flusser@utia.cas.cz

www.utia.cas.cz/people/flusser

Prof. Ing. Jan Flusser, DrSc.

Lecture 4 – Clustering

Unsupervised Classification

(Cluster analysis)

Training set is not available, No. of classes may not be *a priori* known

What are clusters?

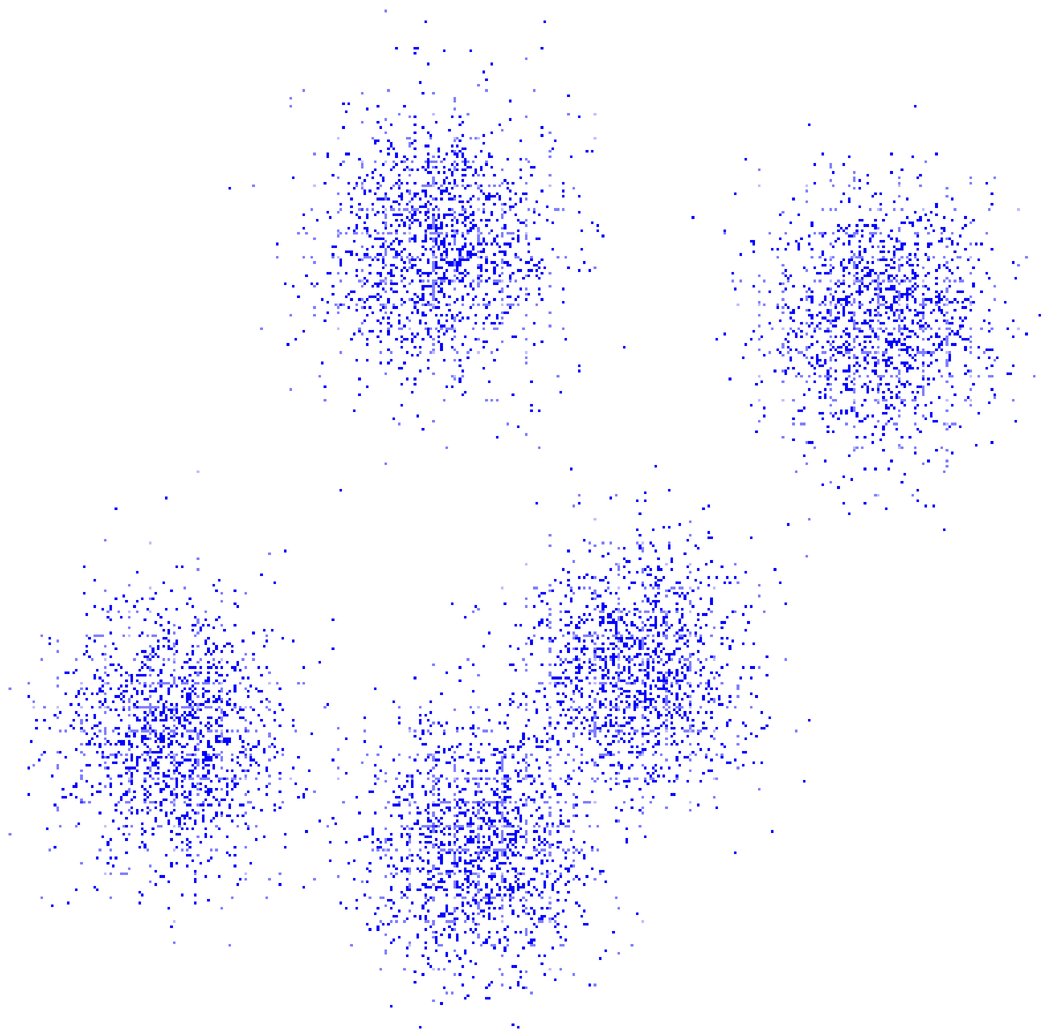
Intuitive meaning

- compact, well-separated subsets

Formal definition

- many attempts, mostly unsatisfactory
  (*t*-connectivity, diameter constraint, ...)
- any partition of the data into disjoint subsets

What are clusters?



How to compare different clusterings? (Ward criterion)

Variance measure J should be minimized

$$J = \sum_{i=1}^{N} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{m}_i\|^2$$

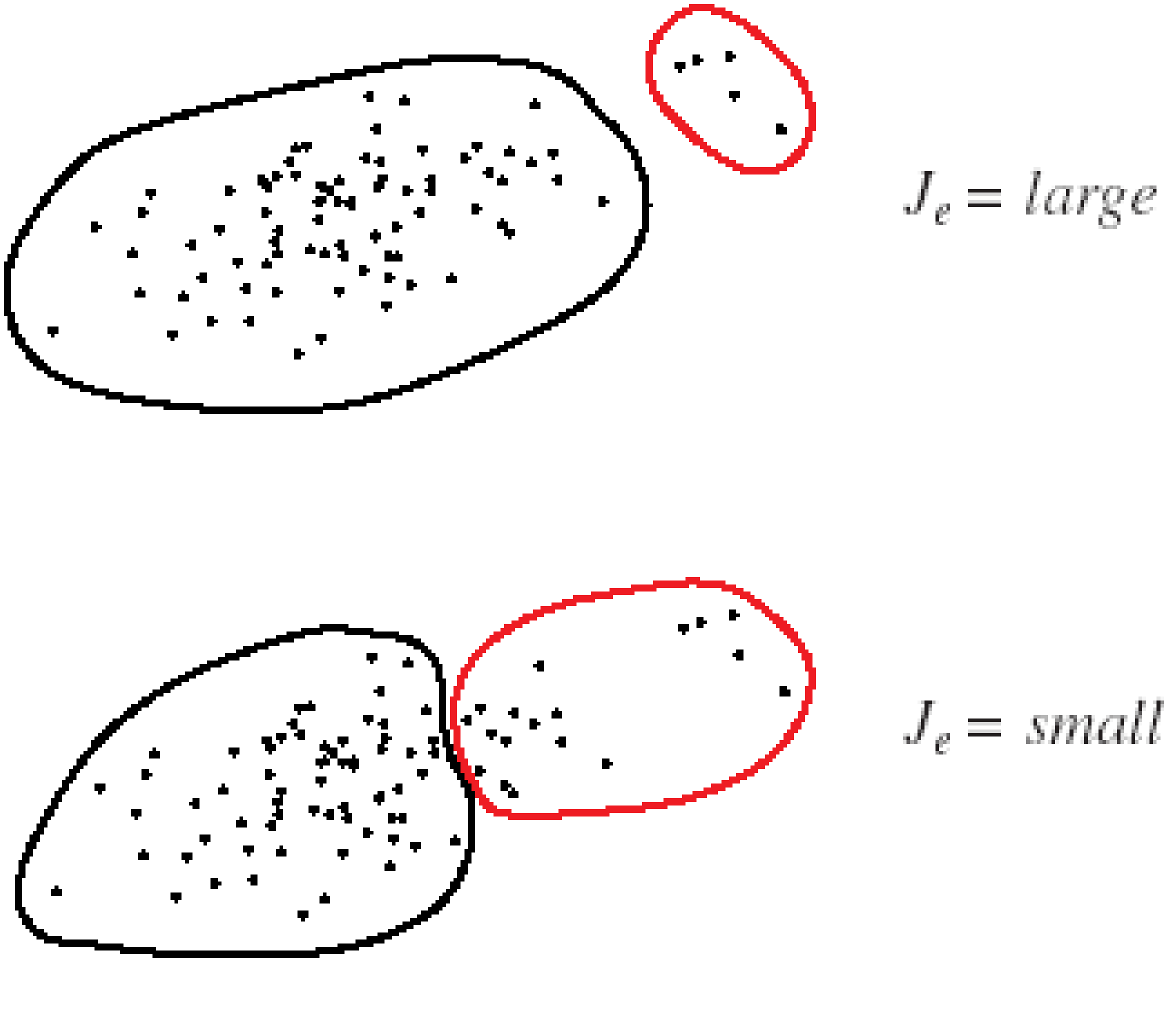
Drawback – only clusterings with the same N can be compared. Global minimum $J = 0$ is reached in the degenerated case.

Minimization of *J*

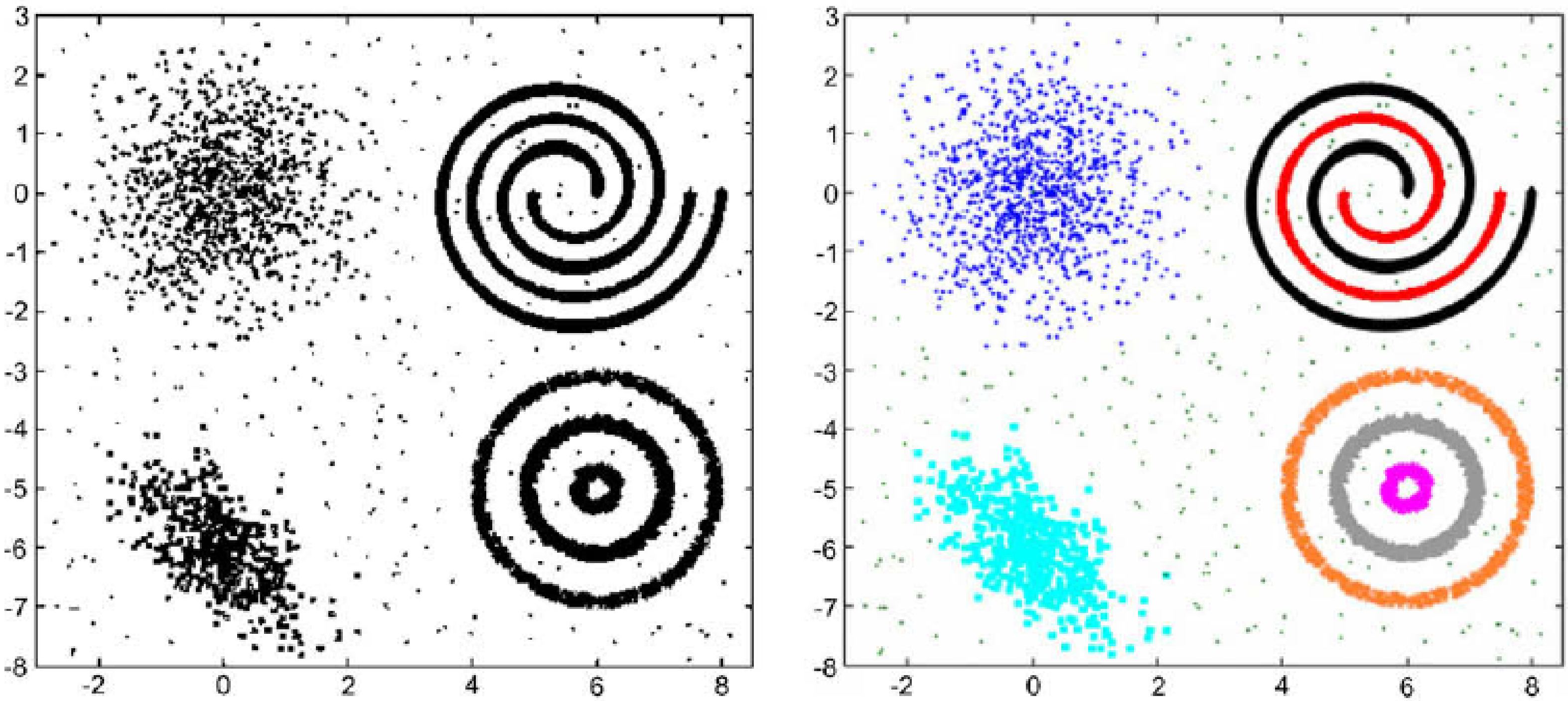
Drawbacks

- The results are sometimes “intuitively wrong” because *J* prefers clusters with approx the same size

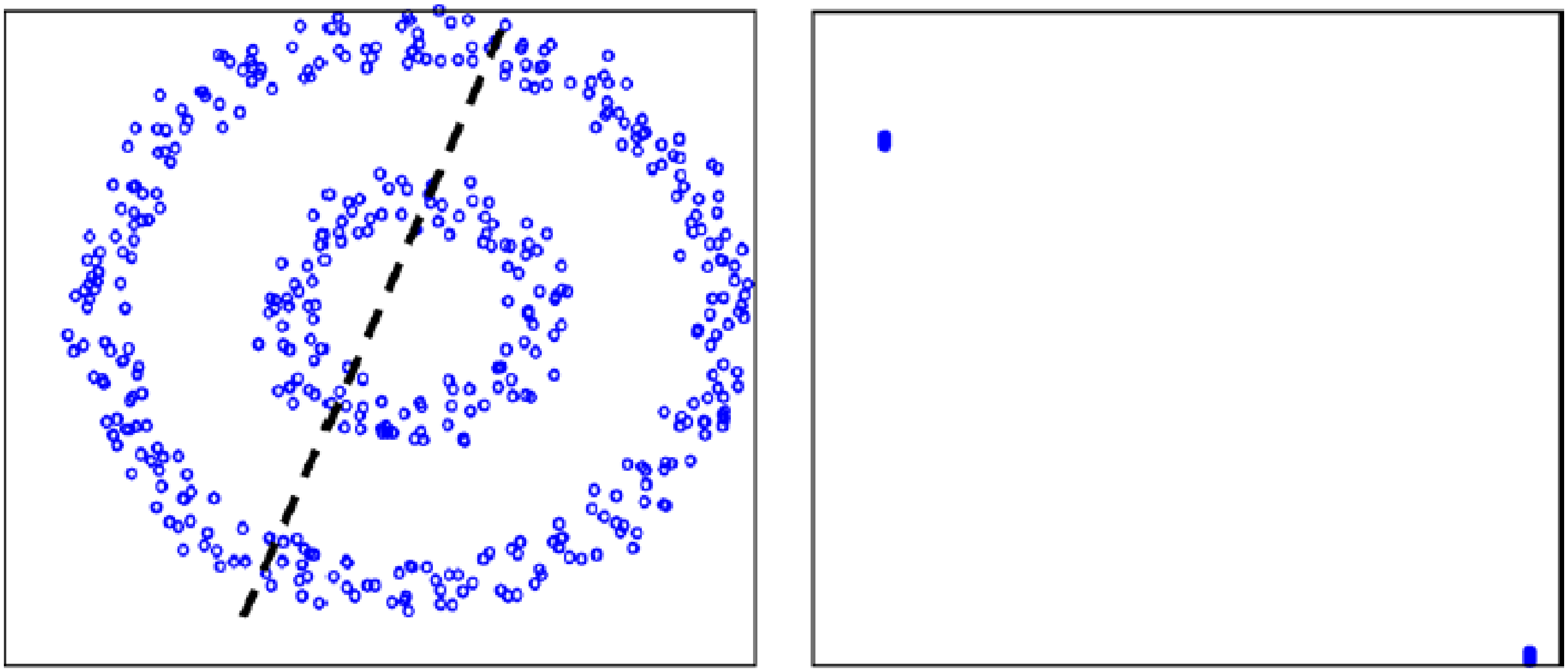
An example of a “wrong” result



Drawback – assumes uncorrelated features and "convex" clusters



Solution – proper transform of the features



Clustering techniques

- **Iterative methods**
  - typically if *N* is given
- **Hierarchical methods**
  - typically if *N* is unknown
- **Other methods**
  - sequential, graph-based, branch & bound, fuzzy, genetic, model-based, etc.

Sequential clustering

- *N* may be unknown
- Very fast but not very good
- Each point is considered only once

Idea: a new point is either added to an existing cluster or it forms a new cluster. The decision is based on the **user-defined distance threshold**.

Sequential clustering

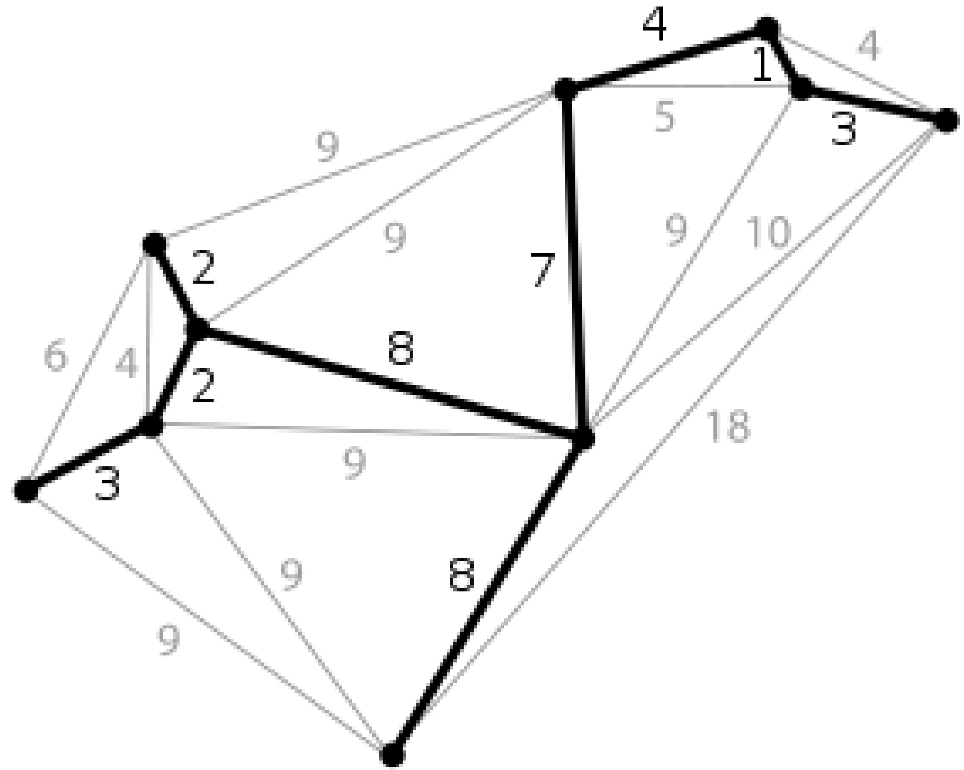
Drawbacks:

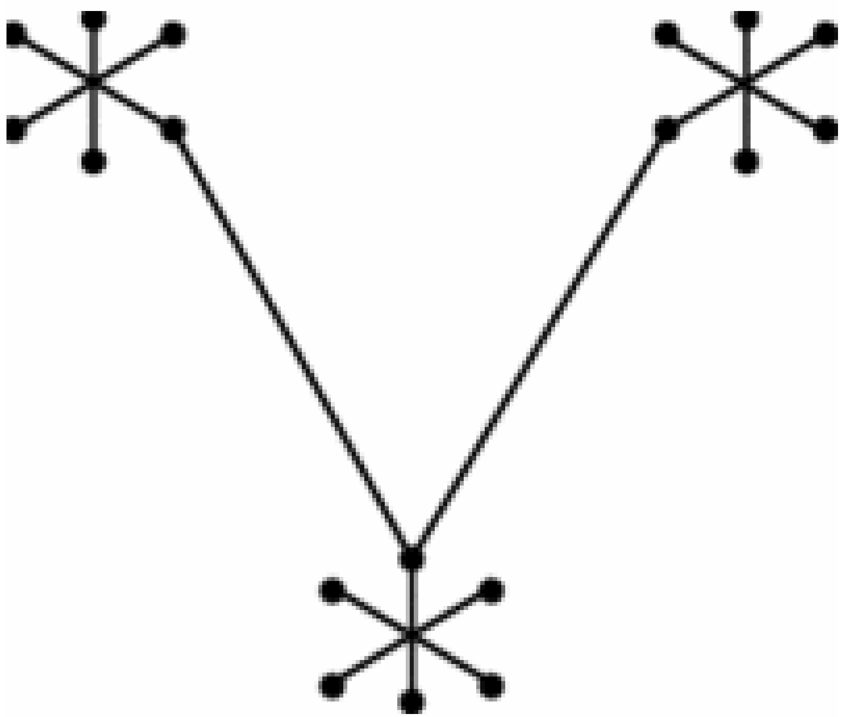
- Dependence on the distance threshold
- Dependence on the order of data points

Graph-based clustering

Idea: Construct a shortest spanning tree and then divide it into clusters by removing some edges.

- Naive approach: remove *N*-1 longest edges
- Better: remove *N*-1 long and **inconsistent** edges


4
1
4
5
3
9
9
10
9
2
7
8
6
4
2
18
9
3
9
8
9



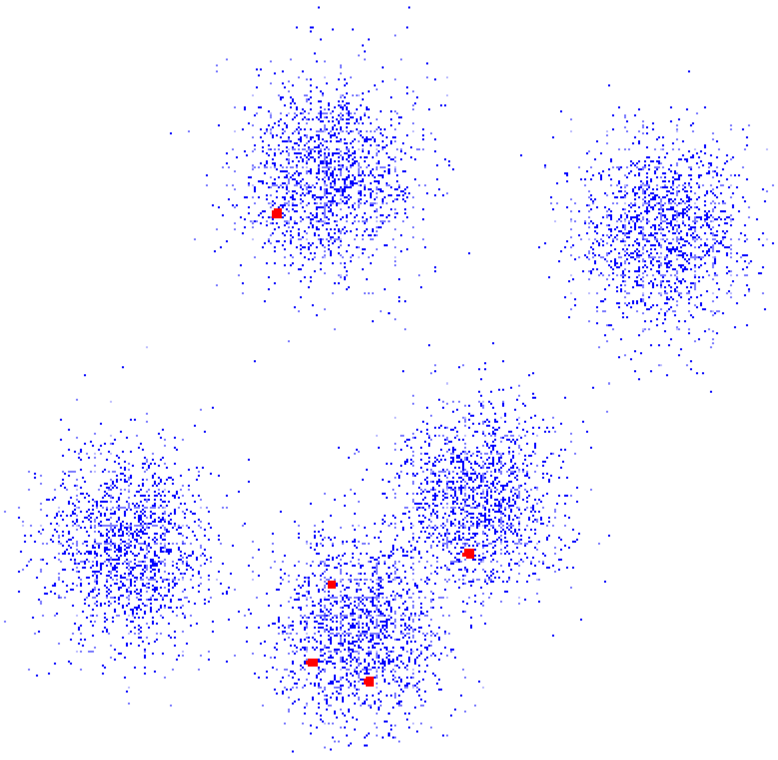
Iterative clustering methods

- ***N*-means clustering**
- **Iterative minimization of *J***
- **ISODATA**

  Iterative Self-Organizing DATa Analysis

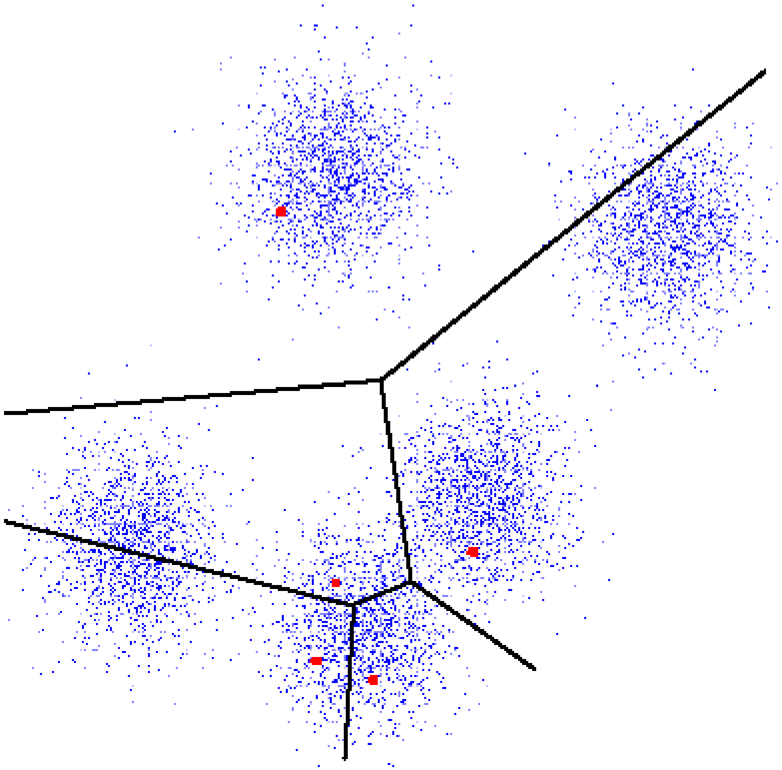
N-means clustering

1. Select *N* initial cluster centroids.



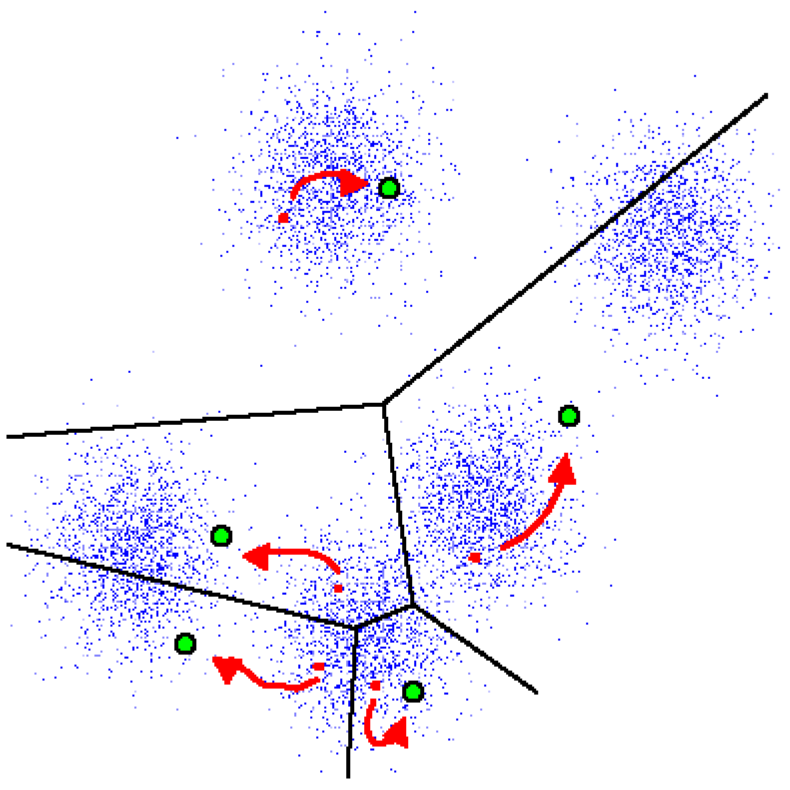
N-means clustering

2. Classify every point *x* according to minimum distance.



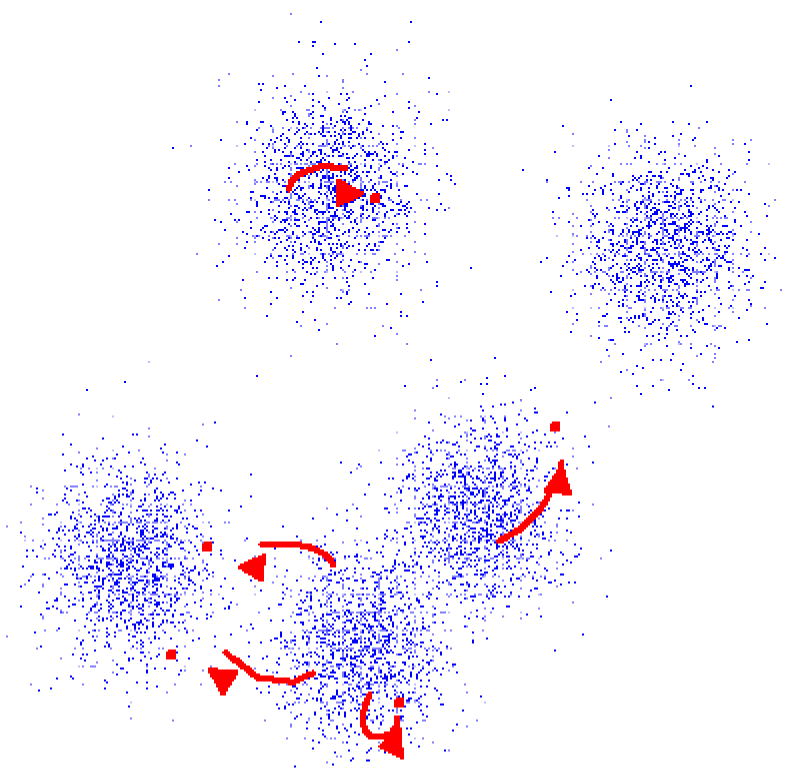
N-means clustering

3. Recalculate the cluster centroids.



N-means clustering

4. If the centroids did not change then STOP else GOTO 2.



N-means clustering

Drawbacks

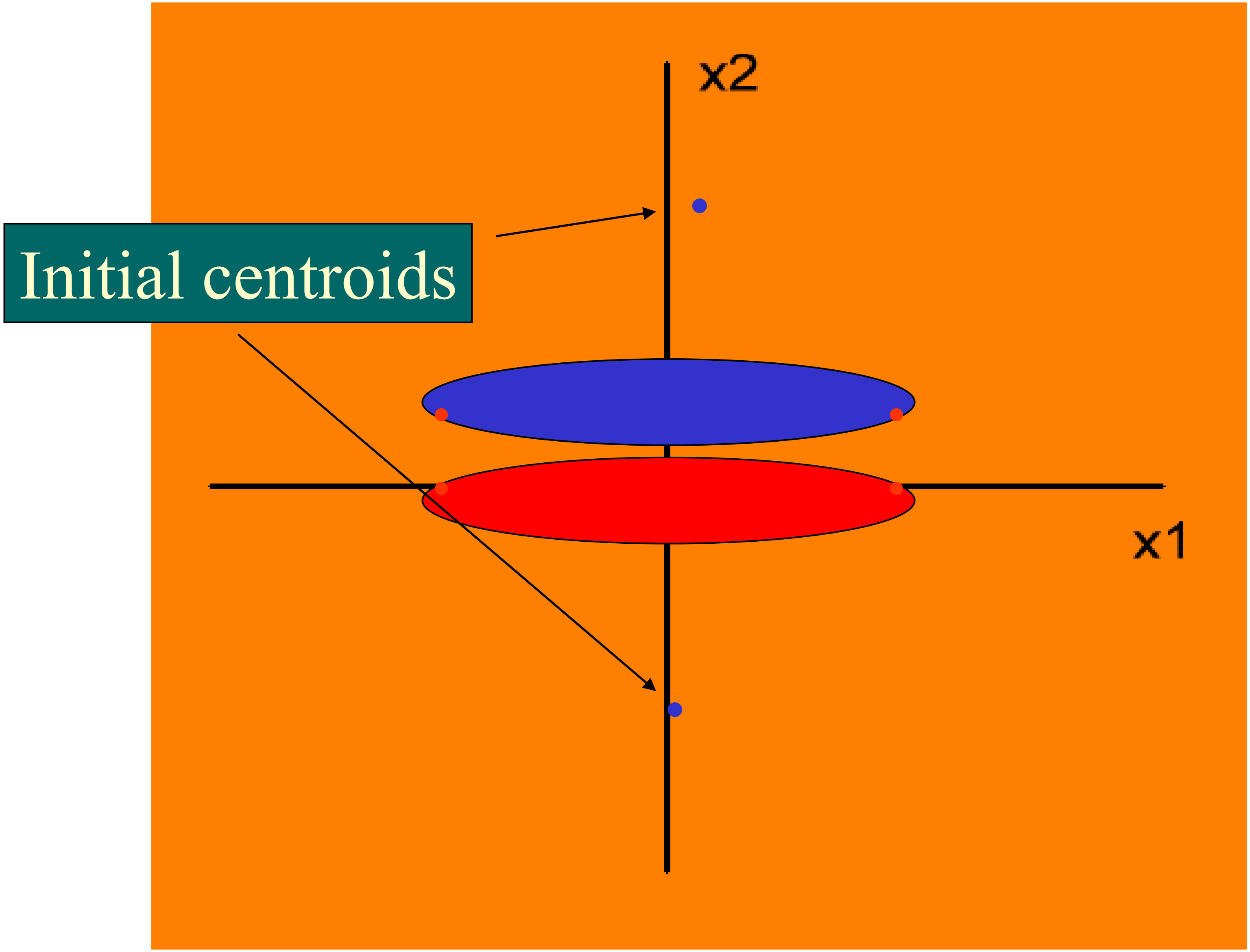
- The result depends on the initialization.
- J is not minimized
- The results are sometimes “intuitively wrong”.

N-means clustering – An example

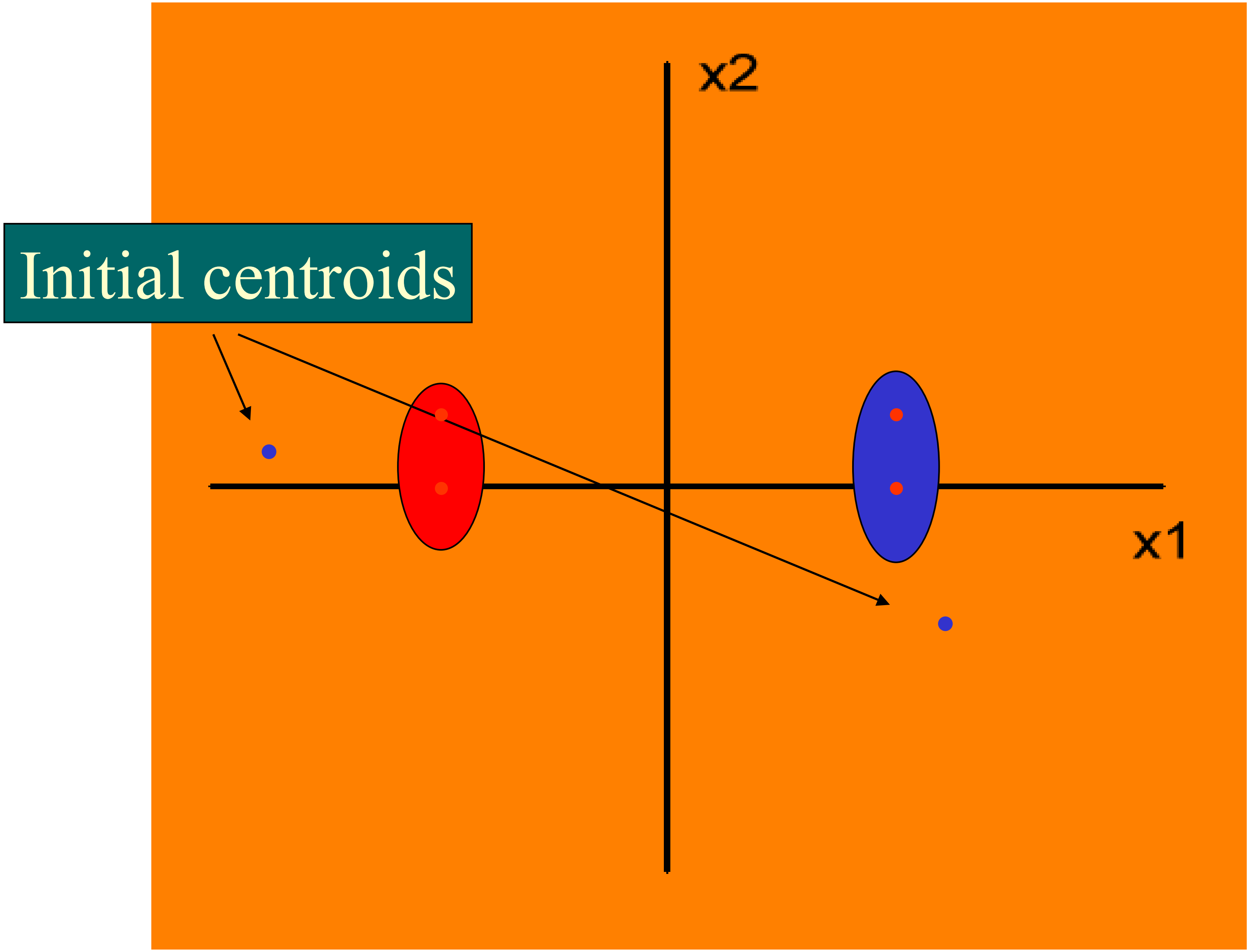
Two features, four points, two clusters ($N = 2$)

Different initializations → different clusterings

N-means clustering – An example



N-means clustering – An example



Iterative minimization of J

1. Let's have an initial clustering (by N-means)
2. For every point x do the following:

   Move x from its current cluster to another cluster, such that the decrease of J is maximized.
3. If all data points do not move, then STOP.

Iterative minimization of *J*

Drawbacks

- The algorithm is optimal in each step but in general global minimum of *J* is not reached.

ISODATA

Iterative clustering, *N* may vary.

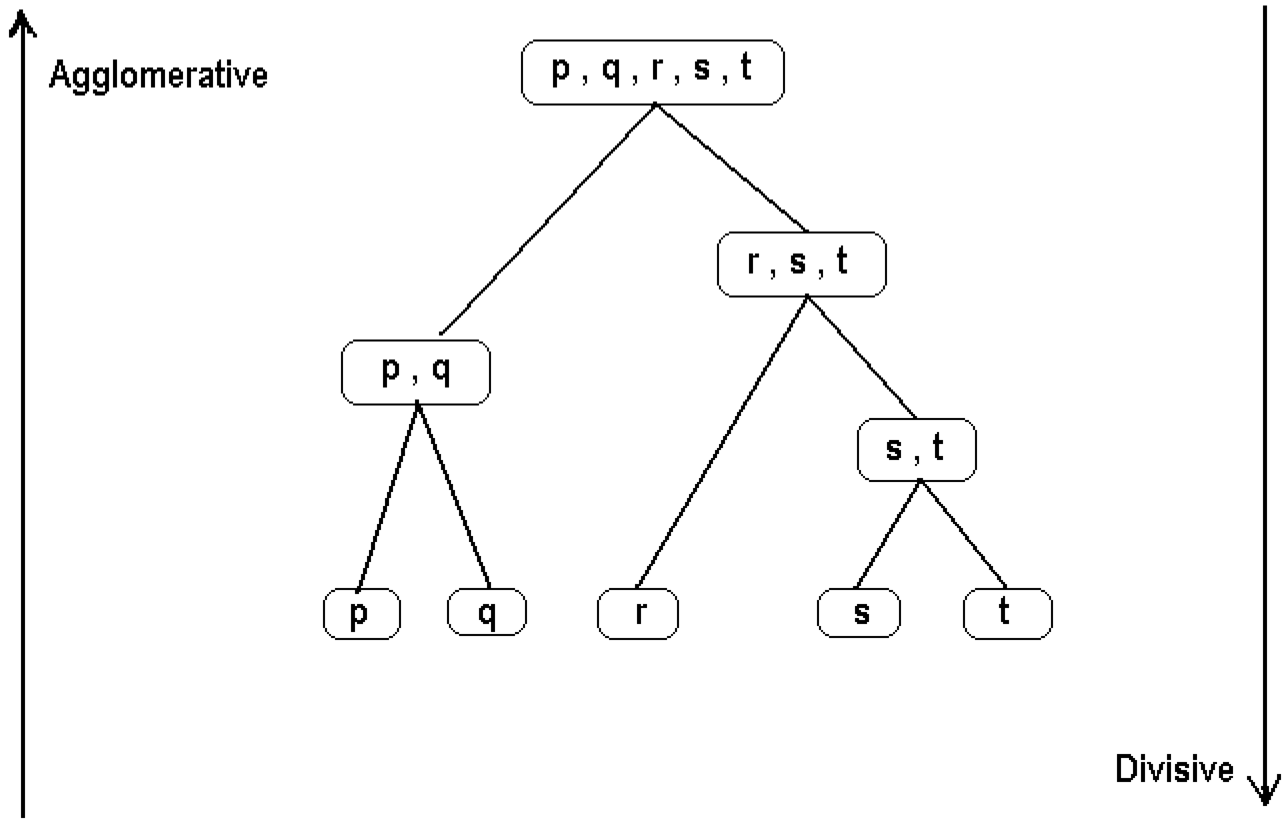
Sophisticated method, a part of many statistical software systems.

Postprocessing after each iteration

- Clusters with few elements are cancelled
- Clusters with big variance are divided
- Other merging and splitting strategies can be implemented

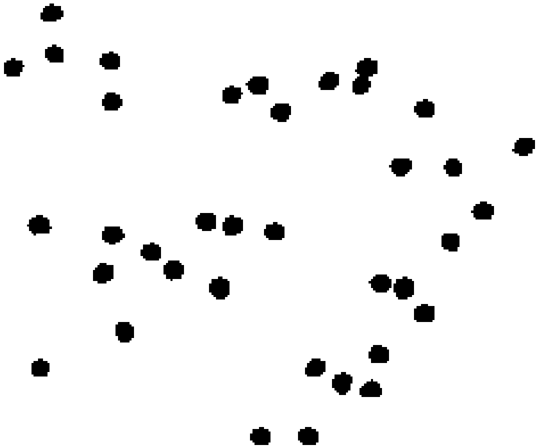
Hierarchical clustering methods

- Agglomerative clustering
- Divisive clustering



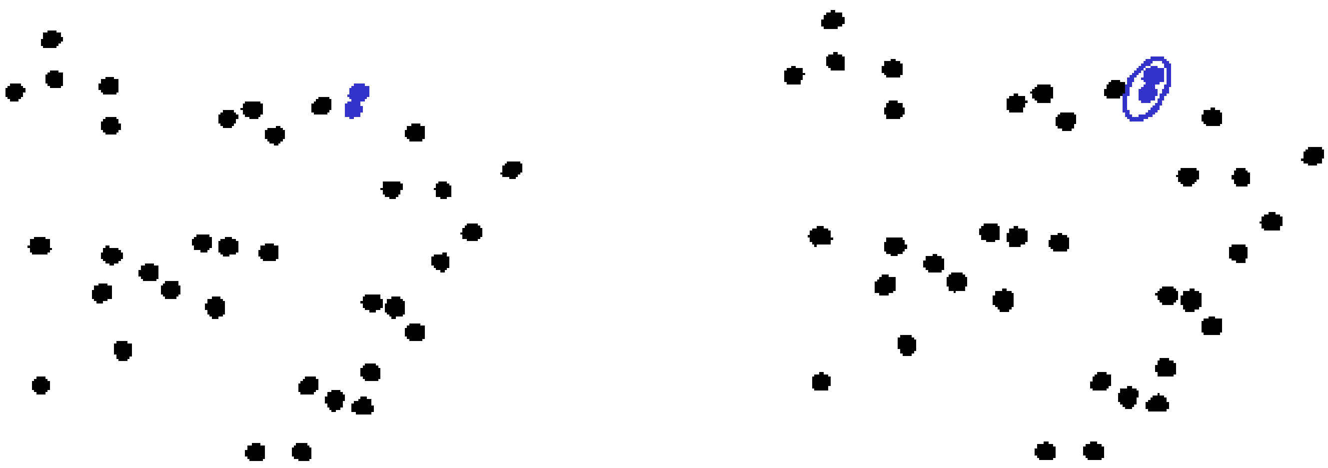
Basic agglomerative clustering

1. Each point = one cluster



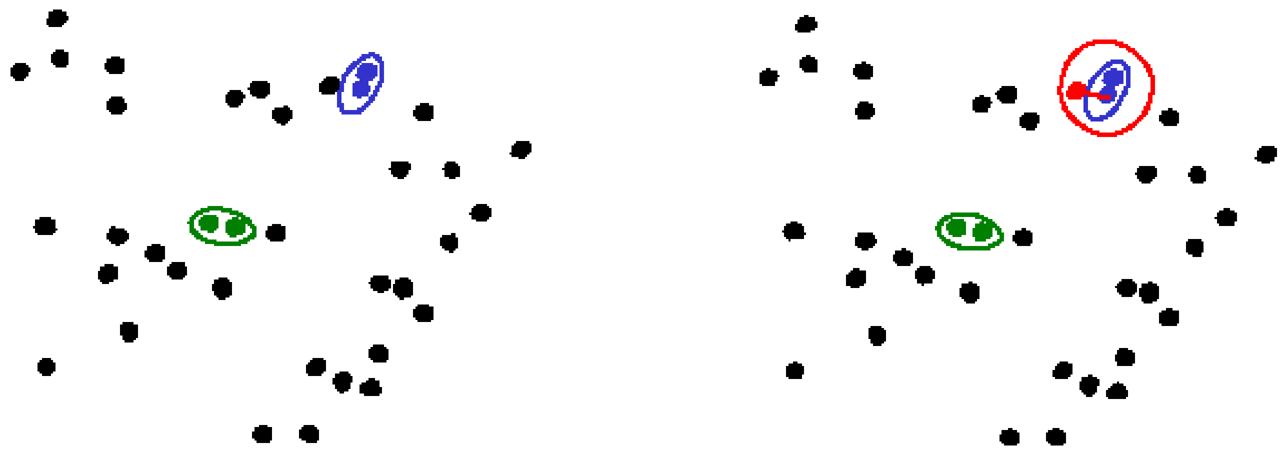
Basic agglomerative clustering

1. Each point = one cluster
2. Find two “nearest” or “most similar” clusters and merge them together



Basic agglomerative clustering

1. Each point = one cluster
2. Find two “nearest” or “most similar” clusters and merge them together
3. Repeat 2 until the stop constraint is reached



Basic agglomerative clustering

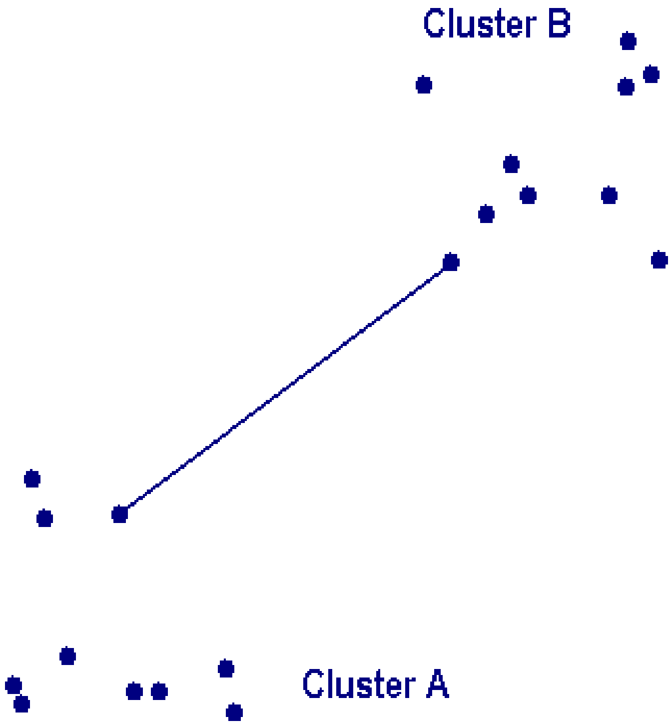
Particular implementations of this method differ from each other by

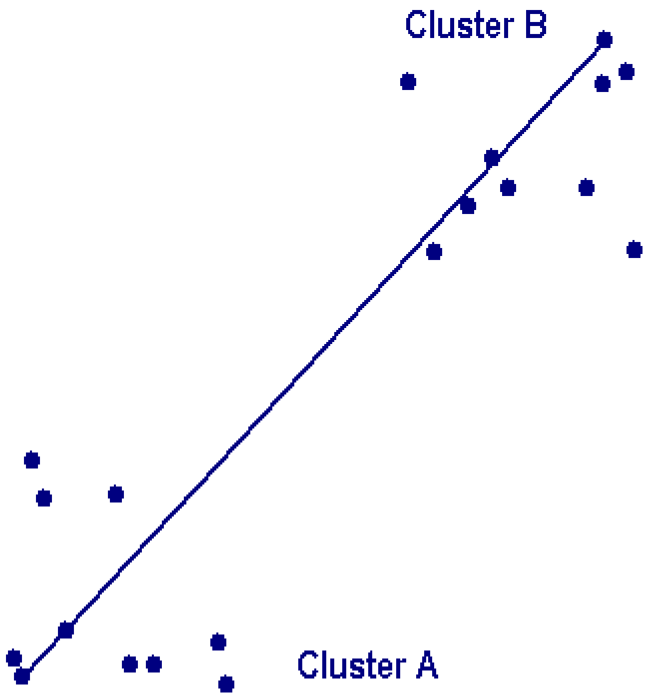
- The STOP constraints
- The distance/similarity measures used

Simple between-cluster distance measures

$$d(A,B) = d(m_1, m_2)$$

$$d(A,B) = \min d(a,b) \qquad\qquad d(A,B) = \max d(a,b)$$





Other between-cluster distance measures

$d(A,B)$ = Hausdorf distance $H(A,B)$

$$d(A,B) = J(A \cup B) - J(A,B)$$

Efficient implementation of hierarchical clustering

At each level, the distances are calculated using the distances from the previous level.

The upgrade is much faster than a complete calculation.

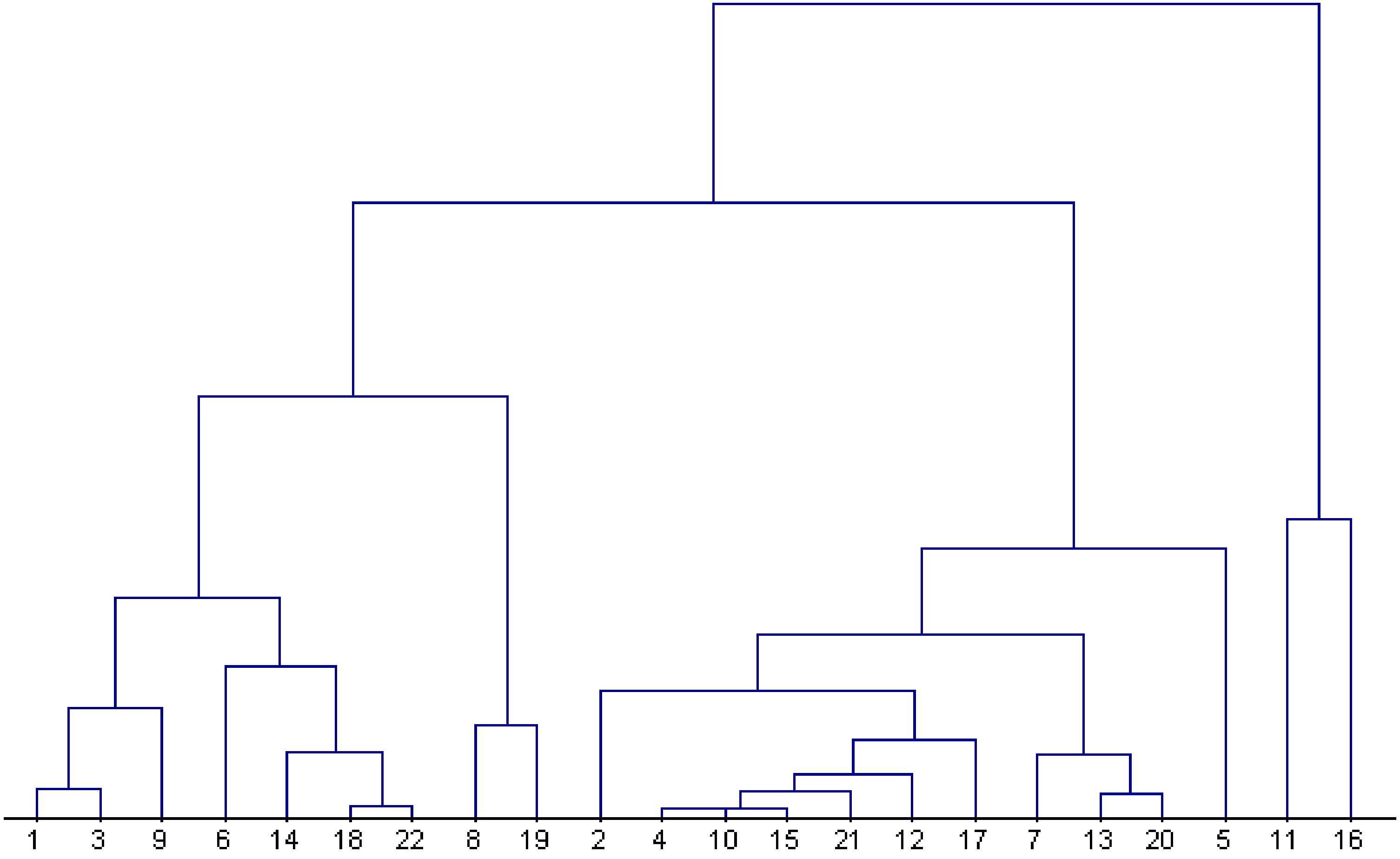
Definite agglomerative clustering

Basic algorithm – at a certain level, there may be multiple candidates of merging.

Random selection leads to ambiguities.

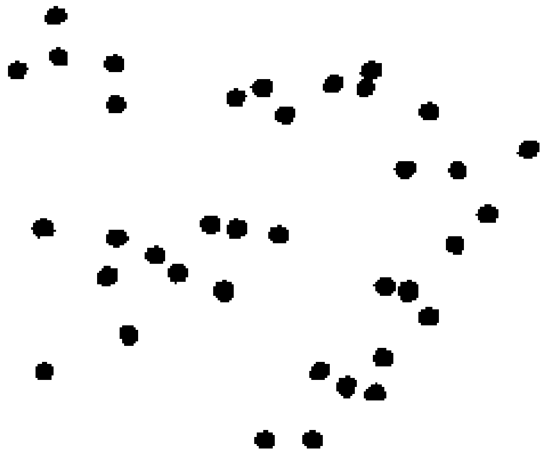
Definite algorithm – **all** such candidates are merged at that level → the number of new clusters emerging at one level may be greater than 1.

Agglomerative clustering – representation by a clustering tree (dendrogram)



Basic divisive clustering

1. All points = one cluster



Basic divisive clustering

1. All points = one cluster
2. Divide the cluster into two parts A,B such that $d(A,B)$ is maximized
3. Select a new cluster to split
4. Apply 2 to the selected cluster
5. Repeat 3-4 until the stop constraint is reached

Full search in Step 2 is very expensive – $O(2^n)$

Suboptimal Step 2 of divisive clustering

1. Find point *p* such that the mean of *d(p,x)* is maximized
2. Divide C into A U B, where B={*p*} (*p* is a seed point of a new cluster B)
3. For *x* from A calculate the mean dist *d(x,A)* and *d(x,B)*. If *d(x,A)* > *d(x,B)* move *x* into B.
4. Repeat 3 for each *x* from A.

Suboptimal Step 2 of divisive clustering

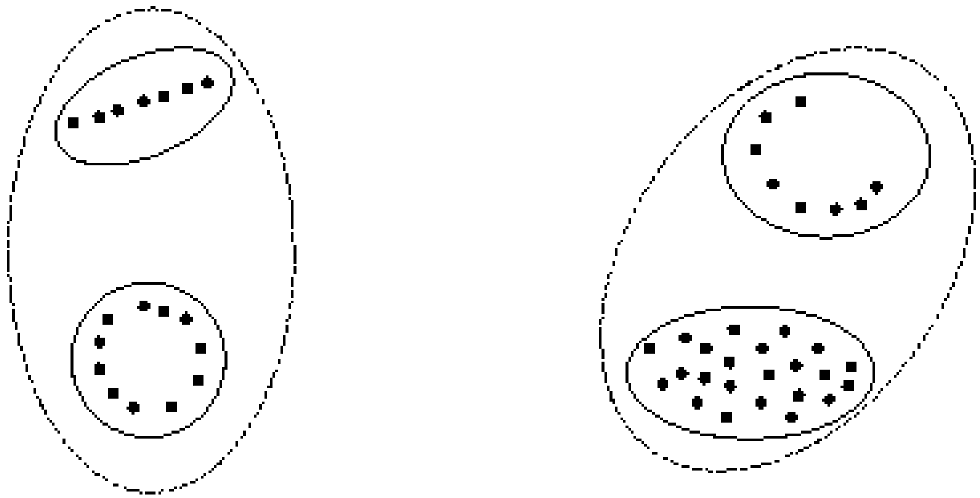
Alternatively, any iterative algorithm for N=2 can be used (N-means, J - minimization, ...)

Step 3 of divisive clustering

Selecting the next cluster to split

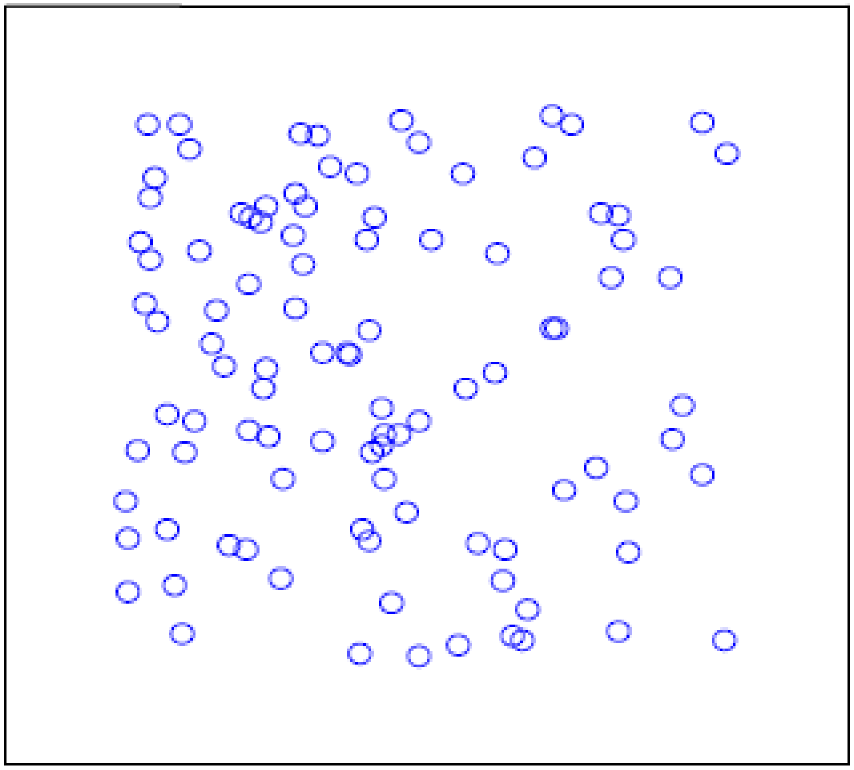
- randomly (when performing complete decomposition)
- by maximum diameter
- by maximum variance
- by maximum *J*

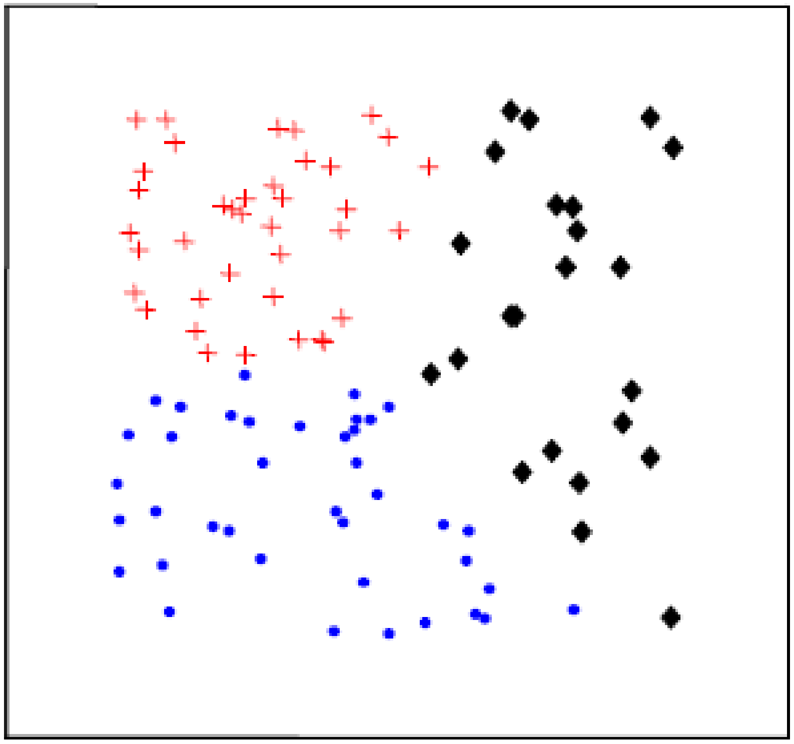
How many clusters are there?



2 or 4 ?

Clustering is a very subjective task

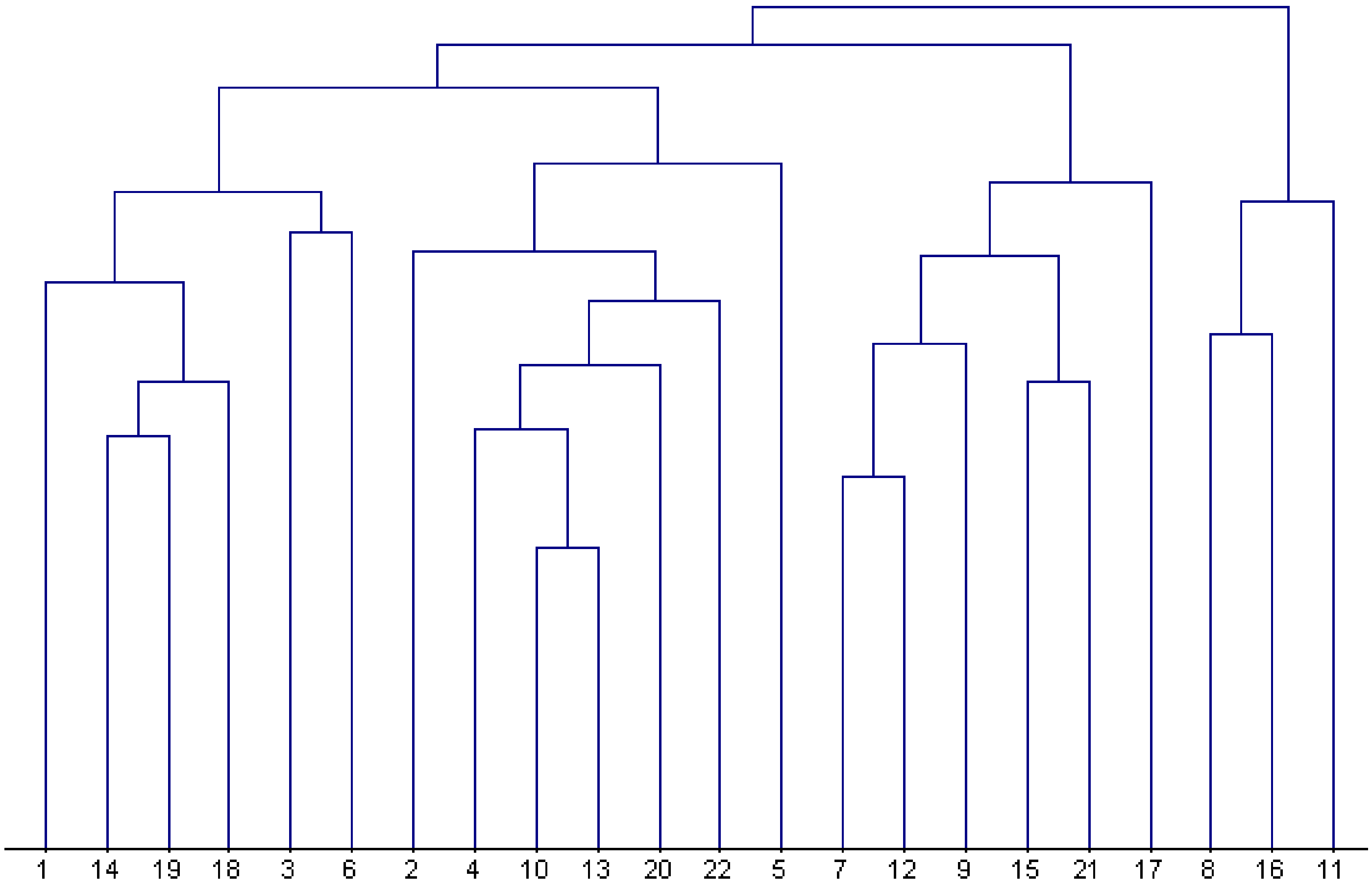




How many clusters are there?

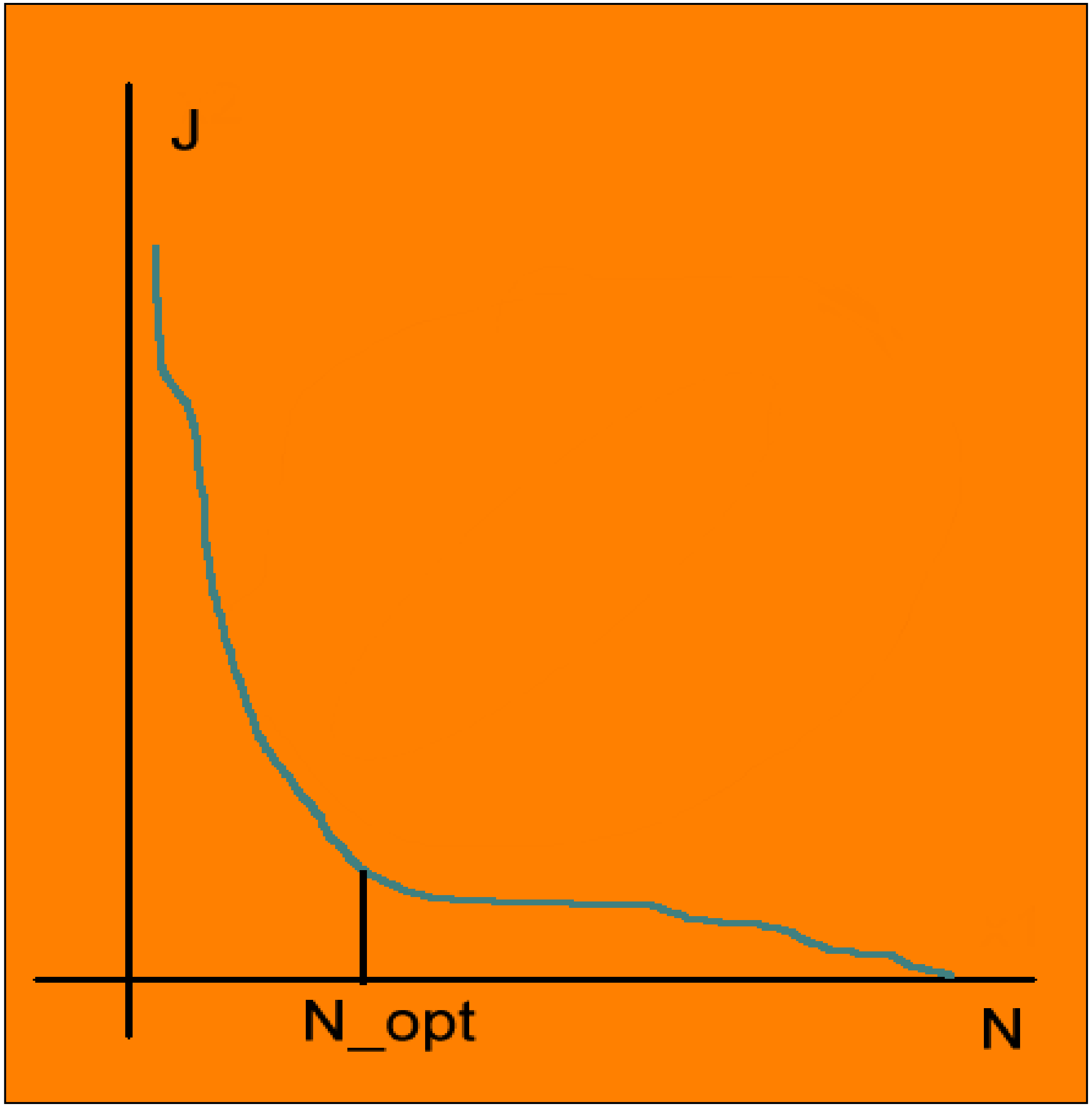
- Difficult to answer even for humans
- "Clustering tendency", "cluster validity"
- Hierarchical methods – *N* can be estimated from the complete dendrogram
- The methods minimizing a cost function – *N* can be estimated from the "knees" in *J*-*N* graph

Life time of the clusters

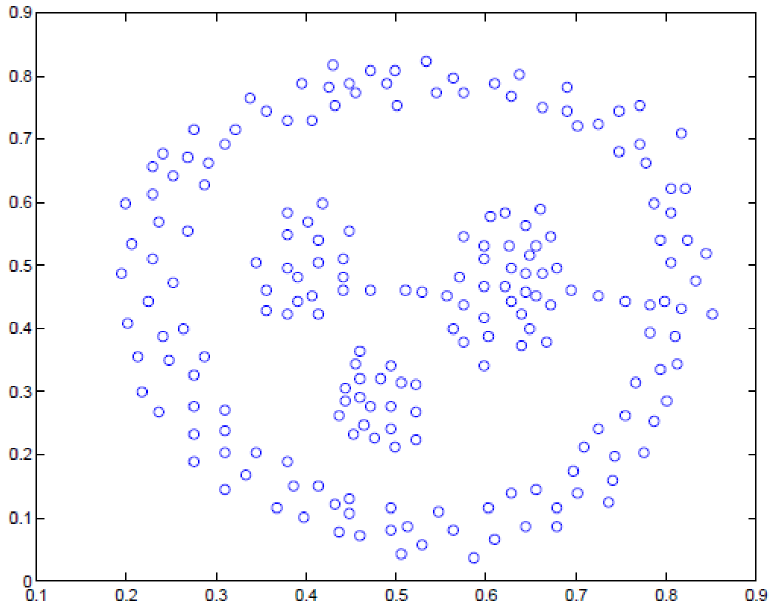


Optimal number of clusters = 4

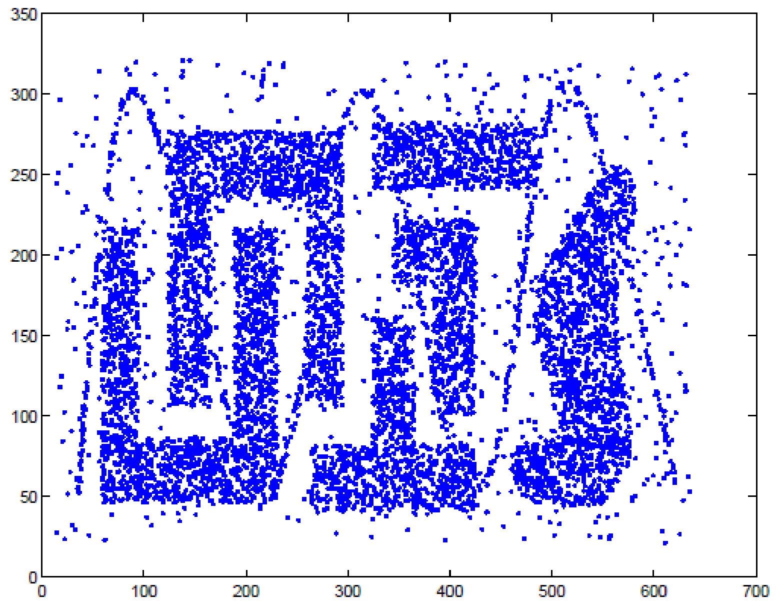
Optimal number of clusters



Hybrid clustering with the choice of N

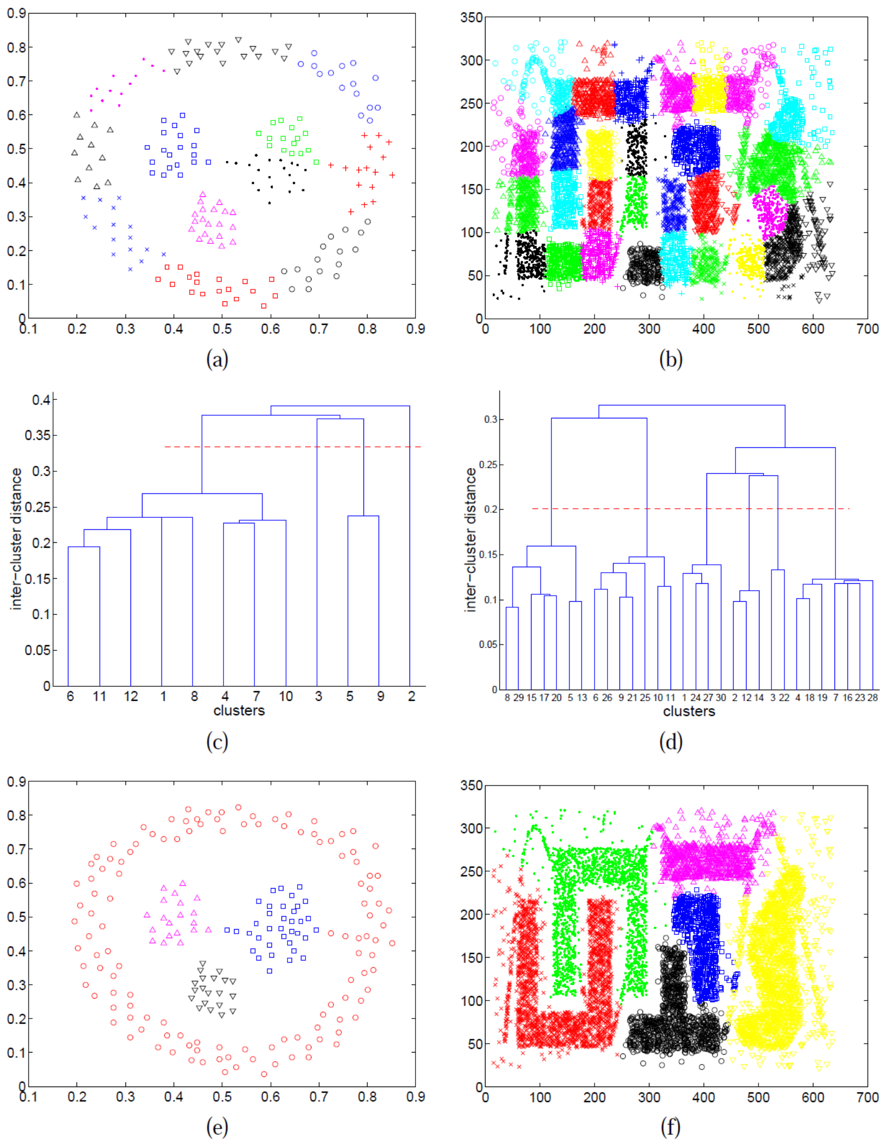


(a)



(b)

Hybrid clustering with the choice of N



Iterative

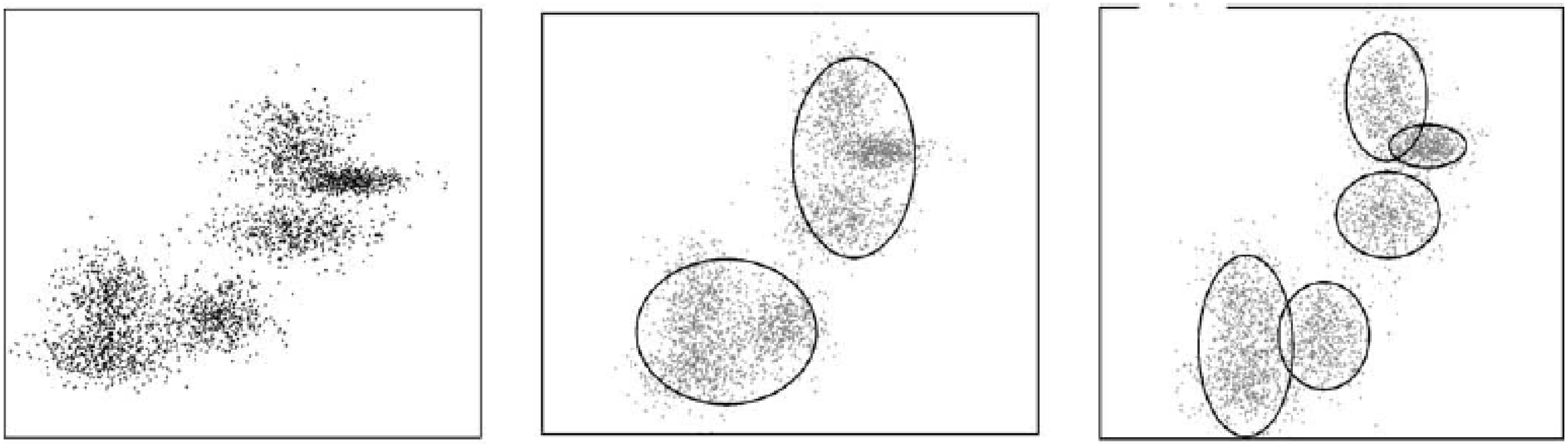
Hierarchical merging

Model-based (parametric) clustering

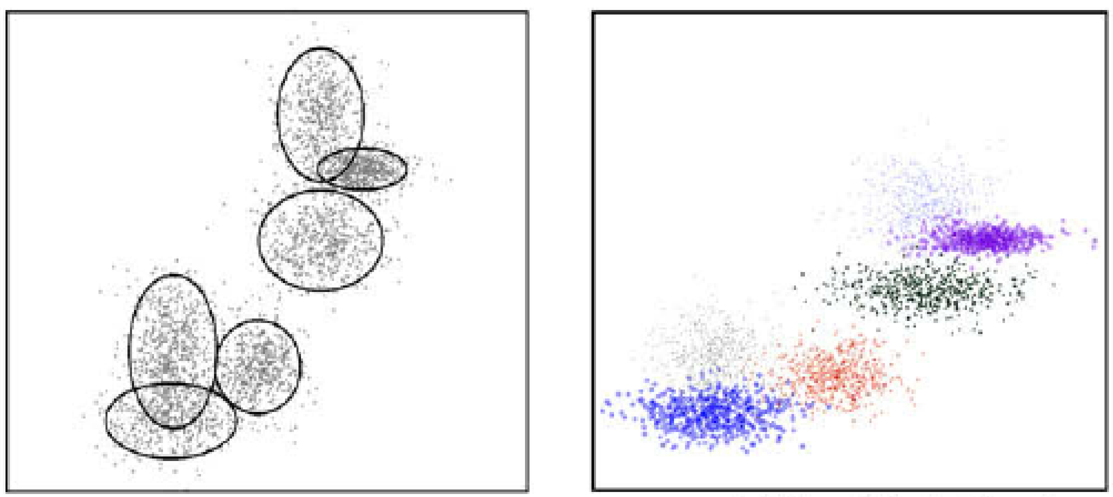
Fitting a Gaussian mixture to the data

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}|\mu_g, \mathbf{\Sigma}_g)$$

Problem: the number of components



(a) Input data (b) GMM (K=2) (c) GMM (K=5)

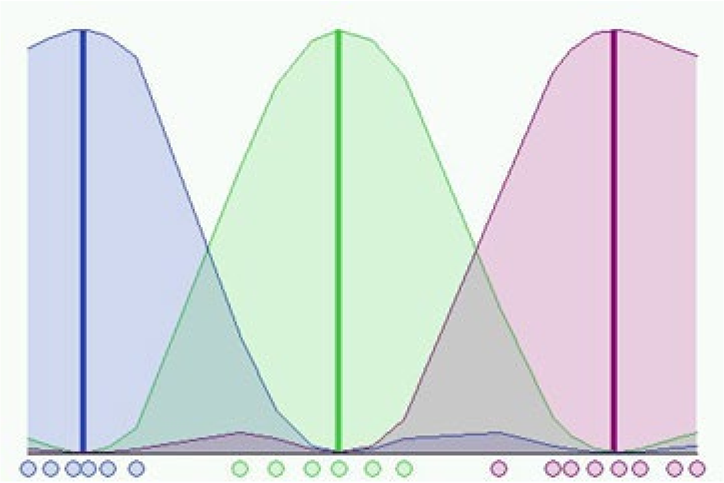


(d) GMM (K=6) (e) True labels, K = 6

Fuzzy clustering

Clusters = Fuzzy sets (Set, Mem. f)

Fuzzy C-means



Other clustering criteria

Scatter matrices

- between cluster matrix

$$B = \sum_{i=1}^{N} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

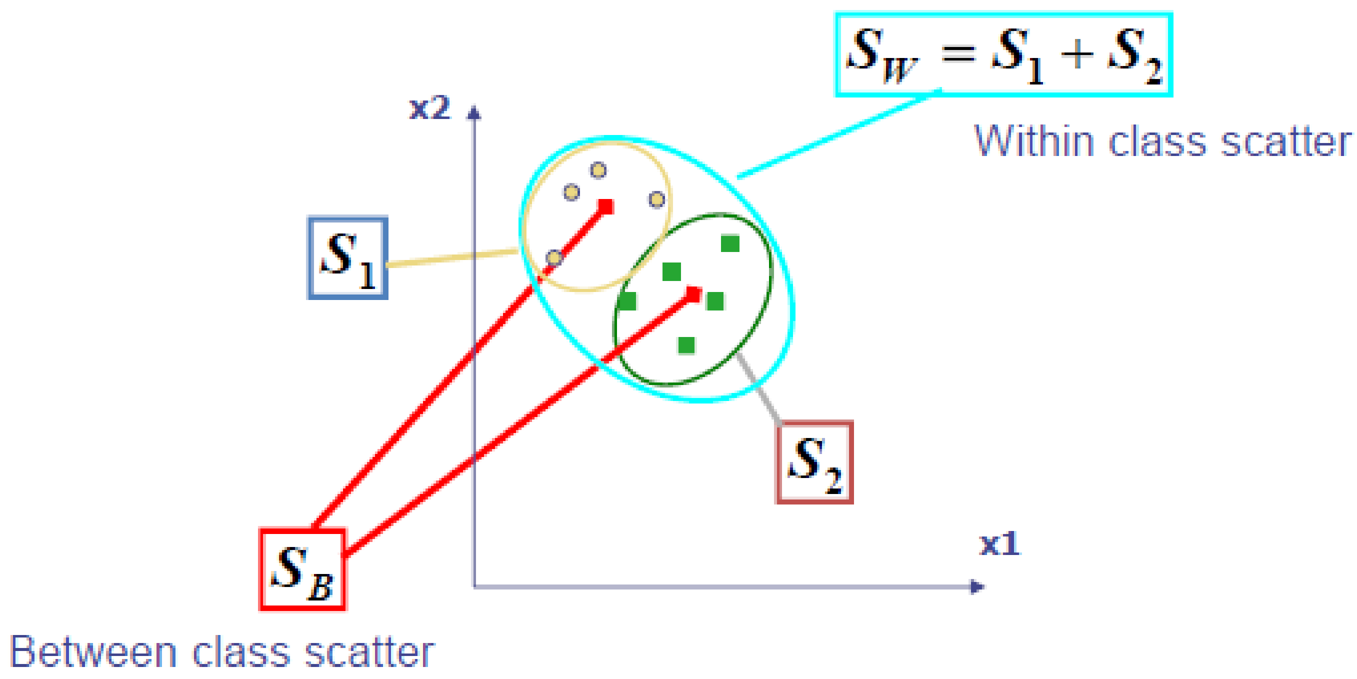
- within cluster matrix

$$W = \sum_{i=1}^{N} W_i$$

$$W_i = \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

- total scatter matrix

$$T = \sum_{\mathbf{x}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$$



$S_W = S_1 + S_2$
Within class scatter
x2
S_1
S_2
x1
S_B
Between class scatter

Other clustering criteria

min $\mathrm{tr}(W) = \sum_{i=1}^{N} \mathrm{tr}(W_i) = \sum_{i=1}^{N} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{m}_i\|^2 = J$

min $\det(W)$

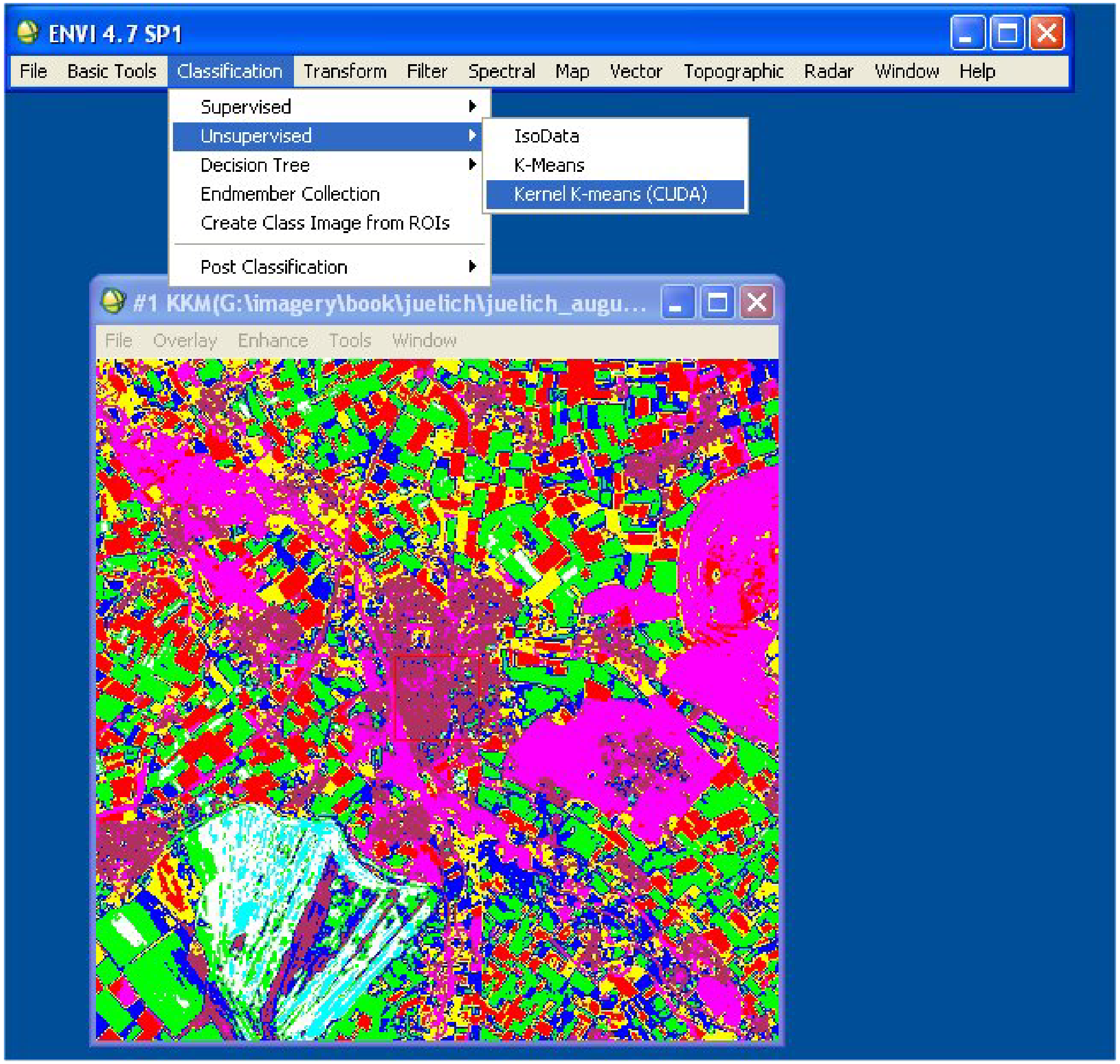
$T = W + B$

max $\mathrm{tr}(W^{-1}B)$

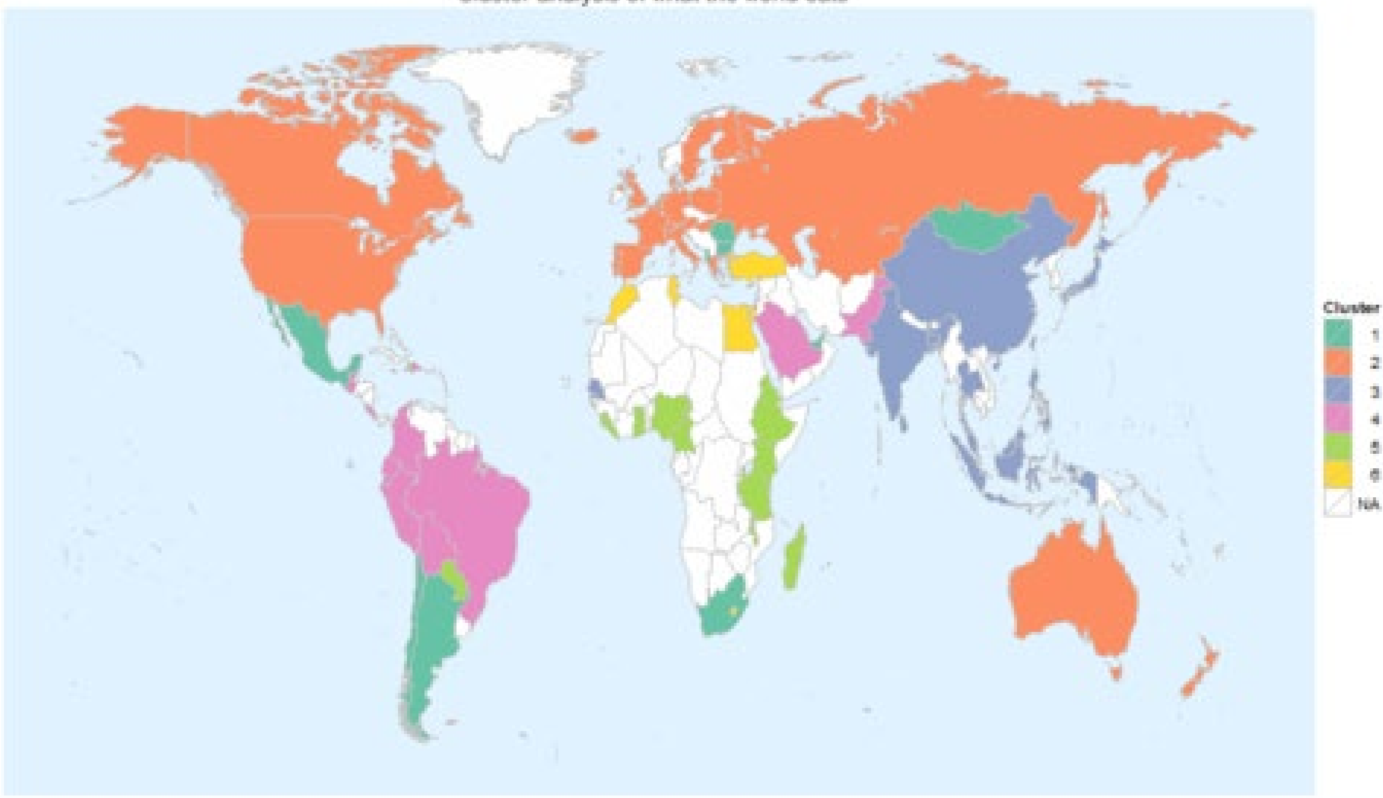
Applications of clustering in image proc.

- Segmentation – clustering in color space
- Preliminary classification of multispectral images
- Clustering in parametric space – RANSAC, image registration and matching

Numerous applications are outside image processing area


ENVI 4.7 SP1
File
Basic Tools
Classification
Transform
Filter
Spectral
Map
Vector
Topographic
Radar
Window
Help
Supervised
Unsupervised
Decision Tree
Endmember Collection
Create Class Image from ROIs
Post Classification
IsoData
K-Means
Kernel K-means (CUDA)
#1 KKM(G:\imagery\book\juelich\juelich_augu...
File
Overlay
Enhance
Tools
Window

Cluster analysis of what the world eats



Thank you !

Any questions ?