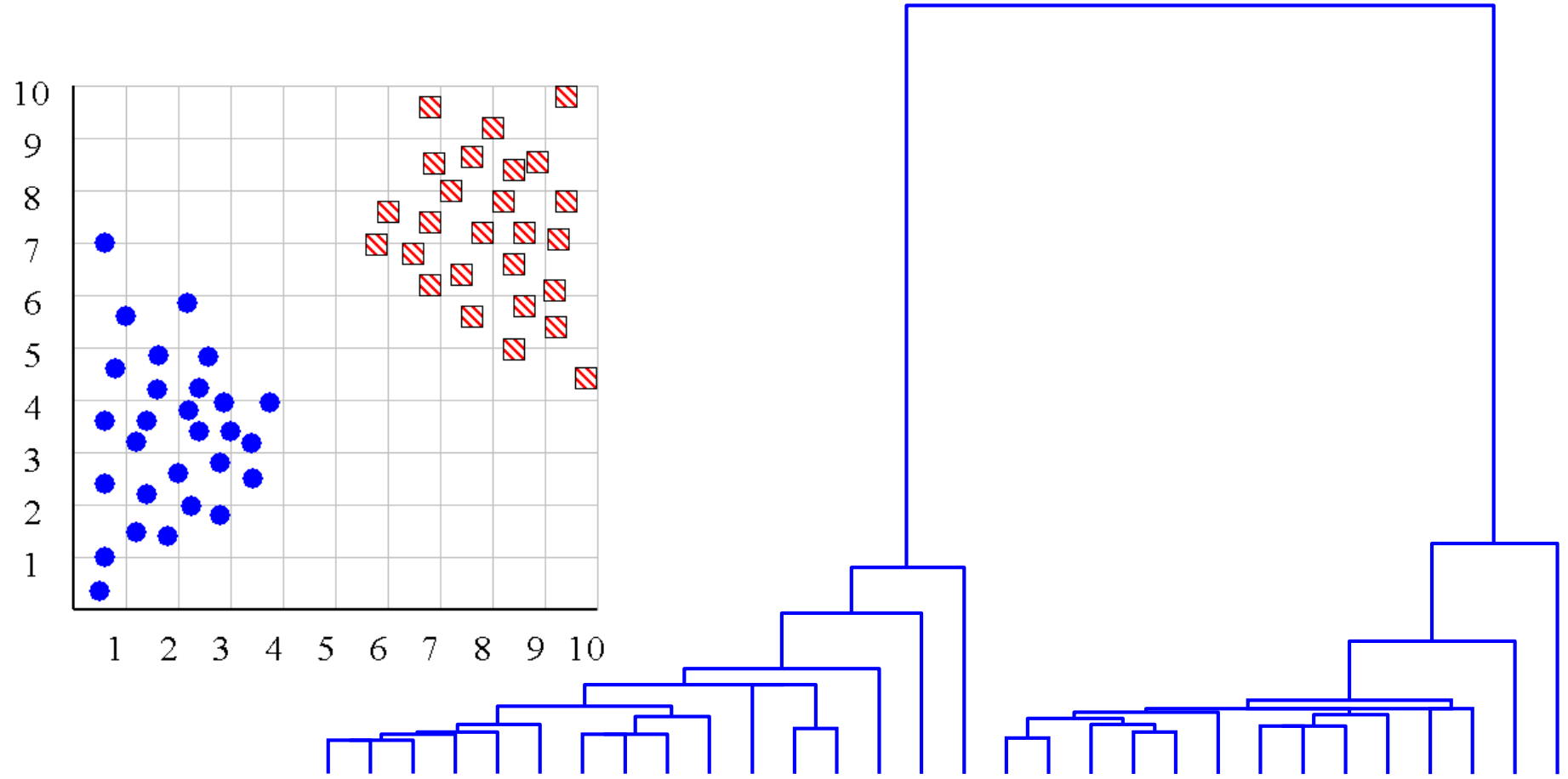
One potential use of a dendrogram is to detect outliers

The single isolated branch is suggestive of a data point that is very different to all others



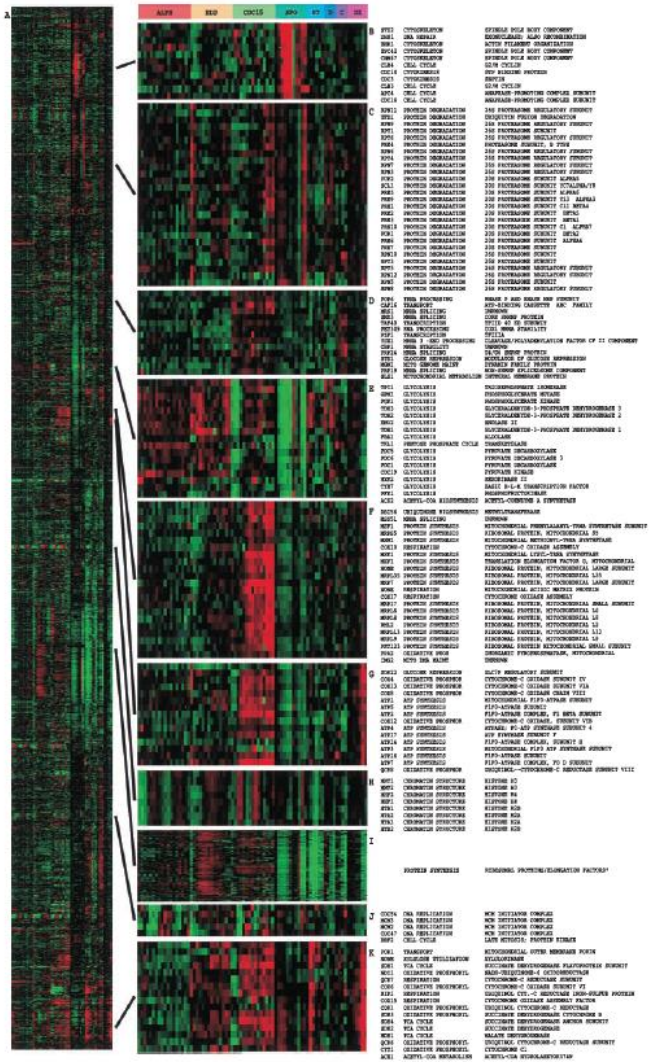
But what are the clusters?

In some cases we can determine the “correct” number of clusters. However, things are rarely this clear cut, unfortunately.



Hierarchical clustering

- As we mentioned, its one of the most popular methods for clustering gene expression data
- One of its main advantages is the global overview of the entire experiment in one figure.
- Biologists often omit the tree and use the figure to determine functional assignments



Partitional Clustering

- Nonhierarchical, each instance is placed in exactly one of K non-overlapping clusters.
- Since the output is only one set of clusters the user has to specify the desired number of clusters K.

Algorithm *k-means*

1. Decide on a value for *K*, the number of clusters.

2. Initialize the *K* cluster centers (randomly, if necessary).

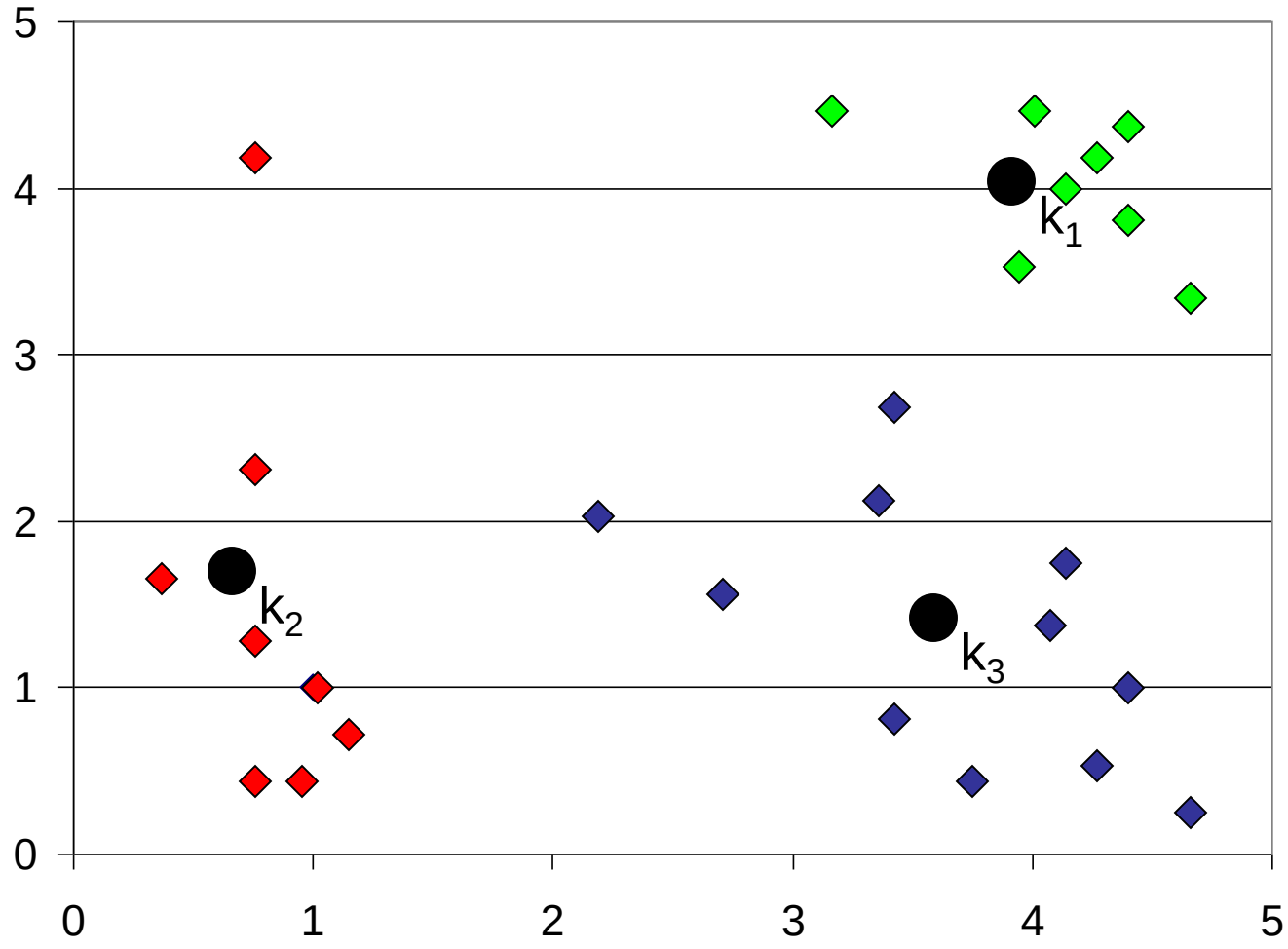
3. Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.

4. Re-estimate the *K* cluster centers, by assuming the memberships found above are correct.

5. Repeat 3 and 4 until none of the *N* objects changed membership in the last iteration.

K-means Clustering: Finished!

Re-assign and move centers, until …
no objects changed membership.



Gaussian Mixture Models

- Decide the number of clusters, K
- Initialize parameters (randomly)
- E-step: assign *probabilistic* membership to all input samples

One for each cluster

$$p_{i,j} = p(C = i \mid x_j) = \frac{p(x_j \mid C = i)\, p(C = i)}{\sum_k p(x_j \mid C = k)\, p(C = k)}$$

$$p_i = \sum_j p_{i,j}$$

- M-step: re-estimate parameters based on *probabilistic* membership

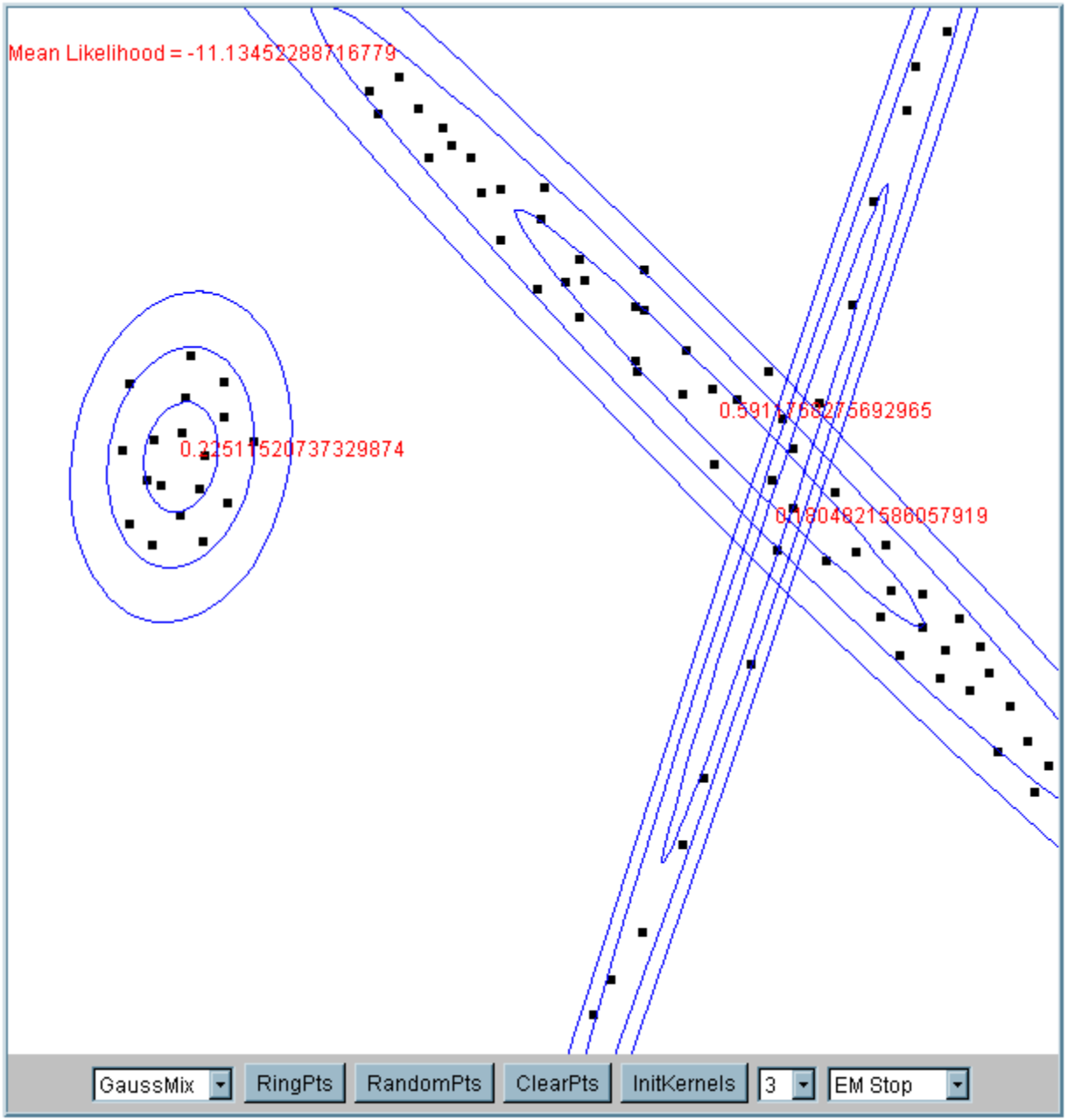
$$\mu_i \leftarrow \sum_j \frac{p_{i,j} \mathrm{x}_j}{p_i}$$

$$\Sigma_i \leftarrow \sum_j \frac{p_{i,j} \mathrm{x}_j \mathrm{x}_j^{\mathrm{T}}}{p_i}$$

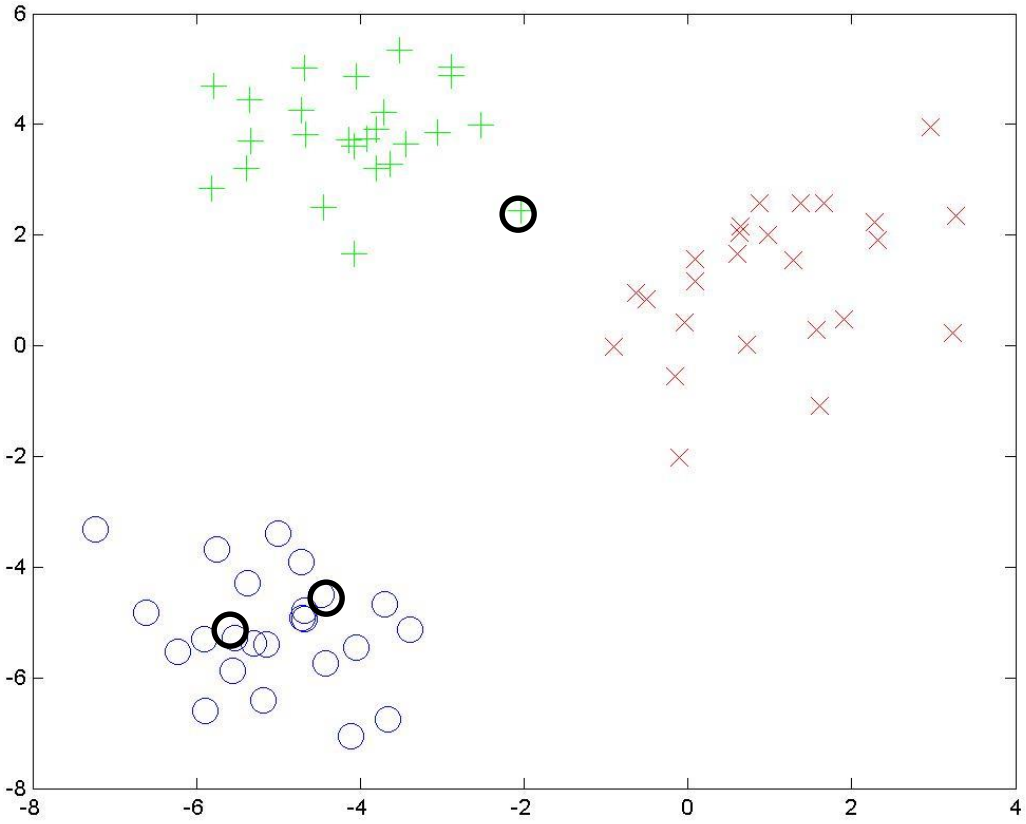
$$w_i = \frac{p_i}{\sum_j p_j}$$

- Repeat until change in parameters is smaller than a threshold

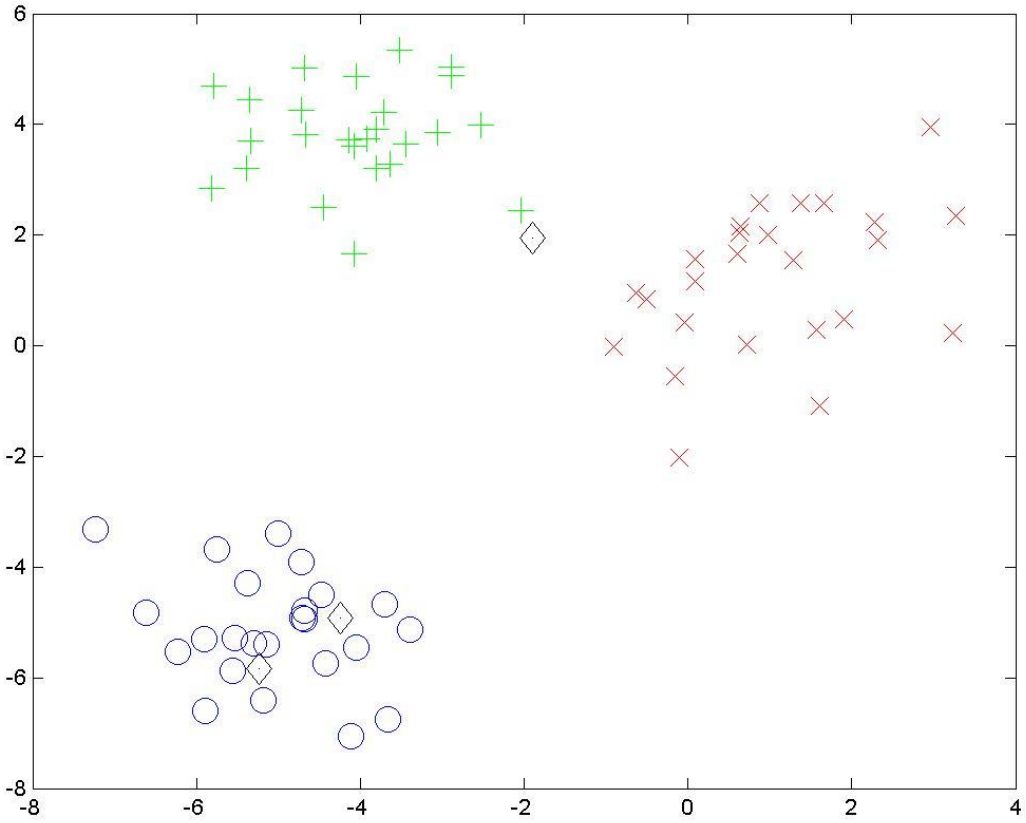
Iteration 25



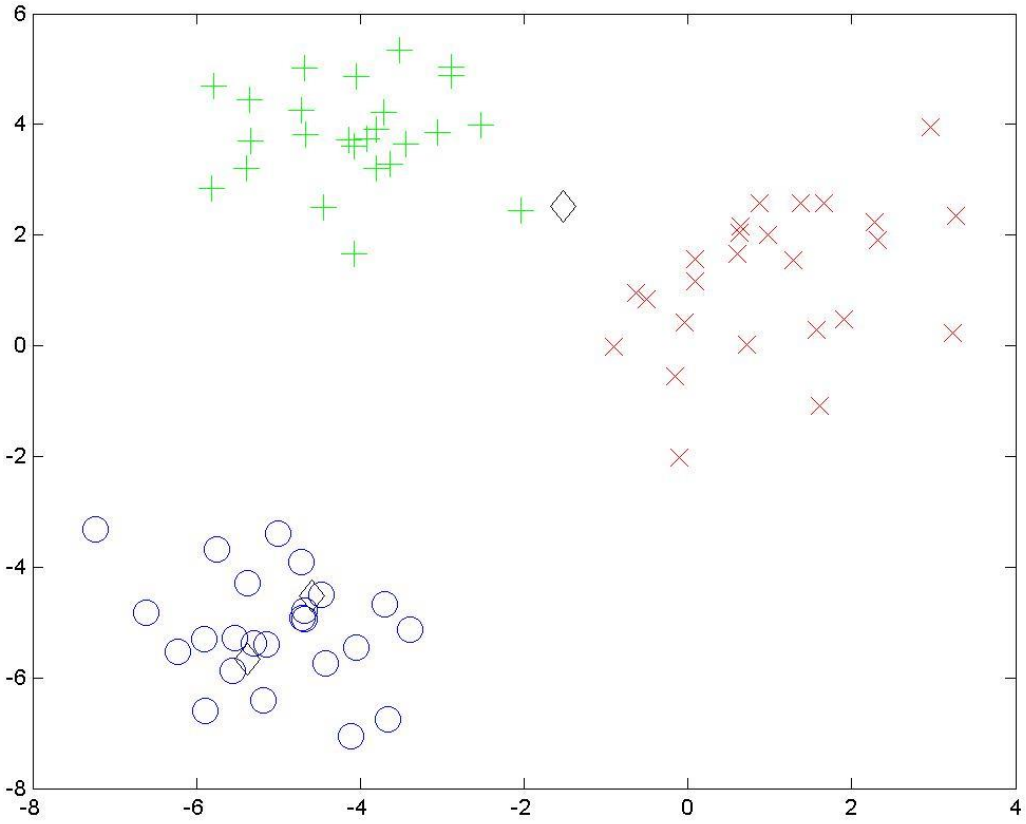
The importance of initializations



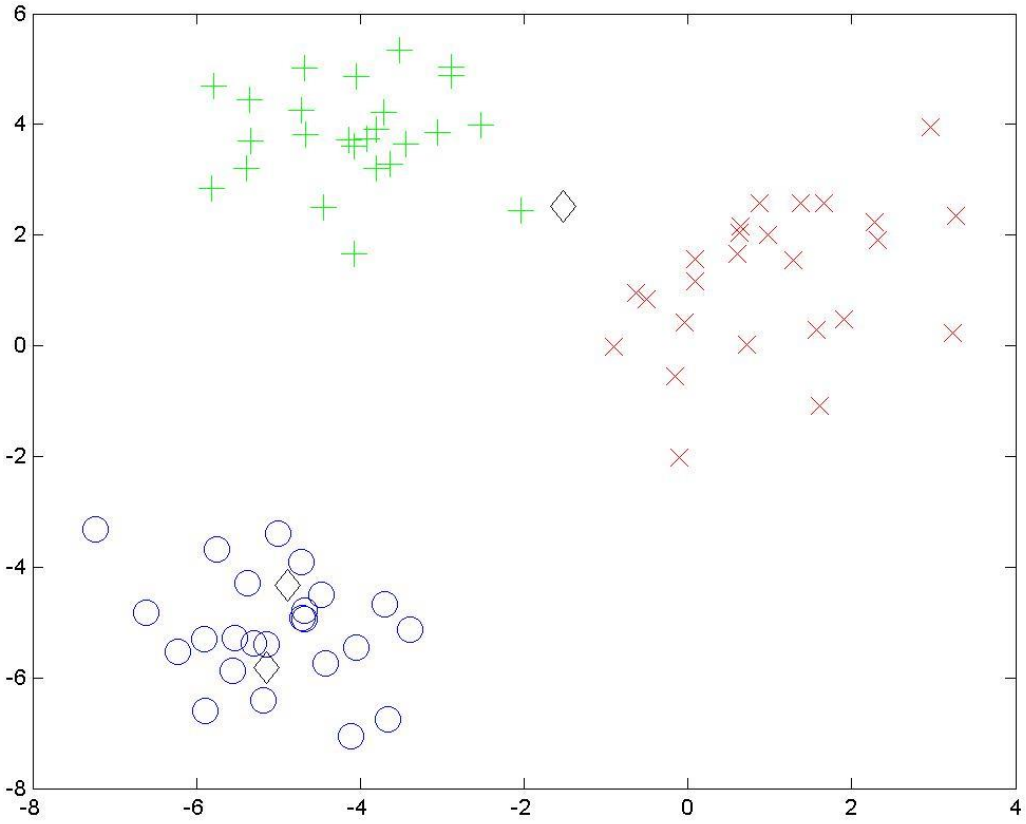
The importance of initializations: Step 1



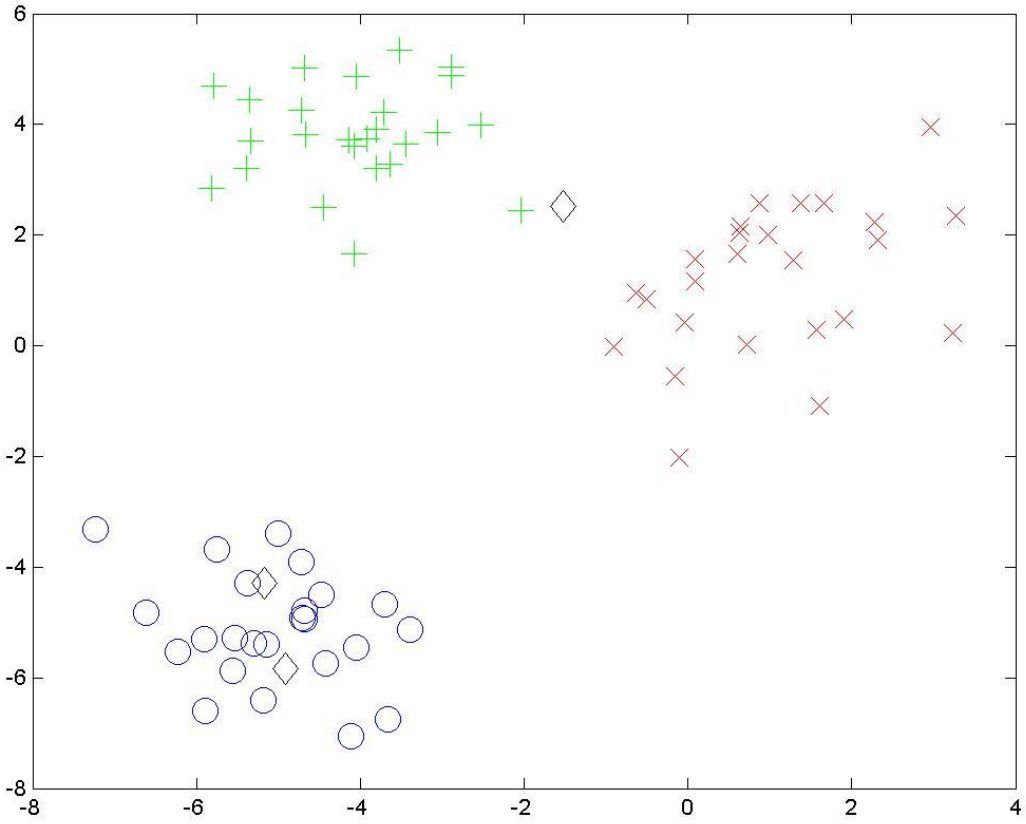
The importance of initializations: Step 2



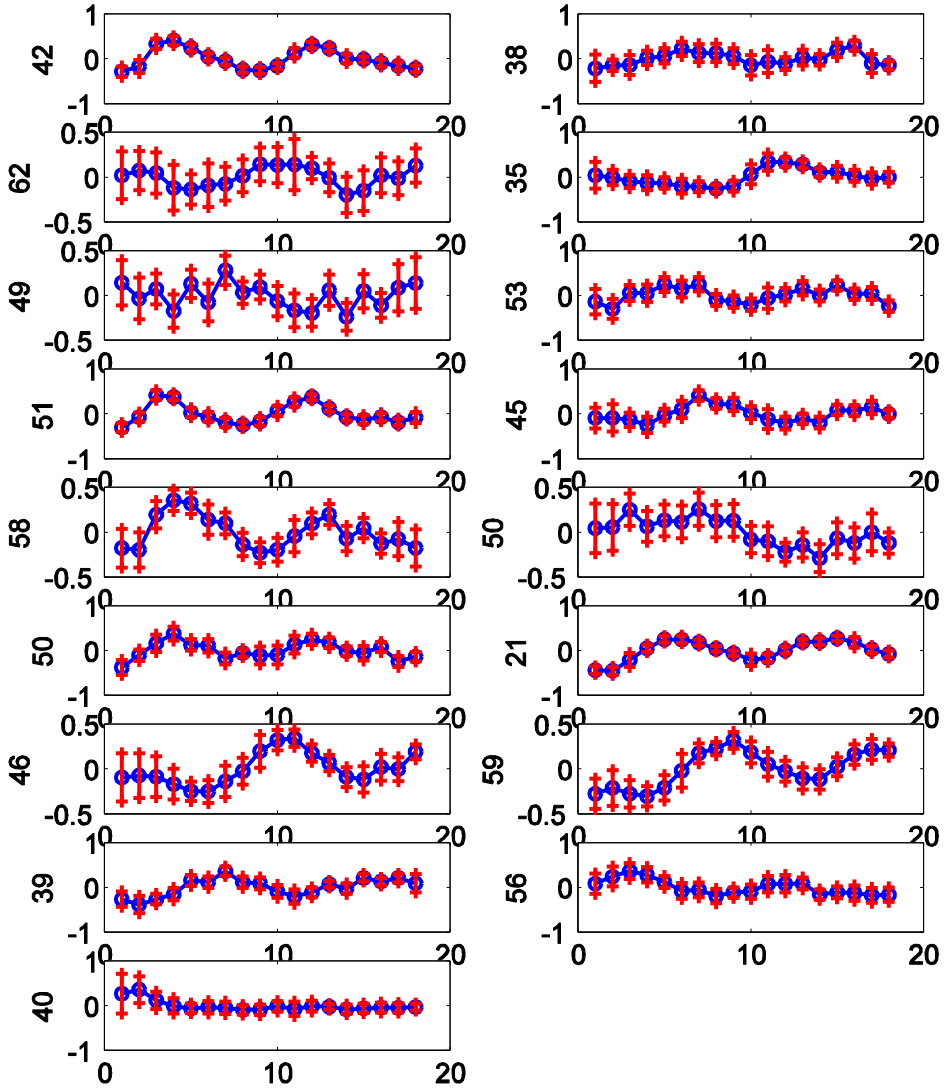
The importance of initializations: Step 5



The importance of initializations; Convergence

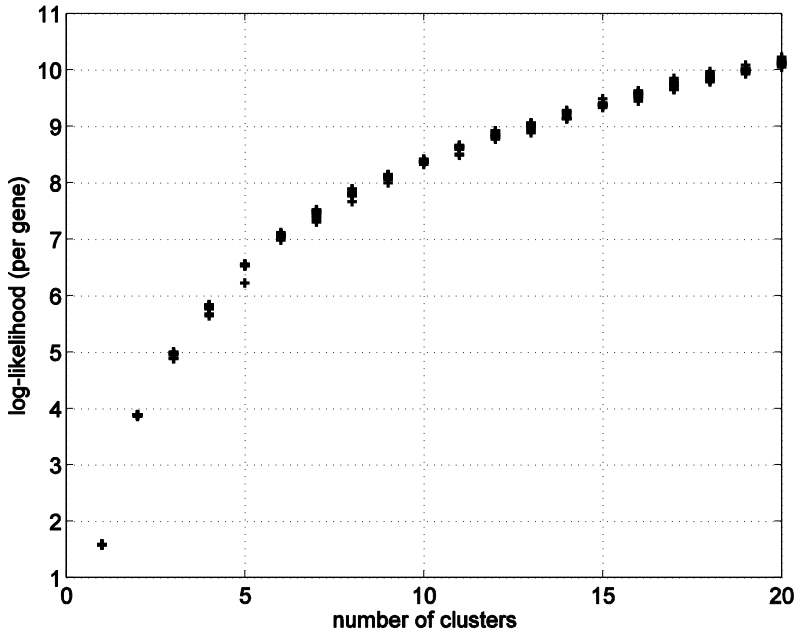


Example of clusters for the cell cycle expression dataset



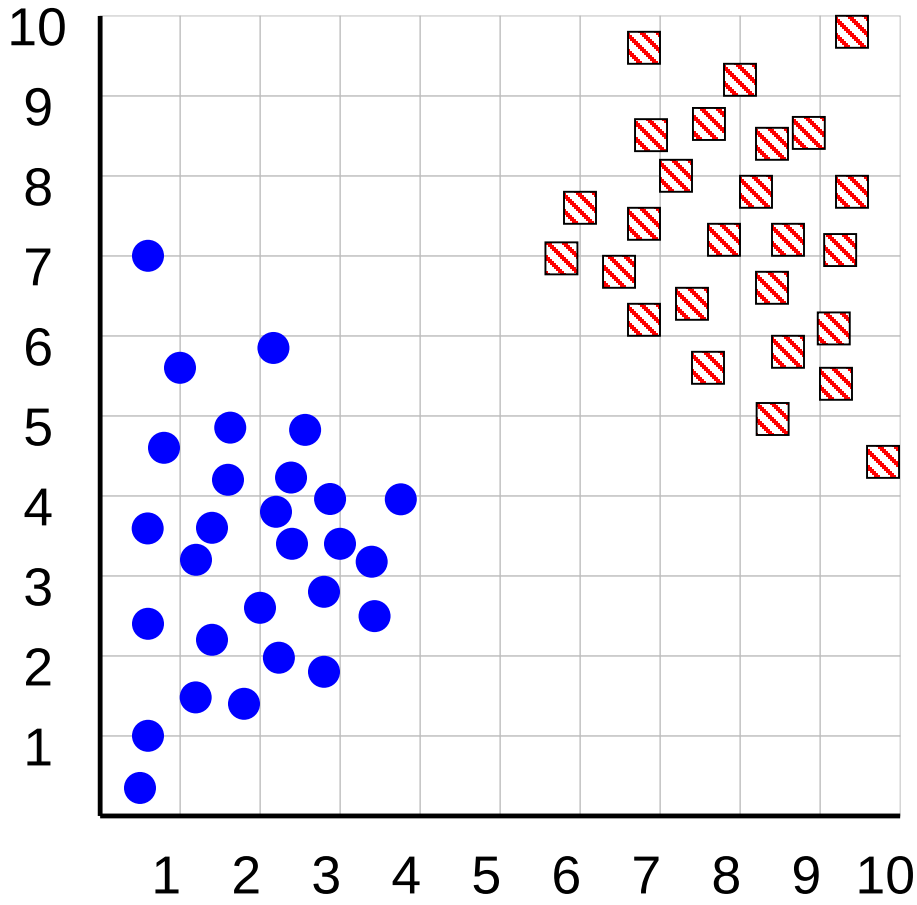
Number of clusters

- How do we find the right number of clusters?
- The overall log-likelihood of the profiles implied by the cluster models goes up as we add clusters
- One way is to use cross validation

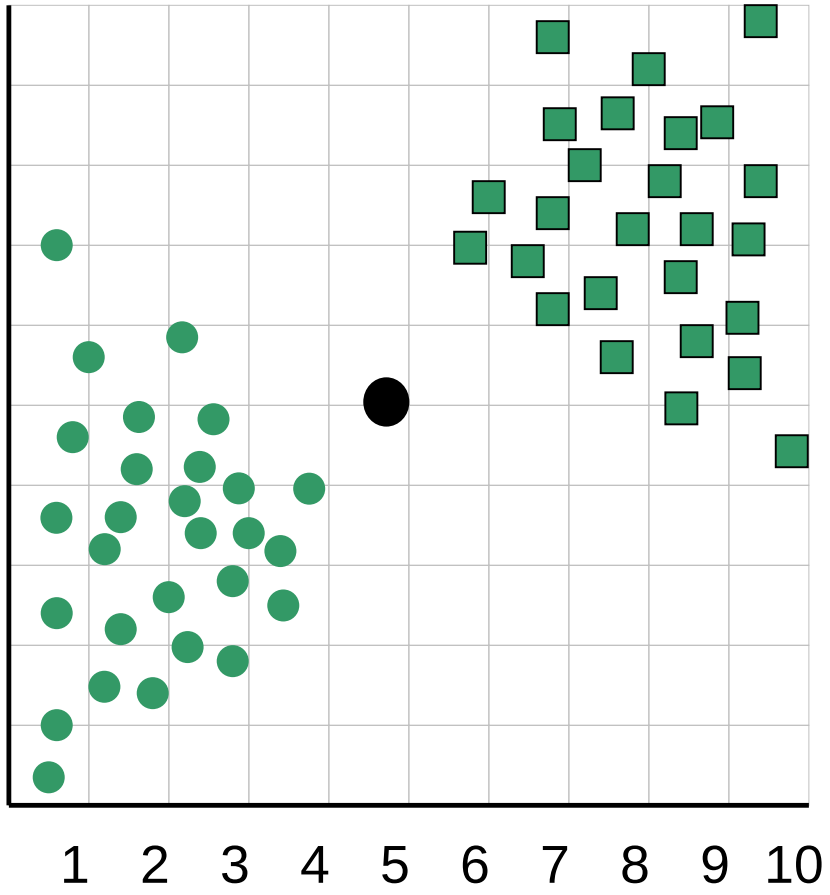


How can we tell the *right* number of clusters?

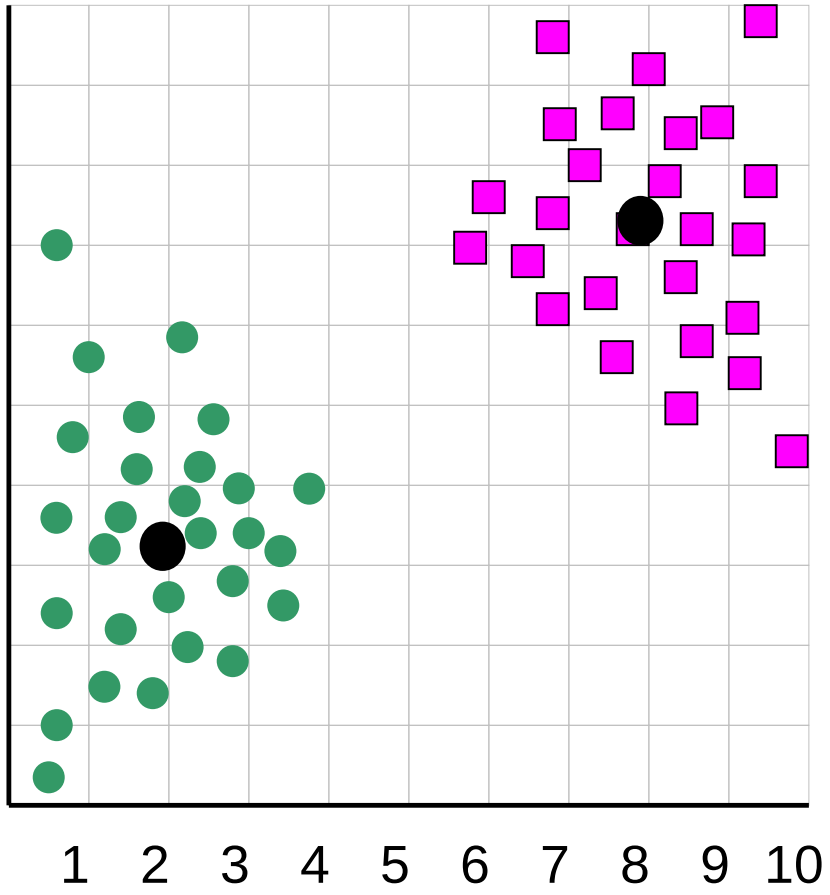
In general, this is a unsolved problem. However there are many approximate methods. In the next few slides we will see an example.



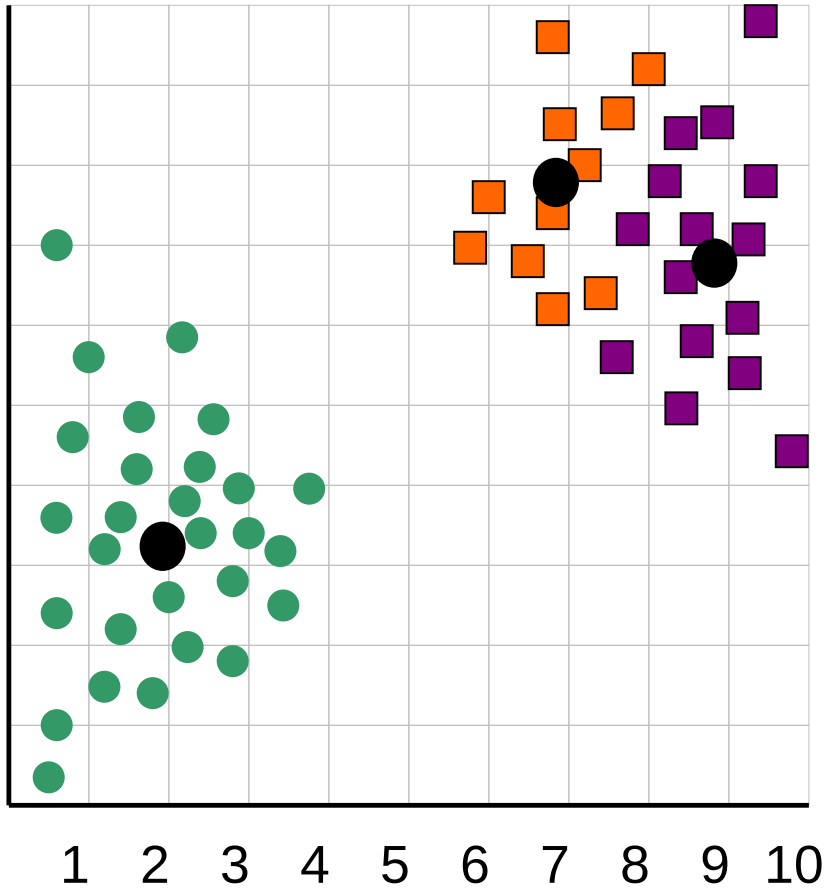
When k = 1, the objective function is 873.0



When k = 2, the objective function is 173.1

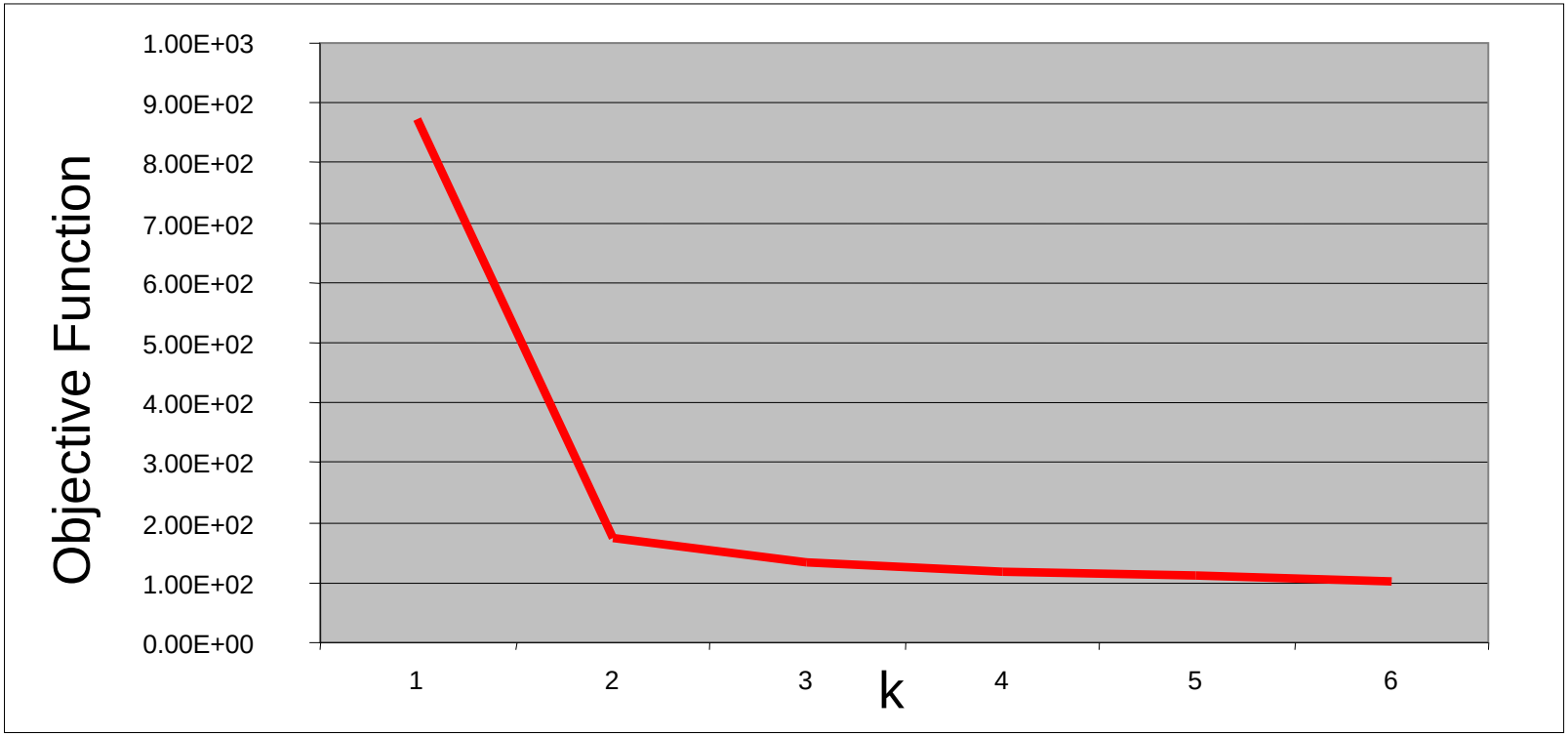


When k = 3, the objective function is 133.6



We can plot the objective function values for k equals 1 to 6…

The abrupt change at k = 2, is highly suggestive of two clusters in the data. This technique for determining the number of clusters is known as “knee finding” or “elbow finding”.

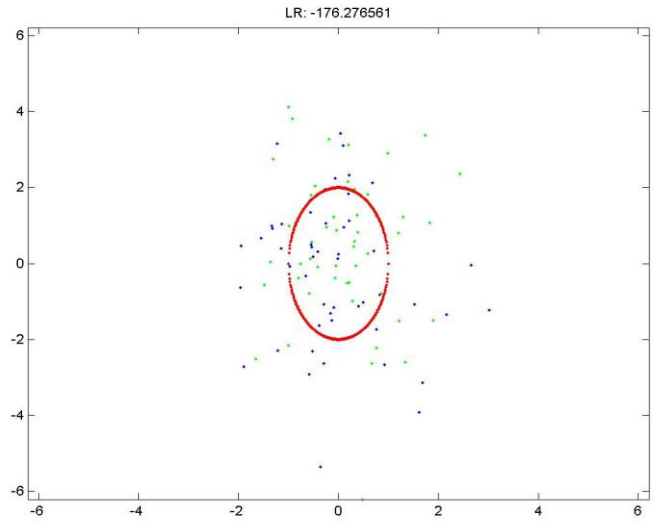


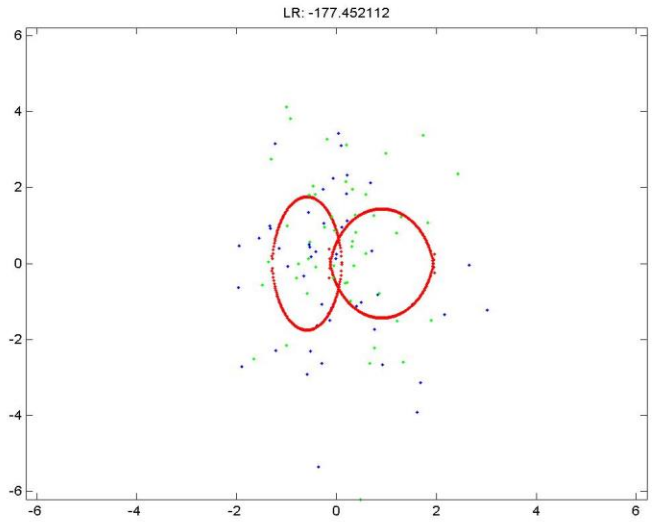
Note that the results are not always as clear cut as in this toy example

Cross validation

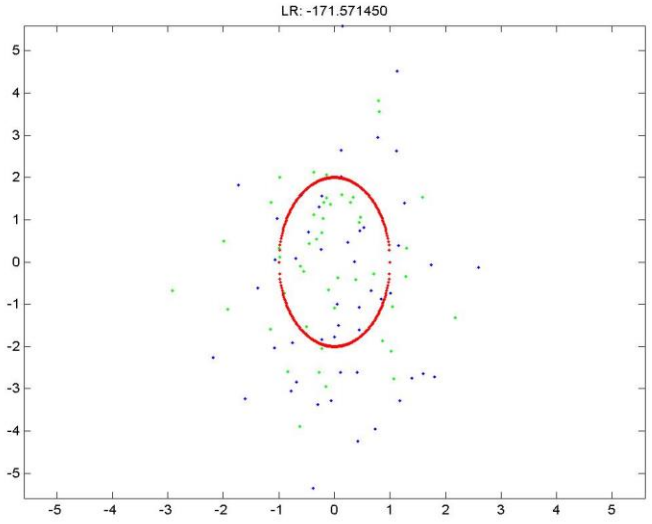
- We can also use cross validation to determine the correct number of classes
- Recall that GMMs is a generative model. We can compute the likelihood of the left out data to determine which model (number of clusters) is more accurate

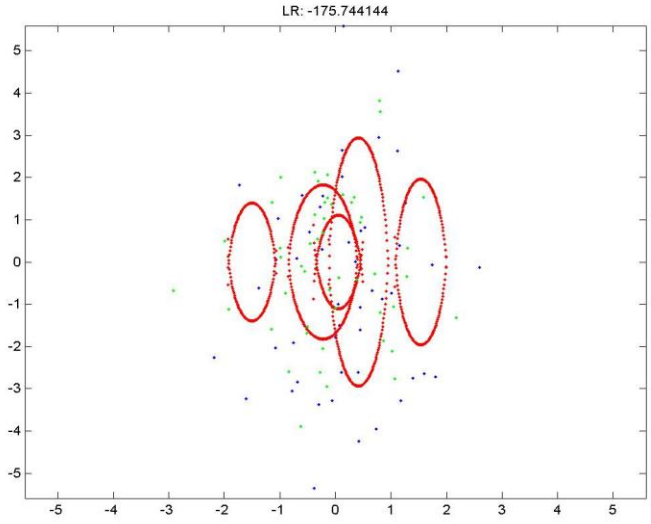
$$p(x_1 \cdots x_n \mid \theta) = \prod_{j=1}^{n} \left(\sum_{i=1}^{k} p(x_j \mid C = i) w_i \right)$$





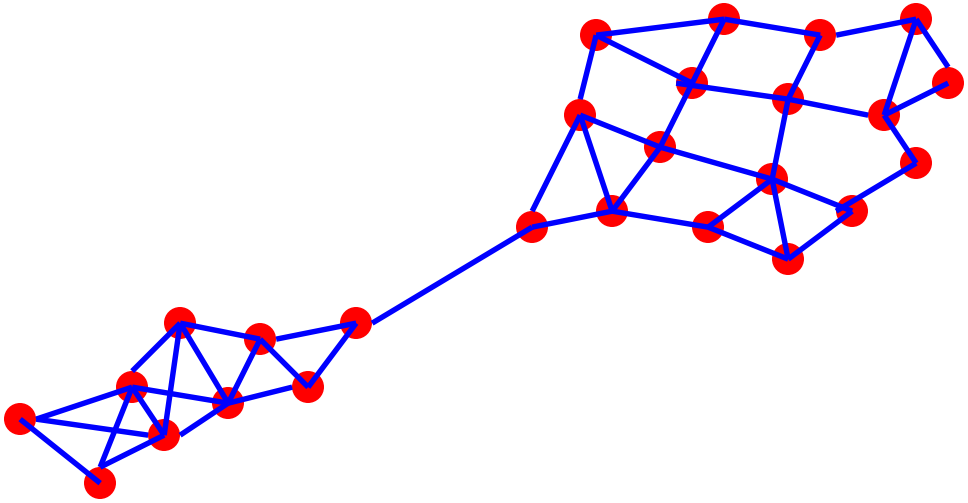
Cross validation





Top down: Graph based clustering

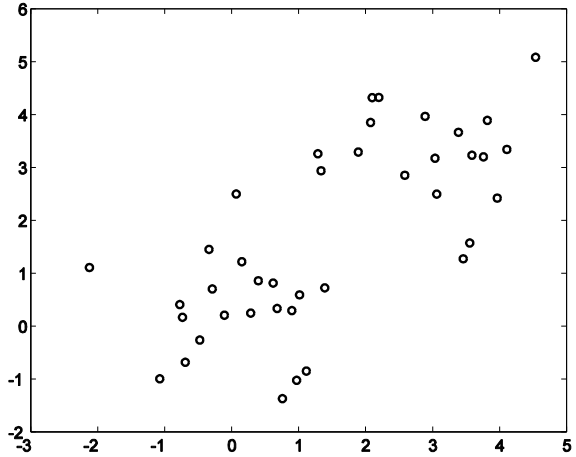
- Many top down clustering algorithms work by first constructing a neighborhood graph and then trying to infer some sort of connected components in that graph

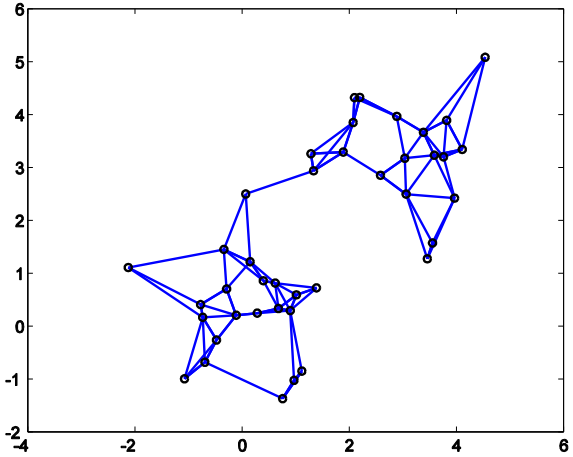


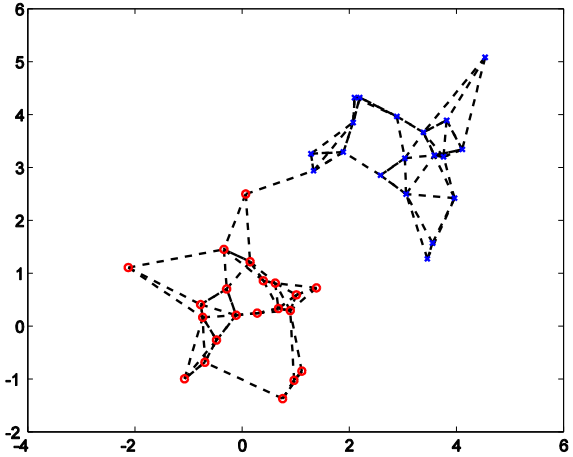
Graph based clustering

- We need to clarify how to perform the following three steps:
  1. construct the neighborhood graph
  2. assign weights to the edges (similarity)
  3. partition the nodes using the graph structure

Example





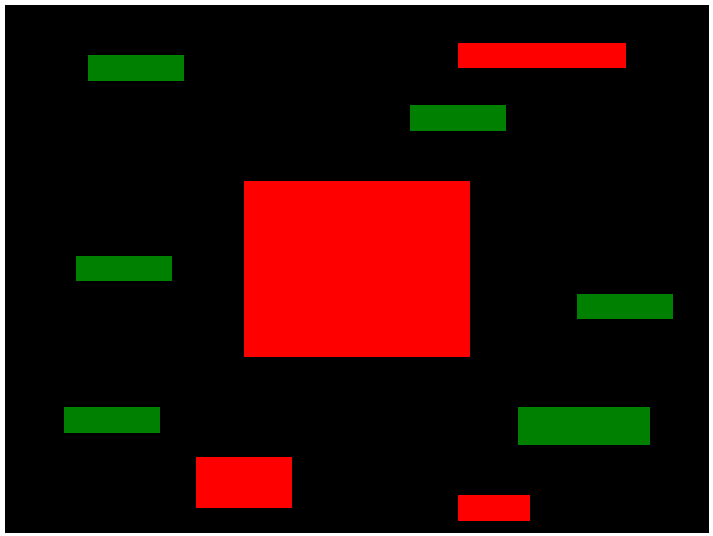


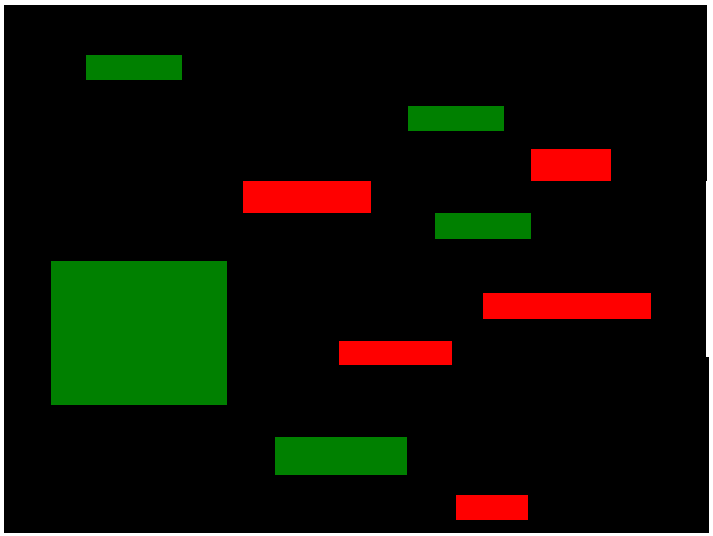
Clustering methods: Comparison

	Bottom up	Model based	Top down
Running time	naively, $O(n^3)$	fast (each iteration is linear)	could be slow (matrix transformation)
Assumptions	requires a similarity / distance measure	strong assumptions	general (except for graph structure)
Input parameters	none	*k* (number of clusters)	either *k* or distance threshold
Clusters	subjective (only a tree is returned)	exactly *k* clusters	depends on the input format

Bi-clustering

- Find subsets of genes and experiments such that the genes in the subset behave similarly across the subset of the experiments





Cluster validation

- We wish to determine whether the clusters are real
  - internal validation (stability, coherence)
  - external validation (match to known categories)

Internal validation: Coherence

- A simple method is to compare clustering algorithm based on the coherence of their results
- We compute the average inter-cluster similarity and the average intra-cluster similarity
- Requires the definition of the similarity / distance metric

Internal validation: Stability

- If the clusters capture real structure in the data they should be stable to minor perturbation (e.g., subsampling) of the data.
- To characterize stability we need a measure of similarity between any two k-clusterings.
- For any set of clusters C we define L(C) as the matrix of 0/1 labels such that $L(C)_{ij}$ =1 if genes i and j belong to the same cluster and zero otherwise.
- We can compare any two k clusterings C and C' by comparing the corresponding label matrices L(C) and L(C').

Validation by subsampling

- C is the set of k clusters based on all the gene profiles
- C' denotes the set of k clusters resulting from a randomly chosen subset (80-90%) of genes
- We have high confidence in the original clustering if Sim(L(C),L(C')) approaches 1 with high probability, where the comparison is done over the genes common to both

External validation

- More common (why ?).
- Suppose we have generated k clusters (sets of gene profiles) $C_1,\ldots,C_k$. How do we assess the significance of their relation to m known (potentially overlapping) categories $G_1,\ldots,G_m$?
- Let's start by comparing a single cluster C with a single category G_j. The p-value for such a match is based on the hyper-geometric distribution.
- Board.
- This is the probability that a randomly chosen $|C_i|$ elements out of n would have l elements in common with G_j.

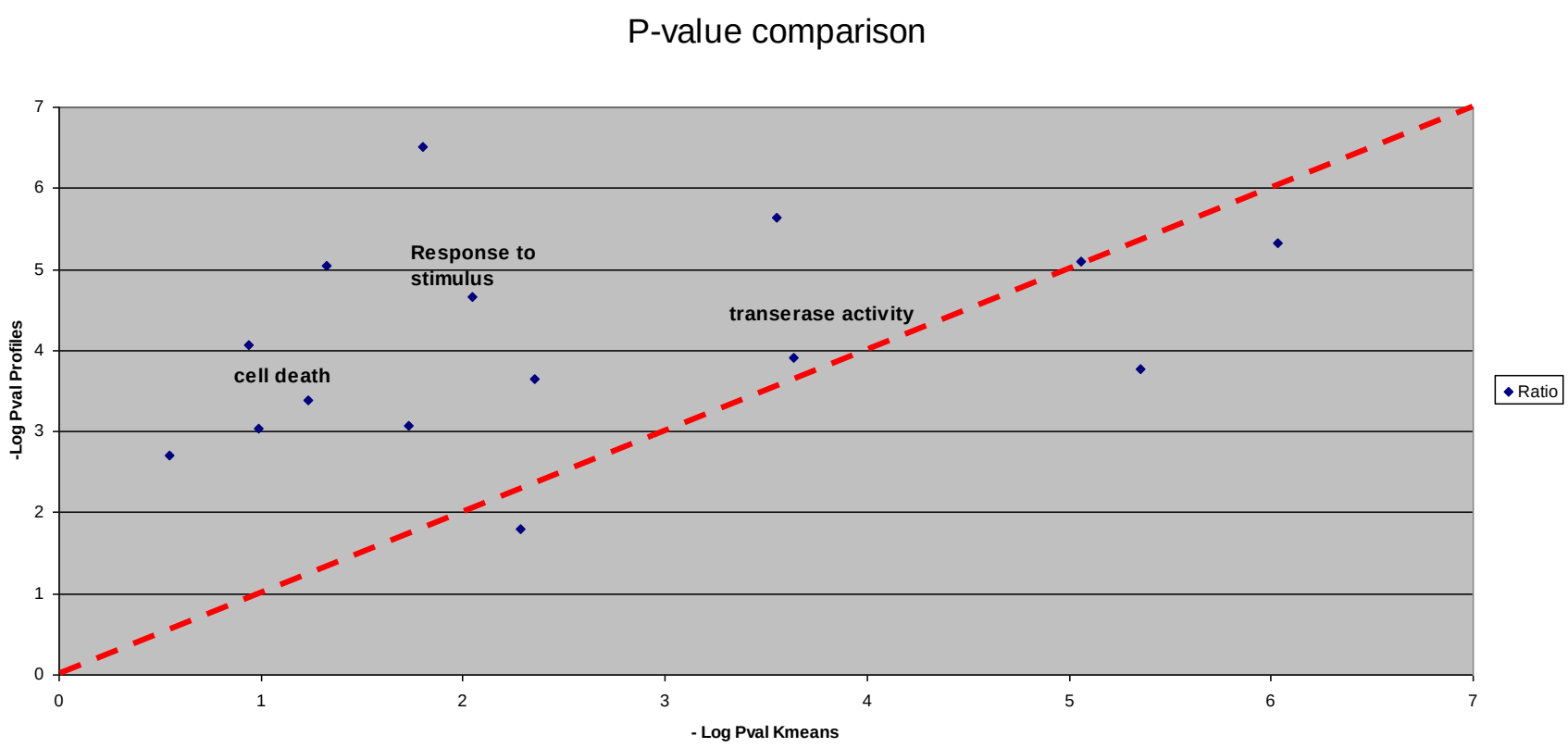
P-value (cont.)

- If the observed overlap between the sets (cluster and category) is l elements (genes), then the p-value is

$$p = prob\ (l \geq \hat{l}) = \sum_{j=l}^{\min(c,m)} prob\ (exactly - j - matches)$$

- Since the categories $G_1,\ldots,G_m$ typically overlap we cannot assume that each cluster-category pair represents an independent comparison
- In addition, we have to account for the multiple hypothesis we are testing.
- Solution ?

External validation: Example



What you should know

- Why is clustering useful
- What are the different types of clustering algorithms
- What are the assumptions we are making for each, and what can we get from them
- Cluster validation: Internal and external