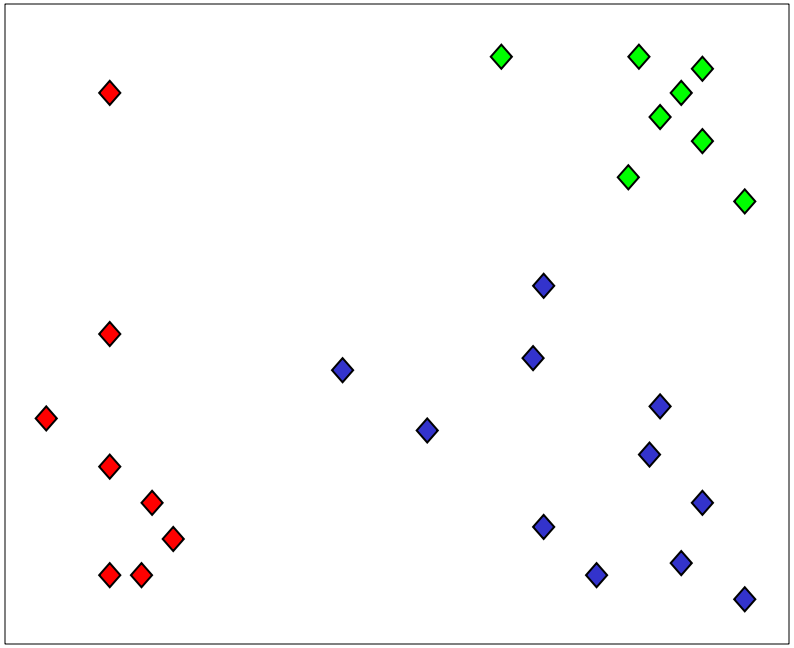
10701
Machine Learning

Clustering

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.



- Why do we want to do that?
- Any REAL application?

Example: clusty

Clusty Search » simpsons - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://clusty.com/search?v%3afile=viv_1023%4019%3akiZm1v&v%3aframe=tree&v%3astate=

web news images wikipedia blogs jobs more »

simpsons Search advanced preferences

clusters sources sites remix

All Results (224)
- Pictures (62)
- Games (21)
- Movie (18)
- Collectibles (14)
- Downloads (15)
- Witness, Trial (10)
  - Bruce Fromong (4)
  - Jurors Hear (3)
  - Alleged robbery (3)
  - Murder, Las Vegas (2)
  - Other Topics (1)
- FOX, Broadcasting Company (7)
- Quotes (12)
- Episode Guides (6)
- Simpson College (10)

more | all clusters

Cluster **Witness, Trial** contains **10** documents.

Search Results

1. Witness contradicts self in O.J. Simpson trial

Sep 17, 2008 - A key **witness** in the O.J. **Simpson** robbery **trial** was confronted with contradictions in his **testimony** Tuesday, including his claim that he didn't try to profit from the casino hotel room confrontation that led to charges against the former football star. Memorabilia dealer Bruce Fromong, who returned to the stand after becoming ill Monday, told defense attorney Gabriel Grasso he didn't have money on his mind while allegedly being robbed of sports collectibles by **Simpson** and a group of other men. "You ...

news.yahoo.com/s/ap/20080917/ap_on_re_us/oj_simpson - [cache] - Yahoo! News

2. Witness in Simpson trial says gun brandished in incident

Sep 16, 2008 - A **witness** who says he was robbed by O.J. **Simpson** testified that a gun was brandished during the incident as the former football star's robbery and kidnapping **trial** opened. Bruce Fromong, 54, one of the two collectibles dealers at the center of the case, told the jury on Monday that someone in the room during the alleged robbery shouted, "Put the gun down," contradicting **Simpson**'s claim he did not know firearms were present. The **witness** said he could not recall which of the six men who burst into the ...

news.yahoo.com/s/afp/20080916/en_afp/entertainmentuscrimetrialsimpson - [cache] - Yahoo! News

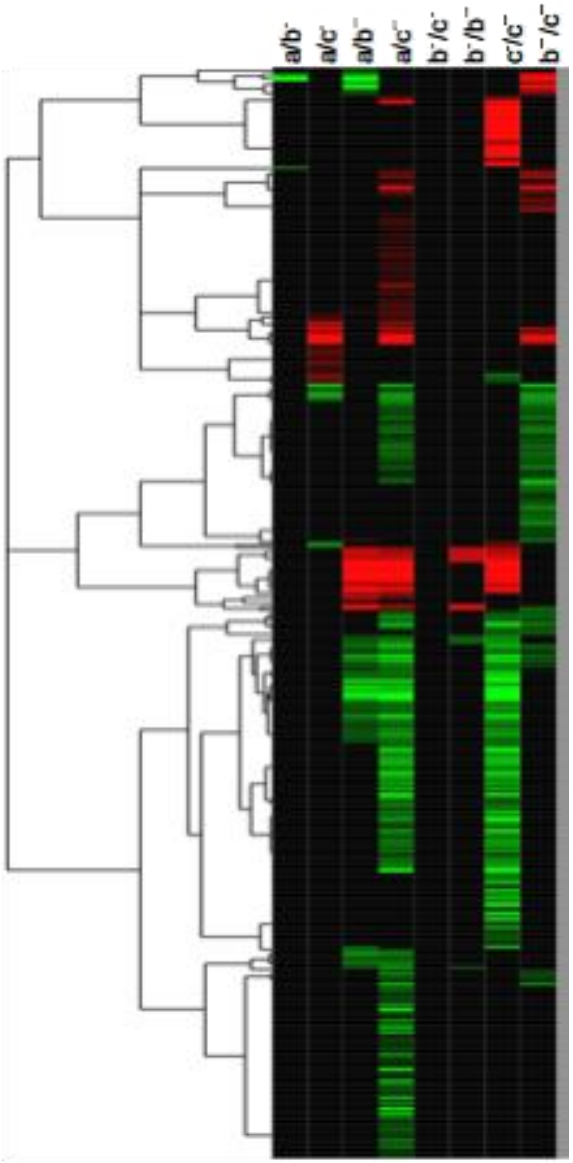
3. Key OJ Simpson witness clutches chest in court

Sep 16, 2008 - A key **witness** in O.J. **Simpson**'s kidnap and robbery **trial** became ill on Monday while testifying about a hotel room confrontation at the heart of the case -- clutching his chest before bailiffs helped him from the **witness** stand.

Done

Example: clustering genes

- Microarrays measures the activities of all genes in different conditions
- Clustering genes can help determine new functions for unknown genes
- An early “killer application” in this area
  - The most cited (11,591) paper in PNAS!



Why clustering?

- Organizing data into clusters provides information about the internal structure of the data
  - Ex. Clusty and clustering genes above
- Sometimes the partitioning is the goal
  - Ex. Image segmentation
- Knowledge discovery in data
  - Ex. Underlying rules, reoccurring patterns, topics, etc.

Unsupervised learning

- Clustering methods are unsupervised learning techniques

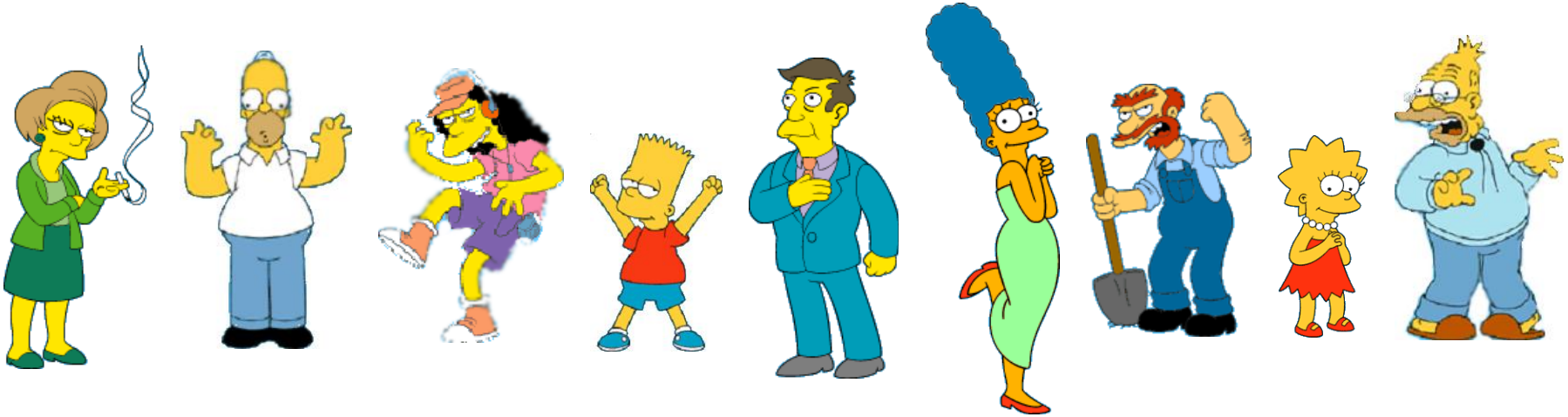
  - We do not have a teacher that provides examples with their labels

- We will also discuss dimensionality reduction, another unsupervised learning method later in the course

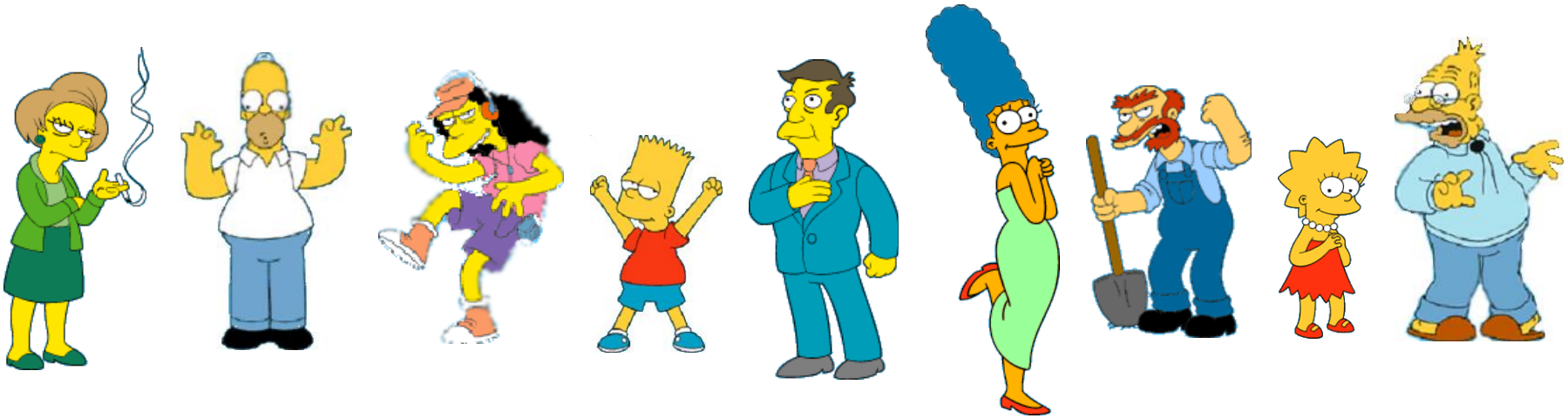
Outline

- Motivation
- Distance functions
- Hierarchical clustering
- Partitional clustering
  - K-means
  - Gaussian Mixture Models
- Number of clusters

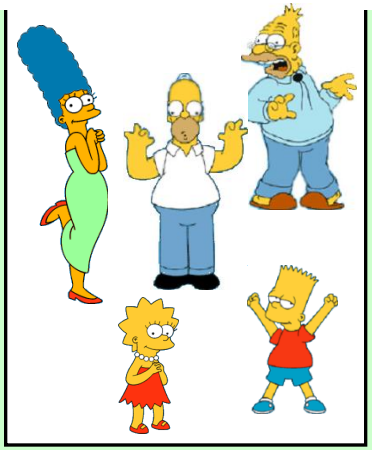
What is a natural grouping among these objects?



What is a natural grouping among these objects?

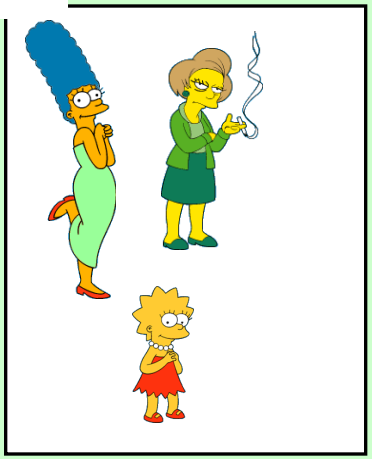


Clustering is subjective



Simpson's Family

School Employees



Females

Males

What is Similarity?

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

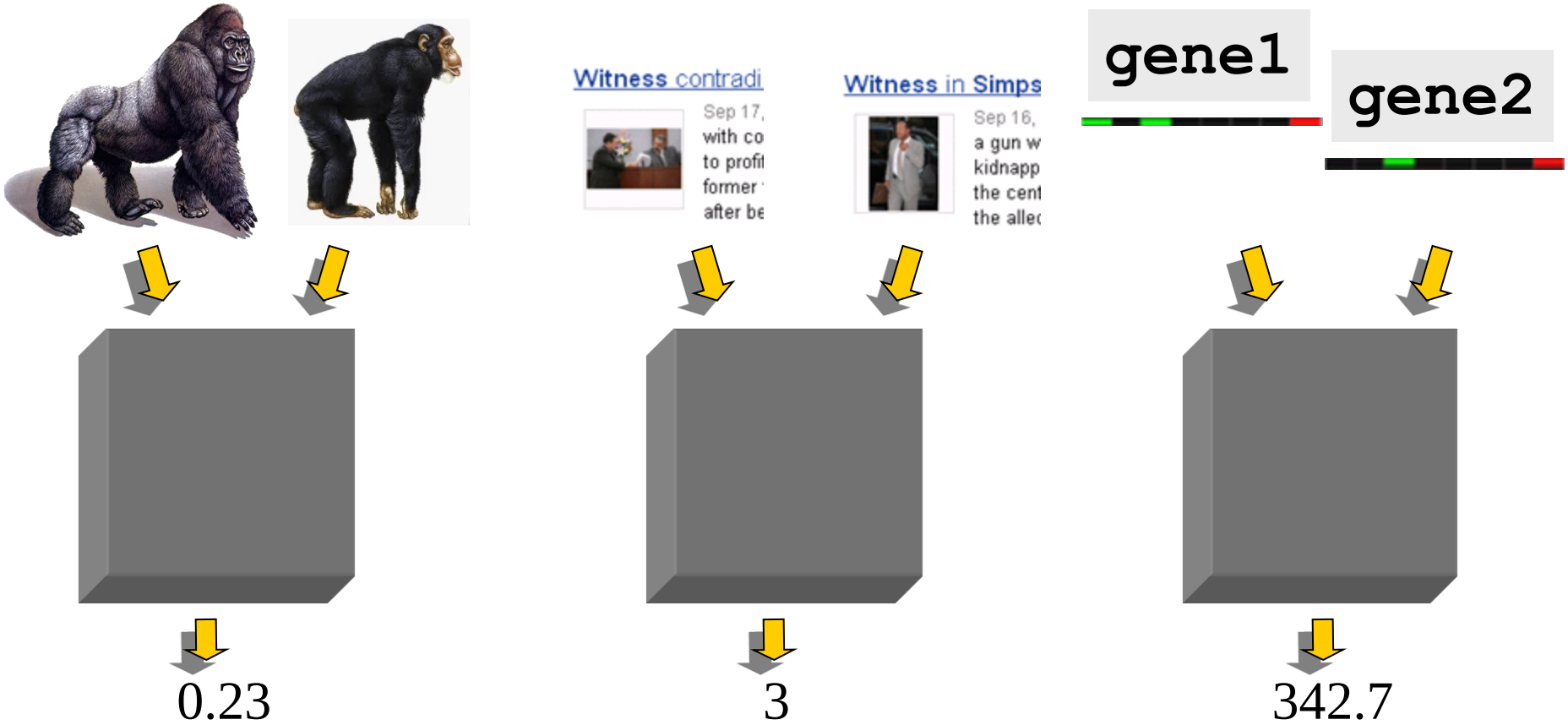
Webster's Dictionary

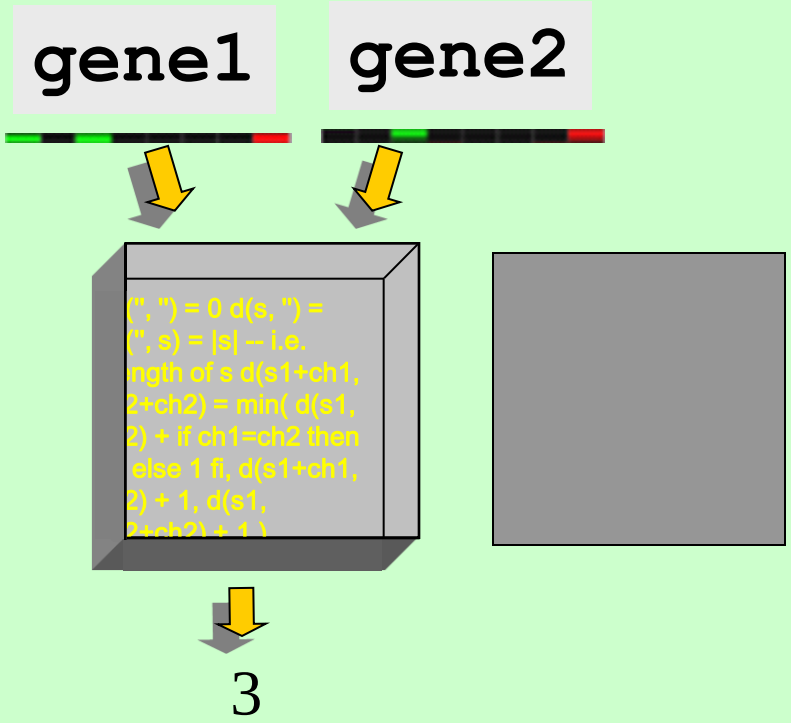
Similarity is hard to define, but…
"*We know it when we see it*"

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

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

- Similarity rather than distance
- Can determine similar trends

- Correlation coefficient

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

Outline

- Motivation
- Distance measure
- Hierarchical clustering
- Partitional clustering
  - K-means
  - Gaussian Mixture Models
- Number of clusters

Desirable Properties of a Clustering Algorithm

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Interpretability and usability

Optional

- Incorporation of user-specified constraints

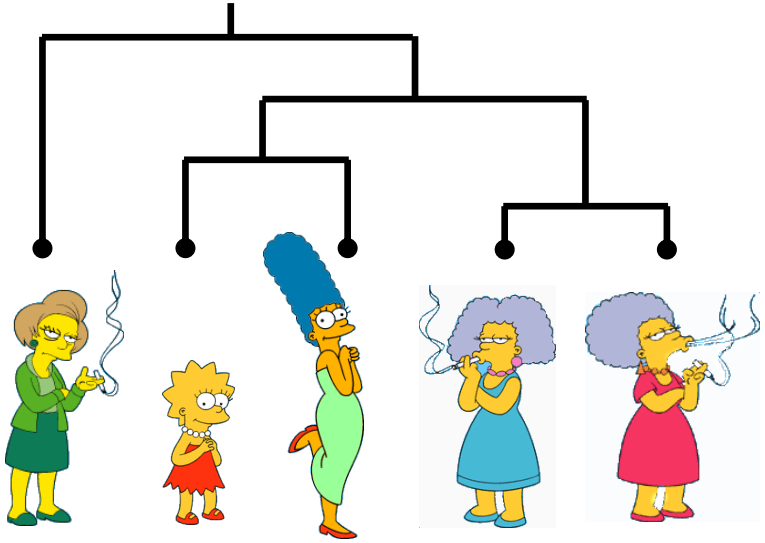
Two Types of Clustering

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion (focus of this class)

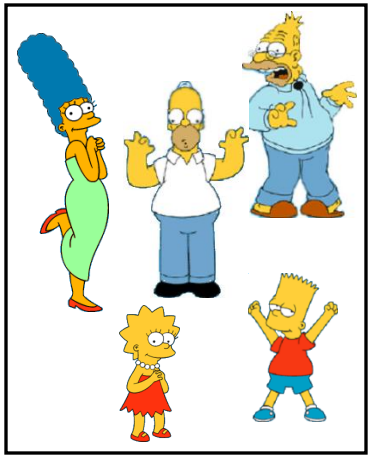
Bottom up or top down

Top down

Hierarchical



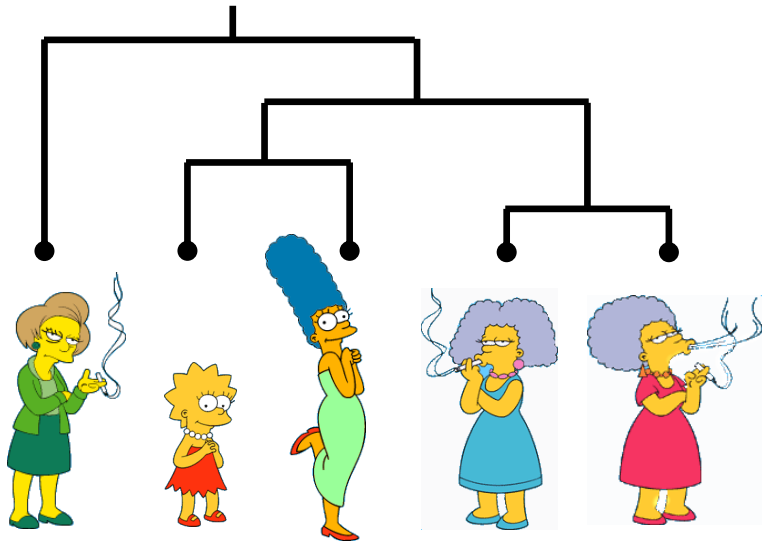
Partitional



(How-to) Hierarchical Clustering

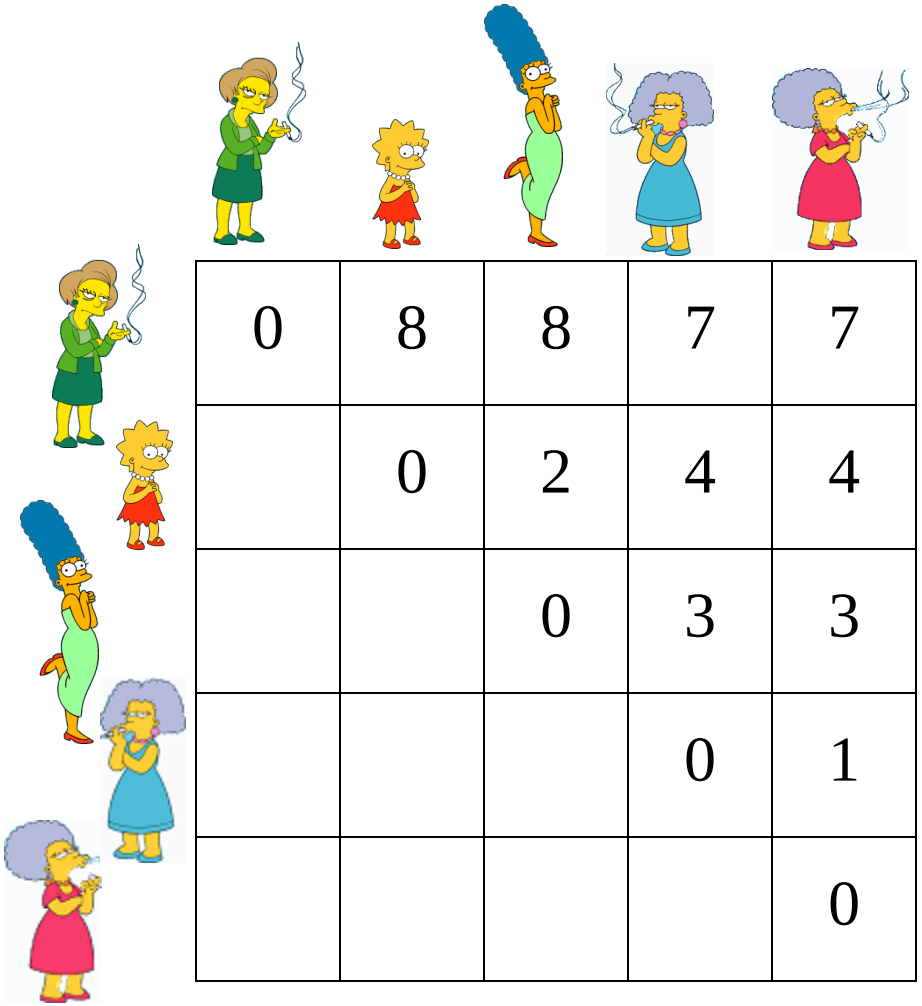
The number of dendrograms with n leafs $= (2n - 3)!/[(2^{(n-2)})\,(n-2)!]$

Number of Leafs	Number of Possible Dendrograms
2	1
3	3
4	15
5	105
...	…
10	34,459,425



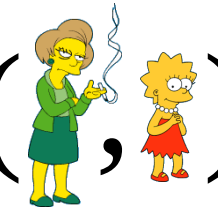
Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

We begin with a distance matrix which contains the distances between every pair of objects in our database.



0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

D(,) = 8

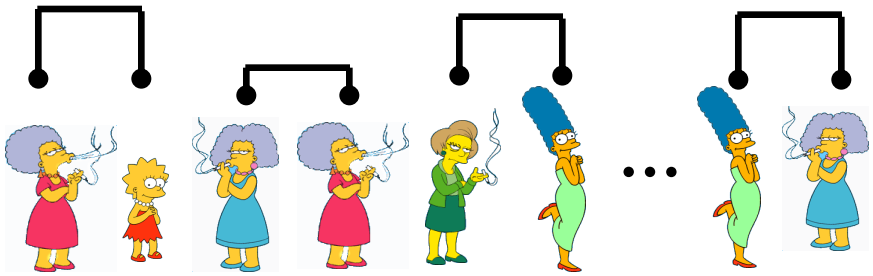


D(,) = 1

Bottom-Up (agglomerative):

Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

Consider all possible merges…

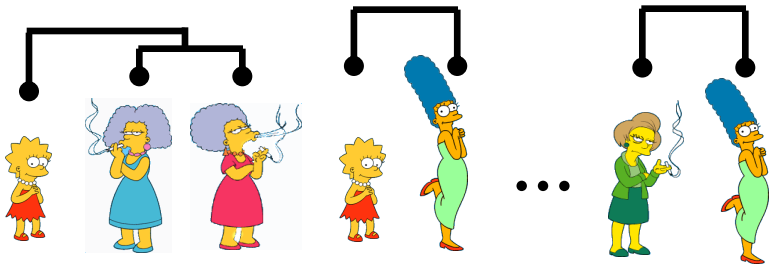


Choose the best

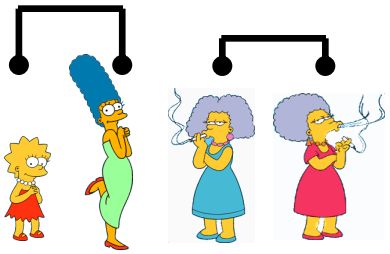
Bottom-Up (agglomerative):

Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

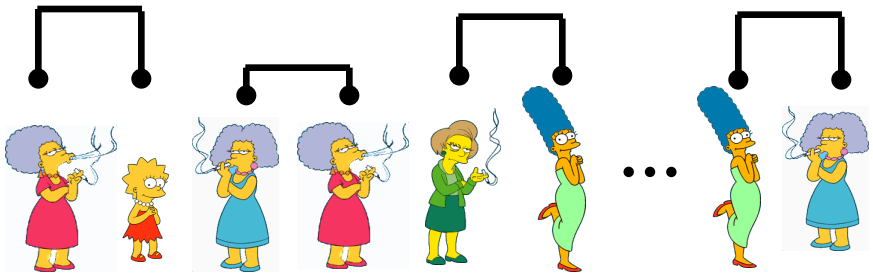
Consider all possible merges…



Choose the best



Consider all possible merges…

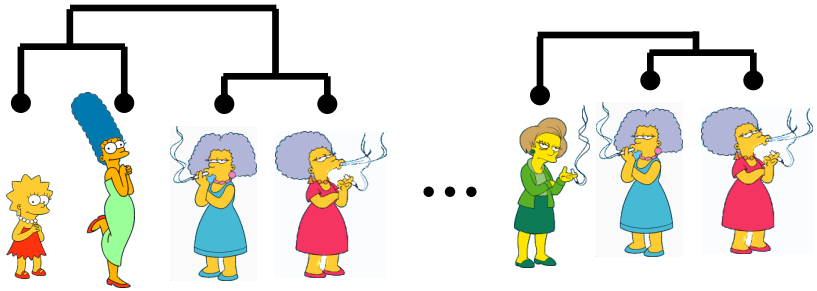


Choose the best

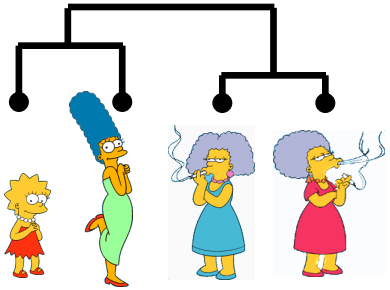
Bottom-Up (agglomerative):

Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

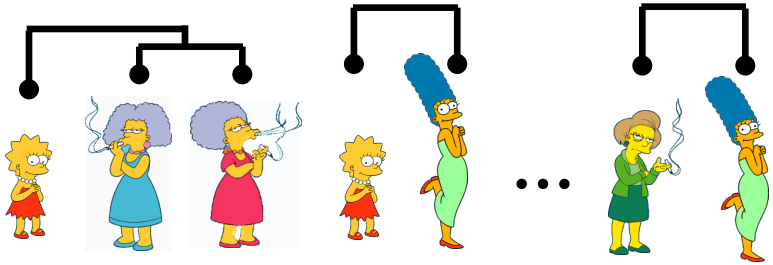
Consider all possible merges…



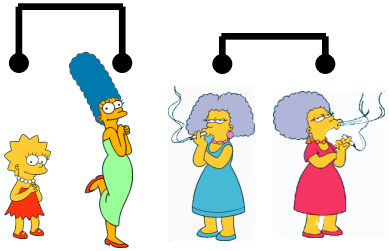
Choose the best



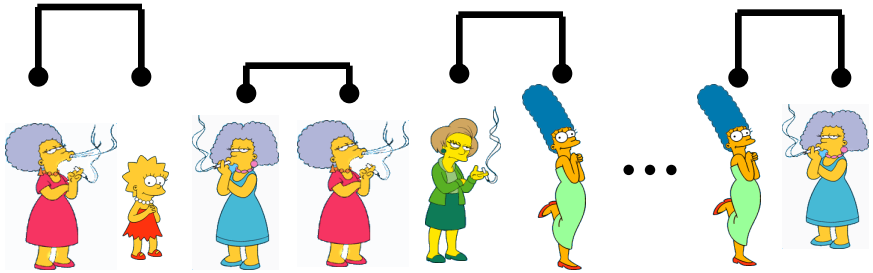
Consider all possible merges…



Choose the best



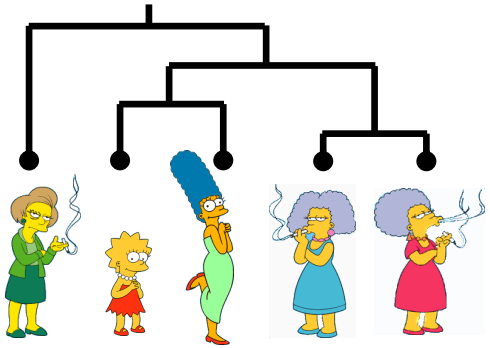
Consider all possible merges…



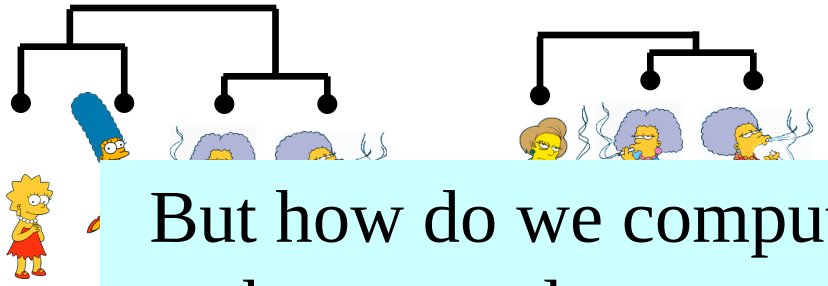
Choose the best

Bottom-Up (agglomerative):

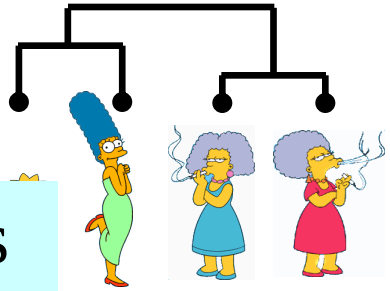
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



Consider all possible merges…

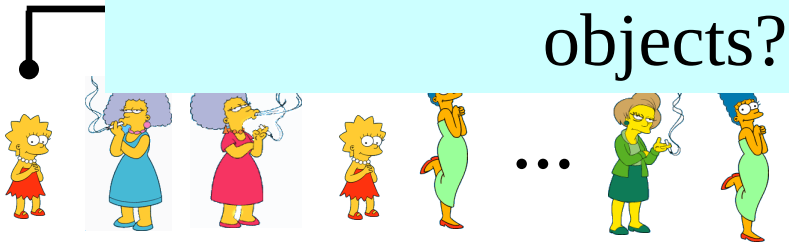


Choose

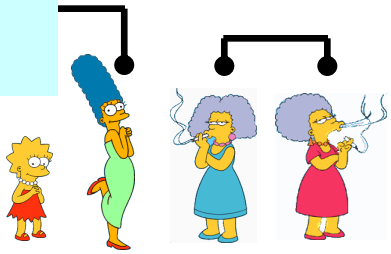


But how do we compute distances between clusters rather than objects?

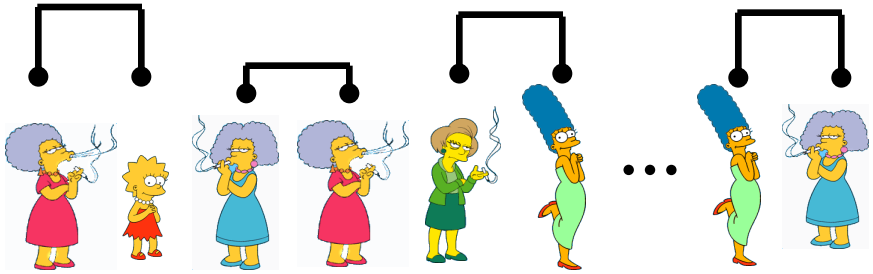
Consider all possible merges…



the best



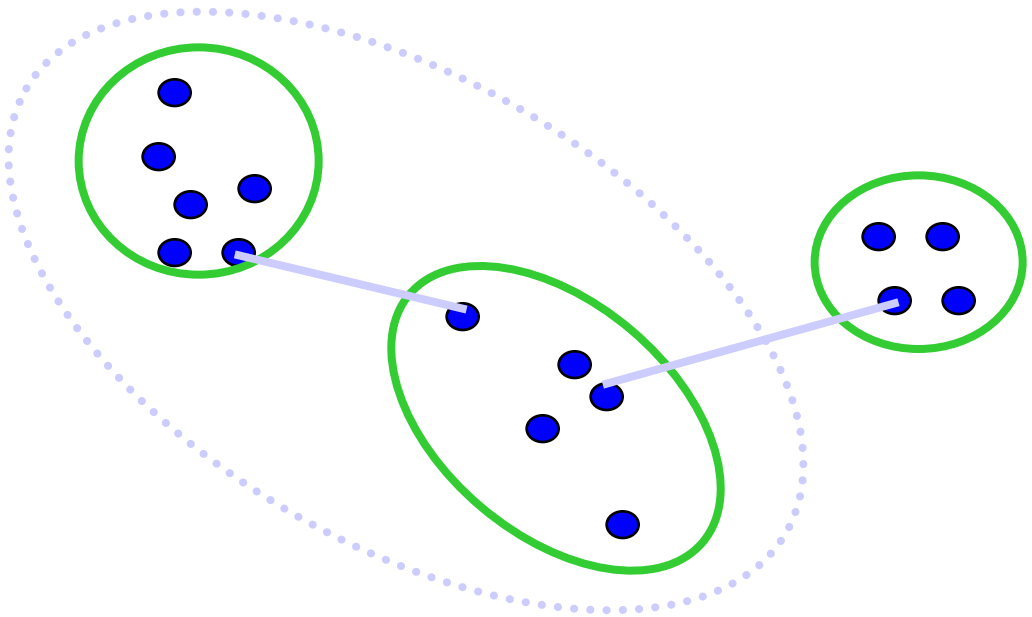
Consider all possible merges…



Choose the best

Computing distance between clusters: Single Link

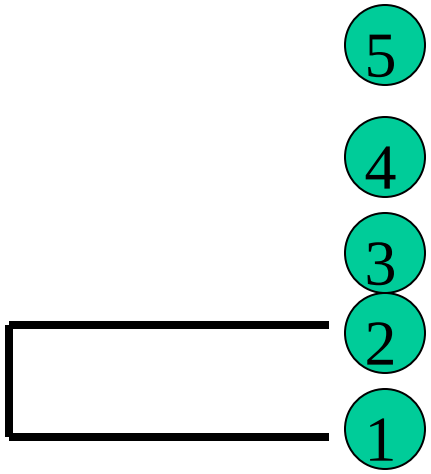
- cluster distance = distance of two closest members in each class



- Potentially long and skinny clusters

Example: single link

$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0 & & & & \\
2 & 2 & 0 & & & \\
3 & 6 & 3 & 0 & & \\
4 & 10 & 9 & 7 & 0 & \\
5 & 9 & 8 & 5 & 4 & 0
\end{array}
$$



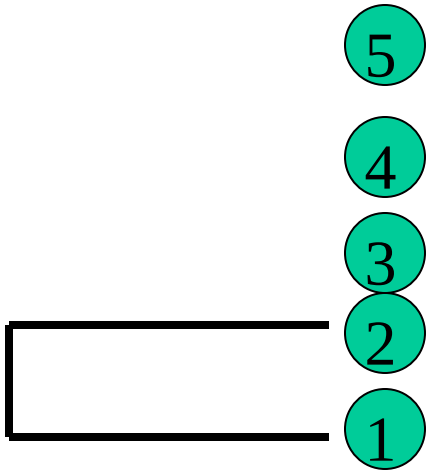
Example: single link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \quad \Rightarrow \quad \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6,3\} = 3$$

$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10,9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9,8\} = 8$$

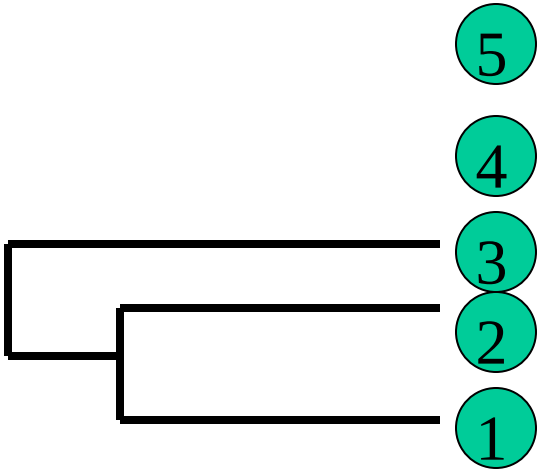


Example: single link

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ \hline (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9,7\} = 7$$

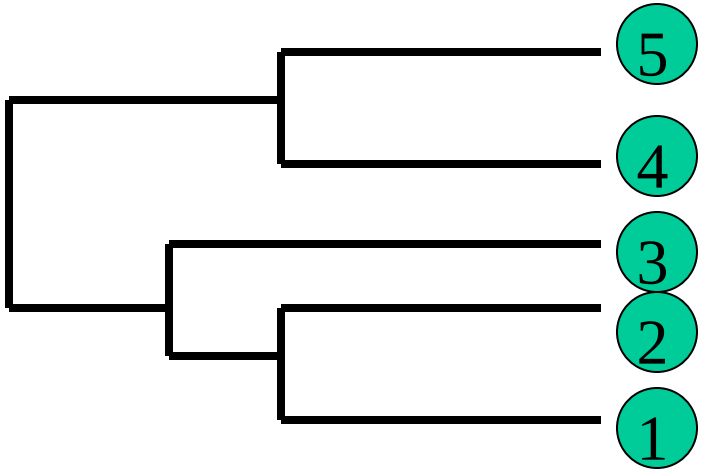
$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8,5\} = 5$$



Example: single link

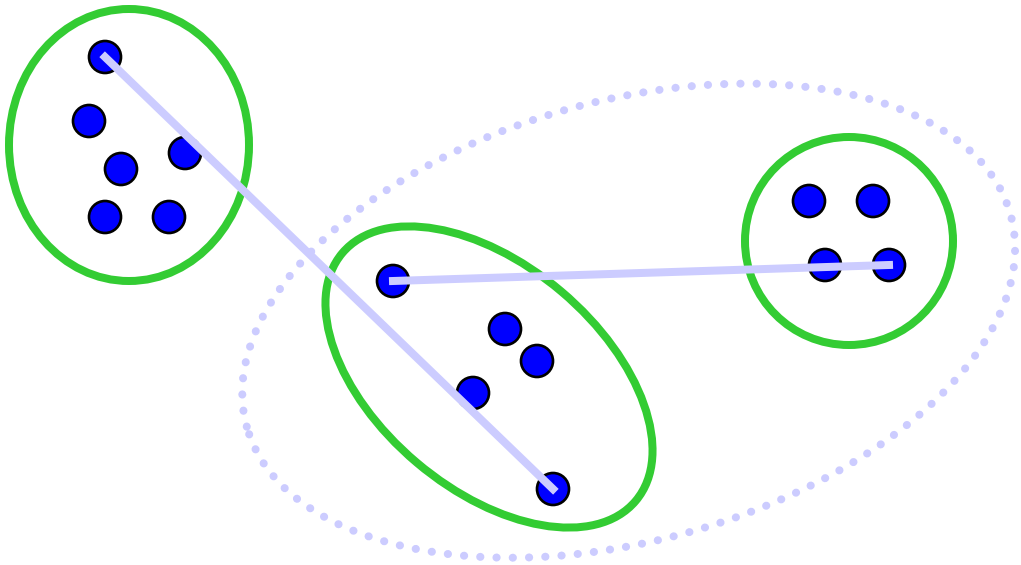
$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ \hline (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$



Computing distance between clusters: : Complete Link

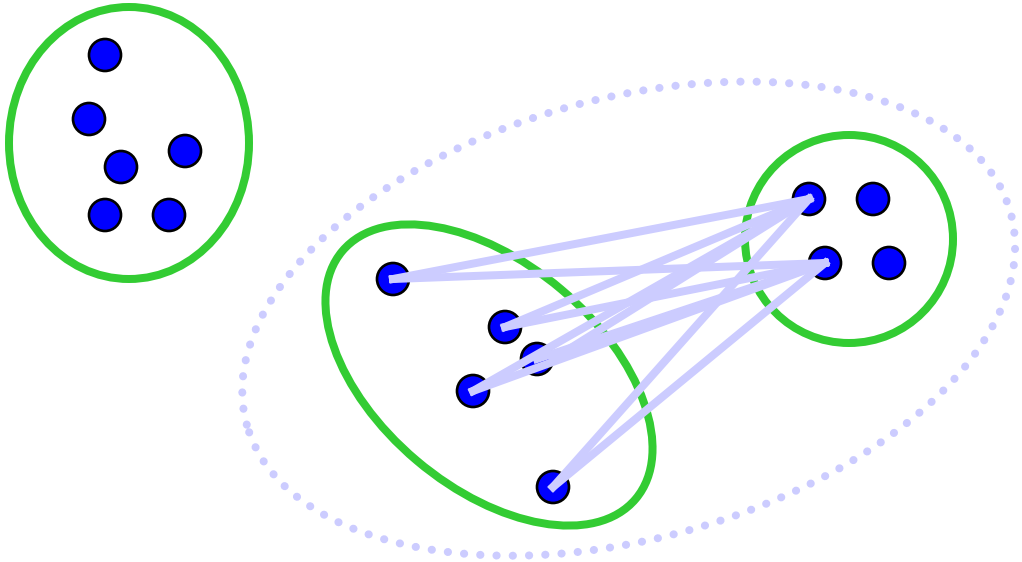
- cluster distance = distance of two farthest members



+ tight clusters

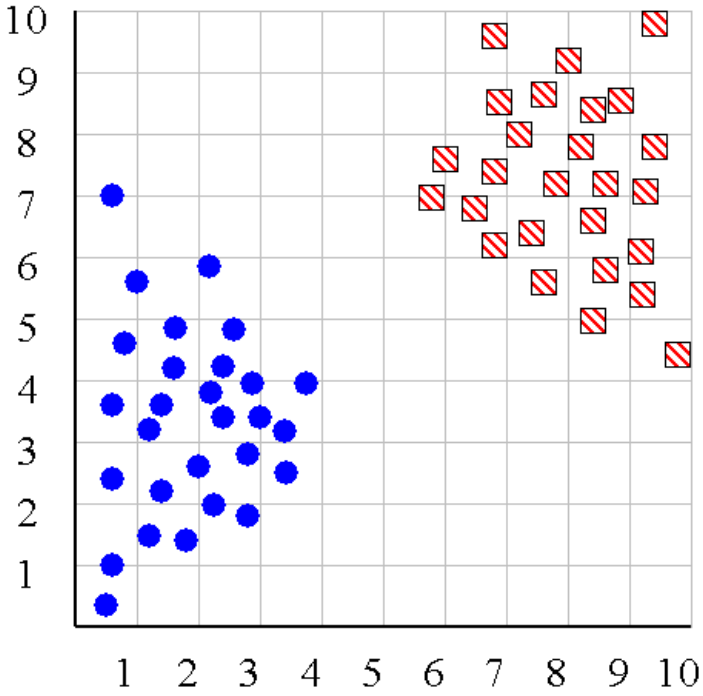
Computing distance between clusters: Average Link

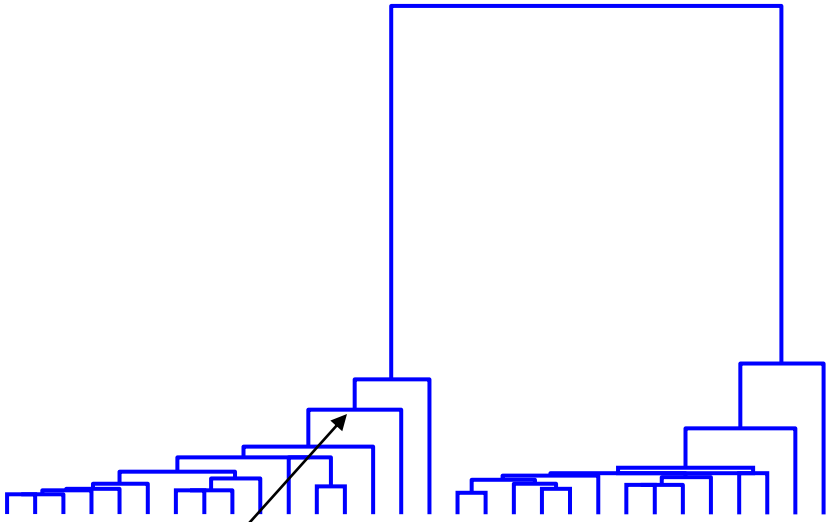
- cluster distance = average distance of all pairs



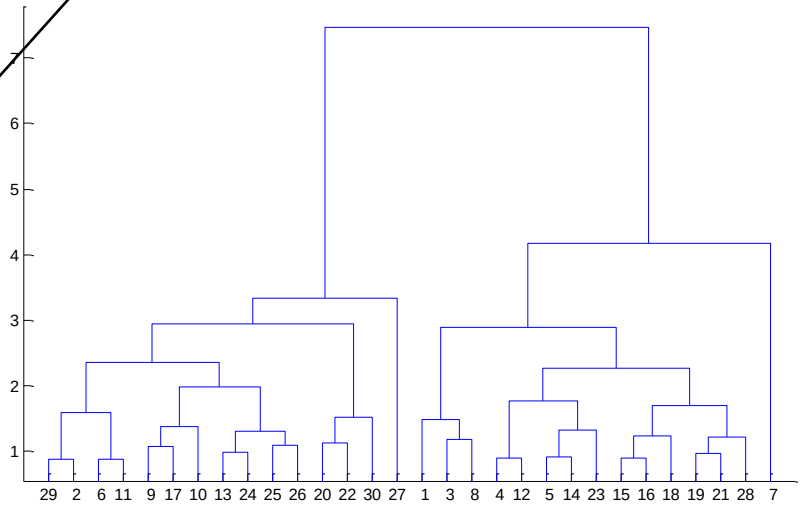
the most widely used measure

Robust against noise





Single linkage



Average linkage

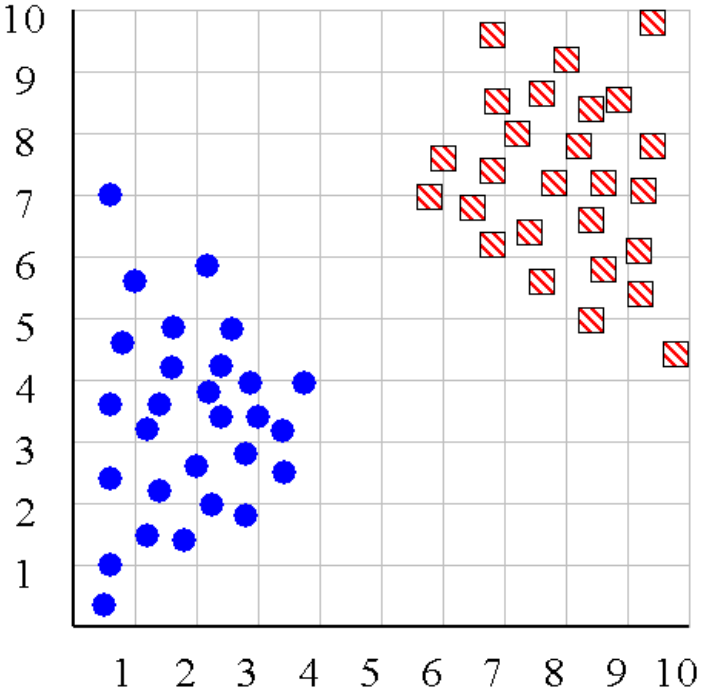
Height represents distance between objects / clusters

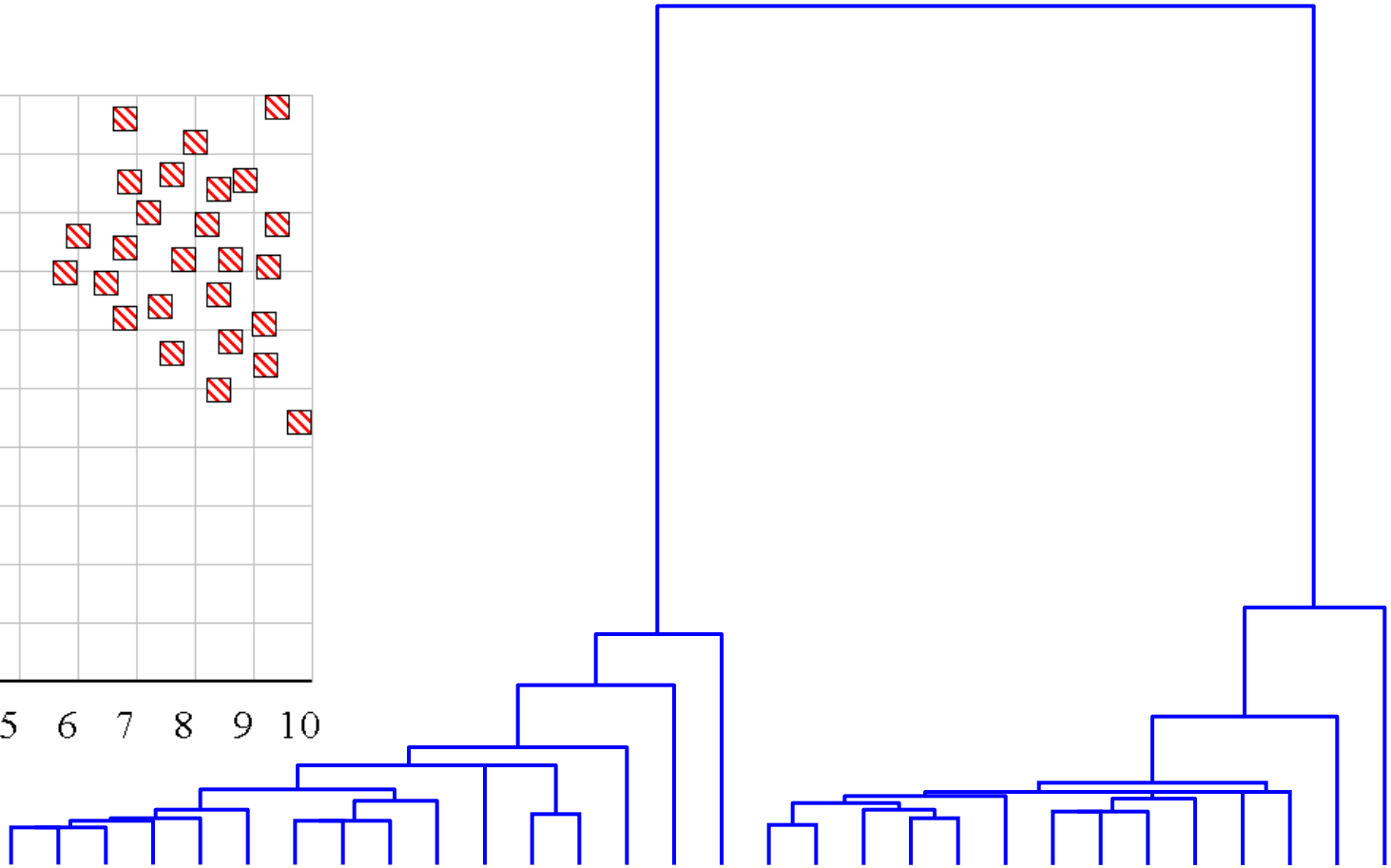
Summary of Hierarchal Clustering Methods

- No need to specify the number of clusters in advance.
- Hierarchical structure maps nicely onto human intuition for some domains
- They do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects.
- Like any heuristic search algorithms, local optima are a problem.
- Interpretation of results is (very) subjective.

But what are the clusters?

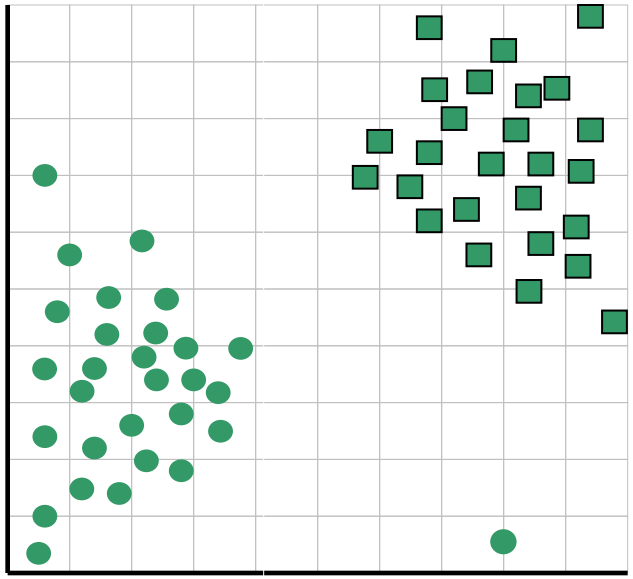
In some cases we can determine the “correct” number of clusters. However, things are rarely this clear cut, unfortunately.



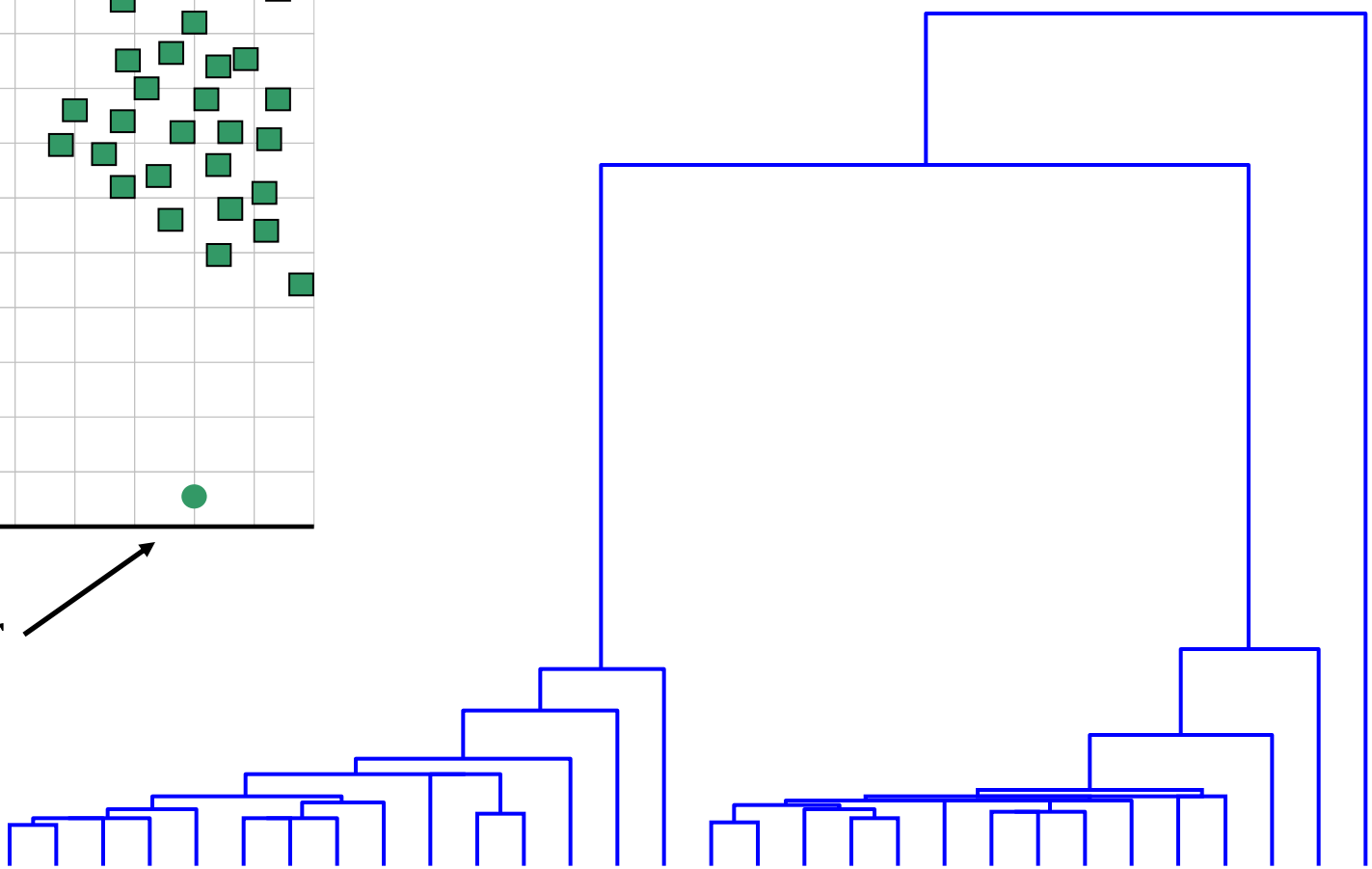


One potential use of a dendrogram is to detect outliers

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



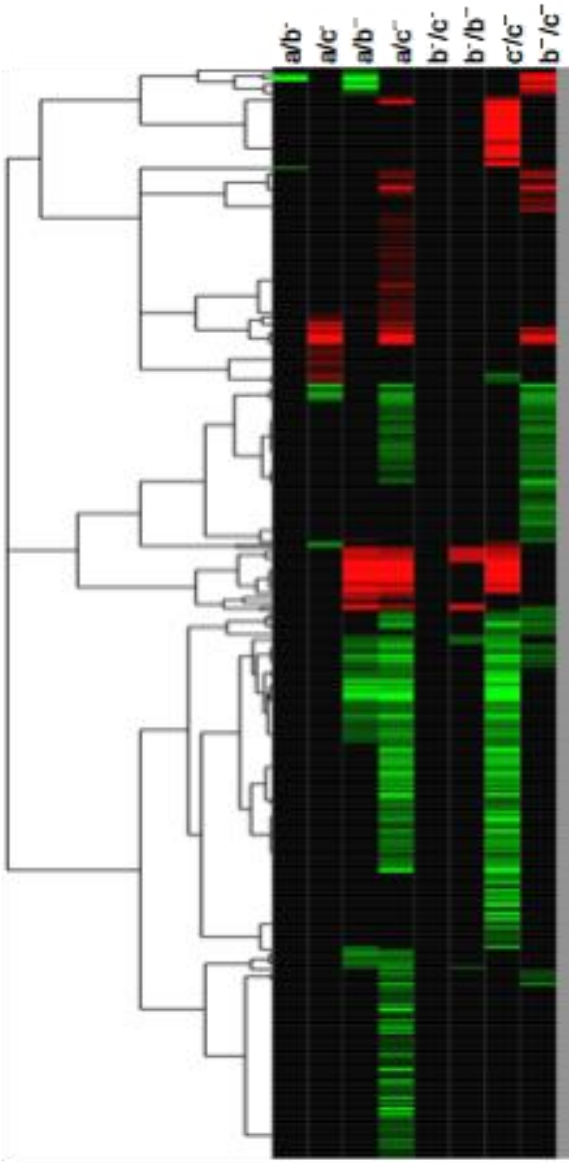
Outlier



Example: clustering genes

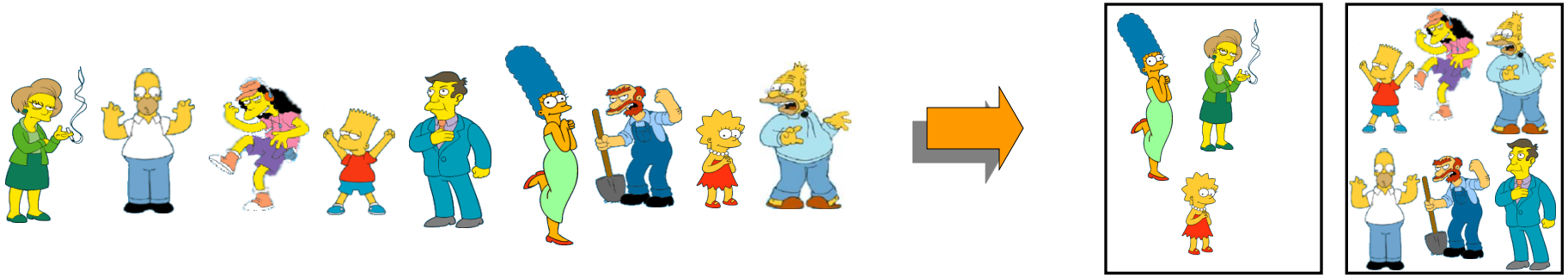
- Microarrays measures the activities of all genes in different conditions

- Clustering genes can help determine new functions for unknown genes



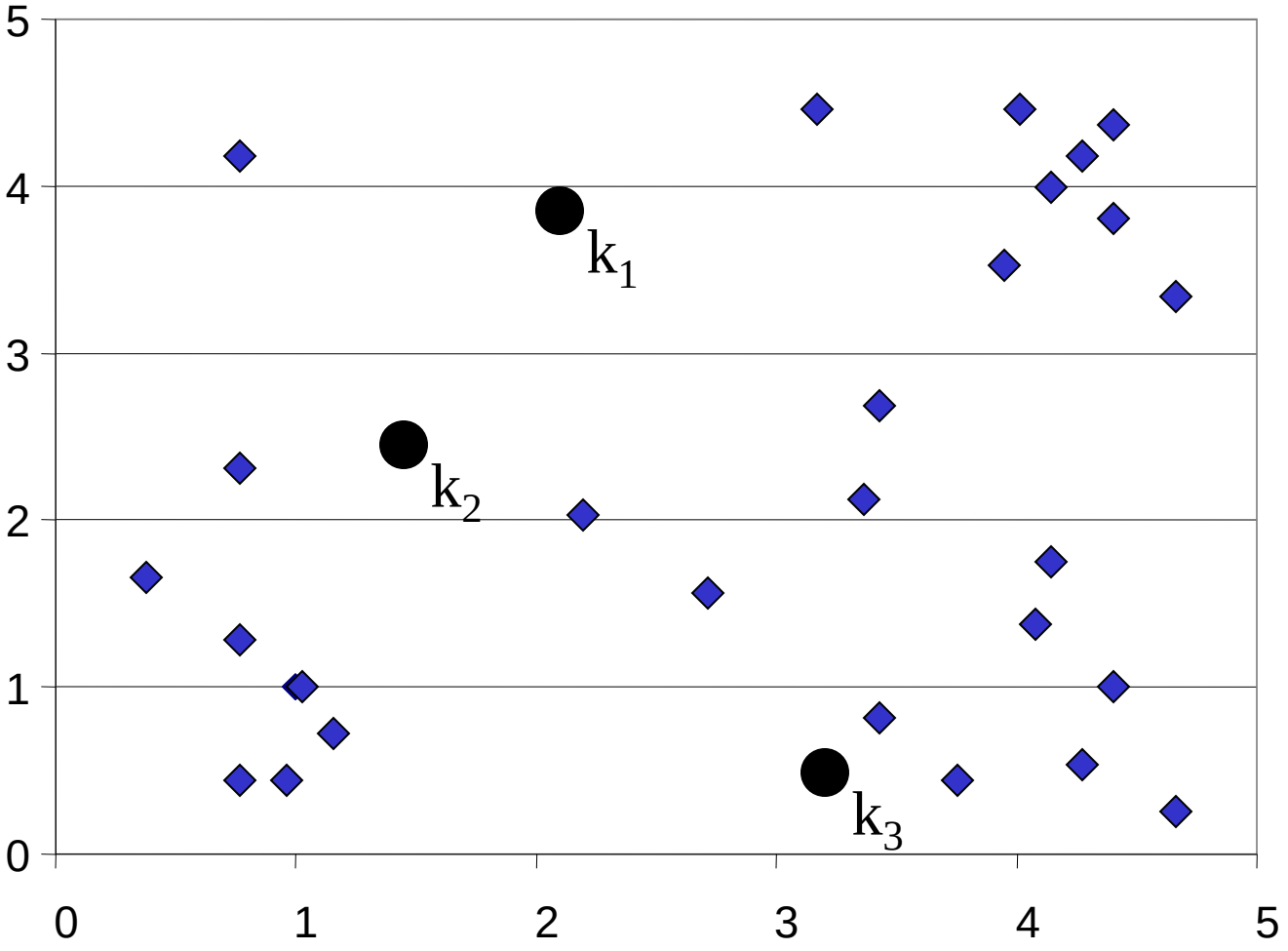
Partitional Clustering

- 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.



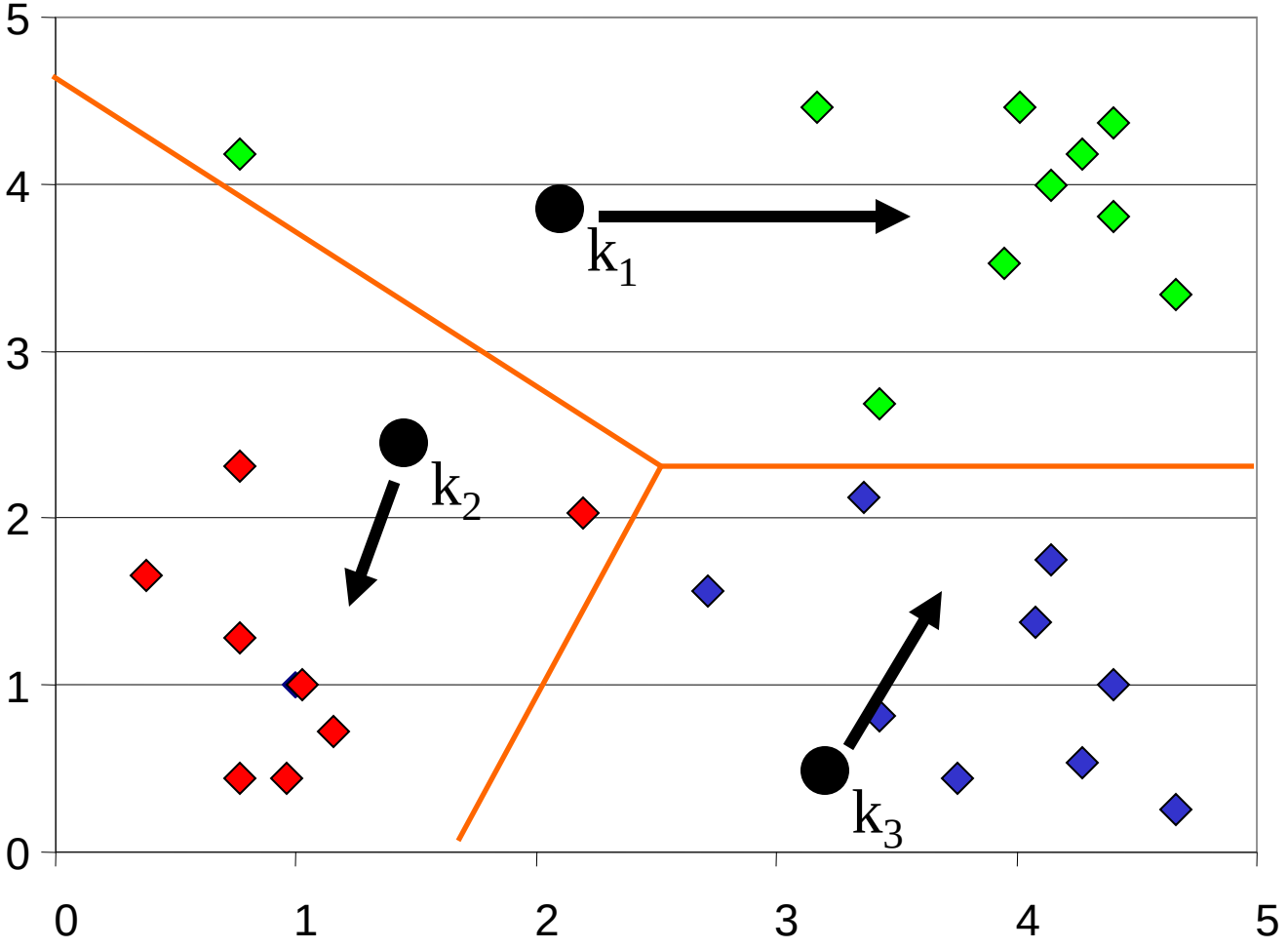
K-means Clustering: Initialization

Decide *K*, and initialize *K* centers (randomly)



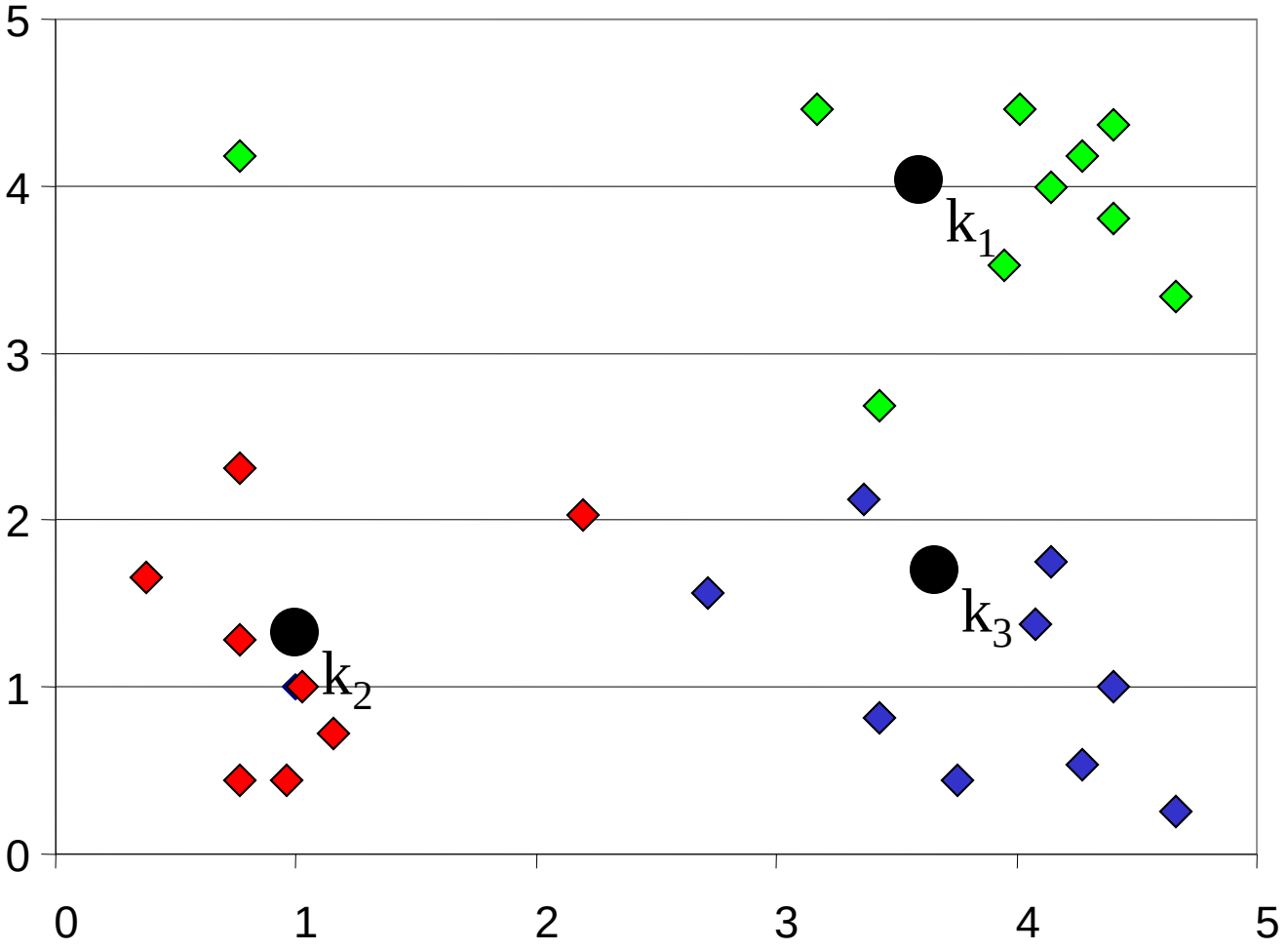
K-means Clustering: Iteration 1

Assign all objects to the nearest center.
Move a center to the mean of its members.



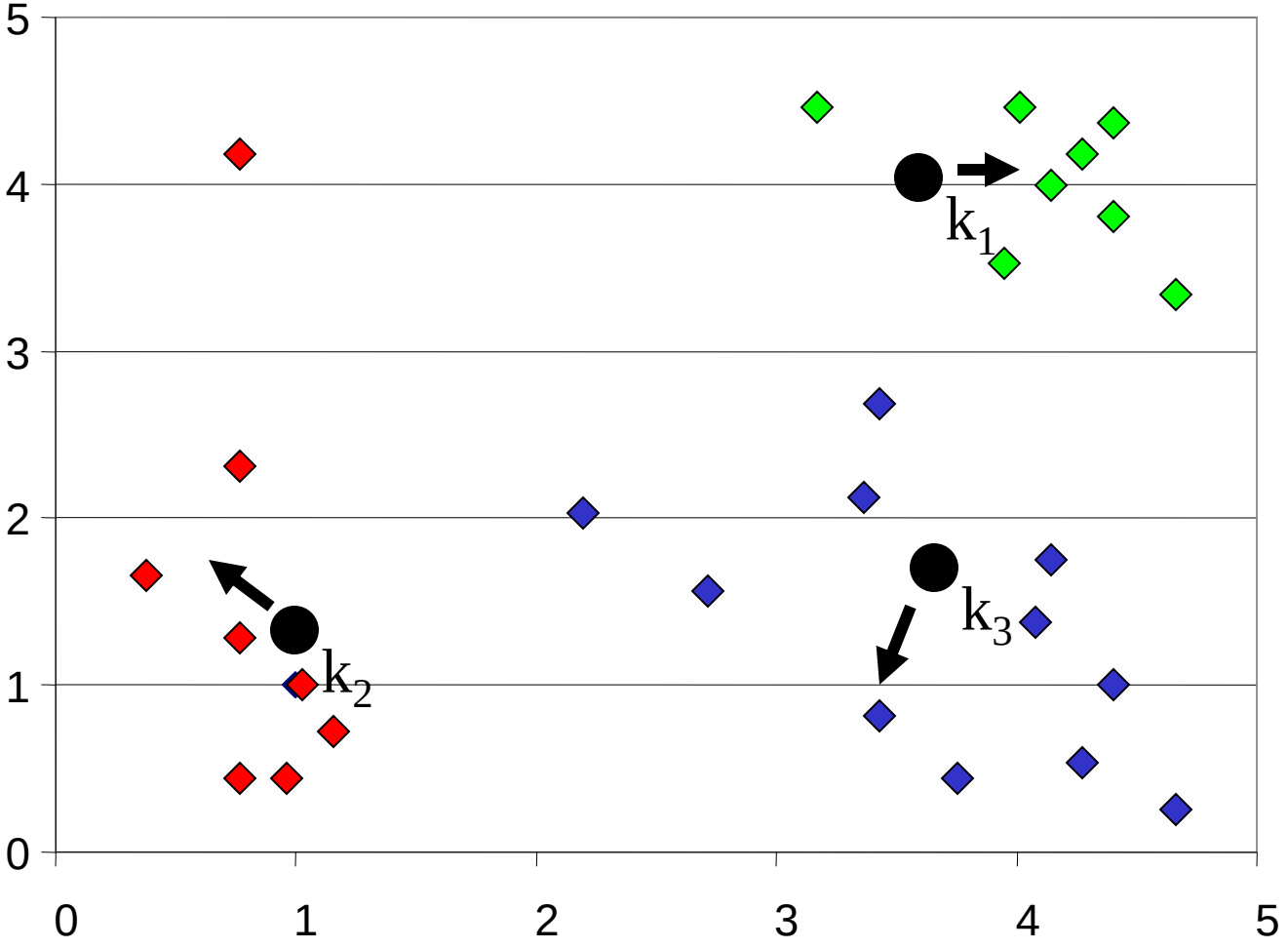
K-means Clustering: Iteration 2

After moving centers, re-assign the objects…



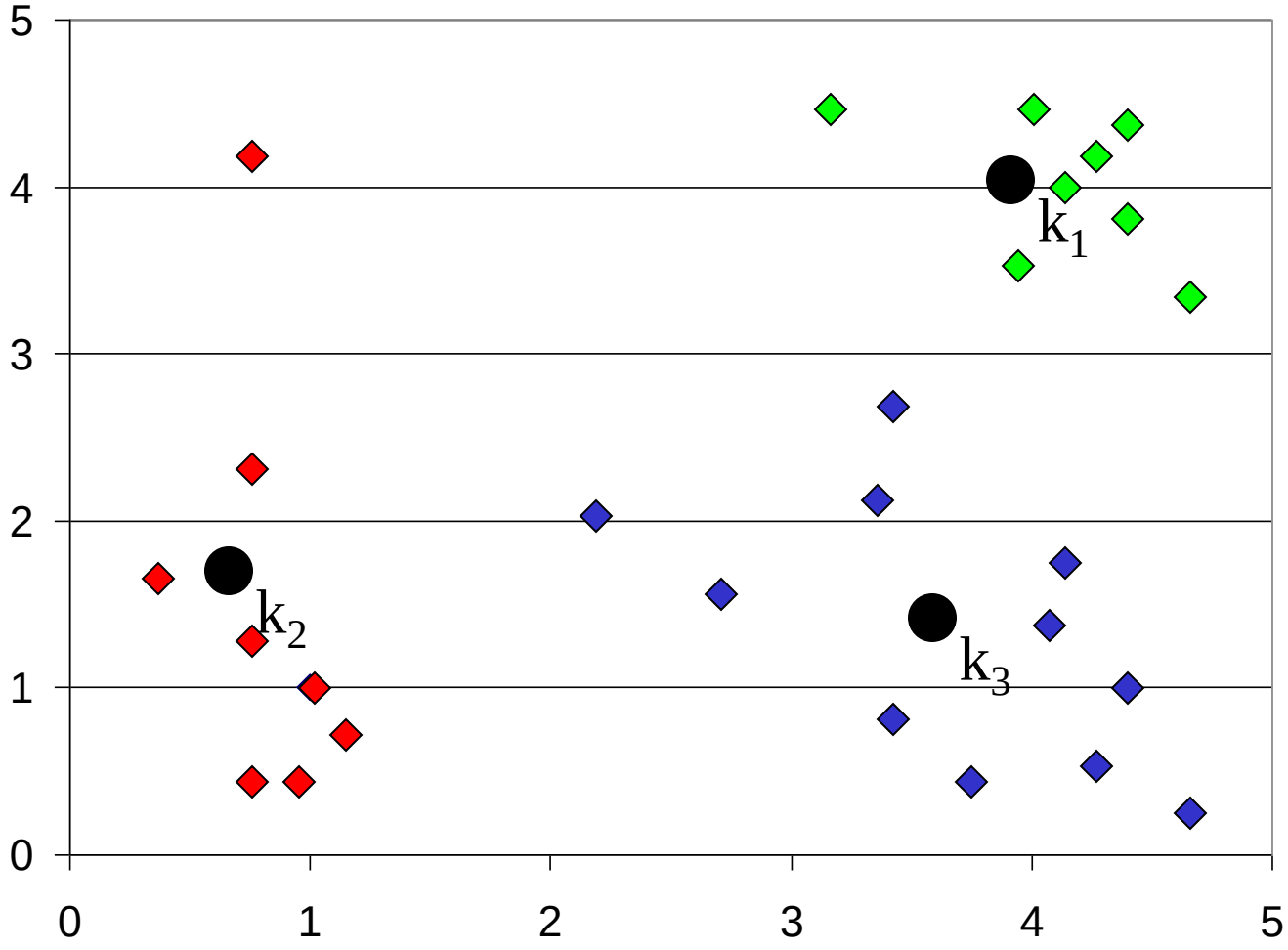
K-means Clustering: Iteration 2

After moving centers, re-assign the objects to nearest centers.
Move a center to the mean of its new members.



K-means Clustering: Finished!

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



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.

Algorithm *k-means*

1. Decide on a value for *K*, the

2. Initialize the *K* cluster cente
necessary).

Use one of the distance / similarity functions we discussed earlier

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

Average / median of class members

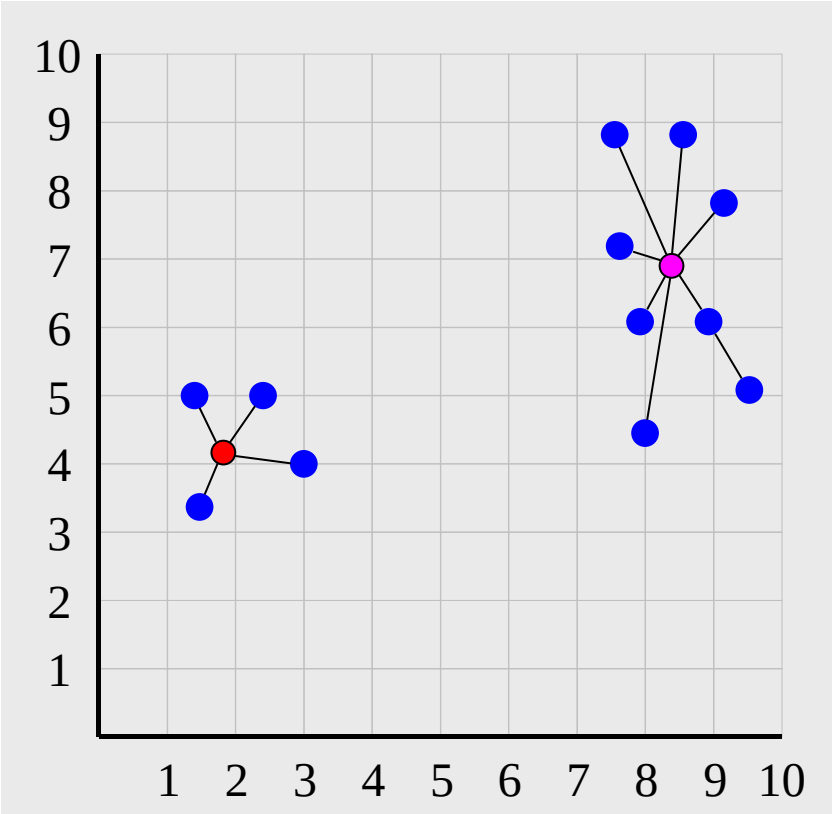
Why K-means Works

- What is a good partition?
- High intra-cluster similarity
- K-means optimizes
  - the average distance to members of the same cluster

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \left\| x_{ki} - x_{kj} \right\|^2$$

  - which is twice the total distance to centers, also called squared error

$$se = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| x_{ki} - \mu_k \right\|^2$$



Summary: *K-Means*

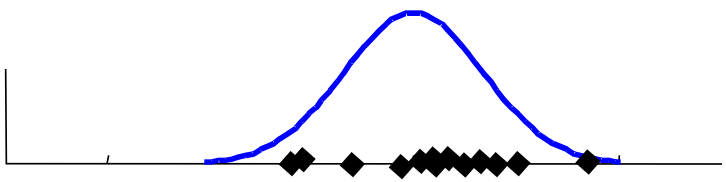
- Strength
  - Simple, easy to implement and debug
  - Intuitive objective function: optimizes intra-cluster similarity
  - *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t << n$.
- Weakness
  - Applicable only when *mean* is defined, what about categorical data?
  - Often terminates at a *local optimum*. Initialization is important.
  - Need to specify K, the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*
- Summary
  - Assign members based on current centers
  - Re-estimate centers based on current assignment

Outline

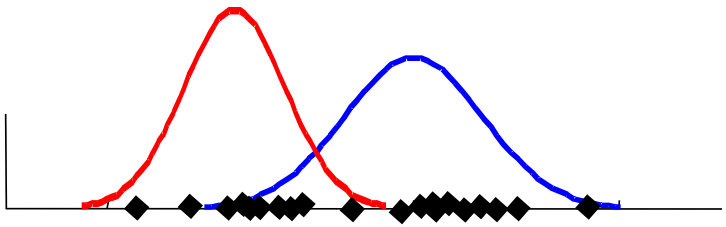
- Motivation
- Distance measure
- Hierarchical clustering
- Partitional clustering
  - K-means
  - Gaussian Mixture Models
  - Number of clusters

Gaussian Mixture Models

- Gaussian $$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\eta)^2}{2\sigma^2}}$$
  - ex. height of one population



- Gaussian Mixture: Generative modeling framework



$$P(C=i) = w_i \quad , \quad P(x \mid C=i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\eta_i)^2}{2\sigma_i^2}}$$

$$P(x \mid \Theta) = \sum_i P(C=i, x \mid \Theta) = \sum_i P(x \mid C=i, \Theta) P(C=i \mid \Theta) =$$

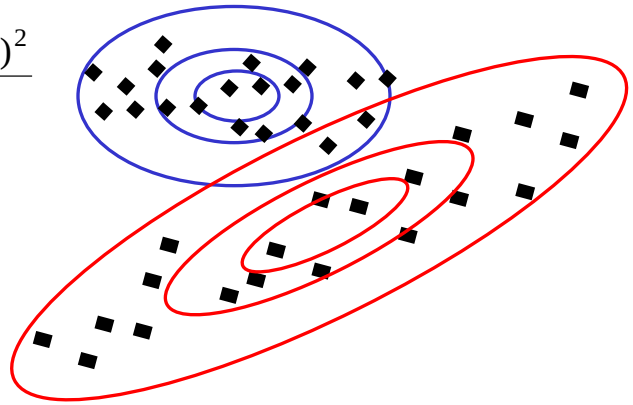
$$\sum_i w_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\eta_i)^2}{2\sigma_i^2}}$$

Likelihood of a data point given the model

Gaussian Mixture Models

- Mixture of Multivariate Gaussian

$$P(C=i)=w_i \qquad P(x\,|\,C=i)=\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x-\eta_i)^2}{2\sigma_i^2}}$$



  – ex. y-axis is blood pressure and x-axis is age

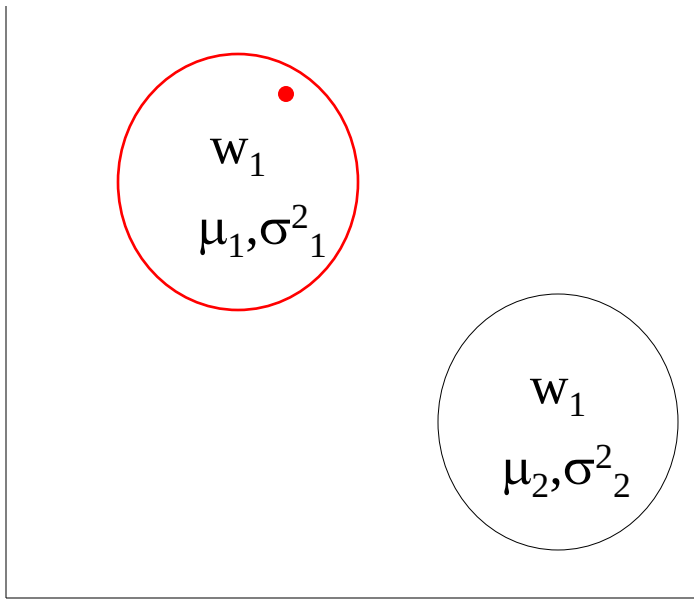
GMM: A generative model

$$\sum_i w_i = 1$$

- Assuming we know the number of components (k), their weights (w_i) and parameters (μ_i, $\sum_i$) we can generate new instances from a GMM in the following way:

  - Pick one component at random with probability w_i for each component

  - Sample a point x from $N(\mu_i, \sum_i)$



Estimating model parameters

- We have a weight, mean and covariance parameters for each class
- As usual we can write the likelihood function for our model

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

GMM+EM = "Soft K-means"

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

One for each cluster

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

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

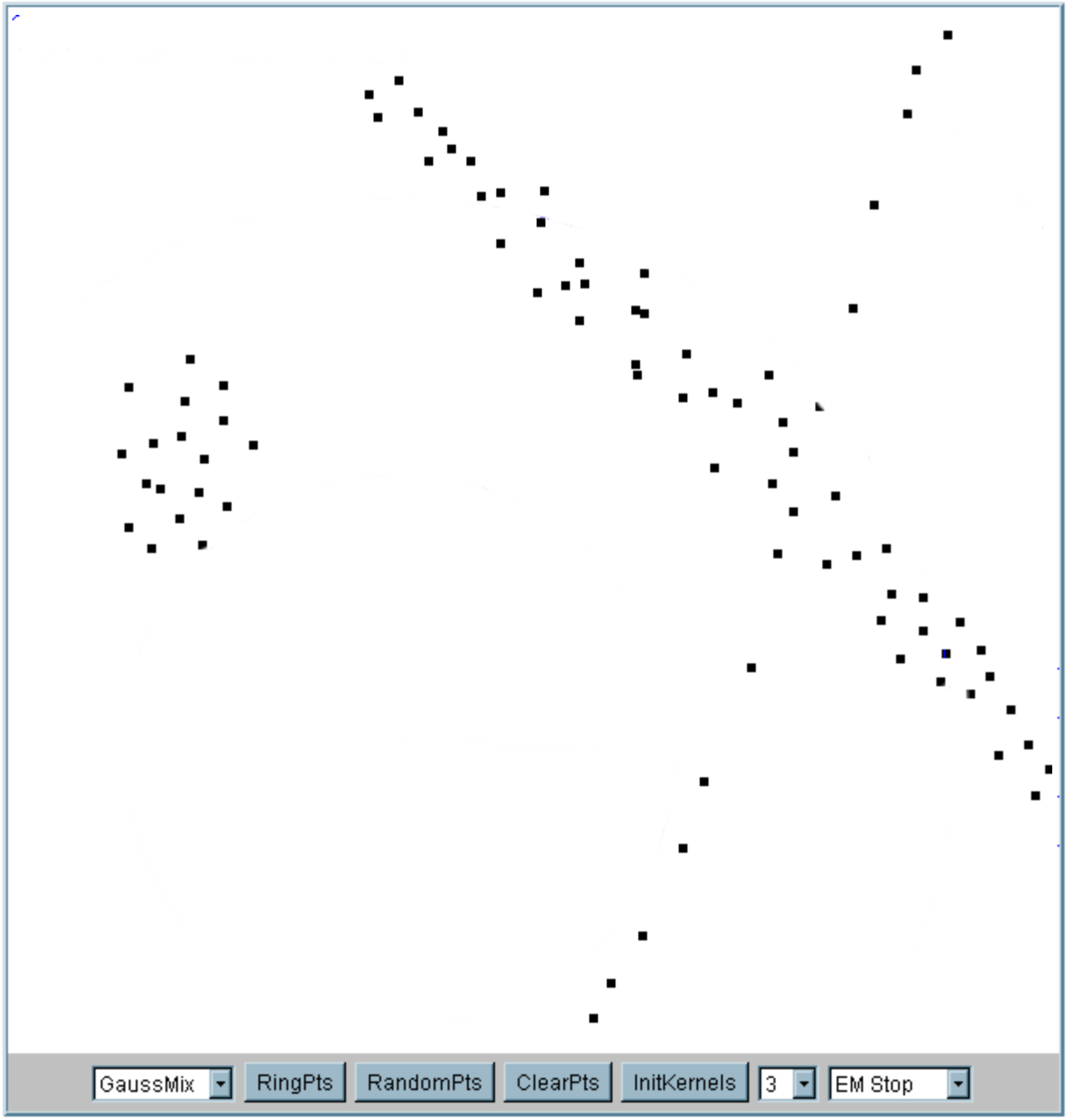
- M-step: re-estimate parameters based on *probabilistic* membership

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

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

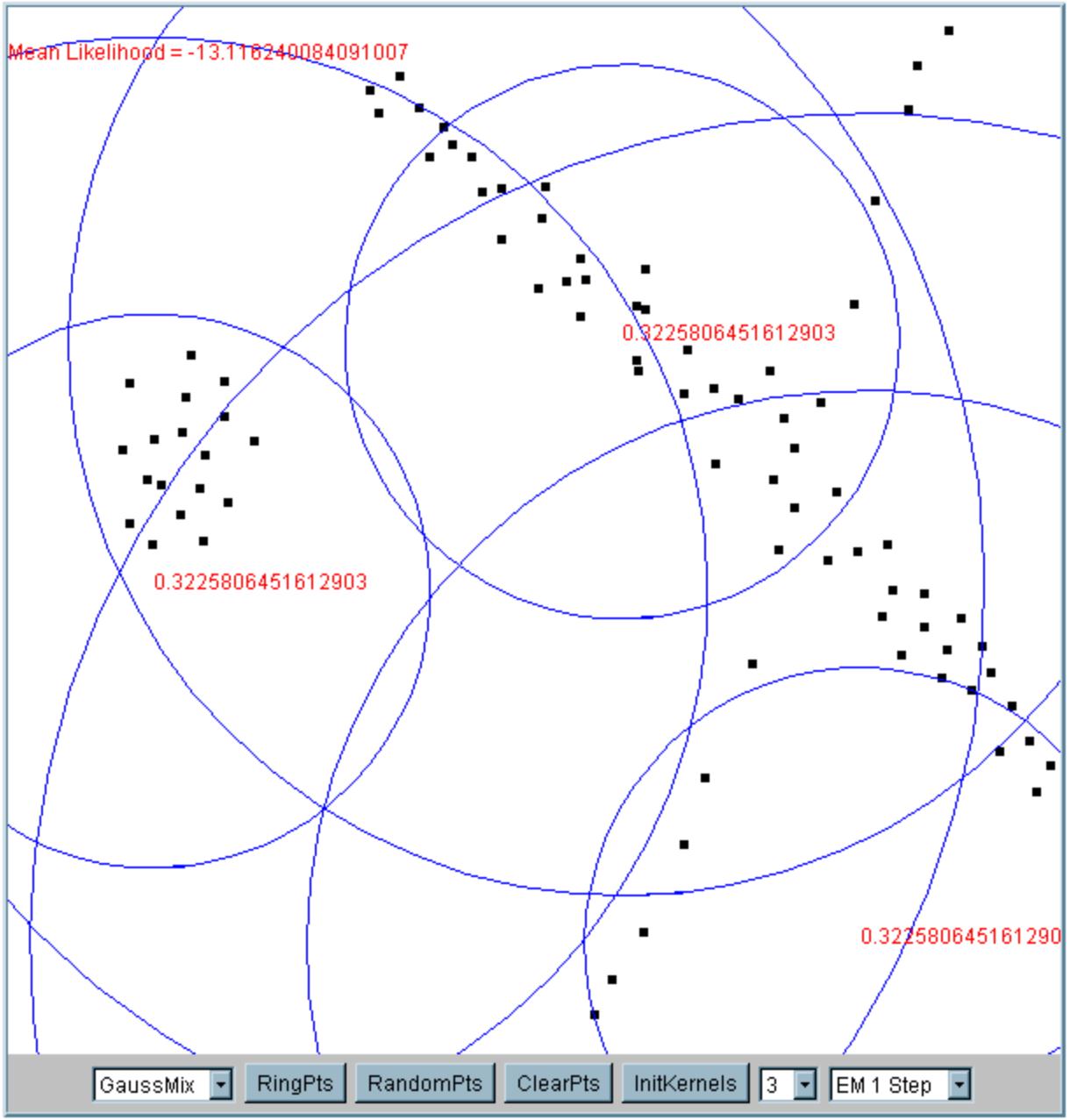
$$w_i = \frac{p_i}{\sum_j p_j}$$

- Repeat until change in parameters are smaller than a threshold

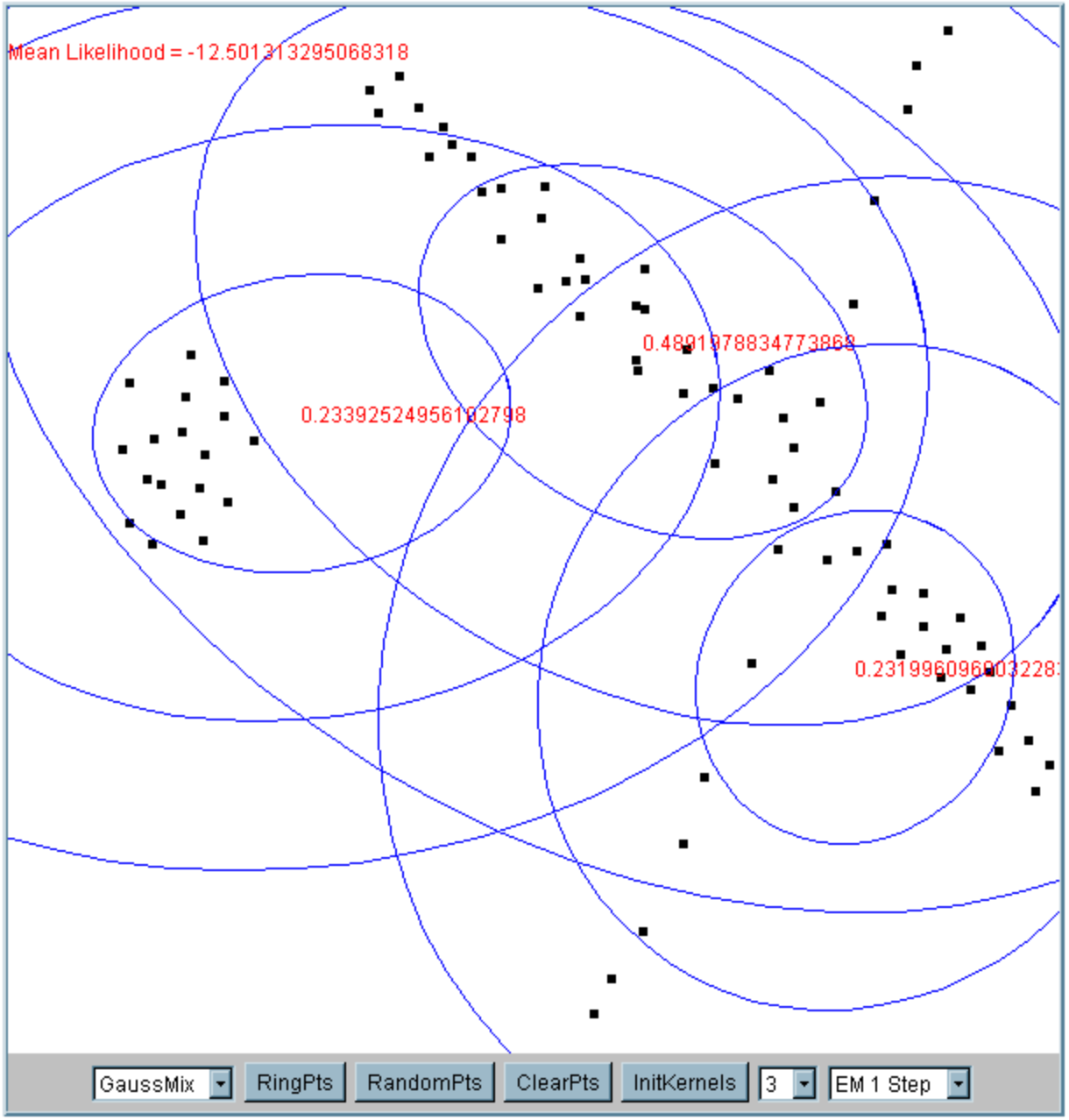

GaussMix
RingPts
RandomPts
ClearPts
InitKernels
3
EM Stop

Iteration 1

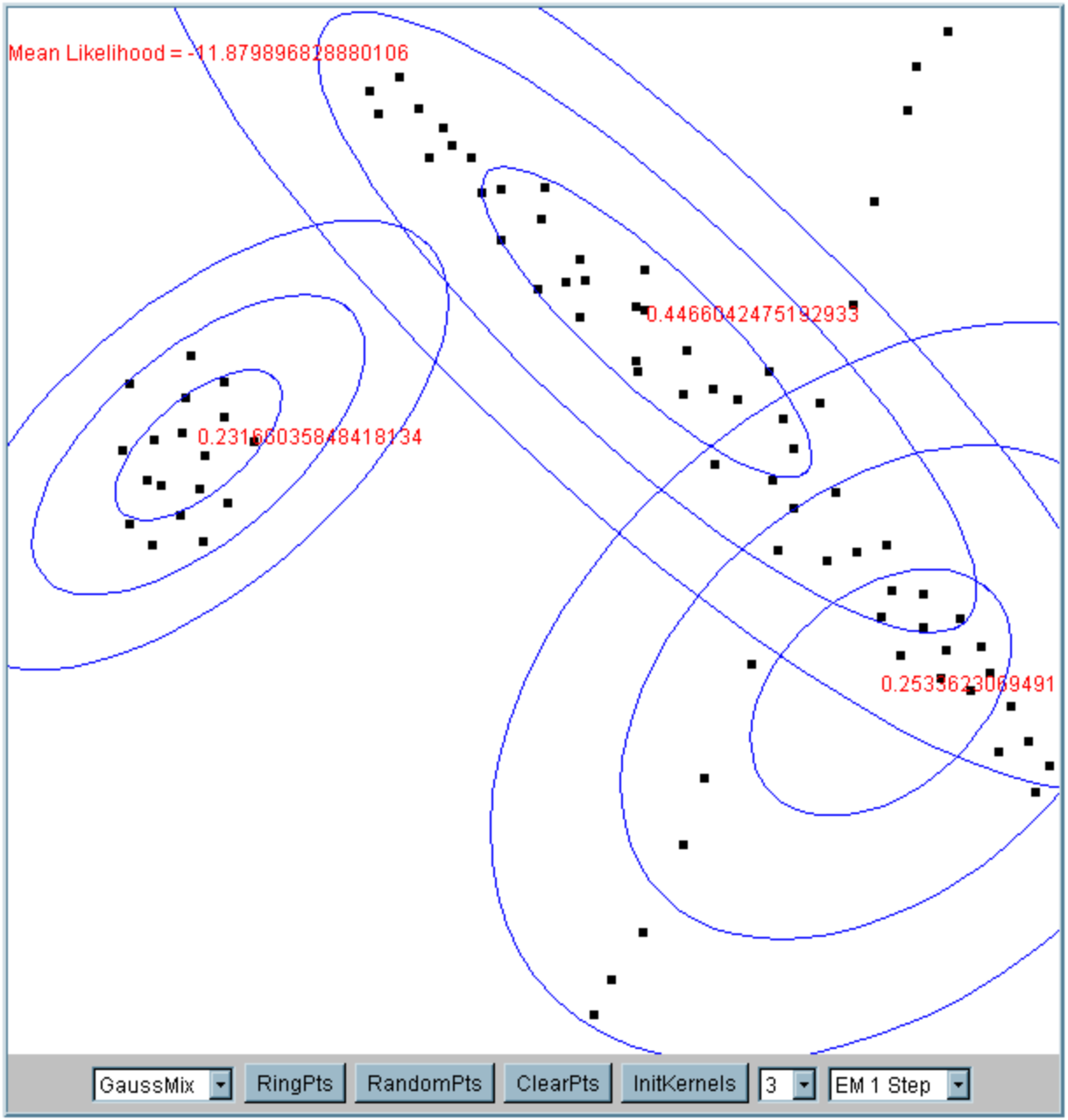
The cluster means are randomly assigned



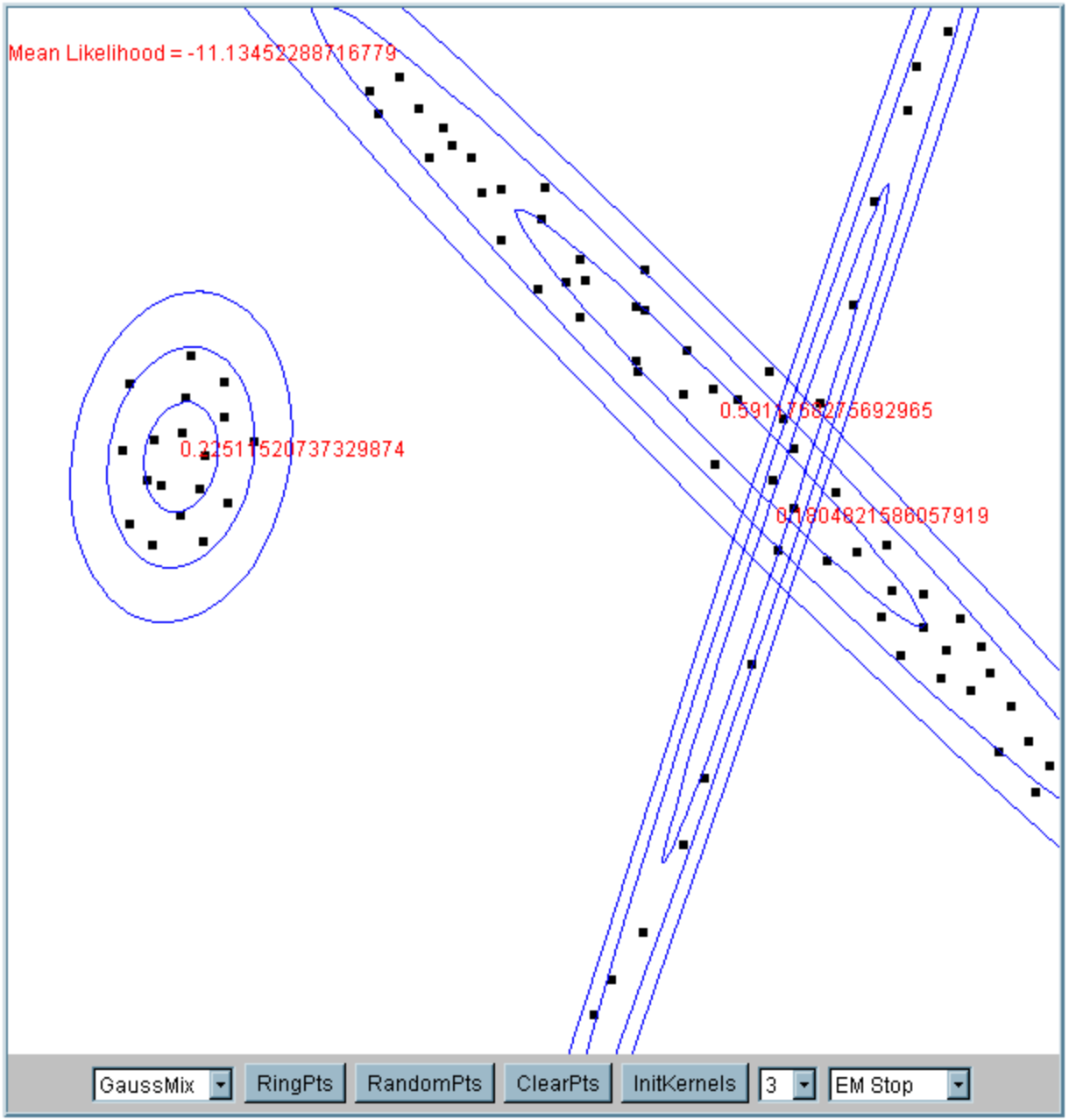
Iteration 2


Mean Likelihood = -12.501313295068318
0.23392524956102798
0.4891878834773868
0.23199609680322811
GaussMix
RingPts
RandomPts
ClearPts
InitKernels
3
EM 1 Step

Iteration 5


Mean Likelihood = -11.879896828880106
0.4466042475192933
0.2316803584841813 4
0.2533623069491
GaussMix
RingPts
RandomPts
ClearPts
InitKernels
3
EM 1 Step

Iteration 25



Strength of Gaussian Mixture Models

- *Interpretability*: learns a generative model of each cluster
  - you can generate new data based on the learned model
- *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t << n$.
- Intuitive (?) objective function: optimizes data likelihood

Weakness of Gaussian Mixture Models

- Often terminates at a *local optimum*. Initialization is important.
- Need to specify *K*, the *number* of clusters, in advance
- Not suitable to discover clusters with *non-convex shapes*

- Summary
  - To learn Gaussian mixture, assign probabilistic membership based on current parameters, and re-estimate parameters based on current membership

Algorithm: K-means and GMM

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

2. Initialize the *K* cluster centers / parameters (randomly).

K-means

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.

GMM

3. E-step: assign *probabilistic* membership

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

5. Repeat 3 and 4 until parameters do not change.

Clustering methods: Comparison

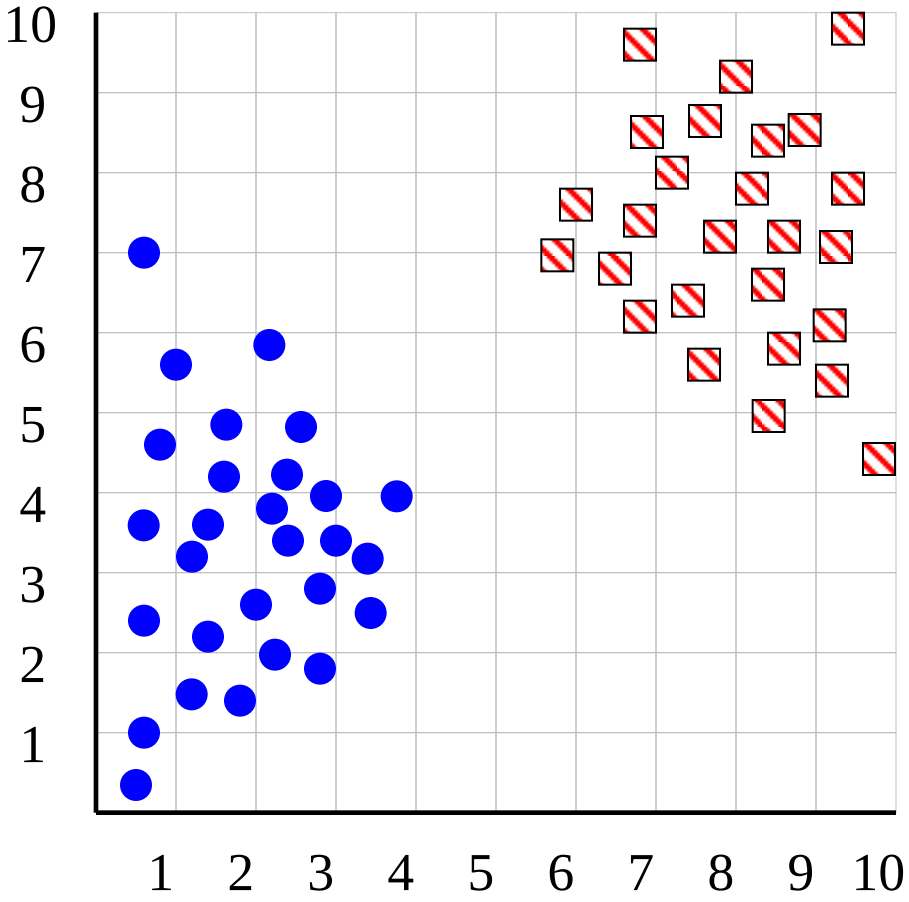
	Hierarchical	**K-means**	**GMM**
Running time	naively, O(N^3)	fastest (each iteration is linear)	fast (each iteration is linear)
Assumptions	requires a similarity / distance measure	strong assumptions	strongest assumptions
Input parameters	none	*K* (number of clusters)	*K* (number of clusters)
Clusters	subjective (only a tree is returned)	exactly *K* clusters	exactly *K* clusters

Outline

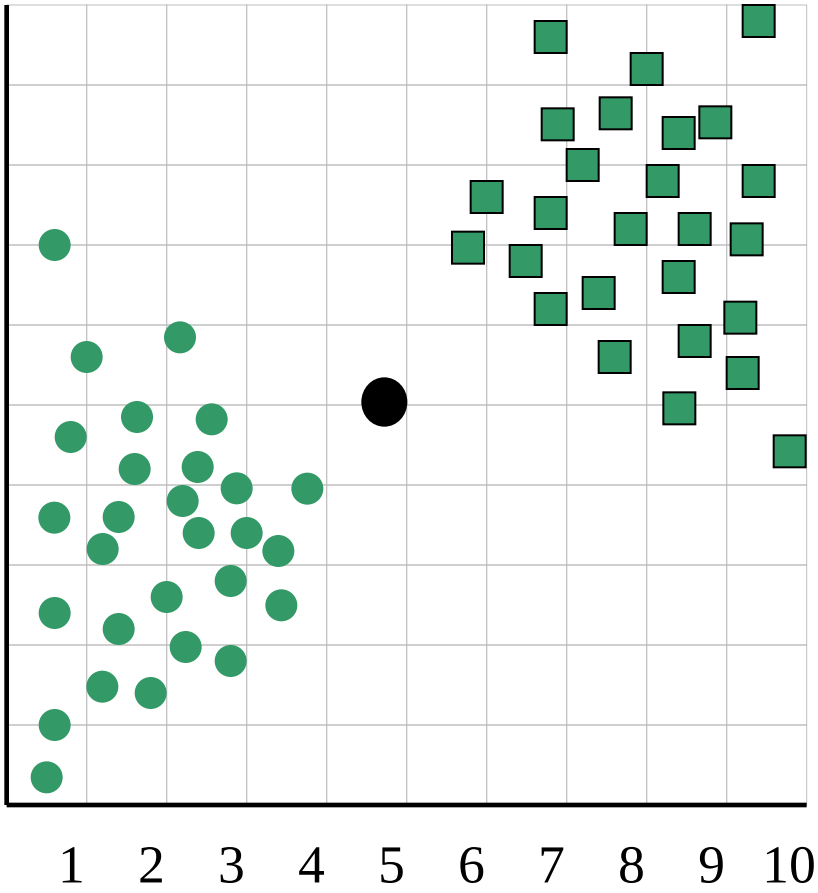
- Motivation
- Distance measure
- Hierarchical clustering
- Partitional clustering
  - K-means
  - Gaussian Mixture Models
  - Number of clusters

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

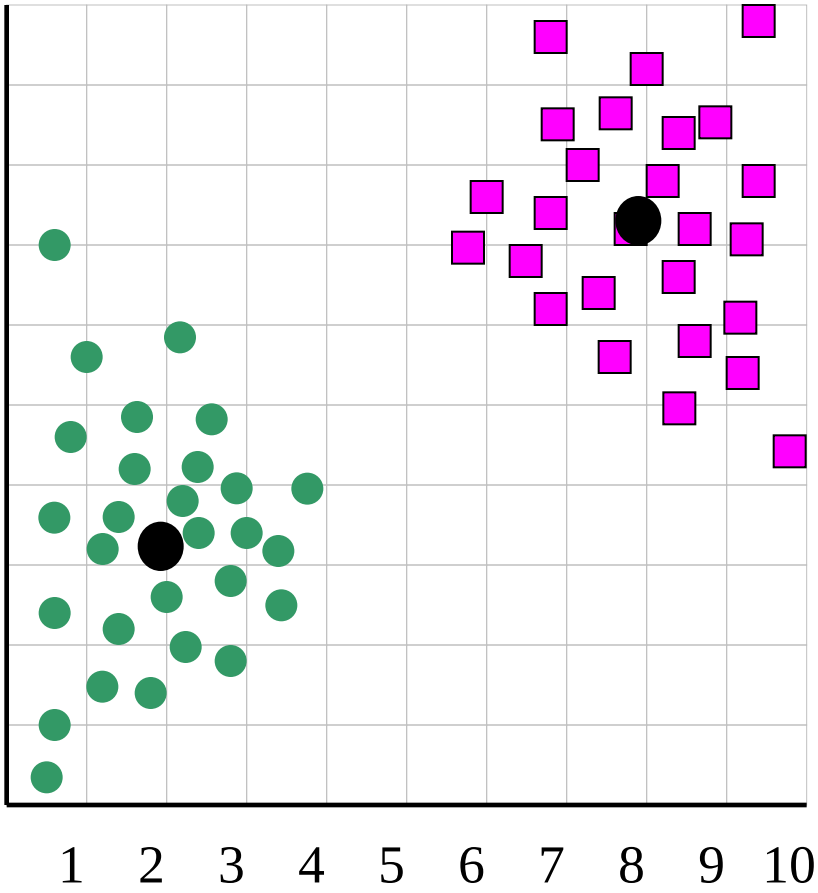
In general, this is a unsolved problem. However there are many approximate methods. In the next few slides we will see an example.



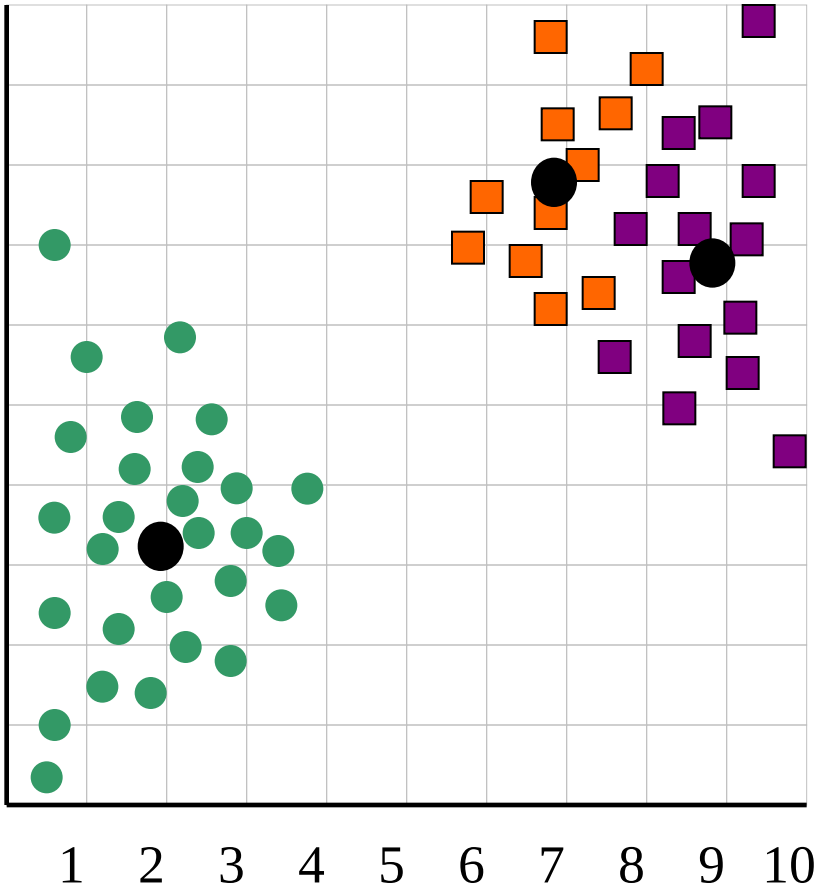
When k = 1, the objective function is 873.0



When k = 2, the objective function is 173.1

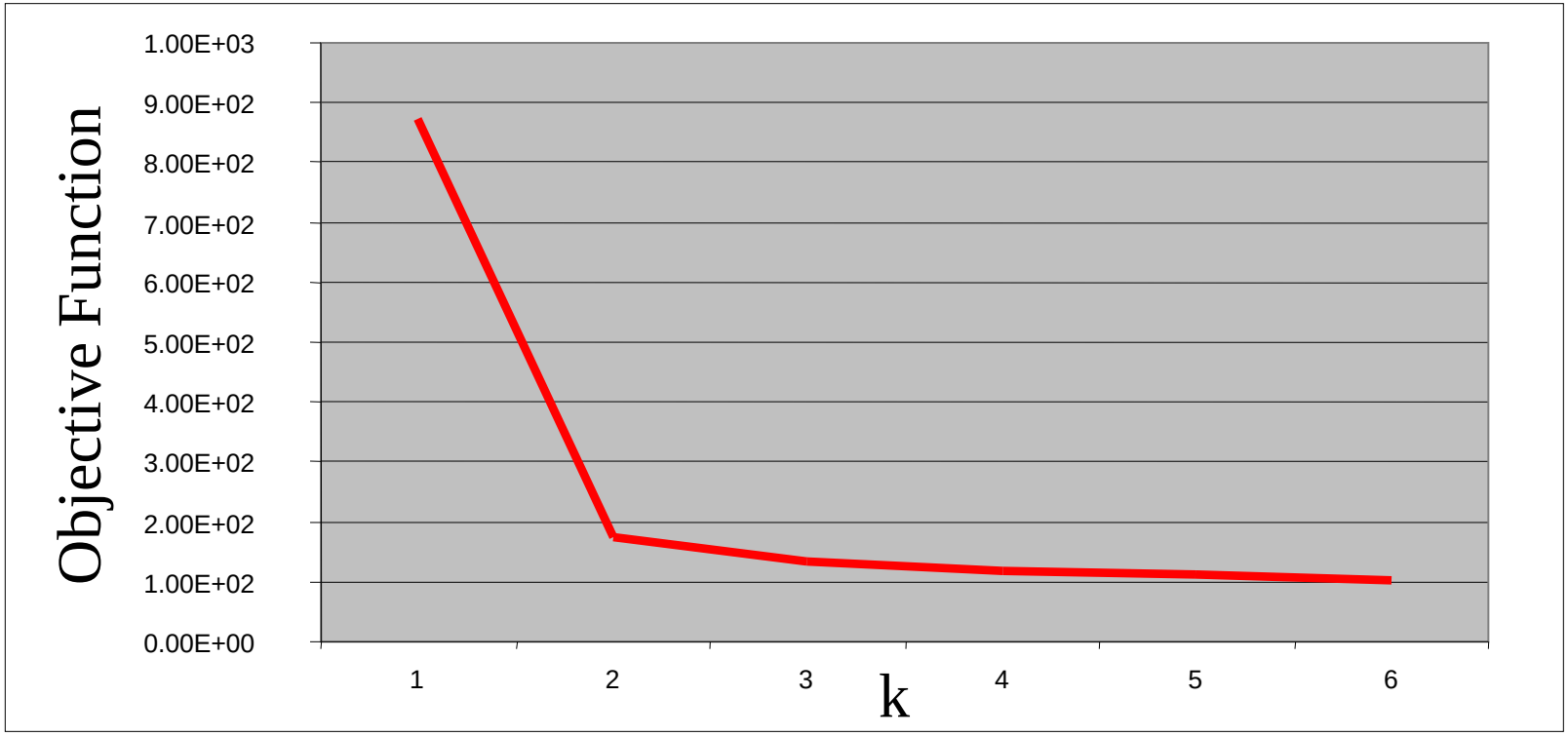


When k = 3, the objective function is 133.6



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”.

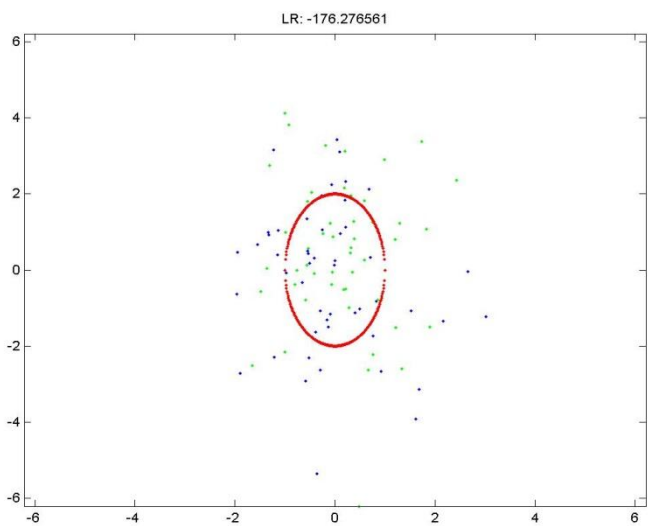


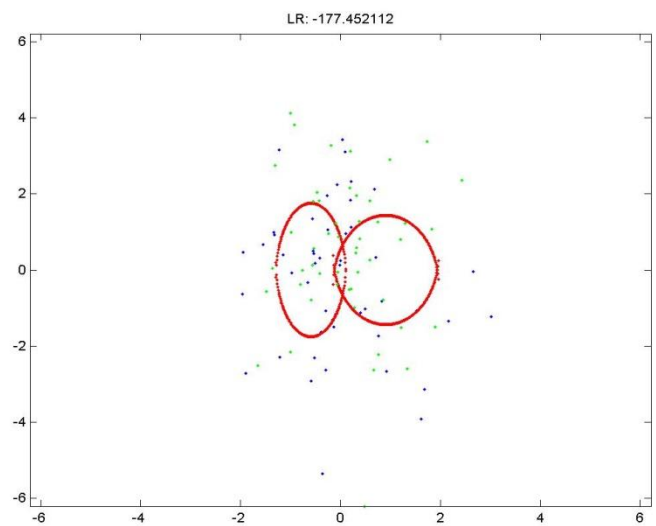
Note that the results are not always as clear cut as in this toy example

Cross validation

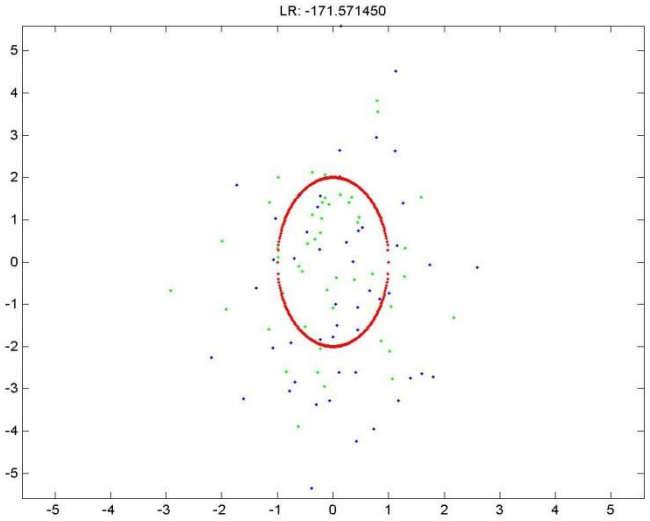
- We can also use cross validation to determine the correct number of classes
- Recall that GMMs is a generative model. We can compute the likelihood of the left out data to determine which model (number of clusters) is more accurate

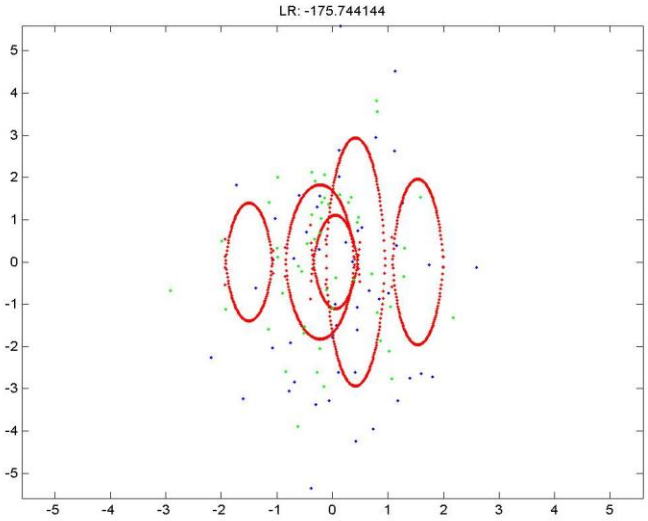
$$p(x_1 \cdots x_n \mid \theta) = \prod_{j=1}^{n} \left(\sum_{i=1}^{k} p(x_j \mid C = i) w_i \right)$$





Cross validation





Cluster validation

- We wish to determine whether the clusters are real or compare different clustering methods.
  - 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 objects 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 objects
- C' denotes the set of k clusters resulting from a randomly chosen subset (80-90%) of objects
- 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 objects common to both

External validation

- For this we need an external source that contains related, but usually not identical information.
- For example, assume we are clustering web pages based on the car pictures they contain.
- We have independently grouped these pages based on the text description they contain.
- Can we use the text based grouping to determine how well our clustering works?

External validation

- Suppose we have generated k clusters $C_1,\ldots,C_k$. 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.

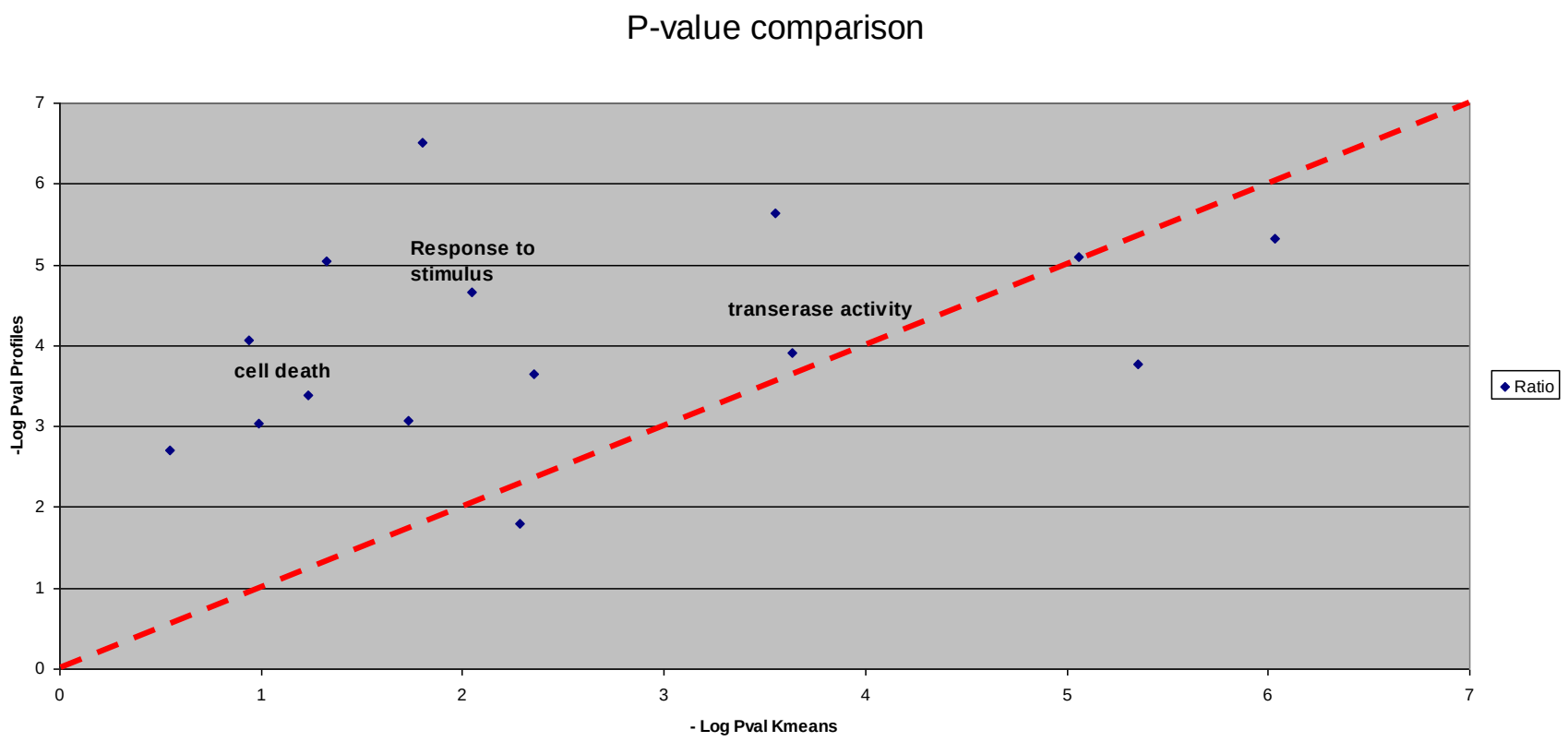
P-value (cont.)

- 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
- Unsolved issues: number of clusters, initialization, etc.