Data Mining - Clustering

Lecturer: JERZY STEFANOWSKI
Institute of Computing Sciences
Poznan University of Technology
Poznan, Poland

Lecture 7
SE Master Course
2008/2009

Aims and Outline of This Module

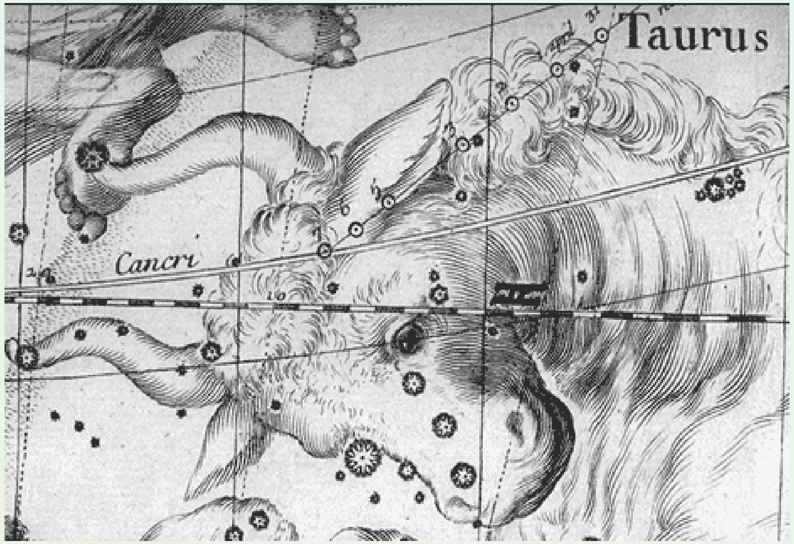
- Discussing the idea of clustering.
- Applications
- Shortly about main algorithms.
- More details on:
  - k-means algorithm/s
  - Hierarchical Agglomerative Clustering
- Evaluation of clusters
- Large data mining perspective
- Practical issues: clustering in Statistica and WEKA.

Acknowledgments:

- As usual I used not only ideas from my older lectures,…
- Smth is borrowed from
    - J.Han course on data mining
    - G.Piatetsky- Shapiro teaching materials
    - WEKA and Statsoft white papers and documentation

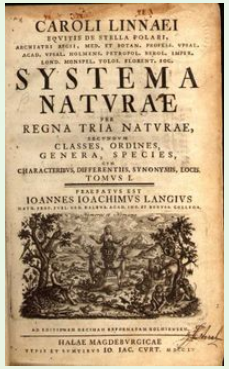
Cluster Analysis

Astronomy - aggregation of stars, galaxies, or super galaxies, …



Historical inspirations – classification of living organisms

- Carl von Linneus → systematic biological taxonomy

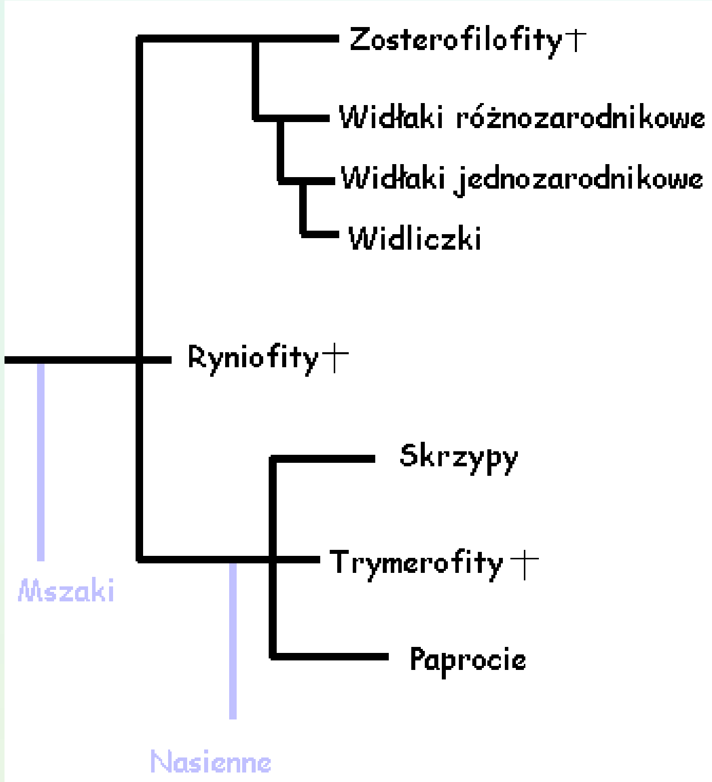


CAROLI LINNAEI
EQVITIS DE STELLA POLARI,
ARCHIATRI REGII, MED. ET BOTAN. PROFESS. VPSAL.
ACAD. VPSAL. HOLMENS. PETROPOL. BEROL. IMPER.
LOND. MONSPEL. TOLOS. FLORENT. SOC.
SYSTEMA
NATVRÆ
PER
REGNA TRIA NATVRAE,
SECVNDVM
CLASSES, ORDINES,
GENERA, SPECIES,
CVM
CHARACTERIBVS, DIFFERENTIIS, SYNONYMIS, LOCIS.
TOMVS I.

PRAEFATVS EST
IOANNES IOACHIMVS LANGIVS
[illegible]

AD EDITIONEM DECIMAM REFORMATAM HOLMIENSEM.

HALAE MAGDEBVRGICAE
TYPIS ET SVMTIBVS IO. IAC. CVRT. MDCCLX

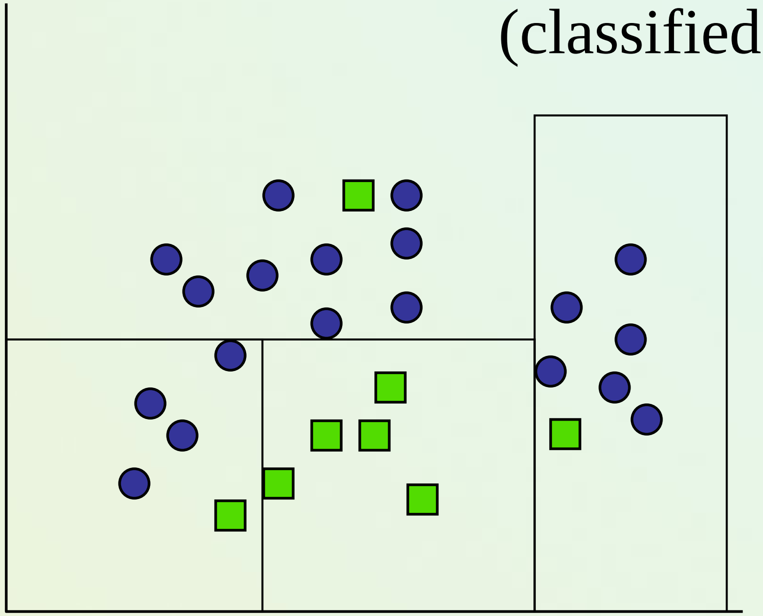


Karol Linneusz znany jest powszechnie jako ojciec systematyki biologicznej i zasad nazewnictwa organizmów. Jego system klasyfikacji organizmów, przede wszystkim tzw. system płciowy roślin, nie był pierwszą próbą uporządkowania świata ożywionego, był jednak nieporównywalny z niczym, co było wcześniej. Linneusz był niezwykle wnikliwym obserwatorem i rzetelnym, skrupulatnym badaczem, który skodyfikował język opisu, sprecyzował terminy naukowe i stworzył podstawy metodologii badawczej systematyki.

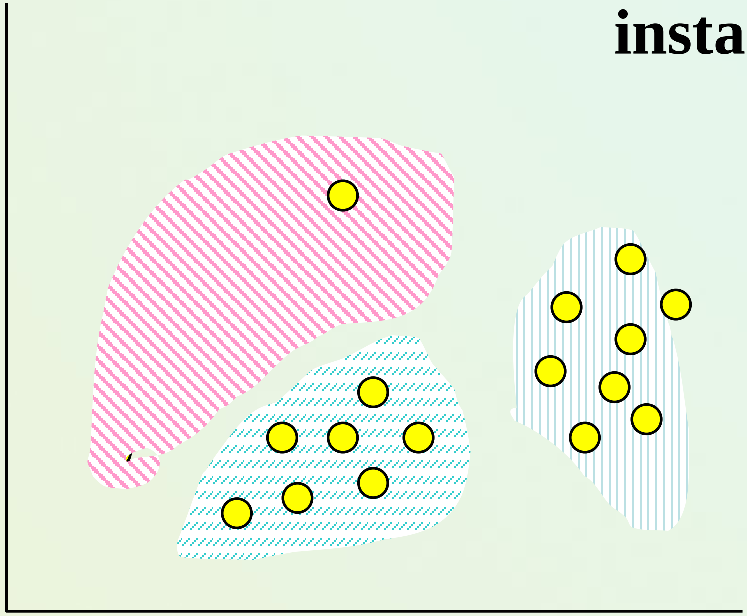
Classification vs. Clustering

Classification: Supervised learning:

Learns a method for predicting the instance class from pre-labeled (classified) instances



Clustering



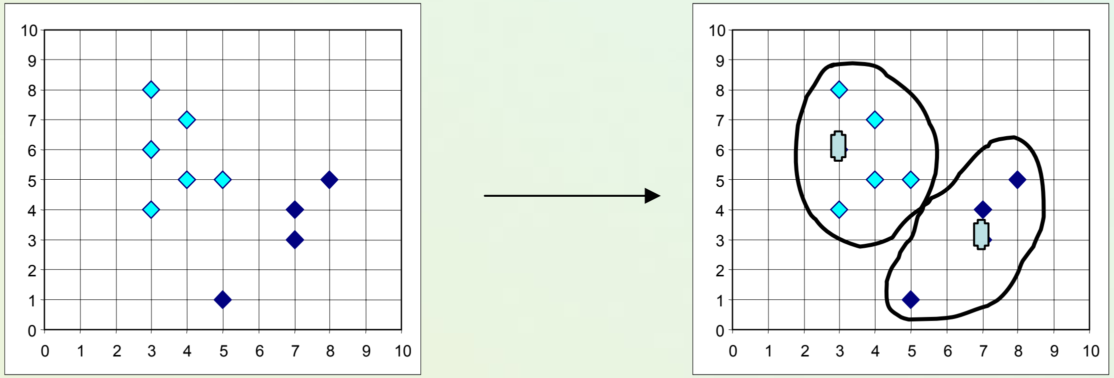
Unsupervised learning:

Finds “natural” grouping of instances given un-labeled data

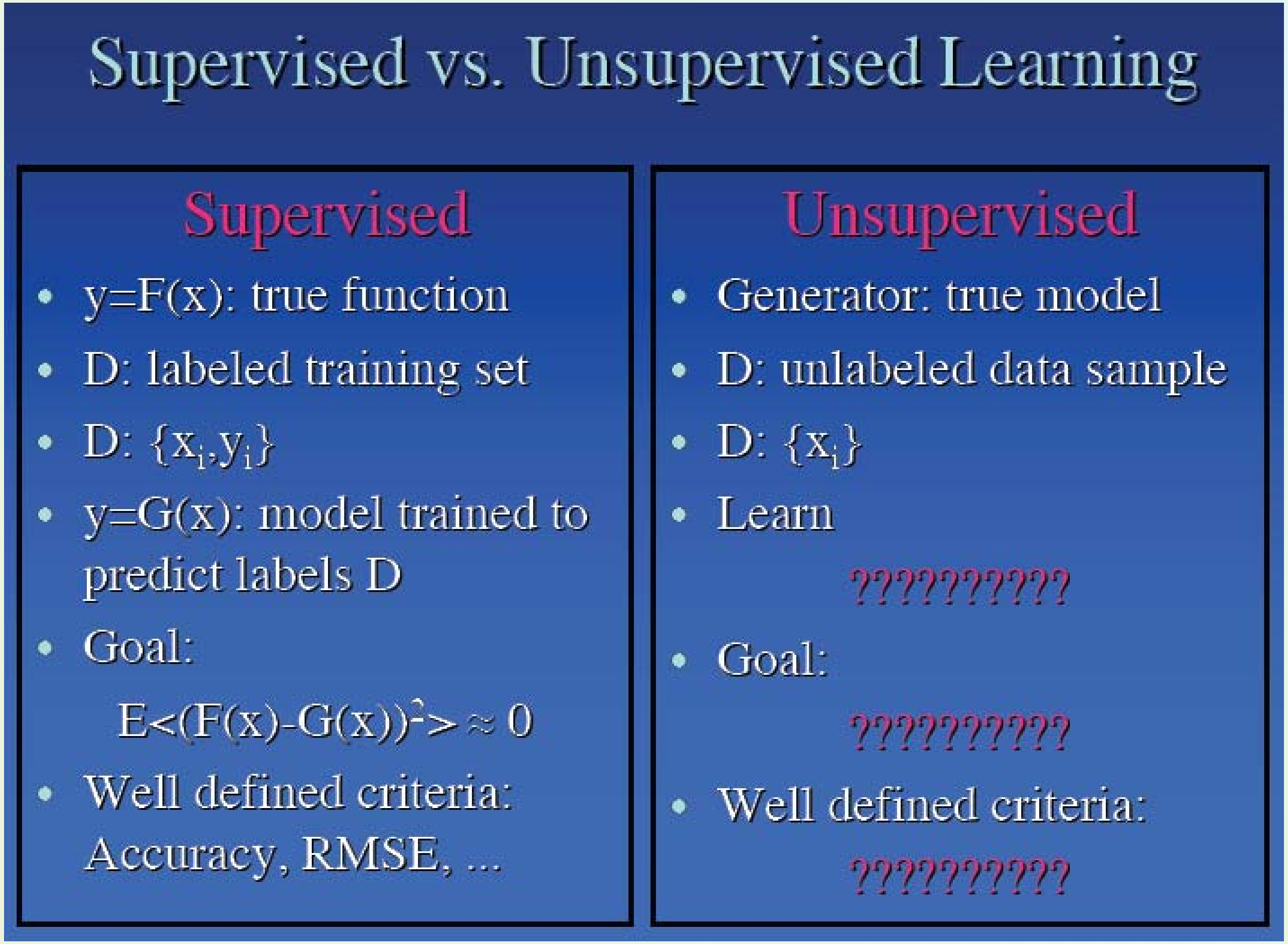
Problem Statement

Given a set of records (instances, examples, objects, observations, …), organize them into clusters (groups, classes)

- Clustering: the process of grouping physical or abstract objects into classes of similar objects



Supervised classification vs. clustering



Supervised vs. Unsupervised Learning

Supervised

- $y=F(x)$: true function
- D: labeled training set
- D: $\{x_i,y_i\}$
- $y=G(x)$: model trained to predict labels D
- Goal:

  $E<(F(x)-G(x))^2> \approx 0$
- Well defined criteria: Accuracy, RMSE, ...

Unsupervised

- Generator: true model
- D: unlabeled data sample
- D: $\{x_i\}$
- Learn

  ??????????
- Goal:

  ??????????
- Well defined criteria:

  ??????????

© Stefanowski 2008

What is a cluster?

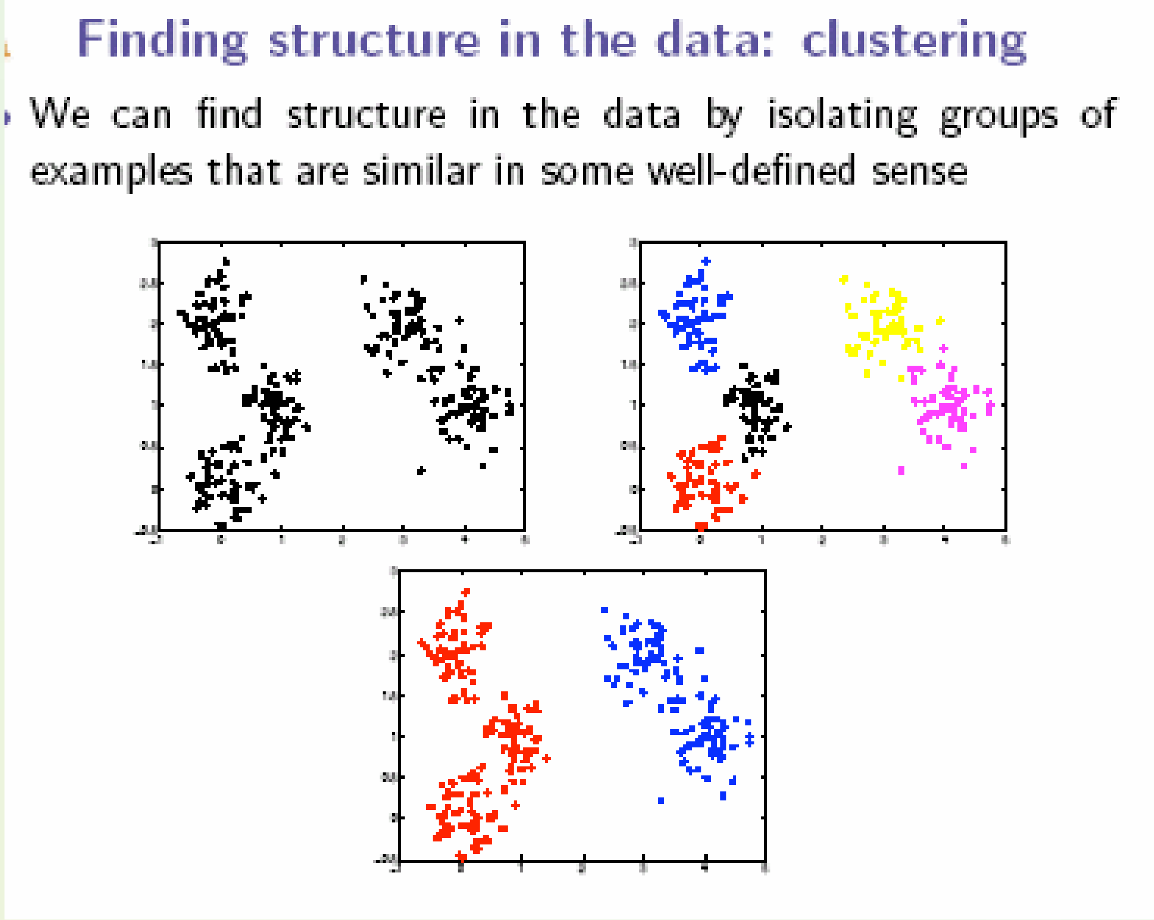
1. A cluster is a subset of objects which are “similar”
2. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it.
3. A connected region of a multidimensional space containing a relatively high density of objects.

What Is Clustering ?

- Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters.
    - Help users understand the natural grouping or structure in a data set.
- Clustering: unsupervised classification: no predefined classes.
- Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.
    - Moreover, data compression, outliers detection, understand human concept formation.

Looking for „comprehensible structures” in data

- Help users to find and try to understand„sth ” in data



Finding structure in the data: clustering

- We can find structure in the data by isolating groups of examples that are similar in some well-defined sense

- Still many possible results

What Is Good Clustering?

- A good clustering method will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high.
  - the inter-class similarity is low.
- The quality of a clustering result also depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.
- However, objective evaluation is problematic: usually done by human / expert inspection.

Polish aspects in clustering

Polish terminology:

- Cluster Analysis → Analiza skupień, Grupowanie.
- Numerical taxonomy → Metody taksonomiczne (ekonomia)
  - Uwaga: znaczenie taksonomii w biologii może mieć inny kontest (podział systematyczny oparty o taksony).
- Cluster→ Skupienie, skupisko, grupa/klasa/pojęcie
- Nigdy nie mów: klaster, klastering, klastrowanie!

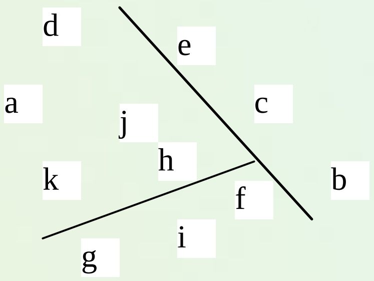
History:

More on Polish History

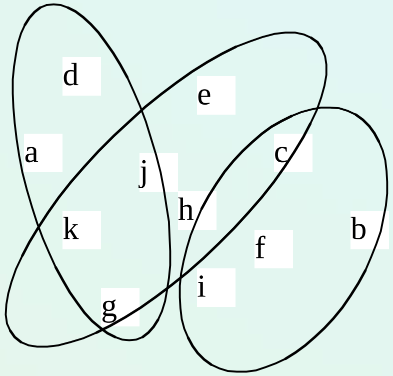
- **Jan Czekanowski** (1882-1965) - wybitny polski antropolog, etnograf, demograf i statystyk, profesor Uniwersytetu Lwowskiego (1913 – 1941) oraz Uniwersytetu Poznańskiego (1946 – 1960).
  - Nowe odległości i metody przetwarzania macierzy odległości w algorytmach, …, tzw. metoda Czekanowskiego.
  - Kontynuacja Jerzy Fierich (1900-1965) Kraków
- **Hugo Steinhaus**, (matematycy Lwów i Wrocław)
  - Wrocławska szkoła taksonomiczna (metoda dendrytowa)
- **Zdzisław Hellwig** (Wrocław)
  - wielowymiarowa analizą porównawcza, i inne …
- Współczesnie …
- „ Sekcja Klasyfikacji i Analizy Danych” (SKAD) Polskiego Towarzystwa Statystycznego

Different ways of representing clusters

(a)



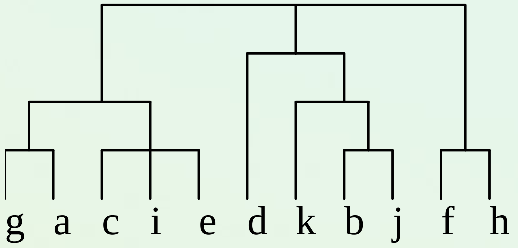
(b)



(c)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1
…			

(d)



Applications of Clustering

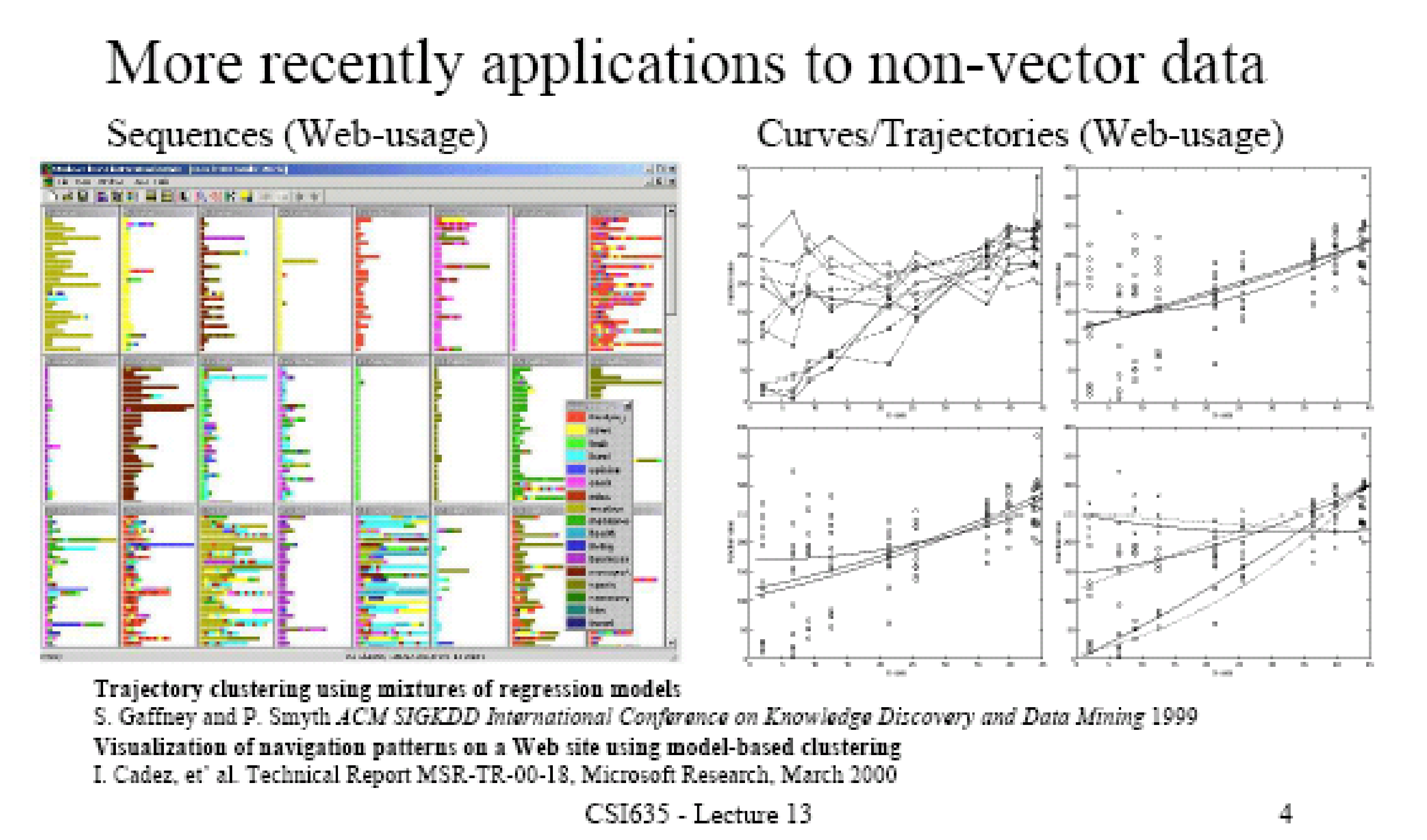
Clustering has wide applications in

- Economic Science (especially market research).
- WWW:
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns
- Pattern Recognition.
- Spatial Data Analysis:
  - create thematic maps in GIS by clustering feature spaces
- Image Processing

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- Land use: Identification of areas of similar land use in an earth observation database.
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.
- and many others,…

Specific data mining applications



More recently applications to non-vector data

Sequences (Web-usage)

Curves/Trajectories (Web-usage)

Trajectory clustering using mixtures of regression models

S. Gaffney and P. Smyth *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1999

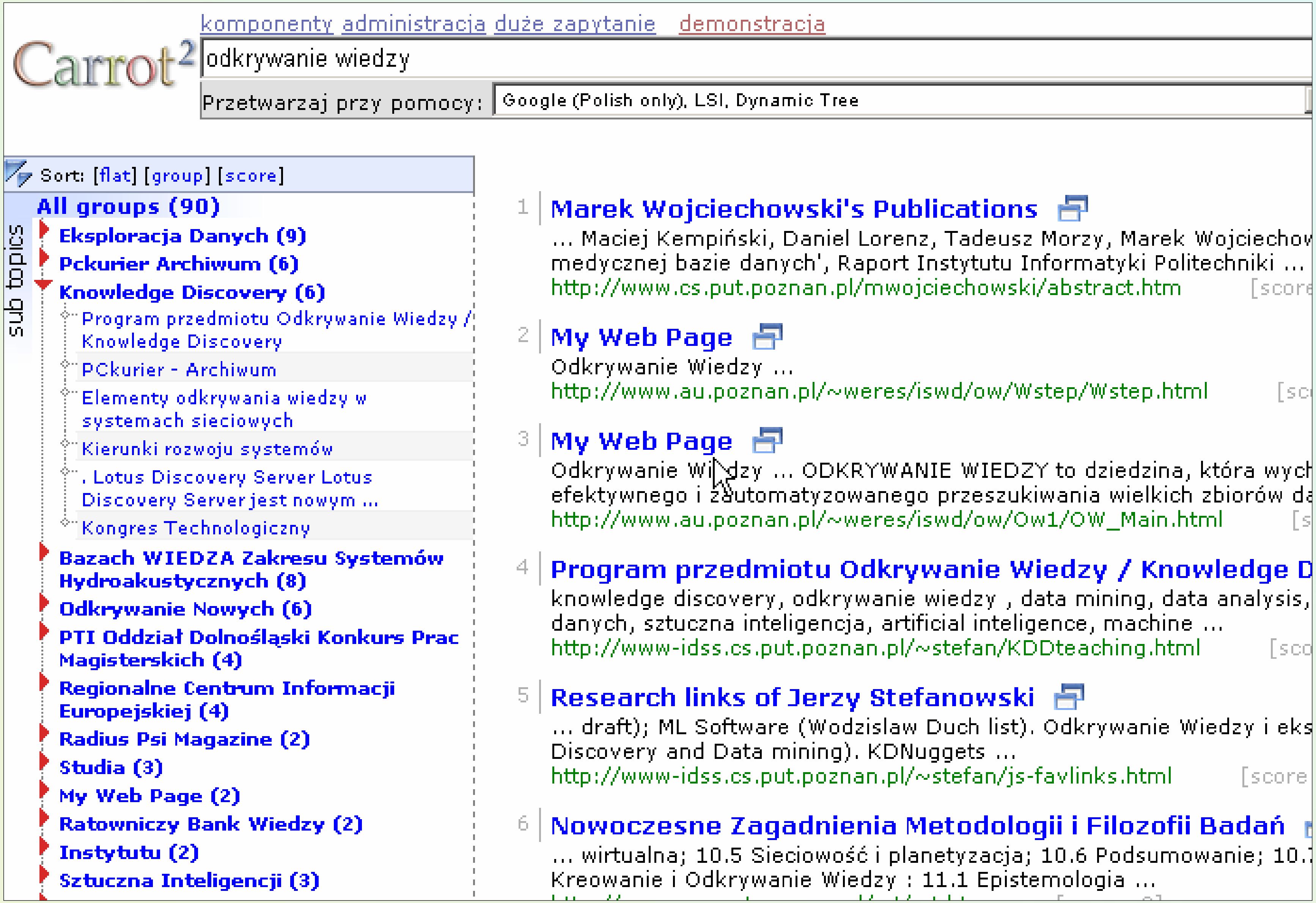
Visualization of navigation patterns on a Web site using model-based clustering

I. Cadez, et' al. Technical Report MSR-TR-00-18, Microsoft Research, March 2000

CSI635 - Lecture 13

4

Web Search Result Clustering



komponenty administracja duże zapytanie demonstracja

Carrot²

odkrywanie wiedzy

Przetwarzaj przy pomocy: Google (Polish only), LSI, Dynamic Tree

Sort: [flat] [group] [score]

sub topics

- All groups (90)
  - Eksploracja Danych (9)
  - Pckurier Archiwum (6)
  - Knowledge Discovery (6)
    - Program przedmiotu Odkrywanie Wiedzy / Knowledge Discovery
    - PCkurier - Archiwum
    - Elementy odkrywania wiedzy w systemach sieciowych
    - Kierunki rozwoju systemów
    - . Lotus Discovery Server Lotus Discovery Server jest nowym ...
    - Kongres Technologiczny
  - Bazach WIEDZA Zakresu Systemów Hydroakustycznych (8)
  - Odkrywanie Nowych (6)
  - PTI Oddział Dolnośląski Konkurs Prac Magisterskich (4)
  - Regionalne Centrum Informacji Europejskiej (4)
  - Radius Psi Magazine (2)
  - Studia (3)
  - My Web Page (2)
  - Ratowniczy Bank Wiedzy (2)
  - Instytutu (2)
  - Sztuczna Inteligencji (3)

1 **Marek Wojciechowski's Publications**
... Maciej Kempiński, Daniel Lorenz, Tadeusz Morzy, Marek Wojciechow
medycznej bazie danych', Raport Instytutu Informatyki Politechniki ...
http://www.cs.put.poznan.pl/mwojciechowski/abstract.htm [score

2 **My Web Page**
Odkrywanie Wiedzy ...
http://www.au.poznan.pl/~weres/iswd/ow/Wstep/Wstep.html [sc

3 **My Web Page**
Odkrywanie Wiedzy ... ODKRYWANIE WIEDZY to dziedzina, która wych
efektywnego i zautomatyzowanego przeszukiwania wielkich zbiorów da
http://www.au.poznan.pl/~weres/iswd/ow/Ow1/OW_Main.html [s

4 **Program przedmiotu Odkrywanie Wiedzy / Knowledge D**
knowledge discovery, odkrywanie wiedzy , data mining, data analysis,
danych, sztuczna inteligencja, artificial inteligence, machine ...
http://www-idss.cs.put.poznan.pl/~stefan/KDDteaching.html [sco

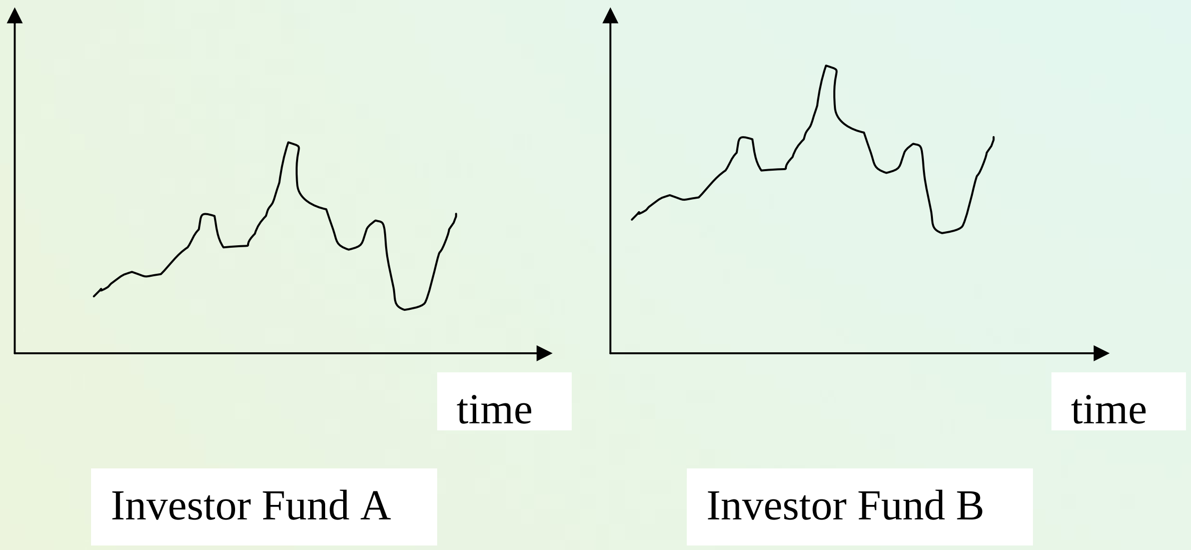
5 **Research links of Jerzy Stefanowski**
... draft); ML Software (Wodzislaw Duch list). Odkrywanie Wiedzy i eks
Discovery and Data mining). KDNuggets ...
http://www-idss.cs.put.poznan.pl/~stefan/js-favlinks.html [score

6 **Nowoczesne Zagadnienia Metodologii i Filozofii Badań**
... wirtualna; 10.5 Sieciowość i planetyzacja; 10.6 Podsumowanie; 10.
Kreowanie i Odkrywanie Wiedzy : 11.1 Epistemologia ...

Time-Series Similarities – specific data mining

Given a database of time-series.

Group “similar” time-series



Clustering Methods

- Many different method and algorithms:
    - For numeric and/or symbolic data
    - Exclusive vs. overlapping
        - Crisp vs. soft computing paradigms
    - Hierarchical vs. flat (non-hierarchical)
    - Access to all data or incremental learning
    - Semi-supervised mode
- Algorithms also vary by:
    - Measures of similarity
    - Linkage methods
    - Computational efficiency

Yet another categorization

- Following Jain’s tutorial

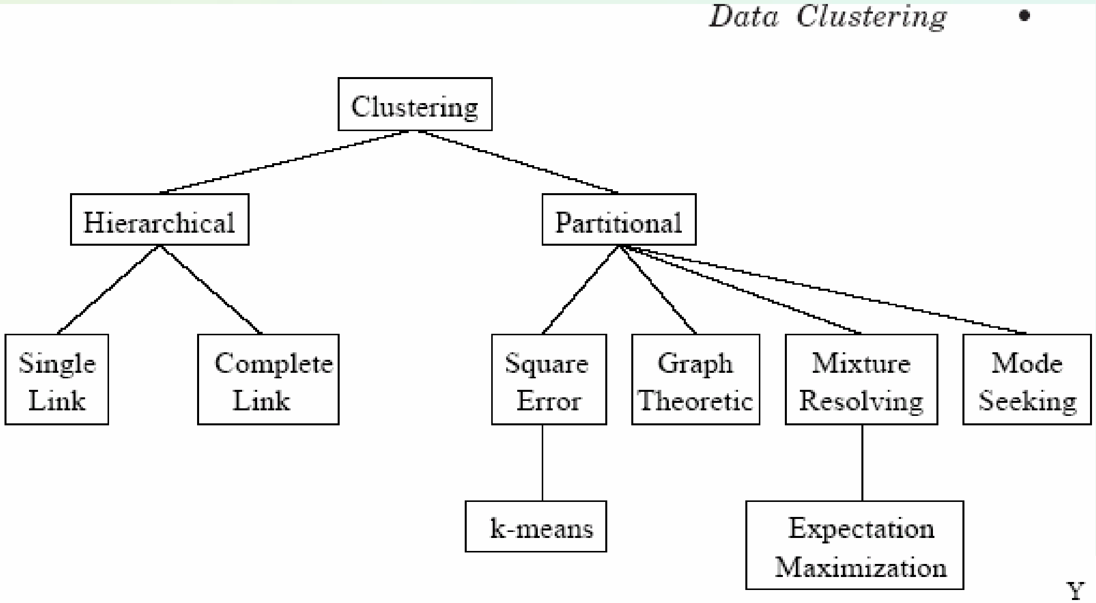


Figure 7. A taxonomy of clustering approaches.

- Furthermore:
  - Crisp vs. Fuzzy
  - Inceremental vs. batch

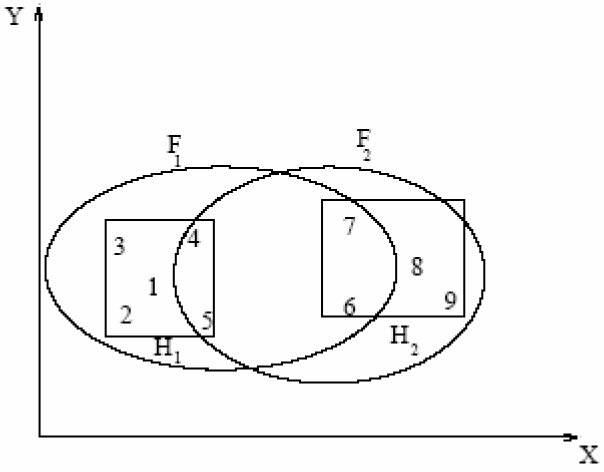


Figure 16. Fuzzy clusters.

© Stefanowski 2008

Data Structures

- Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dis/similarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Measuring Dissimilarity or Similarity in Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric:

  $d(i, j)$

- There are also used in “quality” functions, which estimate the “goodness” of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.

Type of attributes in clustering analysis

- Interval-scaled variables
- Binary variables
- Nominal, ordinal, and ratio variables
- Variables of mixed types

  - Remark: variable vs. attribute

Distance Measures

To discuss whether a set of points is close enough to be considered a cluster, we need a distance measure - D(x, y)

The usual axioms for a distance measure D are:

- $D(x, x) = 0$
- $D(x, y) = D(y, x)$
- $D(x, y) \leq D(x, z) + D(z, y)$ the triangle inequality

Distance Measures (2)

Assume a k-dimensional Euclidean space, the distance between two points, x=[x_1, x_2, ..., x_k] and y=[y_1, y_2, ..., y_k] may be defined using one of the measures:

- Euclidean distance: ("L_2 norm") $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
- Manhattan distance: ("L_1 norm") $\sum_{i=1}^{k}| x_i - y_i |$
- Max of dimensions: ("L_∞ norm") $\max_{i=1}^{k} | x_i - y_i |$

Distance Measures (3)

- Minkowski distance: $$(\sum_{i=1}^{k}(|x_i - y_i|)^q)^{1/q}$$

When there is no Euclidean space in which to place the points, clustering becomes more difficult: Web page accesses, DNA sequences, customer sequences, categorical attributes, documents, etc.

Standarization / Normalization

- If the values of attributes are in different units then it is likely that some of them will take vary large values, and hence the "distance" between two cases, on this variable, can be a big number.
- Other attributes may be small in values, or not vary much between cases, in which case the difference between the two cases will be small.
- The attributes with high variability / range will dominate the metric.
- Overcome this by standardization or normalization

$$z_i = \frac{x_i - \bar{x}_i}{s_{x_i}}$$

Binary variables

- A contingency table for binary data

		Object *j*		
		1	0	*sum*
Object *i*	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Simple matching coefficient (invariant, if the binary variable is *symmetric*):

$$d(i, j) = \frac{b + c}{a + b + c + d}$$

- Jaccard coefficient (noninvariant if the binary variable is *asymmetric*):

$$d(i, j) = \frac{b + c}{a + b + c}$$

Nominal, ordinal and ratio variables

- nominal variables: > 2 states, e.g., red, yellow, blue, green.

$$d(i,j) = \frac{p-u}{p}$$

  - *Simple matching*: *u*: # of matches, *p*: total # of variables.
  - Also, one can use a large number of binary variables.
- ordinal variables: order is important, e.g., rank.
  - Can be treated like interval-scaled, by replacing x_{if} by their rank $r_{if} \in \{1,..., M_f\}$ and replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- ratio variables: a positive measurement on a nonlinear scale, approximately at exponential scale, such as Ae^{Bt} or Ae^{-Bt}
  - One may treat them as continuous ordinal data or perform logarithmic transformation and then treat them as interval-scaled.

Variables of mixed types

- Data sets may contain all types of attrib./variables:
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio.
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - *f* is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$ or, o.w. $d_{ij}^{(f)} = 1$
  - *f* is interval-based: use the normalized distance.
  - *f* is ordinal or ratio-scaled: compute ranks r_{if} and and treat z_{if} as interval-scaled $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

© Stefanowski 2008

Main Categories of Clustering Methods

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion.
- Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion.
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.

Partitioning Algorithms: Basic Concept

- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion.
  - Global optimal: exhaustively enumerate all partitions.
  - Heuristic methods: *k-means* and *k-medoids* algorithms.
  - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster.

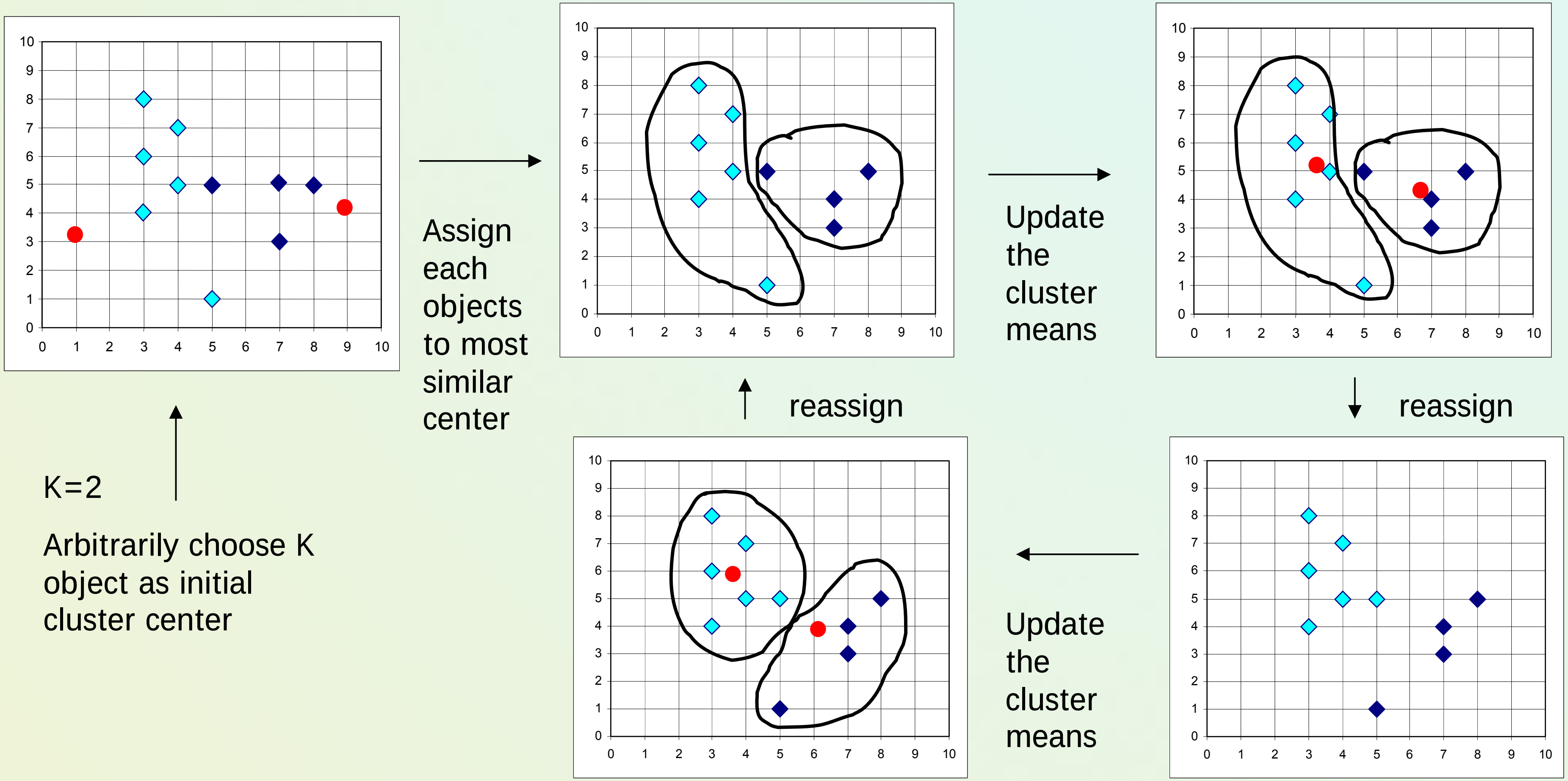
Simple Clustering: K-means

Basic version works with numeric data only

1) Pick a number (K) of cluster centers - *centroids* (at random)
2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
3) Move each cluster center to the mean of its assigned items
4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

Illustrating *K-Means*

- Example



K=2

Arbitrarily choose K object as initial cluster center

Assign each objects to most similar center

Update the cluster means

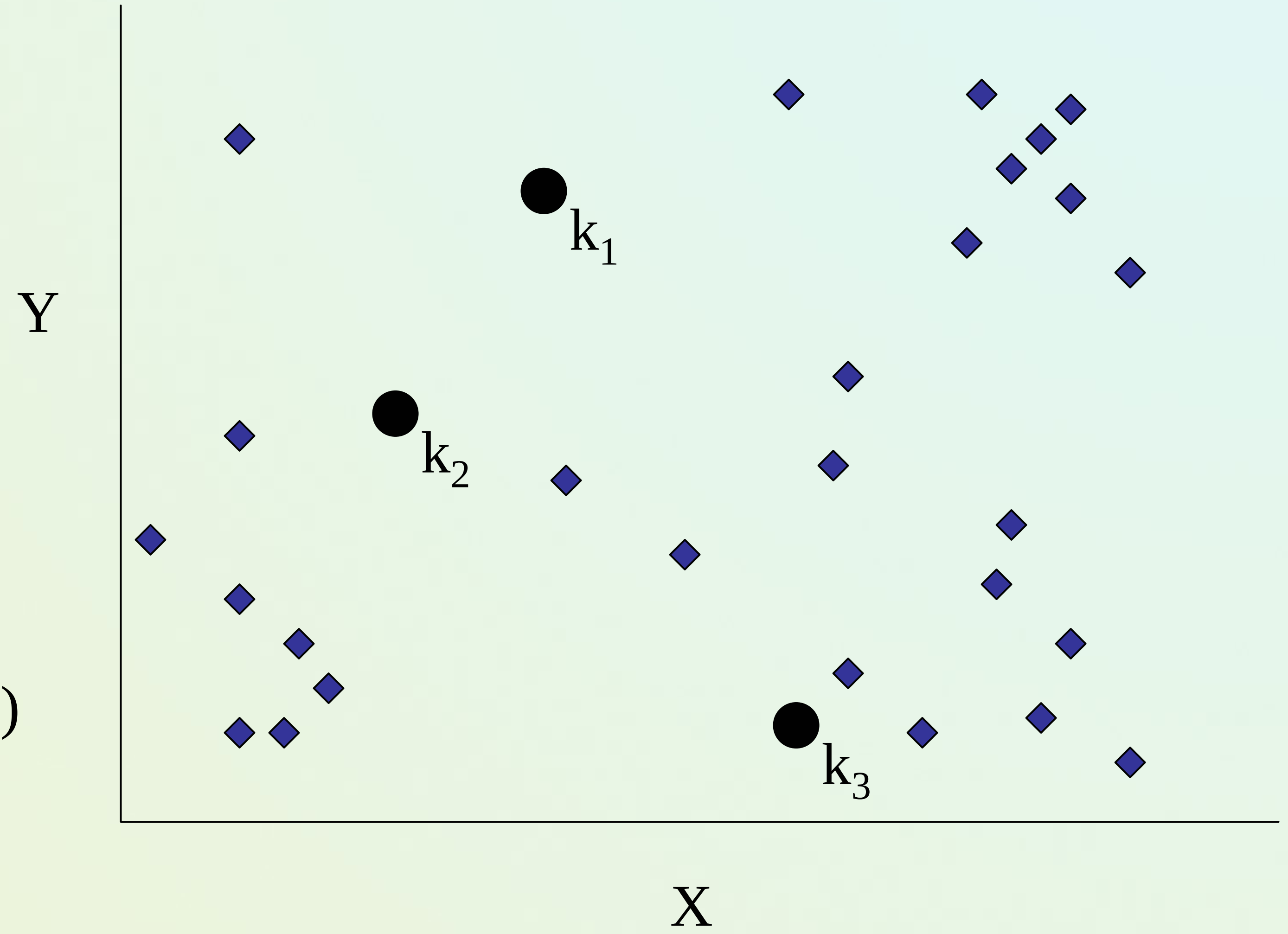
reassign

Update the cluster means

reassign

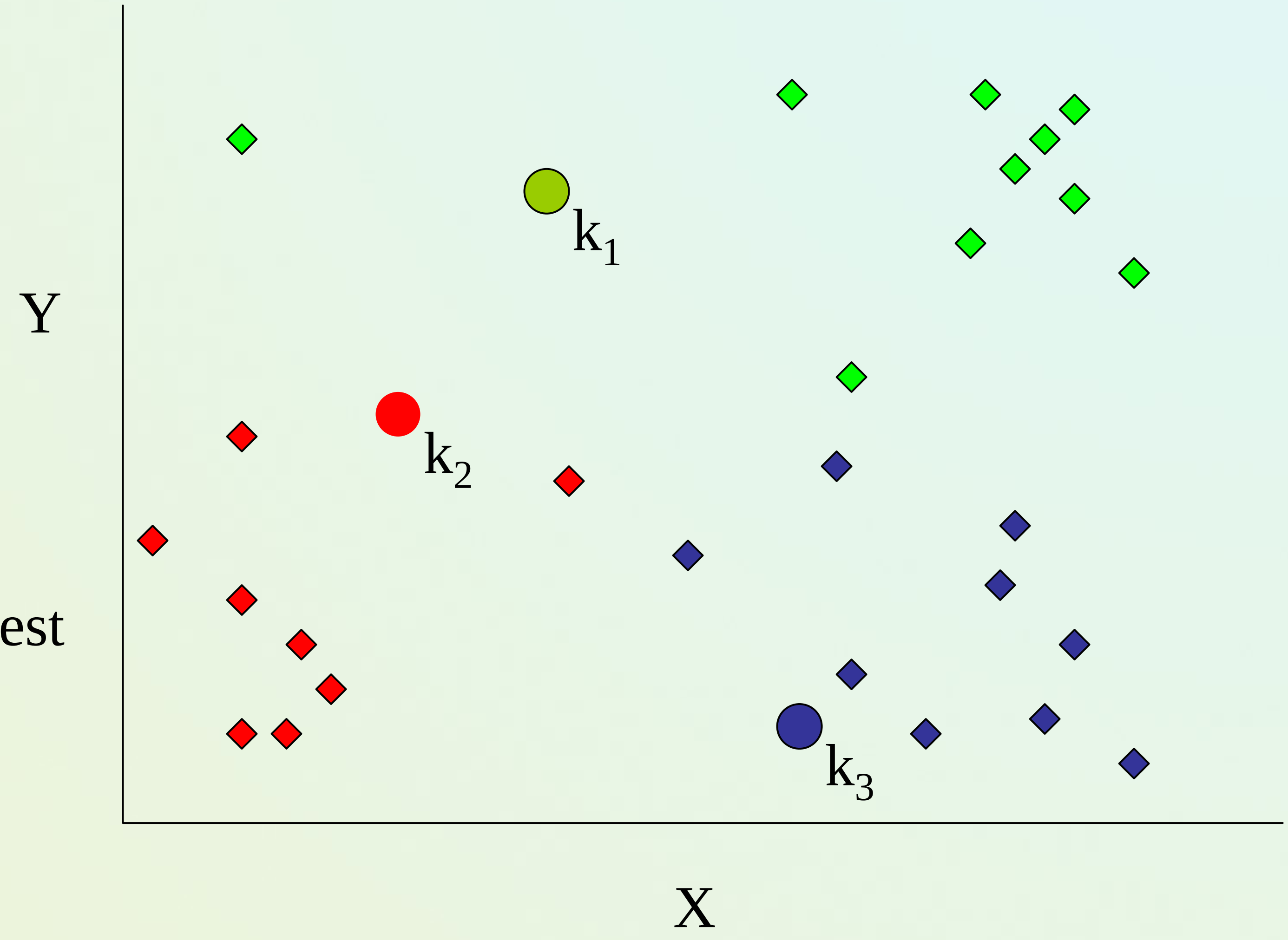
© Stefanowski 2008

K-means example, step 1



Pick 3 initial cluster centers (randomly)

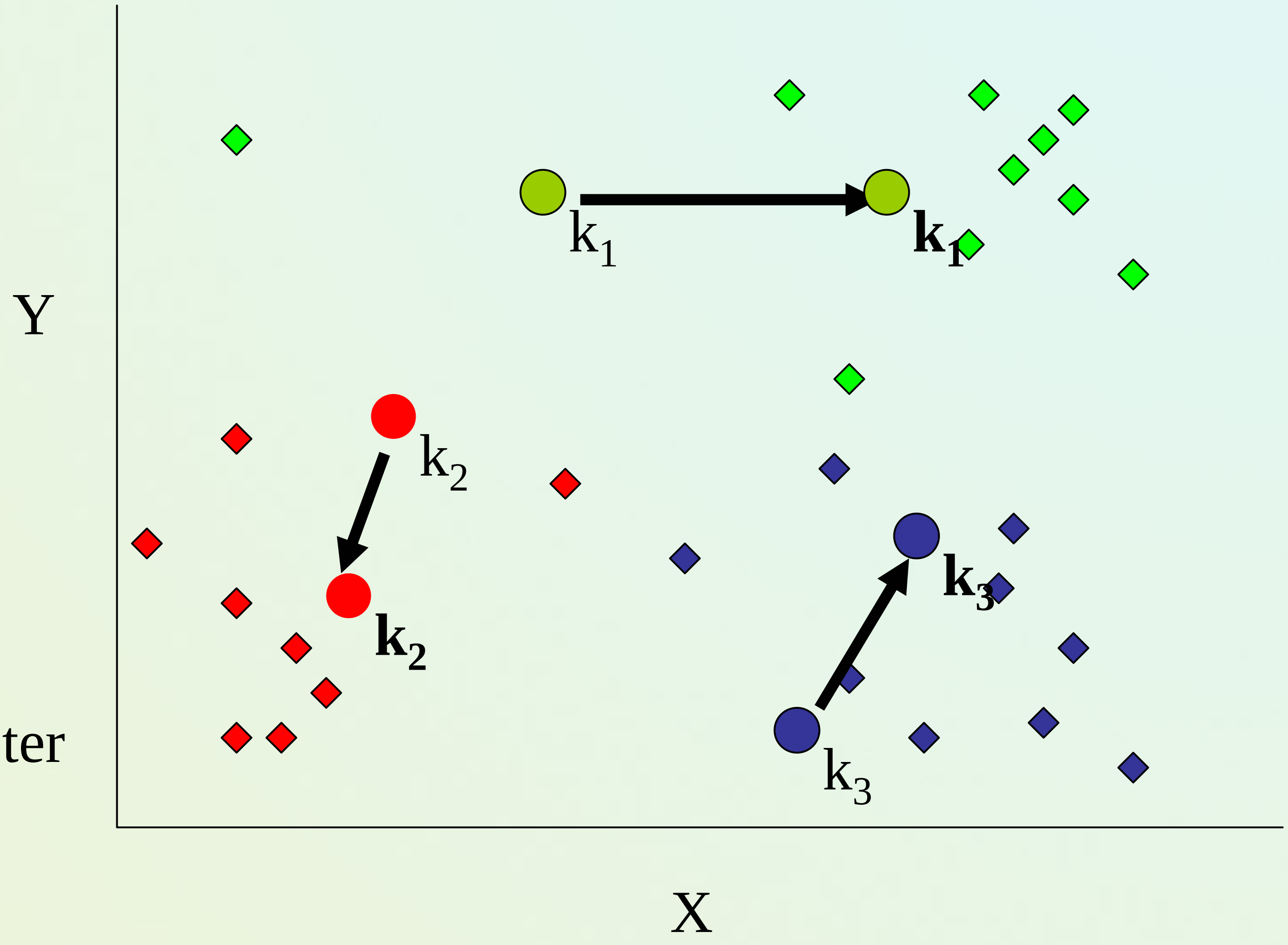
K-means example, step 2



Assign each point to the closest cluster center

© Stefanowski 2008

K-means example, step 3



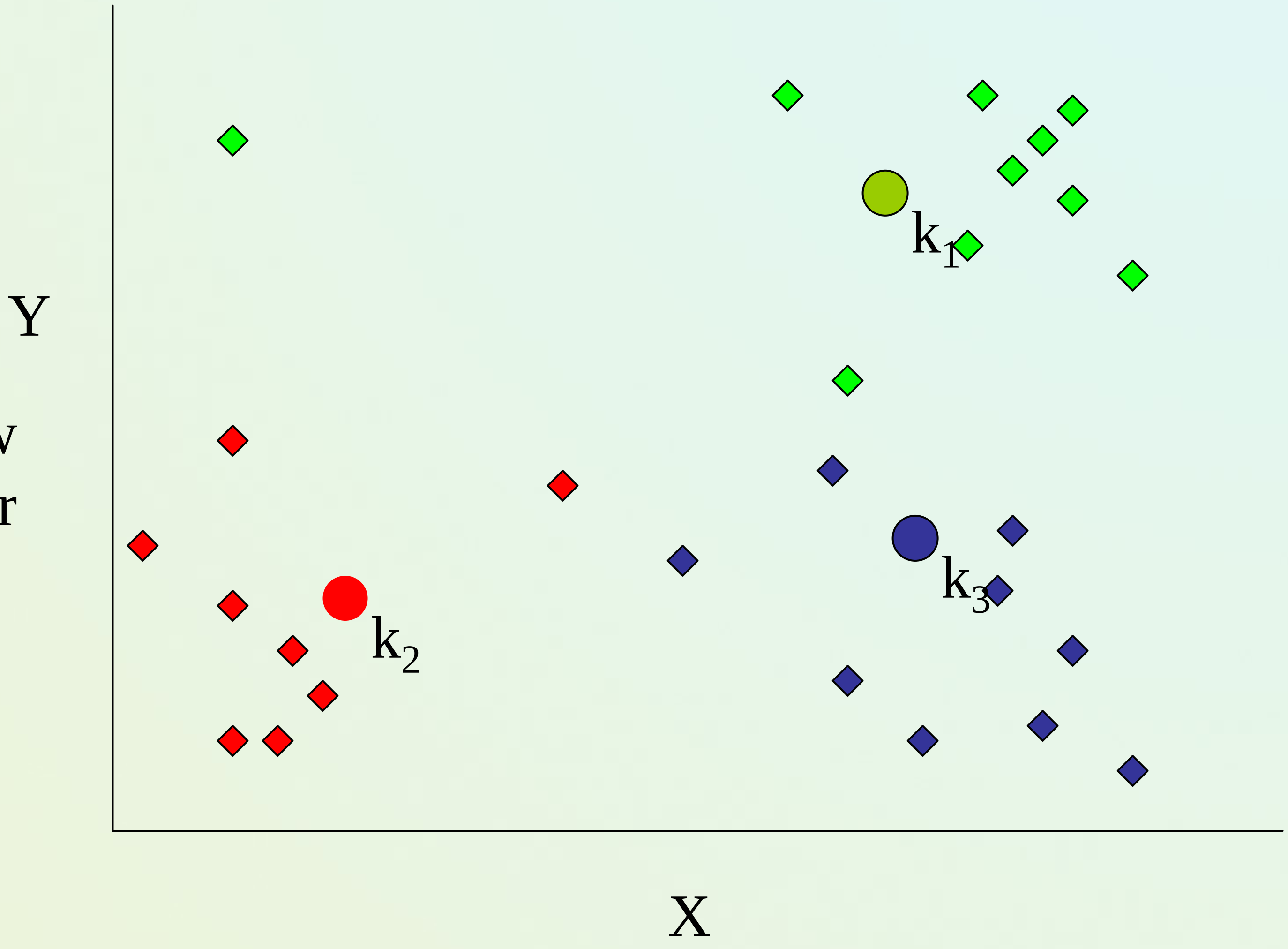
Move
each cluster
center
to the mean
of each cluster

© Stefanowski 2008

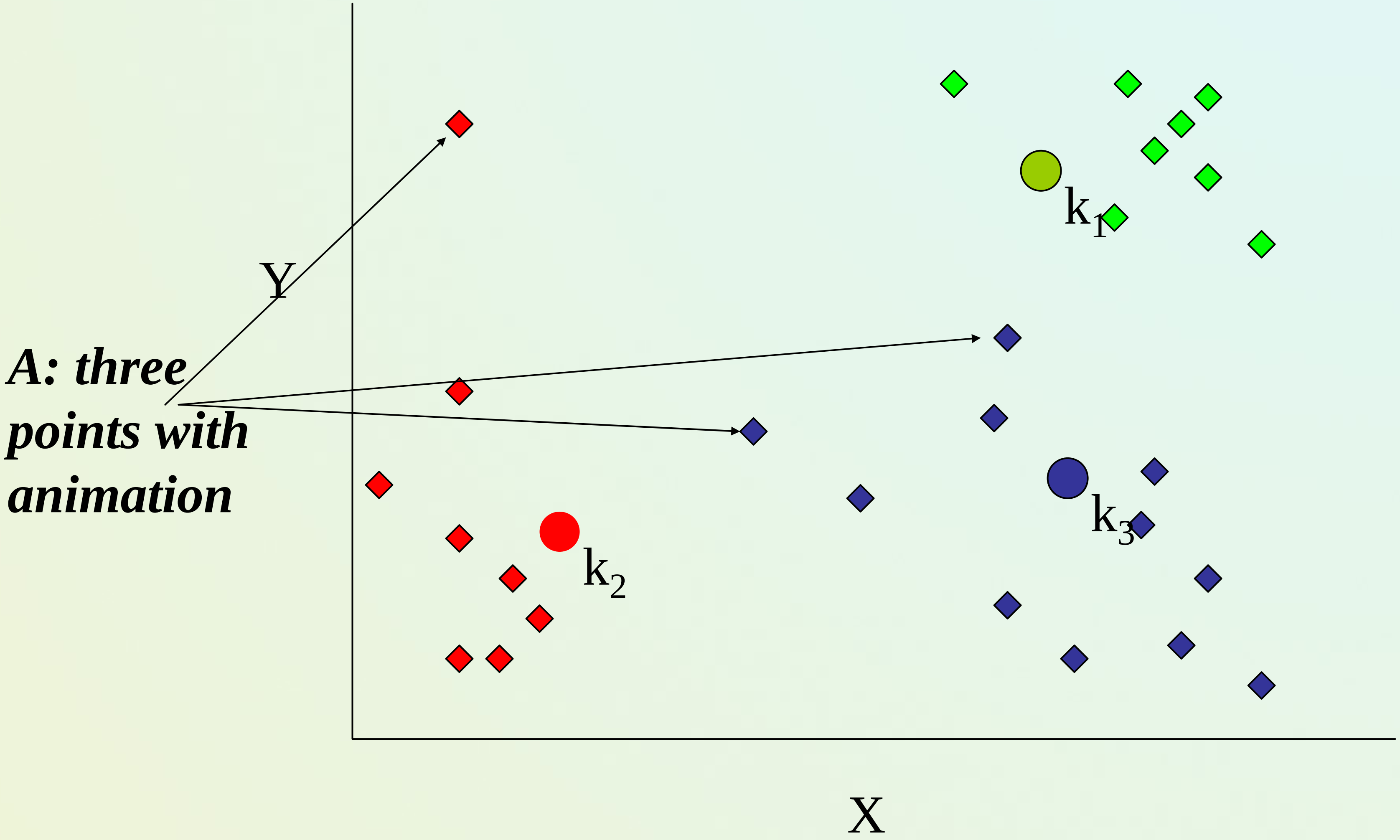
K-means example, step 4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?



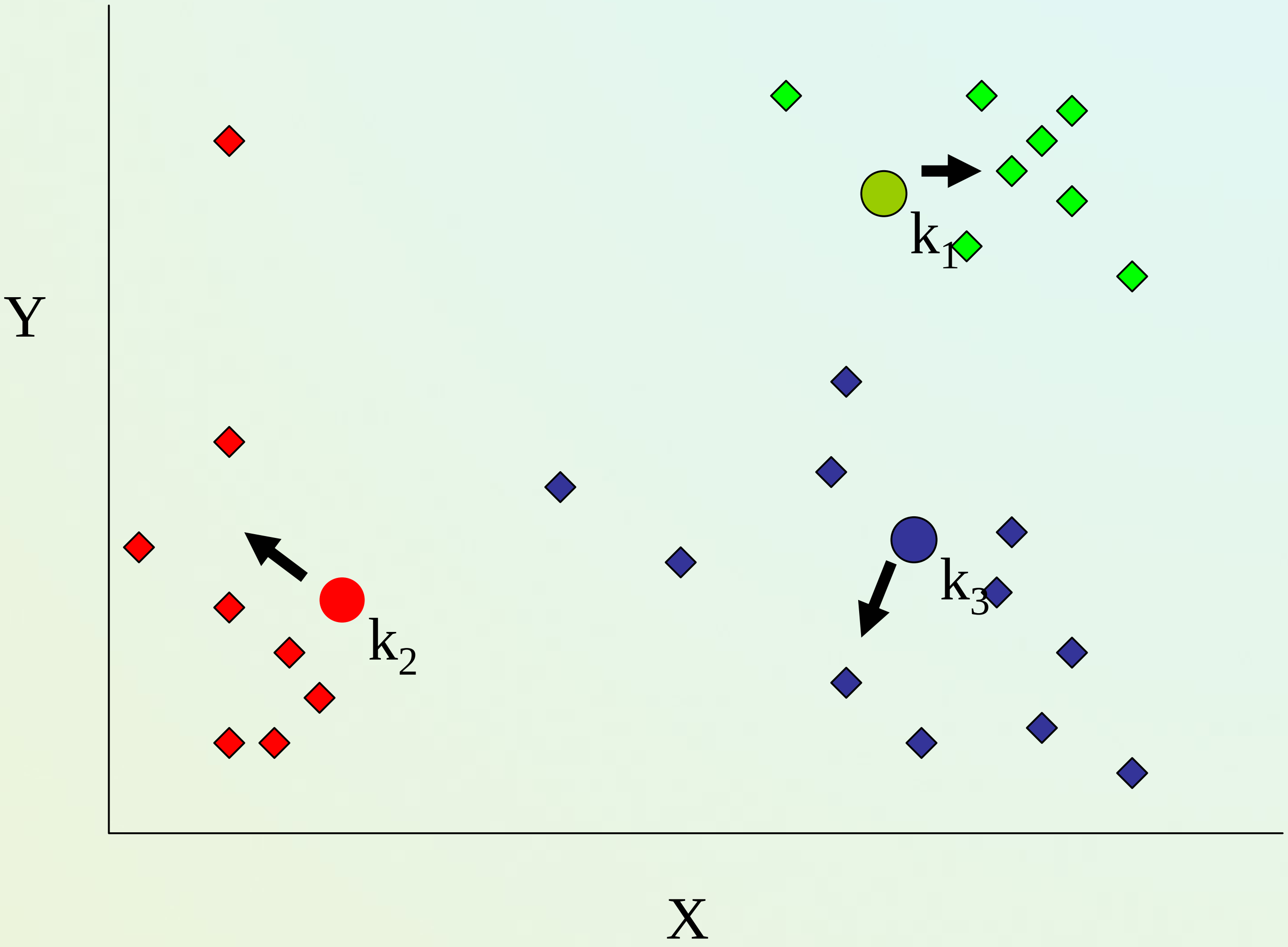
K-means example, step 4 …



A: three points with animation

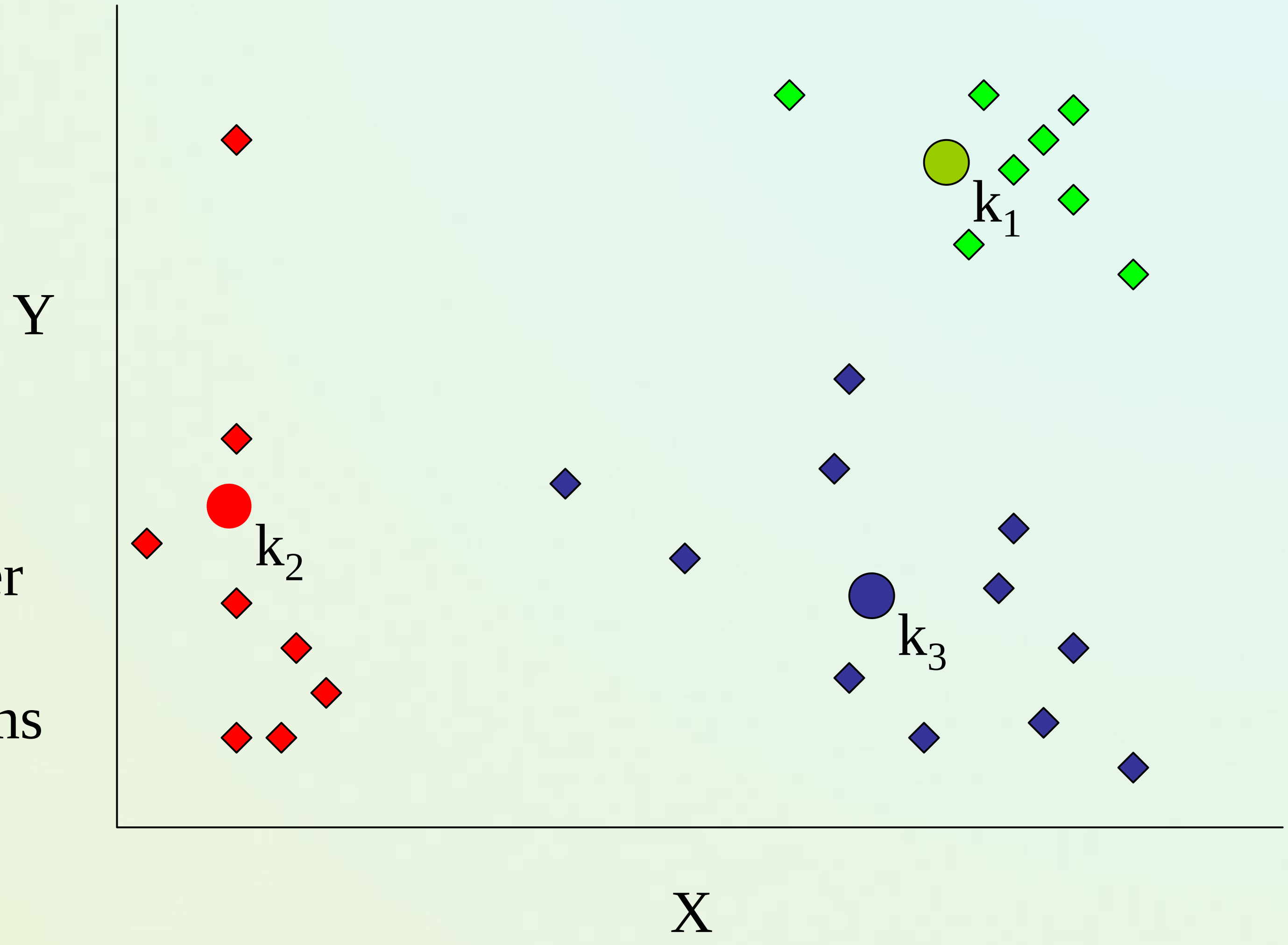
© Stefanowski 2008

K-means example, step 4b



re-compute cluster means

K-means example, step 5



move cluster centers to cluster means

More details on calculatuon (initial matrix)

- Zbiór danych: $x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $x_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ $x_3 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$ $x_4 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $x_5 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$

  $x_6 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$ $x_7 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$ $x_8 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$ $x_9 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$ $x_{10} = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$

- Początkowy przydział

$$B(0) = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

- Przeliczenie centroidów

$$R(0) = \begin{bmatrix} 3 & 3.2 & 2.5 \\ 2 & 2.6 & 2.5 \end{bmatrix}$$

- Przeliczenie odległości

$$D(1) = \begin{bmatrix} 2 & 2.24 & 2.83 & 1.41 & 1 & 1 & 1.41 & 2 & 2.24 & 2.83 \\ 2.28 & 2.24 & 2.61 & 2 & 1.61 & 0.45 & 1.79 & 1.9 & 1.84 & 2.28 \\ 1.58 & 1.58 & 2.12 & 1.58 & 1.58 & 0.71 & 2.12 & 2.55 & 2.55 & 2.91 \end{bmatrix}$$

- Ponowny przydział do skupień:

$$B(1) = \begin{bmatrix} 0 & 0 & 0 & \mathbf{1} & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & \mathbf{1} & 1 \\ \mathbf{1} & \mathbf{1} & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

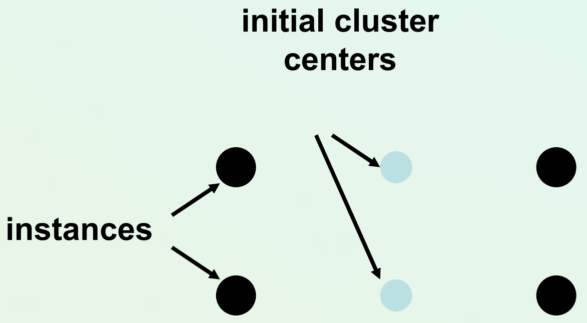
Calculations - more

- Przeliczenie centroidów $R(1) = \begin{bmatrix} 3 & 5 & 1.5 \\ 1 & 3 & 3 \end{bmatrix}$
- Ponowny przydział do skupień: $B(2) = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$
- Warunek końcowy: $B(2) = B(1)$

$$G_1 = \{x_4, x_5, x_7\} \quad G_2 = \{x_8, x_9, x_{10}\} \quad G_3 = \{x_1, x_2, x_3, x_6\}$$

Discussion

- Result can vary significantly depending on initial choice of seeds
- Can get trapped in local minimum
  - Example:



- To increase chance of finding global optimum: restart with different random seeds

K-means clustering summary

Advantages

- Simple, understandable
- items automatically assigned to clusters

Disadvantages

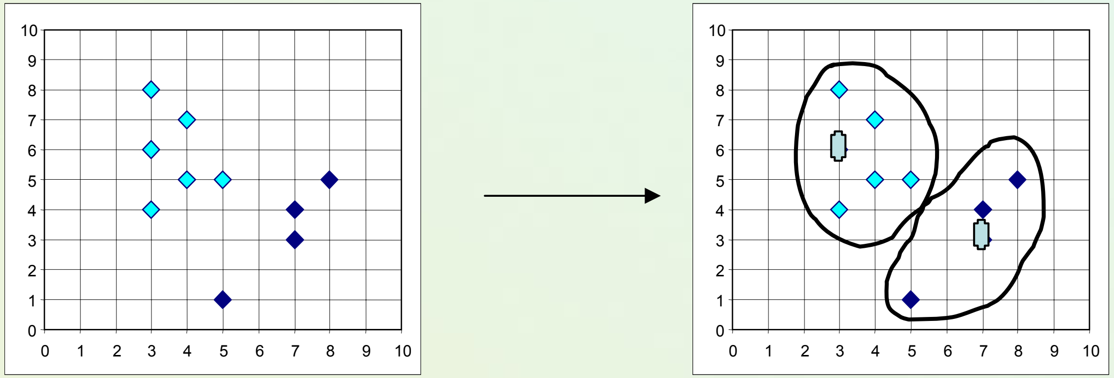
- Must pick number of clusters before hand
- Often terminates at a *local optimum.*
- All items forced into a cluster
- Too sensitive to outliers

Time Complexity

- Assume computing distance between two instances is O(*m*) where *m* is the dimensionality of the vectors.
- Reassigning clusters: O(*kn*) distance computations, or O(*knm*).
- Computing centroids: Each instance vector gets added once to some centroid: O(*nm*).
- Assume these two steps are each done once for *I* iterations: O(*Iknm*).
- Linear in all relevant factors, assuming a fixed number of iterations, more efficient than O(n^2) HAC.

What is the problem of k-Means Method?

- The k-means algorithm is sensitive to outliers !
  - Since an object with an extremely large value may substantially distort the distribution of the data.
- There are other limitations – still a need for reducing costs of calculating distances to centroids.
- **K-Medoids**: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.

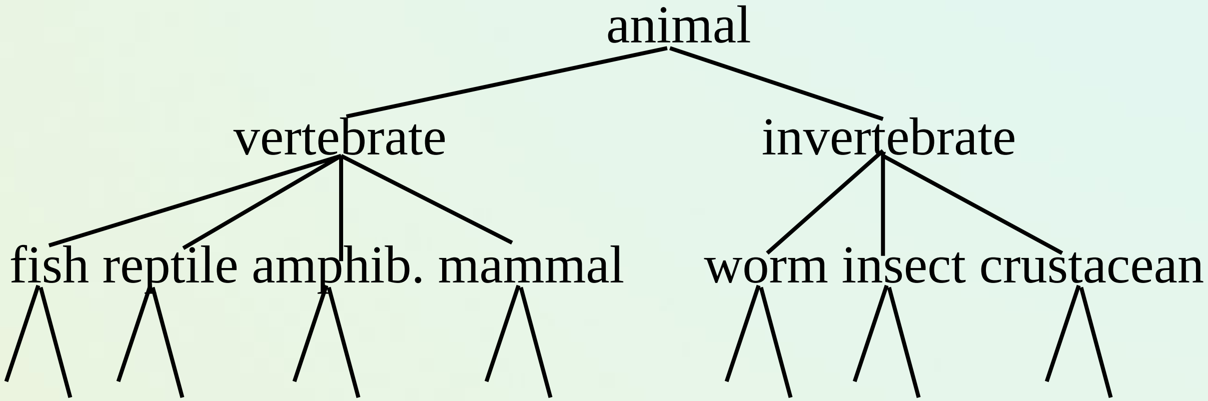


The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
  - To achieve this goal, only the definition of distance from any two objects is needed.
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering.
  - *PAM* works effectively for small data sets, but does not scale well for large data sets.
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): Randomized sampling.
- Focusing + spatial data structure (Ester et al., 1995).

Hierarchical Clustering

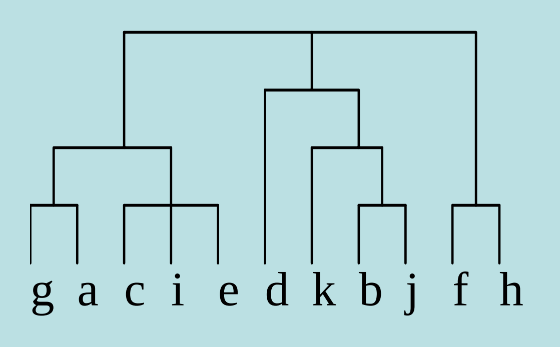
- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of unlabeled examples.



- Recursive application of a standard clustering algorithm can produce a hierarchical clustering.

*Hierarchical clustering

- Bottom up (aglomerative)
  - Start with single-instance clusters
  - At each step, join the two closest clusters
  - Design decision: distance between clusters
    - e.g. two closest instances in clusters vs. distance between means
- Top down (divisive approach / deglomerative)
  - Start with one universal cluster
  - Find two clusters
  - Proceed recursively on each subset
  - Can be very fast
- Both methods produce a *dendrogram*



HAC Algorithm (aglomerative)

```
Start with all instances in their own cluster.
Until there is only one cluster:
    Among the current clusters, determine the two
        clusters, $c_i$ and $c_j$, that are most similar.
    Replace $c_i$ and $c_j$ with a single cluster $c_i \cup c_j$
```

Distance between Clusters

Single linkage
minimum distance:

$$d_{\min}(C_{i,}C_j) = \min_{p\in C_i, p'\in C_j} \left\| p - p' \right\|$$

Complete linkage
maximum distance:

$$d_{\max}(C_{i,}C_j) = \max_{p\in C_i, p'\in C_j} \left\| p - p' \right\|$$

mean distance:

$$d_{mean}(C_{i,}C_j) = \left\| m_i - m_j \right\|$$

average distance:

$$d_{ave}(C_{i,}C_j) = 1/(n_i n_j) \sum_{p\in C_i} \sum_{p'\in C_j} \left\| p - p' \right\|$$

m_i is the mean for cluster C_i n_i is the number of points in C_i

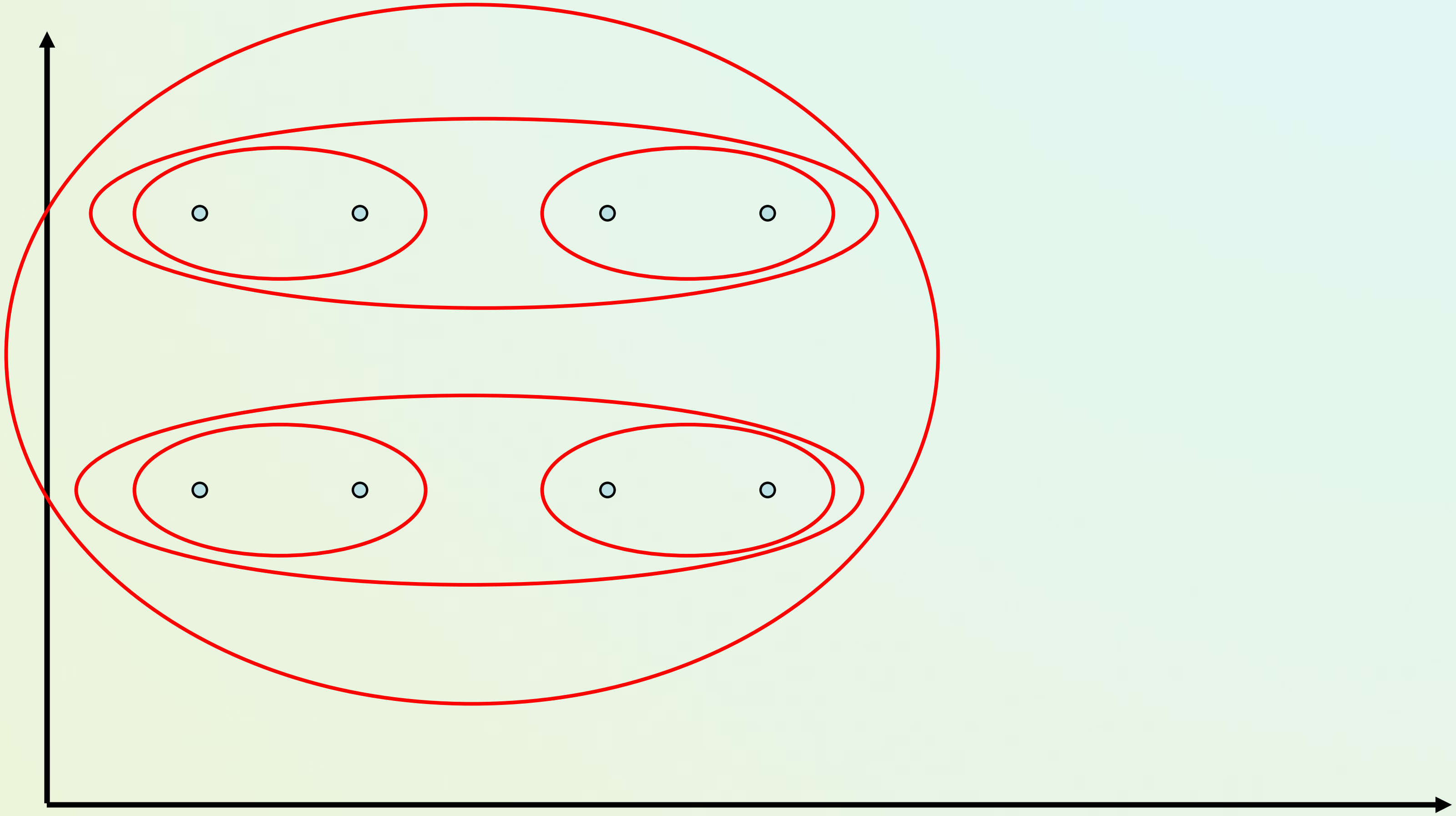
Single Link Agglomerative Clustering

- Use minium similarity of pairs:

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- Can result in “straggly” (long and thin) clusters due to *chaining effect*.
  - Appropriate in some domains, such as clustering islands.

Single Link Example



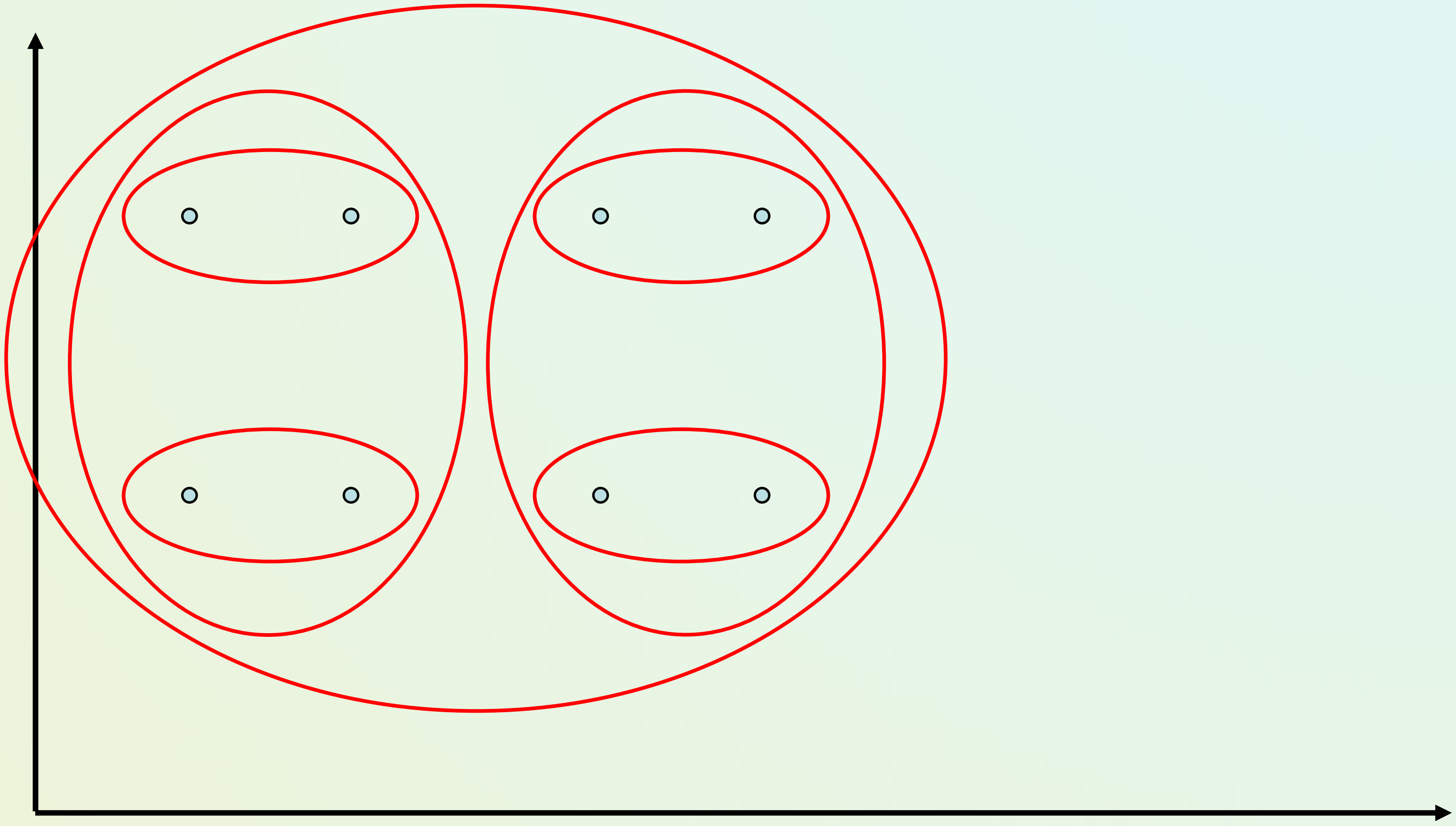
Complete Link Agglomerative Clustering

- Use maximum similarity of pairs:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Makes more “tight,” spherical clusters that are typically preferable.

Complete Link Example



Single vs. Complete Linkage

- A.Jain et al.: Data Clustering. A Review.

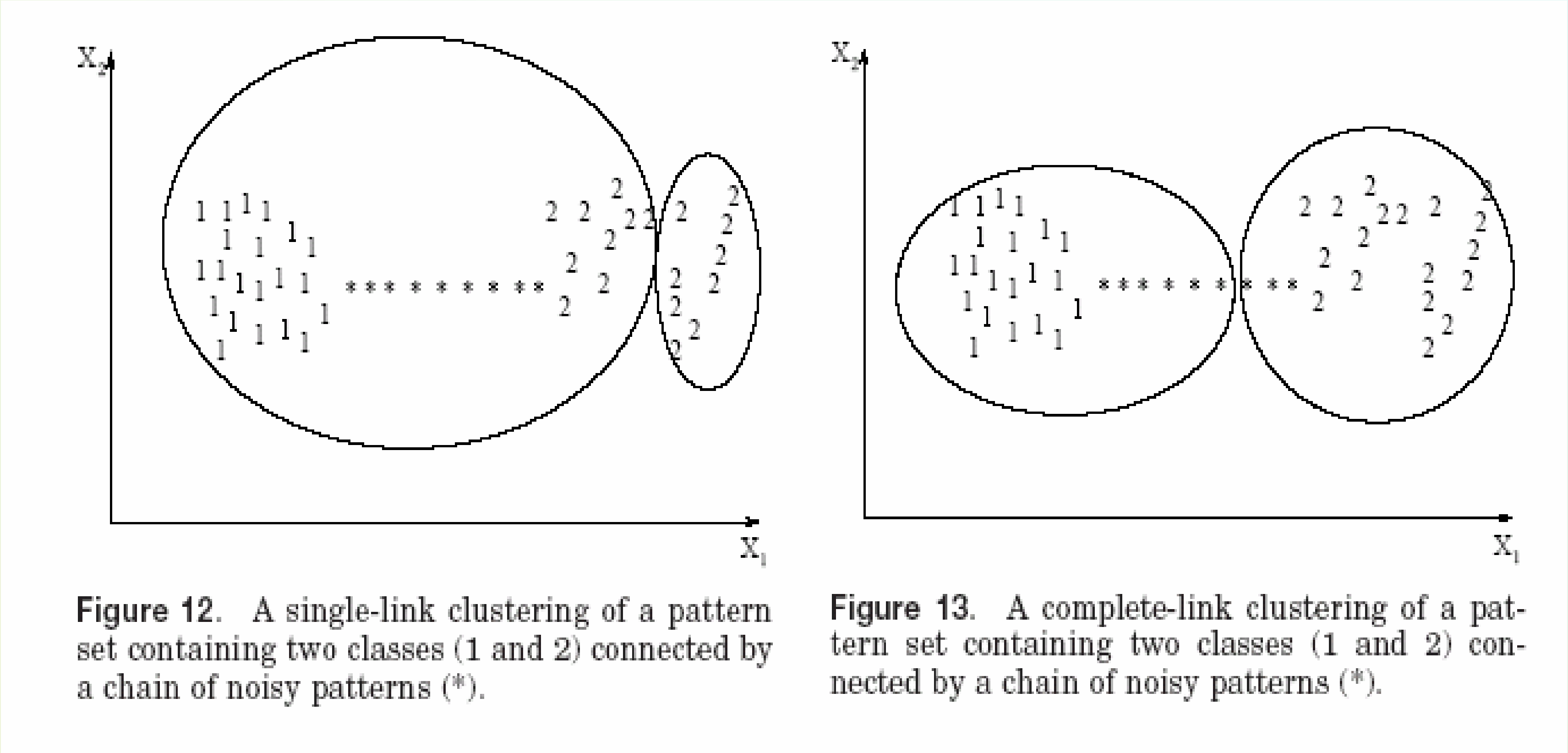
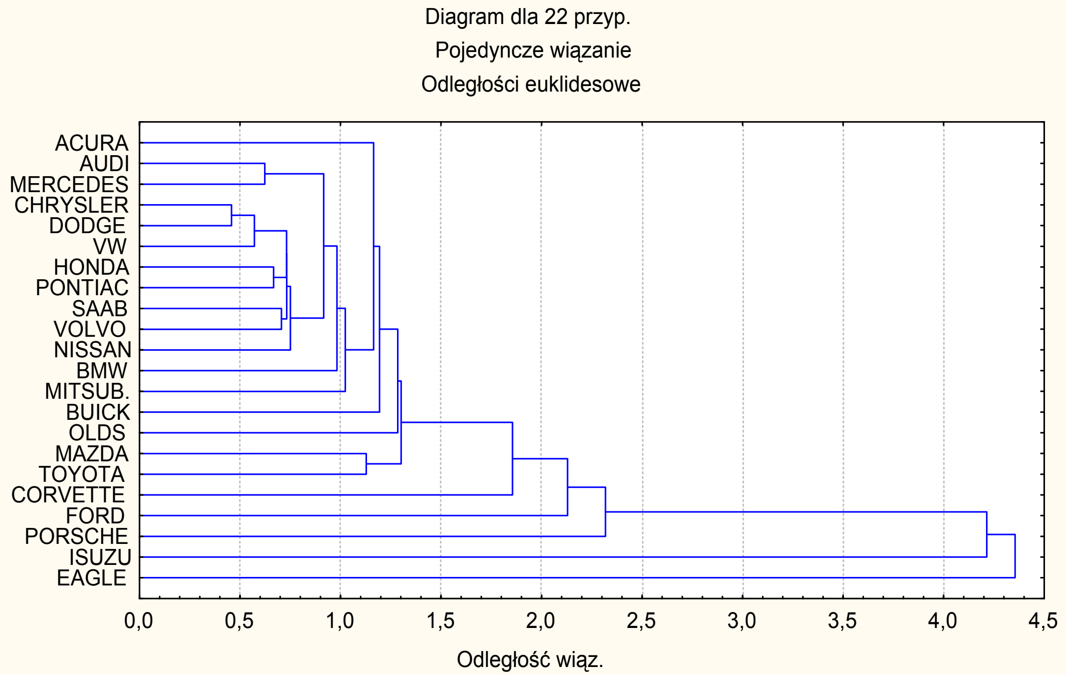
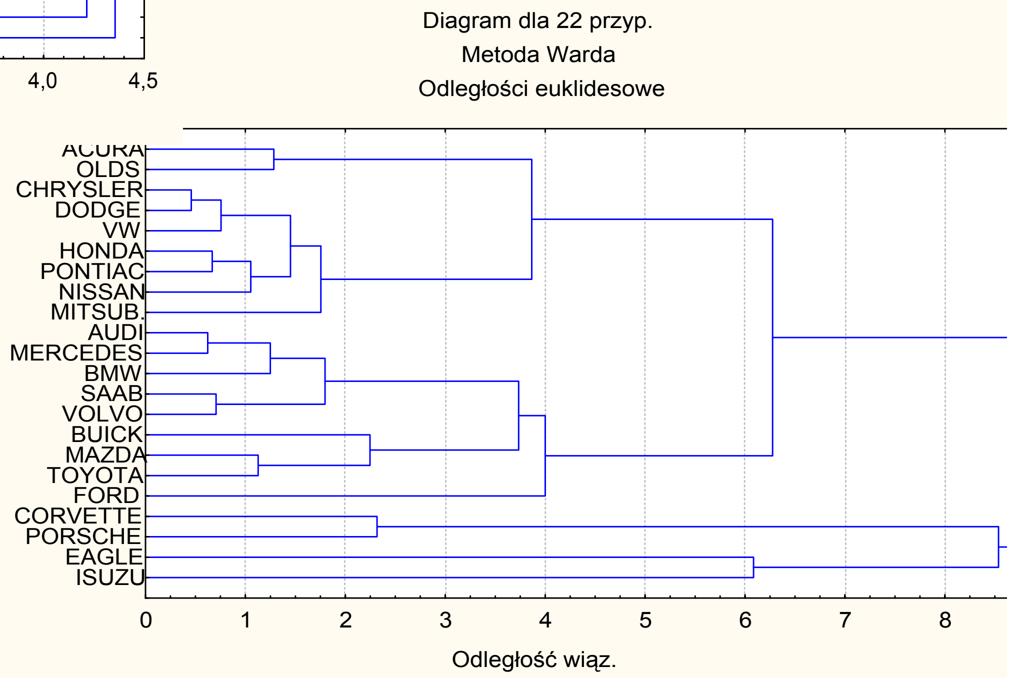


Figure 12. A single-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

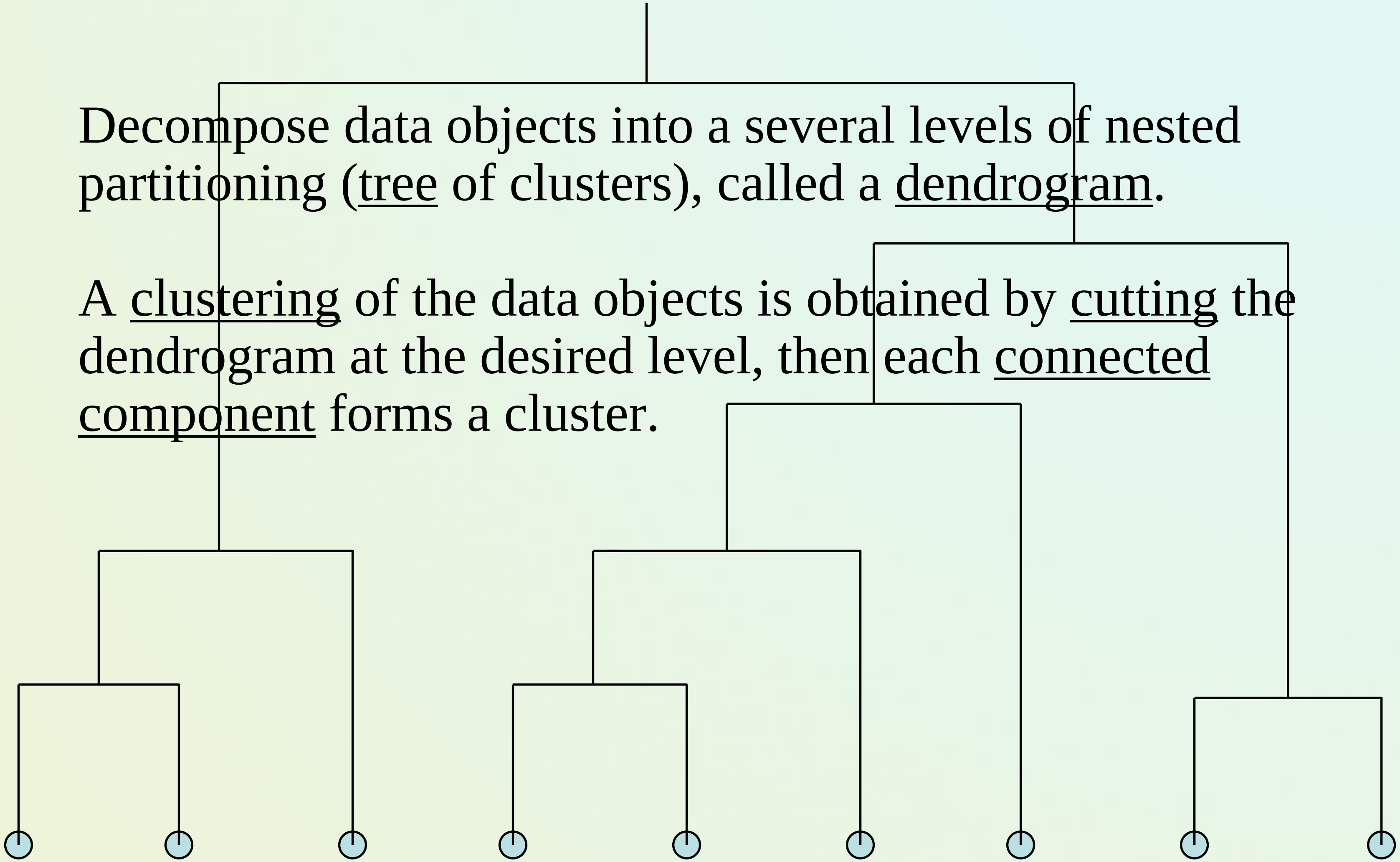
Figure 13. A complete-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

Changing Linkage Methods





Dendrogram: Shows How the Clusters are Merged

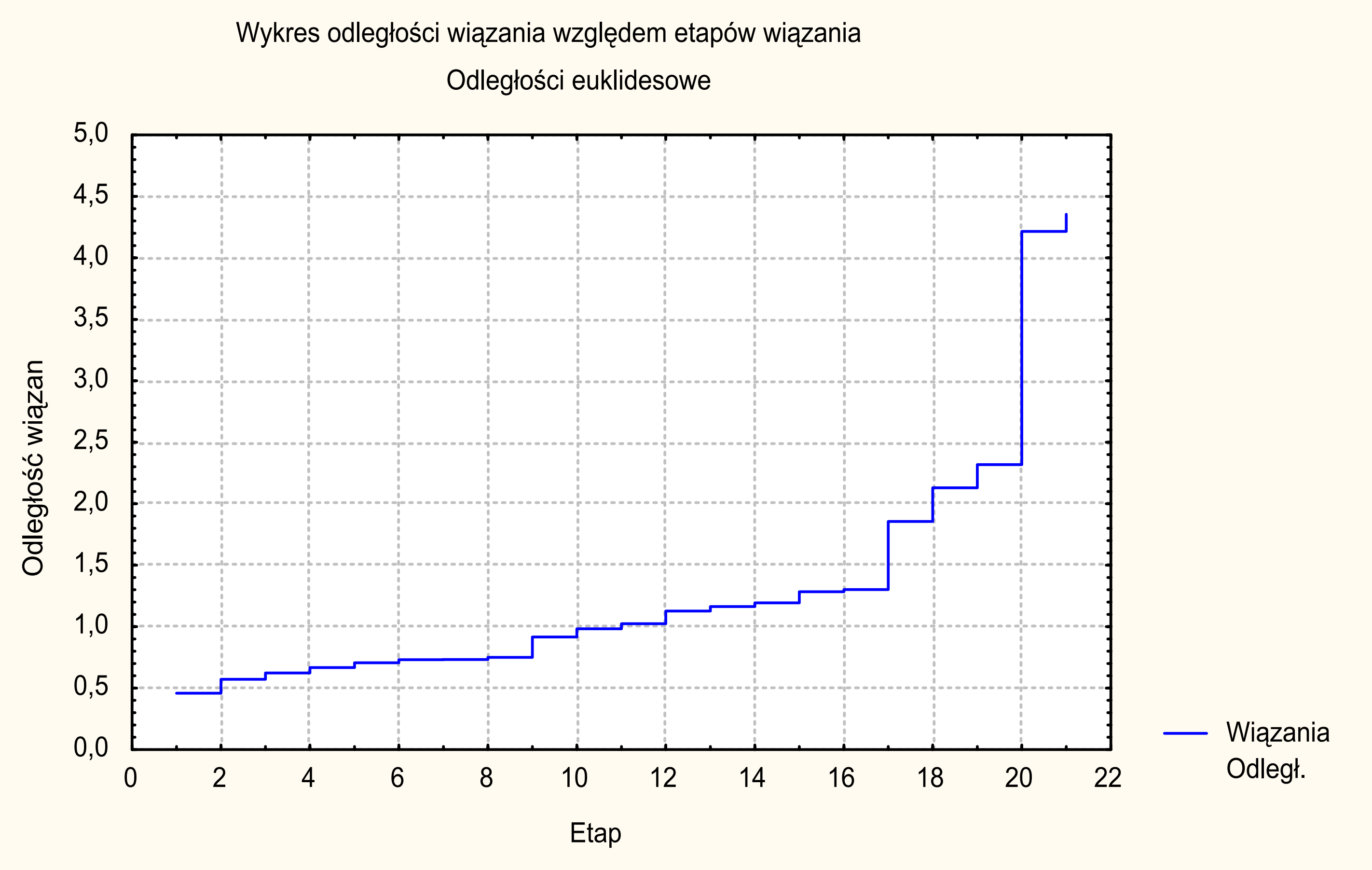


Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

AHC Steps - cutting

- Figure – find a cut point („kolanko" / knee)



Computational Complexity

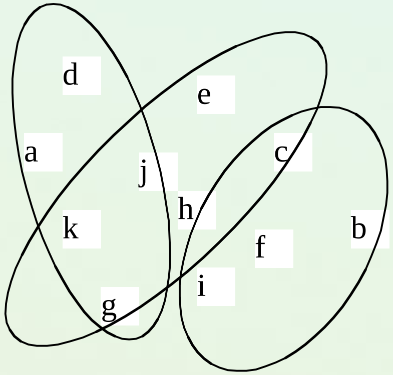
- In the first iteration, all HAC methods need to compute similarity of all pairs of n individual instances which is O(n^2).
- In each of the subsequent n–2 merging iterations, it must compute the distance between the most recently created cluster and all other existing clusters.
- In order to maintain an overall O(n^2) performance, computing similarity to each other cluster must be done in constant time.

More on Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods:
  - do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
  - can never undo what was done previously.
- Integration of hierarchical clustering with distance-based method:
  - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters.
  - CURE (1998): selects well-scattered points from the cluster and then shrinks them towards the center of the cluster by a specified fraction.

Soft Clustering

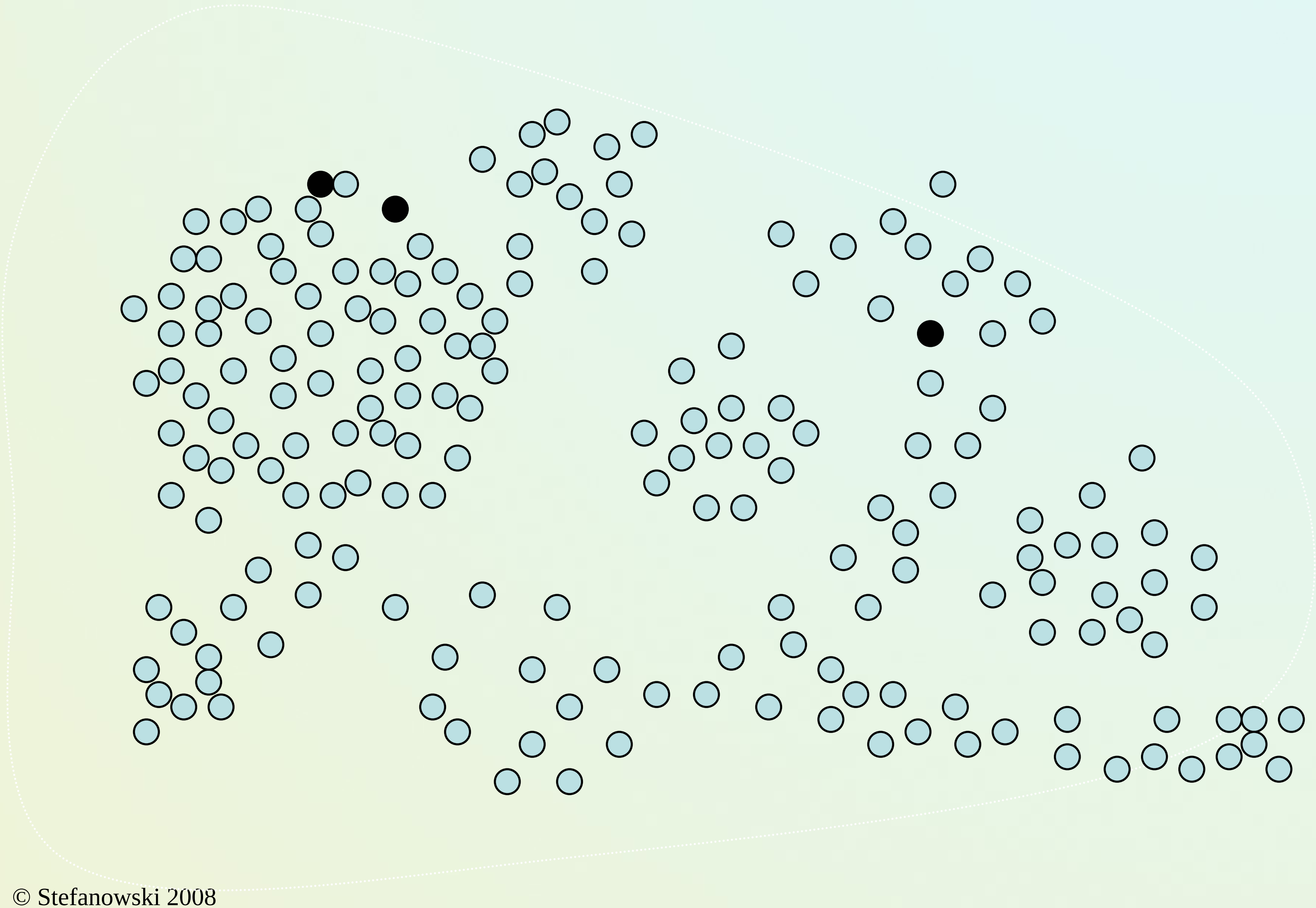
- Clustering typically assumes that each instance is given a "hard" assignment to exactly one cluster.
- Does not allow uncertainty in class membership or for an instance to belong to more than one cluster.
- *Soft clustering* gives probabilities that an instance belongs to each of a set of clusters.
- Each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1).



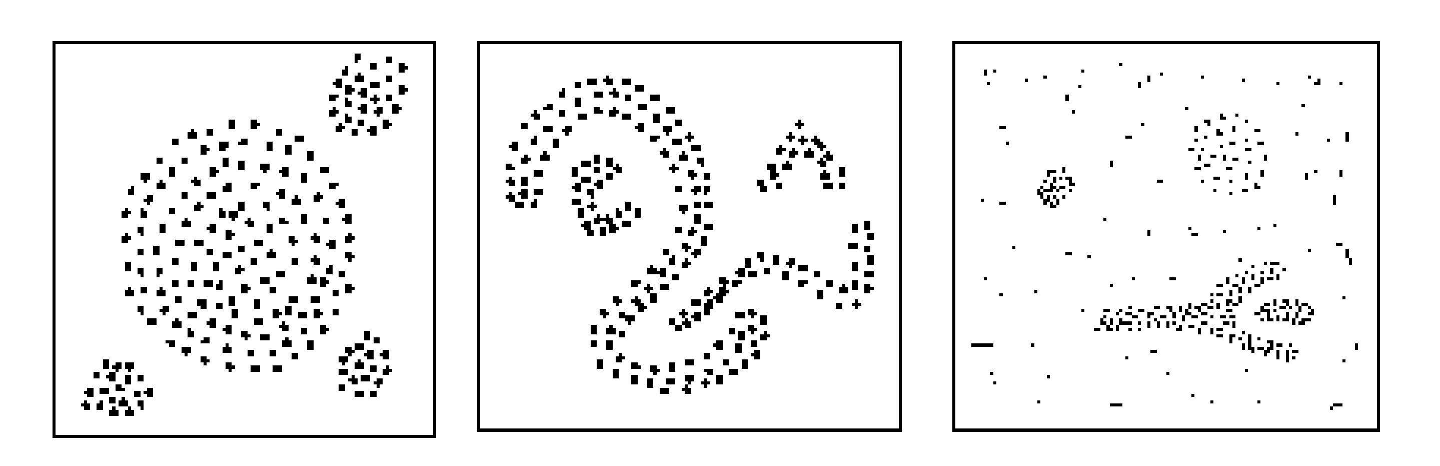
Expectation Maximization (EM Algorithm)

- Probabilistic method for soft clustering.
- Direct method that assumes *k* clusters:$\{c_1, c_2, \ldots c_k\}$
- Soft version of *k*-means.
- Assumes a probabilistic model of categories that allows computing $P(c_i \mid E)$ for each category, c_i, for a given example, *E*.
- For text, typically assume a naïve-Bayes category model.
  - Parameters $\theta = \{P(c_i), P(w_j \mid c_i): i \in \{1,\ldots k\}, j \in \{1,\ldots,|V|\}\}$

Handling Complex Shaped Clusters

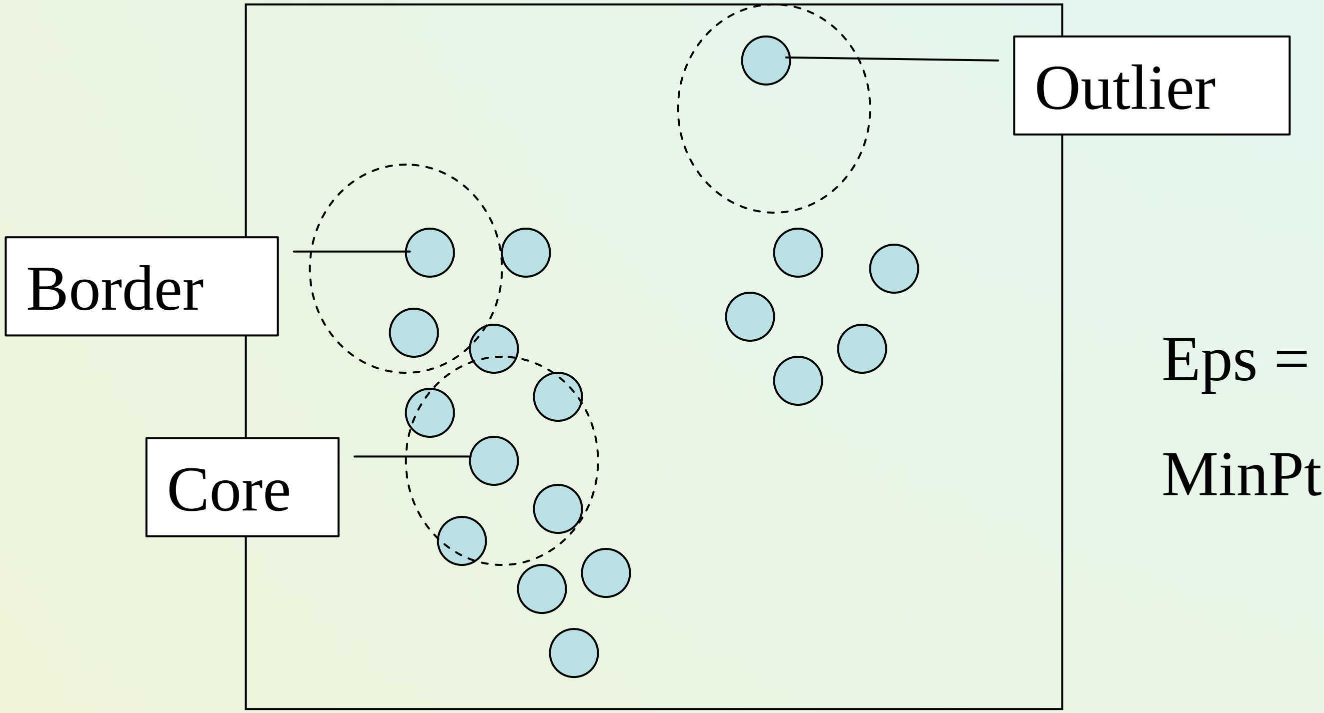


Density-Based Clustering



- Clustering based on density (local cluster criterion), such as density-connected points
- Each cluster has a considerable higher density of points than outside of the cluster

DBSCAN: General Ideas



Outlier

Border

Core

Eps = 1cm

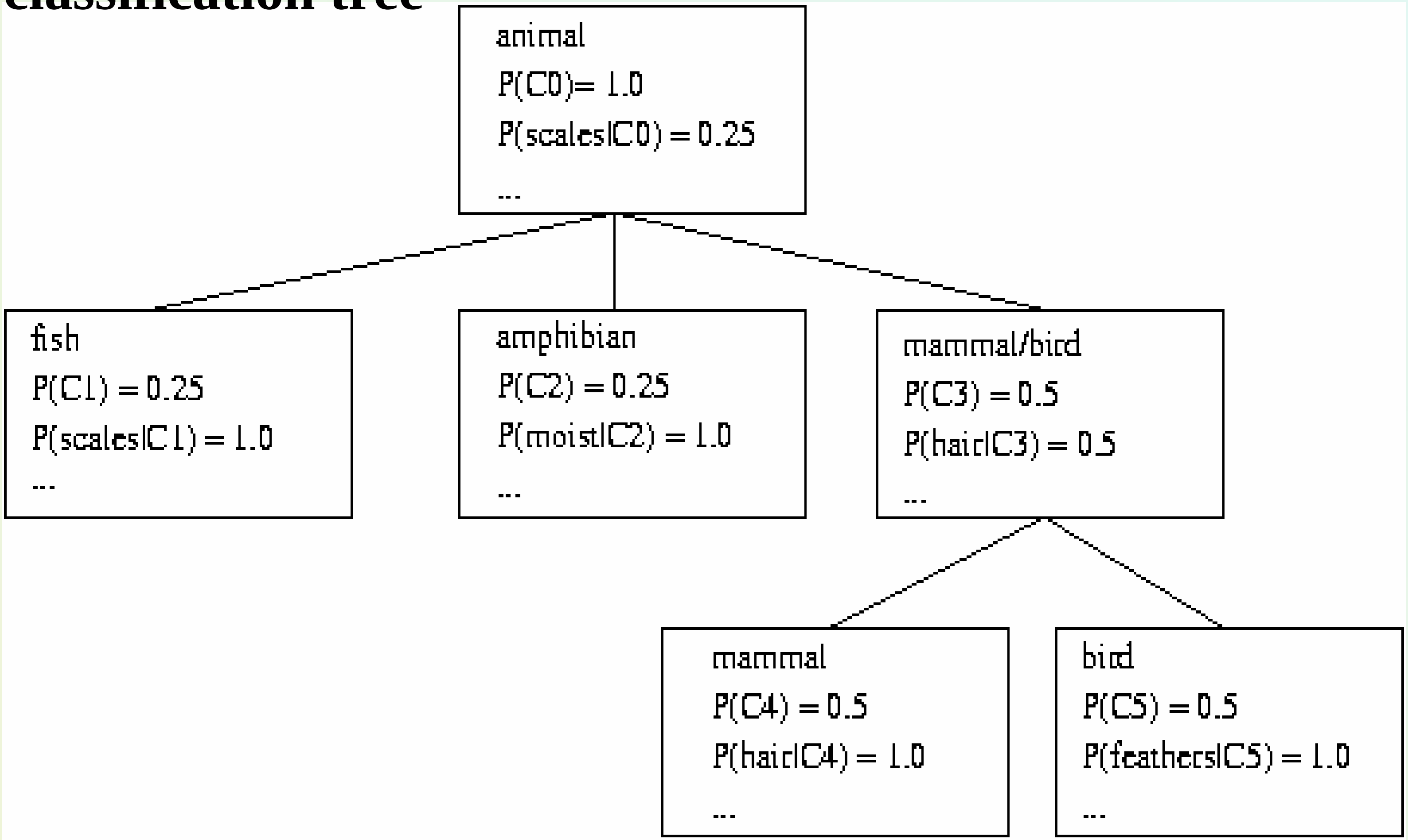
MinPts = 5

Model-Based Clustering Methods

- Attempt to optimize the fit between the data and some mathematical model
- Statistical and AI approach
  - Conceptual clustering
    - A form of clustering in machine learning
    - Produces a classification scheme for a set of unlabeled objects
    - Finds characteristic description for each concept (class)
  - COBWEB (Fisher'87)
    - A popular a simple method of incremental conceptual learning
    - Creates a hierarchical clustering in the form of a classification tree
    - Each node refers to a concept and contains a probabilistic description of that concept

COBWEB Clustering Method

A classification tree



animal
P(C0)= 1.0
P(scales|C0) = 0.25
...

fish
P(C1) = 0.25
P(scales|C1) = 1.0
...

amphibian
P(C2) = 0.25
P(moist|C2) = 1.0
...

mammal/bird
P(C3) = 0.5
P(hair|C3) = 0.5
...

mammal
P(C4) = 0.5
P(hair|C4) = 1.0
...

bird
P(C5) = 0.5
P(feathers|C5) = 1.0
...

*Incremental clustering (COBWEB based)

- Heuristic approach (COBWEB/CLASSIT)
- Form a hierarchy of clusters incrementally
- Start:
  - tree consists of empty root node
- Then:
  - add instances one by one
  - update tree appropriately at each stage
  - to update, find the right leaf for an instance
  - May involve restructuring the tree
- Base update decisions on *category utility*

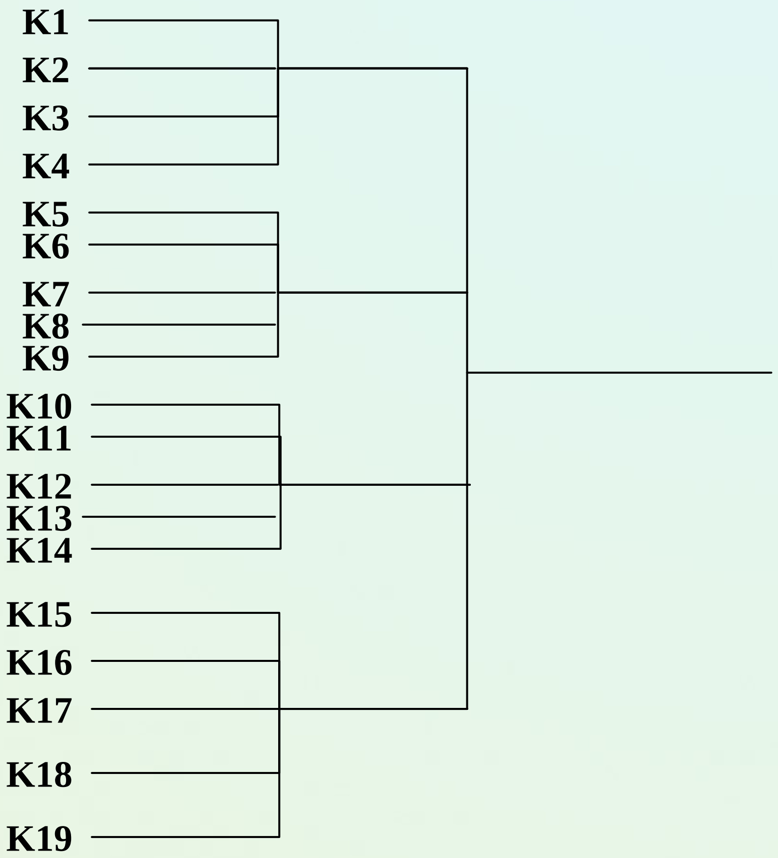
World countries data

Kraj	C1	C2	C3	C4	C5	C6	C7	C8	C9
Afganistan	M	AP	S	N	N	N	N	N	S
Argentyna	K	AL.	U	N	S	W	W	W	N
Armenia	O	SW	SM	S	S	W	W	W	N
Australia	P	OECD	S	N	S	W	W	W	N
Austria	K	OECD	U	N	W	W	W	W	N
Azerbejdżar	M	SW	S	N	W	W	W	W	N
Belgia	K	OCED	U	W	S	W	W	W	N
Białoruś	O	EW	U	N	W	W	S	S	N
Boliwia	K	A	SM	N	W	S	S	S	S
...	...	...	...	...	...	...	...	...	...

COBWB results

Selected classes

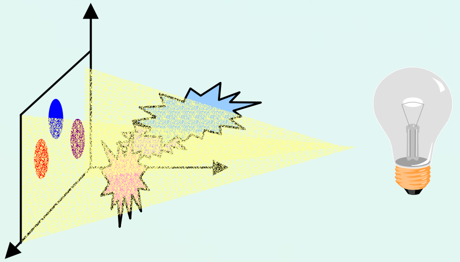
- K1: Rosja, Portugalia, Polska, Litwa, Łotwa, Węgry, Grecja, Gruzja, Estonia, Czechy, Chorwacja
- K2: USA, Szwajcaria, Hiszpania, Norwegia, Holandia, Włochy, Irlandia, Niemcy, Francja, Dania, Belgia, Austria
- K3: Szwecja, Korea Płd., Nowa Zelandia, Finlandia, Kanada, Australia, Islandia
- ...
- K17: Somalia, Gambia, Etiopia, Kambodża
- K18: Uganda, Tanzania, Ruanda, Haiti, Burundi
- ...



Other Model-Based Clustering Methods

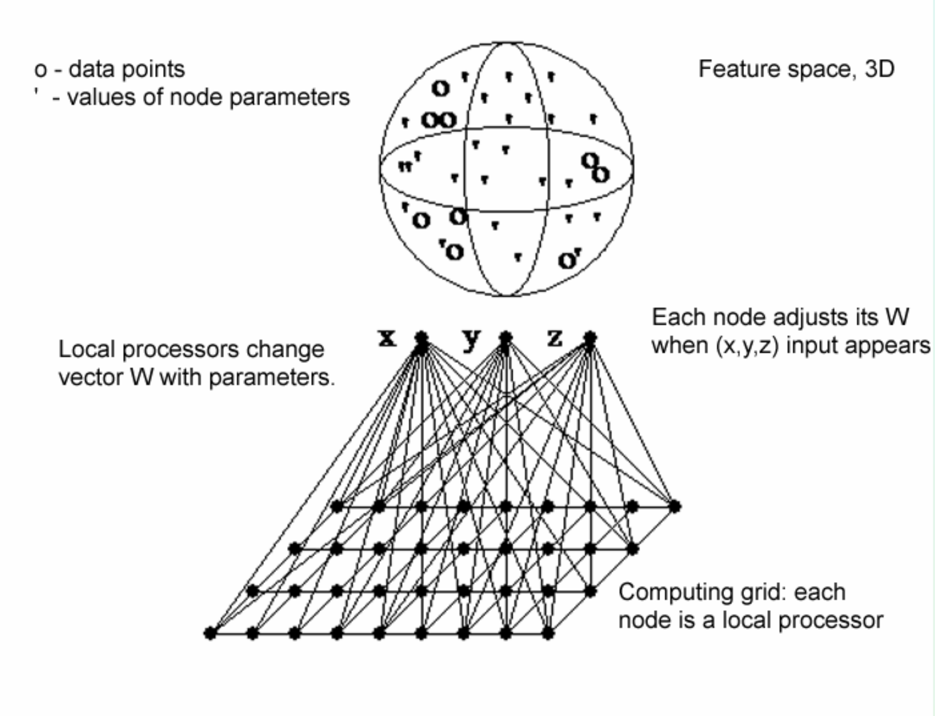
- Neural network approaches
  - Represent each cluster as an exemplar, acting as a “prototype” of the cluster
  - New objects are distributed to the cluster whose exemplar is the most similar according to some dostance measure
- Competitive learning
  - Involves a hierarchical architecture of several units (neurons)
  - Neurons compete in a “winner-takes-all” fashion for the object currently being presented

Kohonen's Self-Organizing Map (SOM)



- Another Clustering Algorithm
  - *aka* Self-Organizing Feature Map (SOFM)
  - Given: vectors of attribute values (x_1, x_2, …, x_n)
  - Returns: vectors of attribute values (x_1', x_2', …, x_k')
    - Typically, *n* >> *k* (*n* is high, *k* = 1, 2, or 3; hence "dimensionality reducing")
    - Output: vectors *x'*, the projections of input points *x*; also get $P(x_j' \mid x_i)$
    - Mapping from *x* to *x'* is topology preserving
- Topology Preserving Networks
  - Intuitive idea: similar input vectors will map to similar clusters
  - Recall: informal definition of cluster (isolated set of mutually similar entities)
  - Restatement: "clusters of *X* (high-D) will still be clusters of *X'* (low-D)"
- Representation of Node Clusters
  - Group of neighboring artificial neural network units (neighborhood of nodes)
  - SOMs: combine ideas of topology-preserving networks, unsupervised learning
- Implementation: http://www.cis.hut.fi/nnrc/ and *MATLAB* NN Toolkit

Self-Organizing Maps - more


o - data points
' - values of node parameters
Feature space, 3D
Local processors change vector W with parameters.
x
y
z
Each node adjusts its W when (x,y,z) input appears
Computing grid: each node is a local processor

Data: vectors $\mathbf{X}^T = (X_1, \ldots X_d)$ from d-dimensional space.

Grid of nodes, with local processor (called neuron) in each node.

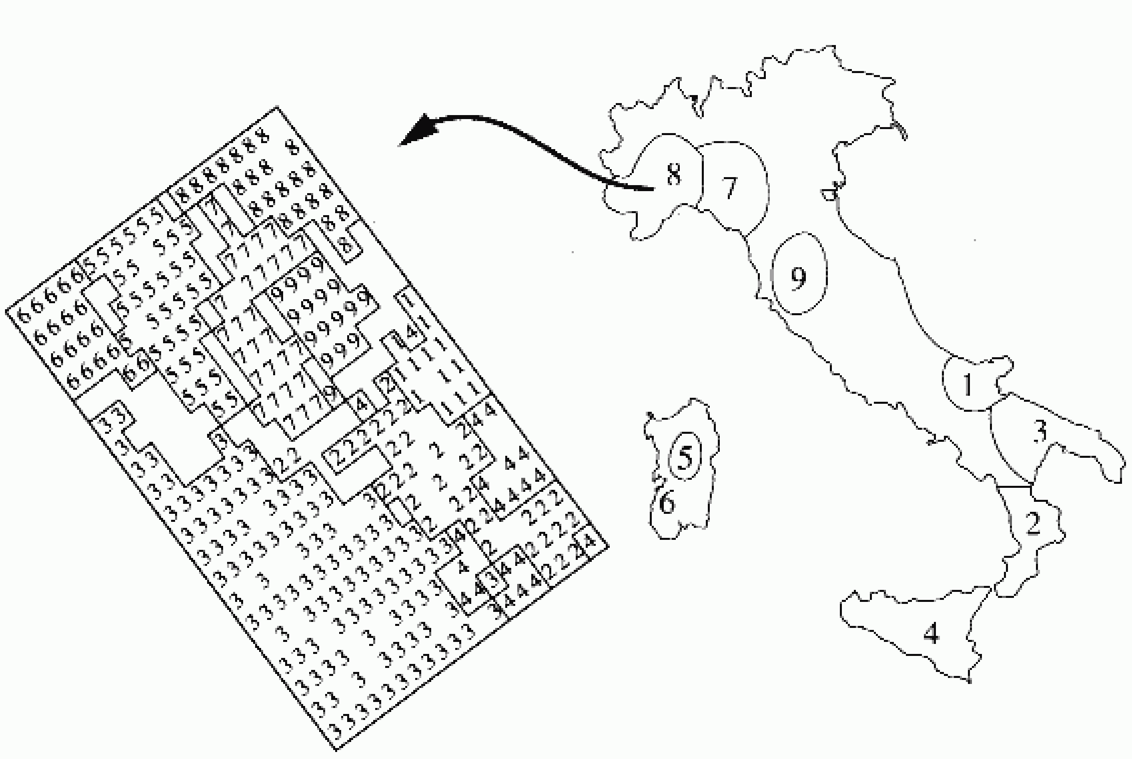
Local processor # *j* has *d* adaptive parameters $\mathbf{W}^{(j)}$.

Goal: change $\mathbf{W}^{(j)}$ parameters to recover data clusters in **X** space.

An example of analysing olive oil in Italy

An example of SOM application:

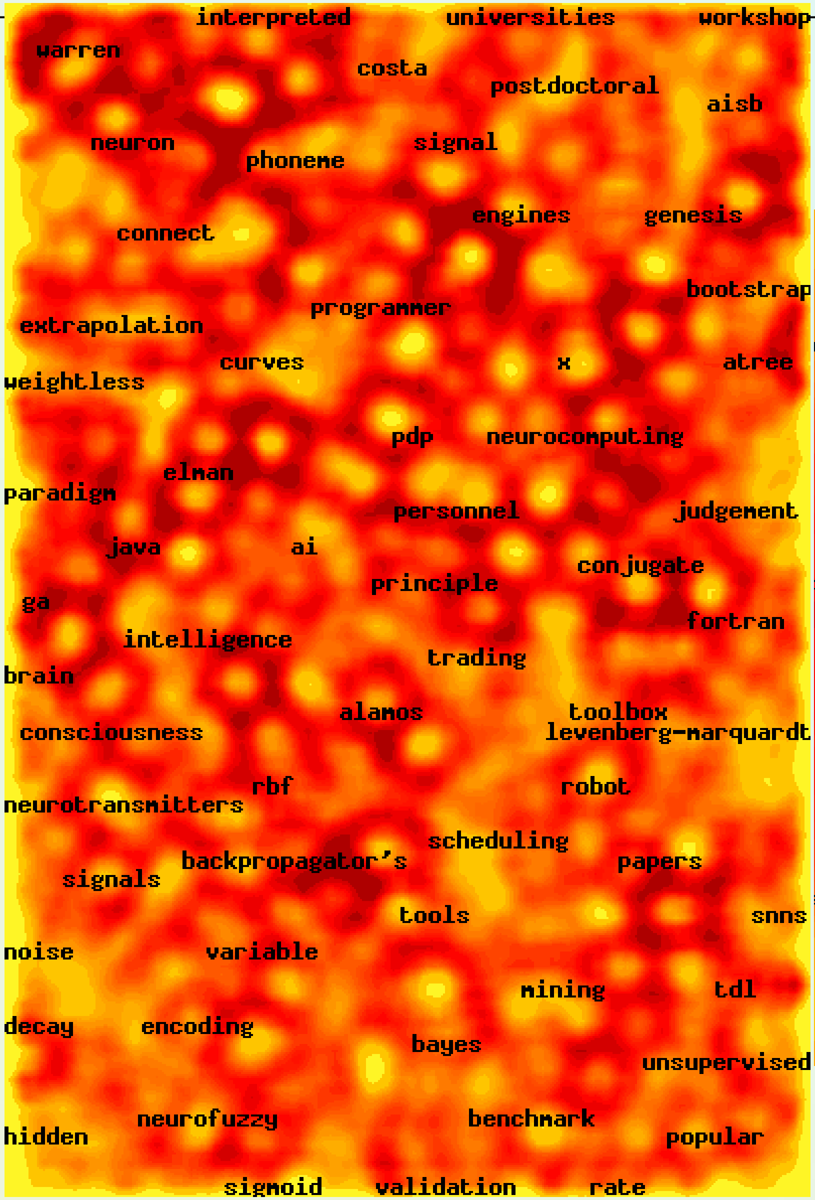
- 572 samples of olive oil were collected from 9 Italian provinces. Content of 8 fats was determine for each oil.
- SOM 20 x 20 network,
- Maps 8D => 2D.
- Classification accuracy was around 95-97%.

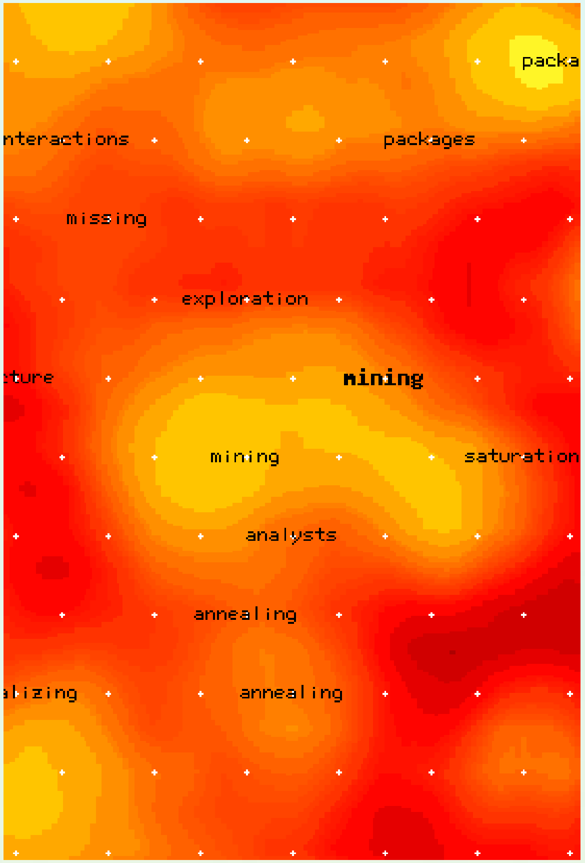


Note that topographical relations are preserved, region 3 is most diverse.

Web Document Clustering Using SOM

- The result of SOM clustering of 12088 Web articles
- The picture on the right: drilling down on the keyword “mining”
- Based on websom.hut.fi Web page





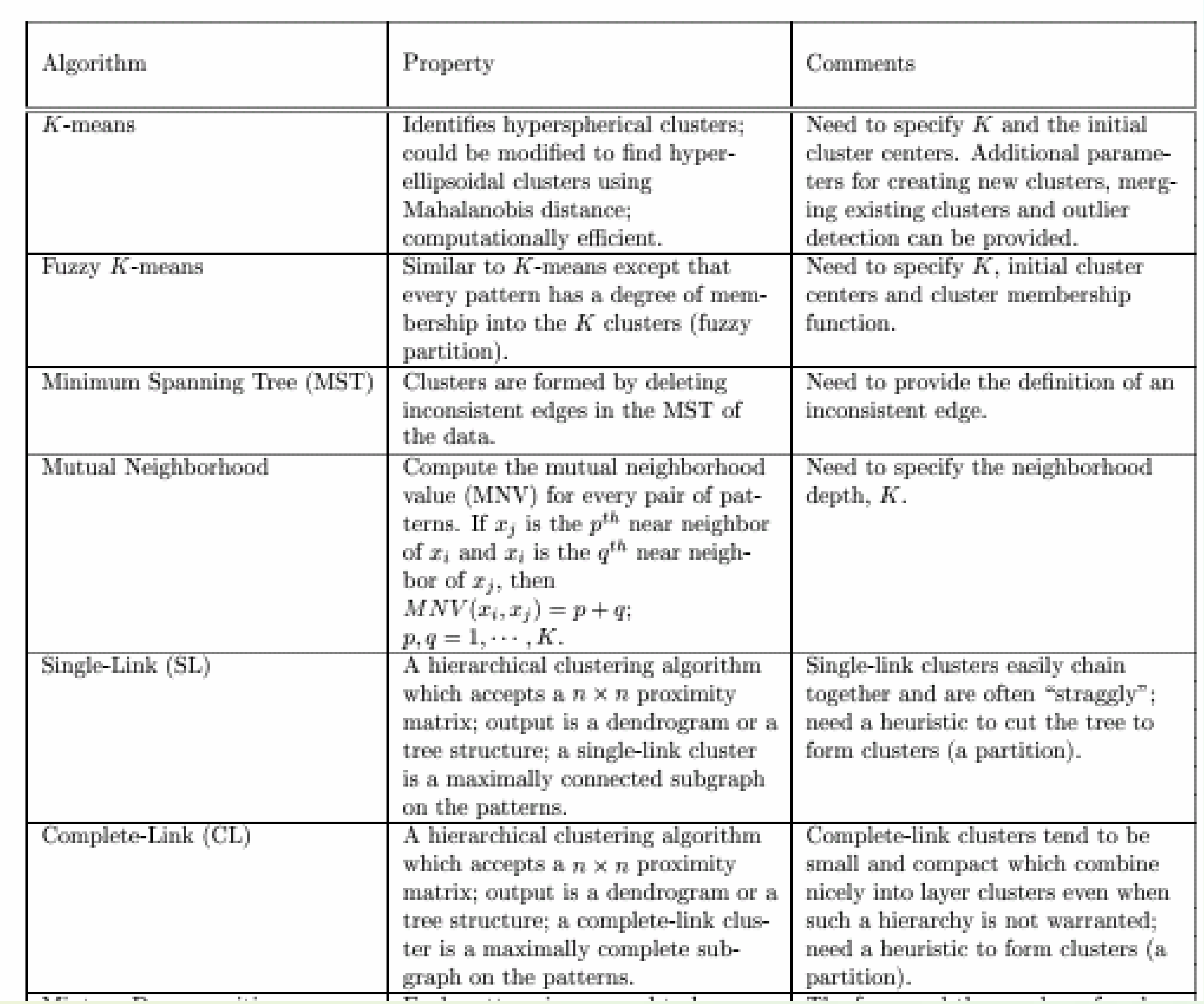
Other examples

- Natural language processing: linguistic analysis, parsing, learning languages, hyphenation patterns.
- Optimization: configuration of telephone connections, VLSI design, time series prediction, scheduling algorithms.
- Signal processing: adaptive filters, real-time signal analysis, radar, sonar seismic, USG, EKG, EEG and other medical signals ...
- Image recognition and processing: segmentation, object recognition, texture recognition ...
- Content-based retrieval: examples of WebSOM, Cartia, VisierPicSom – similarity based image retrieval.

Clustering High-Dimensional Data

- Clustering high-dimensional data
    - Many applications: text documents, DNA micro-array data
    - Major challenges:
        - Many irrelevant dimensions may mask clusters
        - Distance measure becomes meaningless—due to equi-distance
        - Clusters may exist only in some subspaces
- Methods
    - Feature transformation: only effective if most dimensions are relevant
        - PCA & SVD useful only when features are highly correlated/redundant
    - Feature selection: wrapper or filter approaches
        - useful to find a subspace where the data have nice clusters
    - Subspace-clustering: find clusters in all the possible subspaces
        - CLIQUE, ProClus, and frequent pattern-based clustering

Jain – comparing various algorithms



Algorithm	Property	Comments
K-means	Identifies hyperspherical clusters; could be modified to find hyper-ellipsoidal clusters using Mahalanobis distance; computationally efficient.	Need to specify K and the initial cluster centers. Additional parameters for creating new clusters, merging existing clusters and outlier detection can be provided.
Fuzzy K-means	Similar to K-means except that every pattern has a degree of membership into the K clusters (fuzzy partition).	Need to specify K, initial cluster centers and cluster membership function.
Minimum Spanning Tree (MST)	Clusters are formed by deleting inconsistent edges in the MST of the data.	Need to provide the definition of an inconsistent edge.
Mutual Neighborhood	Compute the mutual neighborhood value (MNV) for every pair of patterns. If x_j is the p^{th} near neighbor of x_i and x_i is the q^{th} near neighbor of x_j, then $MNV(x_i, x_j) = p + q$; $p, q = 1, \cdots, K$.	Need to specify the neighborhood depth, K.
Single-Link (SL)	A hierarchical clustering algorithm which accepts a $n \times n$ proximity matrix; output is a dendrogram or a tree structure; a single-link cluster is a maximally connected subgraph on the patterns.	Single-link clusters easily chain together and are often "straggly"; need a heuristic to cut the tree to form clusters (a partition).
Complete-Link (CL)	A hierarchical clustering algorithm which accepts a $n \times n$ proximity matrix; output is a dendrogram or a tree structure; a complete-link cluster is a maximally complete subgraph on the patterns.	Complete-link clusters tend to be small and compact which combine nicely into layer clusters even when such a hierarchy is not warranted; need a heuristic to form clusters (a partition).

Clustering Evaluation

- Manual inspection
- Benchmarking on existing labels
  - Comparing clusters with ground-truth categories
- Cluster quality measures
  - distance measures
  - high similarity within a cluster, low across clusters

Evaluating variability of clusters

- Homogenuous clusters!
- Intuition → „zmienność wewnątrzskupieniowa” intra-class variability *wc*($\mathcal{C}$) i „zmienność międzyskupieniowa” inter-class distances *bc*($\mathcal{C}$)
  - May be defined in many ways
  - Take average of clusters $\mathbf{r}_k$ (centroids)
  - Then

$$wc(C) = \sum_{k=1}^{K} \sum\nolimits_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{r}_k)^2$$

$$\mathbf{r}_k = \frac{1}{n_k} \sum\nolimits_{\mathbf{x} \in C_k} \mathbf{x}$$

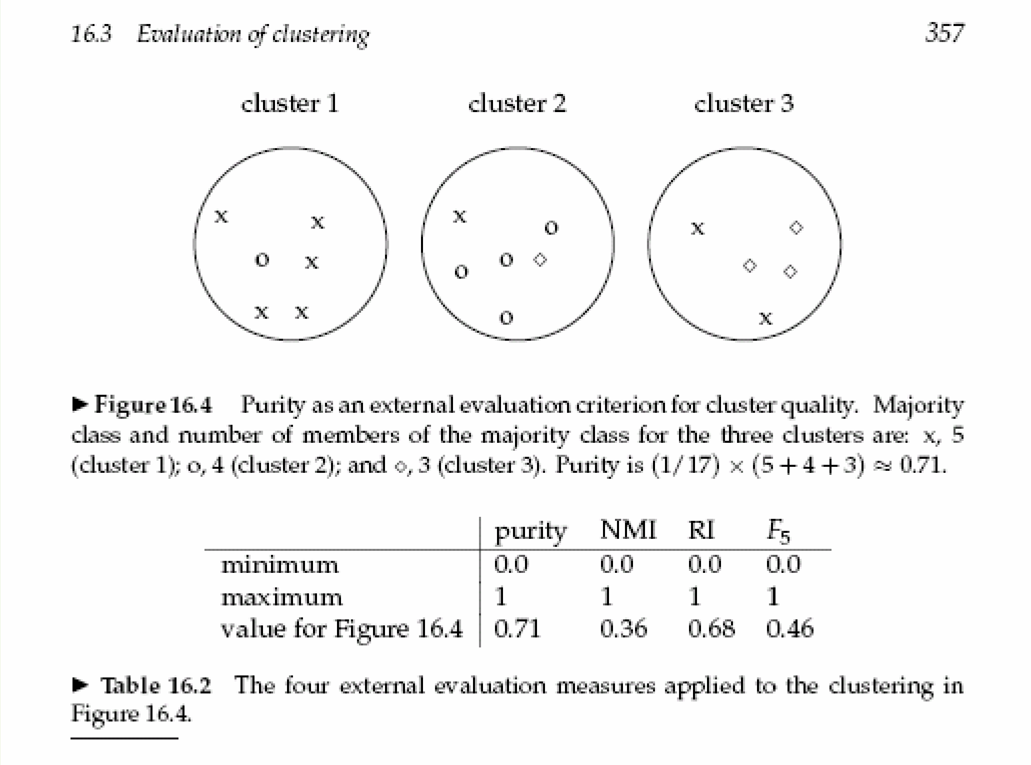
$$bc(C) = \sum\nolimits_{1 \le j < k \le K} d(\mathbf{r}_j, \mathbf{r}_k)^2$$

Measure of Clustering Accuracy

- Accuracy
  - Measured by manually labeled data
    - We manually assign tuples into clusters according to their properties (e.g., professors in different research areas)
  - Accuracy of clustering: Percentage of pairs of tuples in the same cluster that share common label
    - This measure favors many small clusters
    - We let each approach generate the same number of clusters

Testing class assignment (ground truth)

- Jain’s example



cluster 1
cluster 2
cluster 3

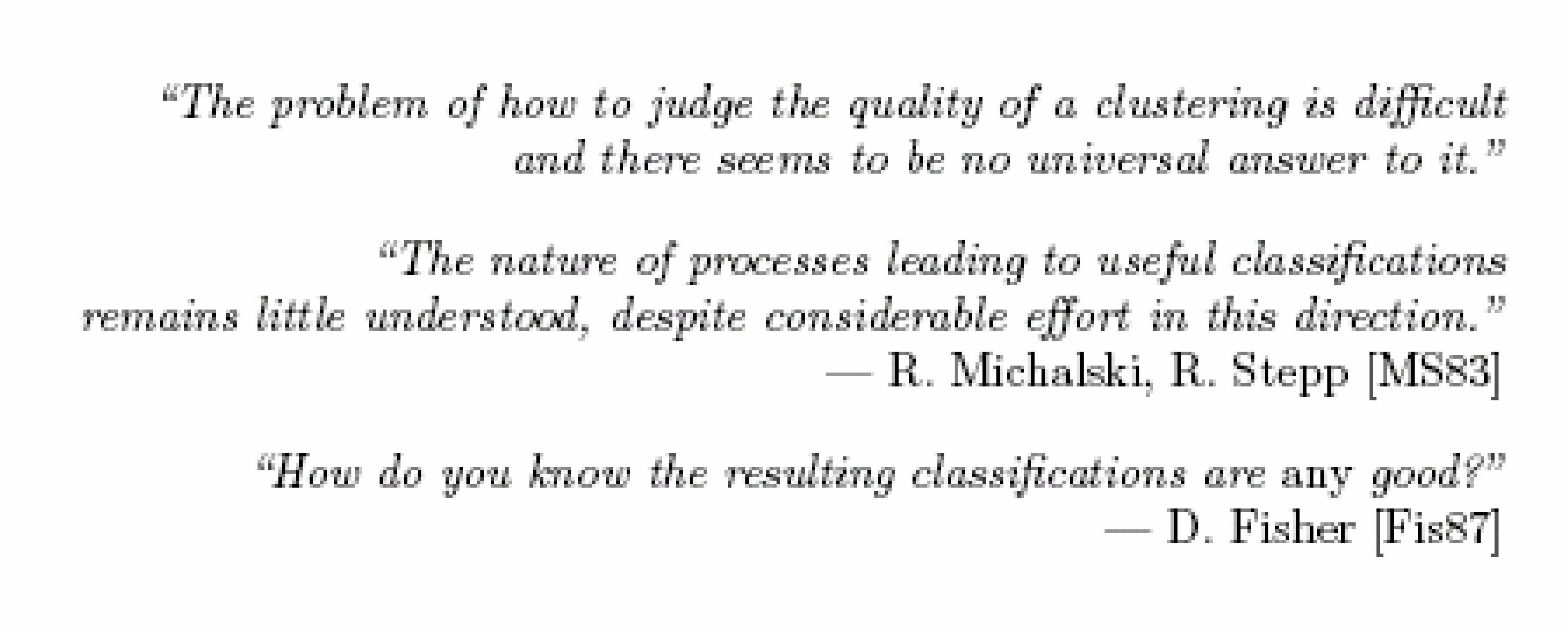
► **Figure 16.4** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

	purity	NMI	RI	F_5
minimum	0.0	0.0	0.0	0.0
maximum	1	1	1	1
value for Figure 16.4	0.71	0.36	0.68	0.46

► **Table 16.2** The four external evaluation measures applied to the clustering in Figure 16.4.

Could we analyse an objective single measure?

- Some opinions



"The problem of how to judge the quality of a clustering is difficult and there seems to be no universal answer to it."

"The nature of processes leading to useful classifications remains little understood, despite considerable effort in this direction."
— R. Michalski, R. Stepp [MS83]

"How do you know the resulting classifications are any *good?"*
— D. Fisher [Fis87]

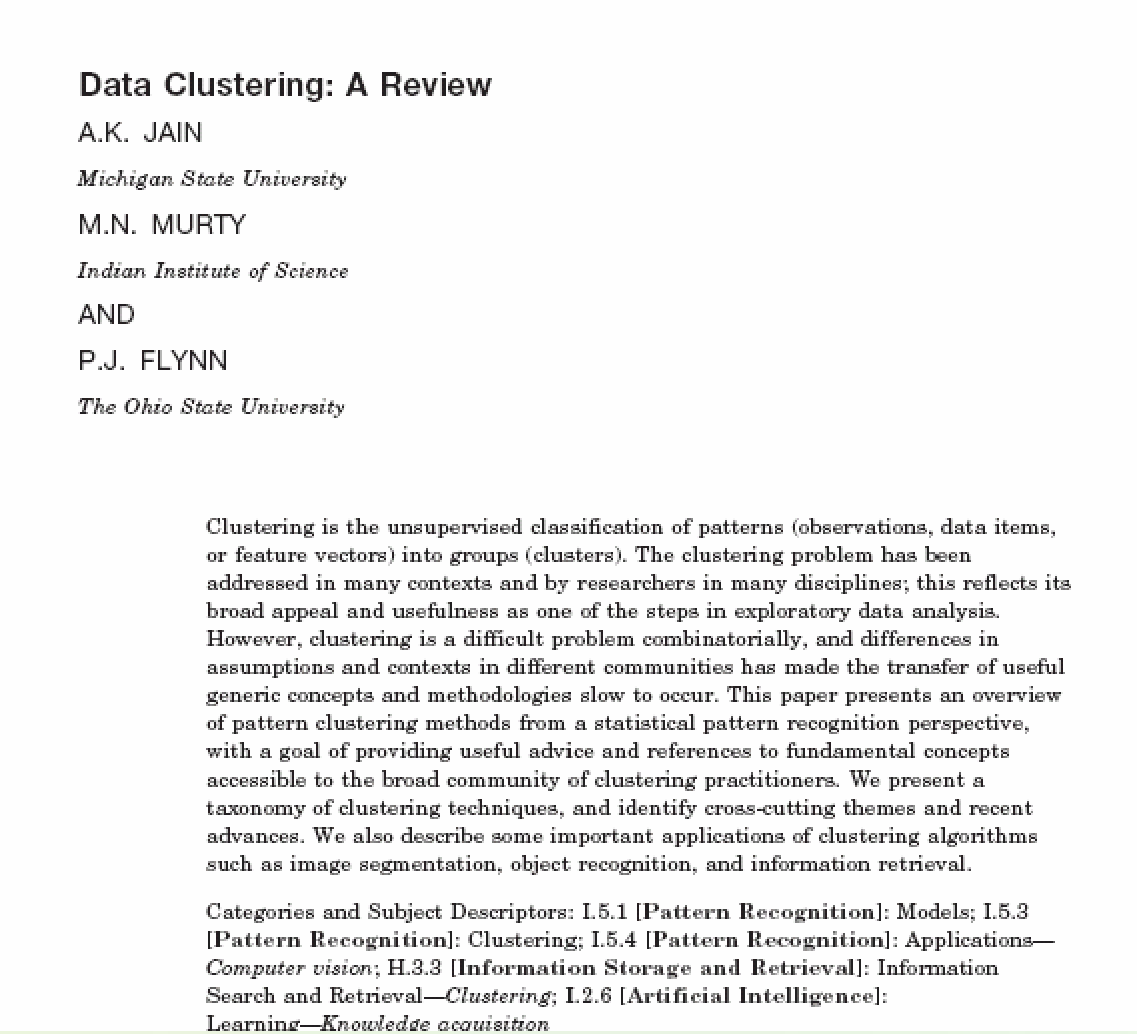
Requirements of Clustering in Data Mining

- Scalability
- Dealing with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Interpretability and usability.

Data-Intensive Clustering Methods: Recent Works

- CLARANS (Ng & Han’94): An extension to k-medoid algorithm based on randomized search + Focusing techniques with spatial access methods (EKX95).
- BIRCH (Zhang et al’96): CF tree + hierarchical clustering
- DBSCAN (EKXS96): density + connected dense regions
- STING (WYM97): A grid-based hierarchical cell structure that store statistical information
- CLIQUE (Agrawal et al’98): Cluster high dimensional data
- CURE (Guha, et al’98): hierarchical + partitioning
- WaveCluster (SCZ‘98): grid-clustering + wavelet analysis
- OPTICS (Ankerst et al’99): ordering points to identify clustering structures

Clustering in Data Mining – read more



Data Clustering: A Review

A.K. JAIN

Michigan State University

M.N. MURTY

Indian Institute of Science

AND

P.J. FLYNN

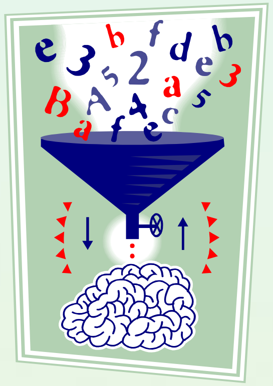
The Ohio State University

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur. This paper presents an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. We present a taxonomy of clustering techniques, and identify cross-cutting themes and recent advances. We also describe some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

Categories and Subject Descriptors: I.5.1 [**Pattern Recognition**]: Models; I.5.3 [**Pattern Recognition**]: Clustering; I.5.4 [**Pattern Recognition**]: Applications—*Computer vision*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*

Clustering Summary

- unsupervised
- many approaches
  - K-means – simple, sometimes useful
    - K-medoids is less sensitive to outliers
  - Hierarchical clustering – works for symbolic attributes
- Evaluation is a problem



Polish bibliography

- Koronacki J. Statystyczne systemy uczące się, WNT 2005.
- Pociecha J., Podolec B., Sokołowski A., Zając K. „Metody taksonomiczne w badaniach społeczno-ekonomicznych”. PWN, Warszawa 1988,
- Stąpor K. „Automatyczna klasyfikacja obiektów” Akademicka Oficyna Wydawnicza EXIT, Warszawa 2005.
- Hand, Mannila, Smyth, „Eksploracja danych”, WNT 2005.
- Larose D: „Odkrywania wiedzy z danych”, PWN 2006.
- Kucharczyk J. „Algorytmy analizy skupień w języku ALGOL 60” PWN Warszawa, 1982,
- Materiały szkoleniowe firmy Statsoft.

References

- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96
- B.S. Everitt, S. Landau, M. Leese. Cluster Analysis. Oxford University Press,fourth edition, 2001.
- D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.
- Allan D. Gordon. Classification. Chapman & Hall, London, second edition, 1999.
- J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, March 2006.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.
- A. K. Jain, M. Narasimha Murty, and P.J. Flynn. Data Clustering: A Review. ACM Computing Surveys, 31(3):264–323, 1999.
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : an efficient data clustering method for very large databases. SIGMOD'96

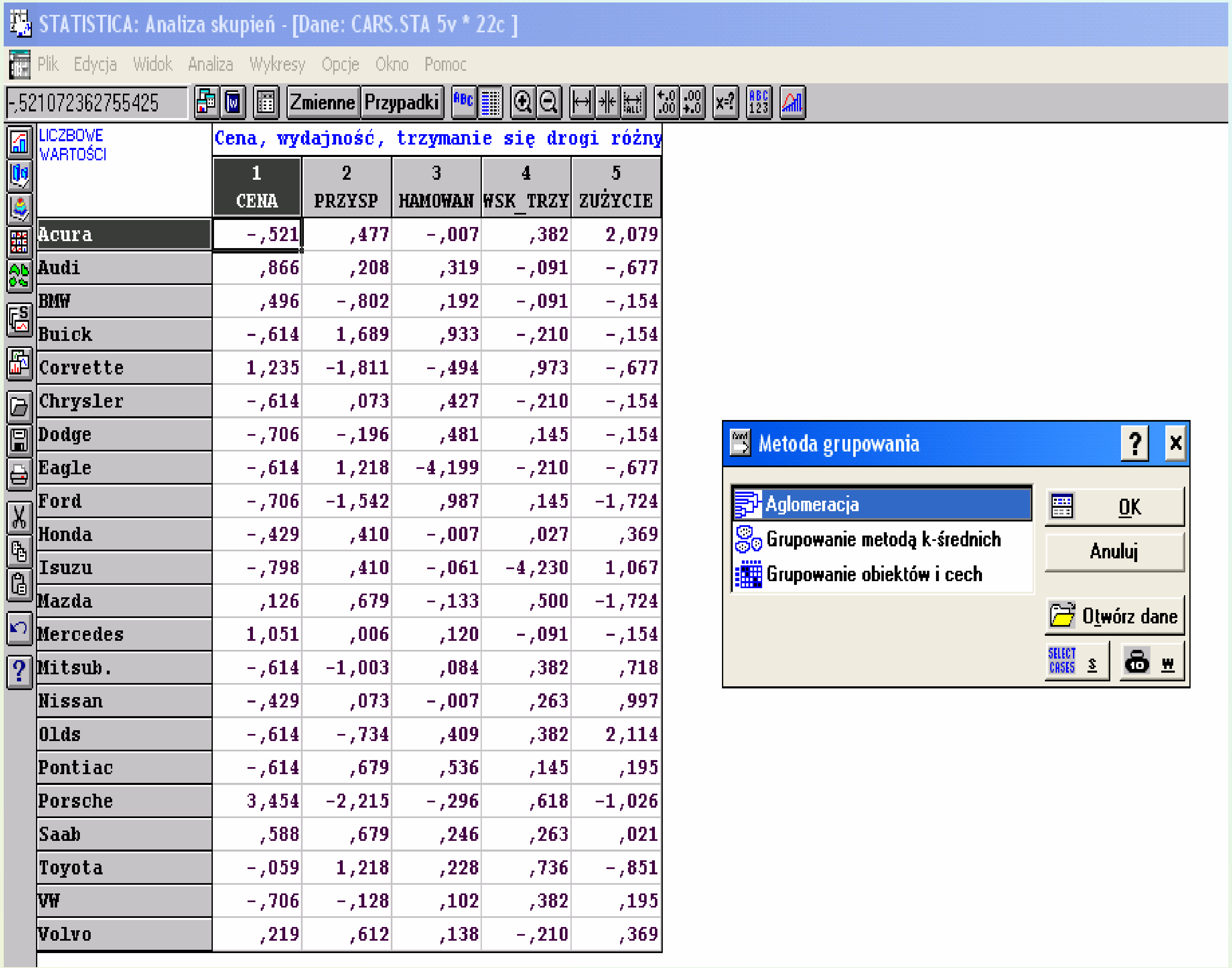
Any questions, remarks?

?

Exercise:

Clustering
in Statsoft -Statistica

Analiza Skupień – Statistica; więcej na www.statsoft.com. Przykład analizy danych o parametrach samochodów



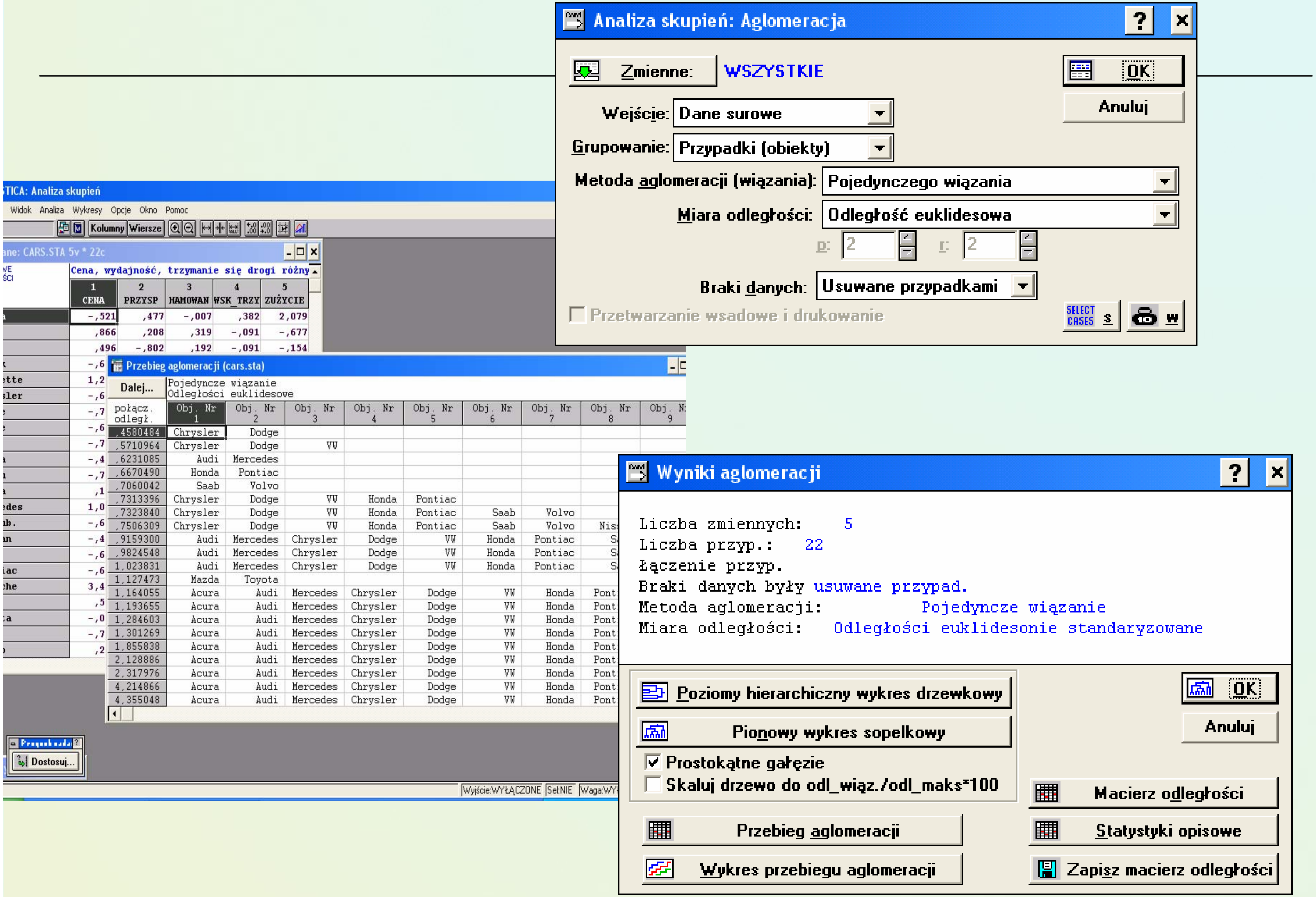
STATISTICA: Analiza skupień - [Dane: CARS.STA 5v * 22c]
Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc
-,521072362755425
Zmienne
Przypadki
LICZBOWE WARTOŚCI

Cena, wydajność, trzymanie się drogi różny

	1 CENA	2 PRZYSP	3 HAMOWAN	4 WSK_TRZY	5 ZUŻYCIE
Acura	-,521	,477	-,007	,382	2,079
Audi	,866	,208	,319	-,091	-,677
BMW	,496	-,802	,192	-,091	-,154
Buick	-,614	1,689	,933	-,210	-,154
Corvette	1,235	-1,811	-,494	,973	-,677
Chrysler	-,614	,073	,427	-,210	-,154
Dodge	-,706	-,196	,481	,145	-,154
Eagle	-,614	1,218	-4,199	-,210	-,677
Ford	-,706	-1,542	,987	,145	-1,724
Honda	-,429	,410	-,007	,027	,369
Isuzu	-,798	,410	-,061	-4,230	1,067
Mazda	,126	,679	-,133	,500	-1,724
Mercedes	1,051	,006	,120	-,091	-,154
Mitsub.	-,614	-1,003	,084	,382	,718
Nissan	-,429	,073	-,007	,263	,997
Olds	-,614	-,734	,409	,382	2,114
Pontiac	-,614	,679	,536	,145	,195
Porsche	3,454	-2,215	-,296	,618	-1,026
Saab	,588	,679	,246	,263	,021
Toyota	-,059	1,218	,228	,736	-,851
VW	-,706	-,128	,102	,382	,195
Volvo	,219	,612	,138	-,210	,369

Metoda grupowania

- Aglomeracja
- Grupowanie metodą k-średnich
- Grupowanie obiektów i cech



Cena, wydajność, trzymanie się drogi różny

1 CENA	2 PRZYSP	3 HAMOWAN	4 WSK_TRZY	5 ZUŻYCIE
-,521	,477	-,007	,382	2,079
,866	,208	,319	-,091	-,677
,496	-,802	,192	-,091	-,154

Analiza skupień: Aglomeracja

Zmienne: WSZYSTKIE

Wejście: Dane surowe

Grupowanie: Przypadki (obiekty)

Metoda aglomeracji (wiązania): Pojedynczego wiązania

Miara odległości: Odległość euklidesowa

p: 2 r: 2

Braki danych: Usuwane przypadkami

□ Przetwarzanie wsadowe i drukowanie

Przebieg aglomeracji (cars.sta)

Dalej...

Pojedyncze wiązanie
Odległości euklidesowe

połącz. odległ.	Obj. Nr 1	Obj. Nr 2	Obj. Nr 3	Obj. Nr 4	Obj. Nr 5	Obj. Nr 6	Obj. Nr 7	Obj. Nr 8	Obj. Nr 9
,4580484	Chrysler	Dodge							
,5710964	Chrysler	Dodge	VW						
,6231085	Audi	Mercedes							
,6670490	Honda	Pontiac							
,7060042	Saab	Volvo							
,7313396	Chrysler	Dodge	VW	Honda	Pontiac				
,7323840	Chrysler	Dodge	VW	Honda	Pontiac	Saab	Volvo		
,7506309	Chrysler	Dodge	VW	Honda	Pontiac	Saab	Volvo	Nis	
,9159300	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pontiac	S	
,9824548	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pontiac	S	
1,023831	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pontiac	S	
1,127473	Mazda	Toyota							
1,164055	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
1,193655	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
1,284603	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
1,301269	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
1,855838	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
2,128886	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
2,317976	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
4,214866	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	
4,355048	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont	

Wyniki aglomeracji

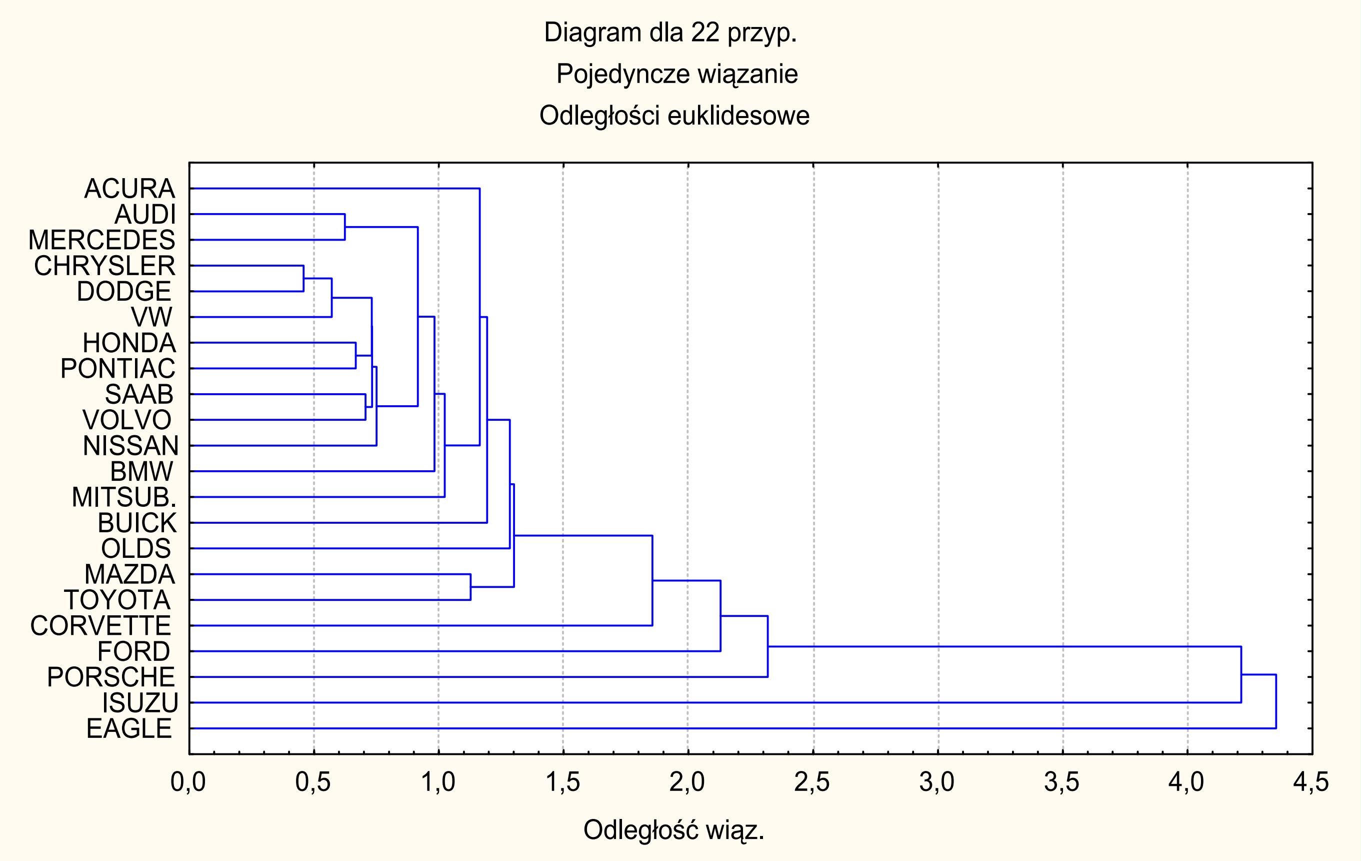
```
Liczba zmiennych:    5
Liczba przyp.:    22
Łączenie przyp.
Braki danych były usuwane przypad.
Metoda aglomeracji:         Pojedyncze wiązanie
Miara odległości:    Odległości euklidesowe standaryzowane
```

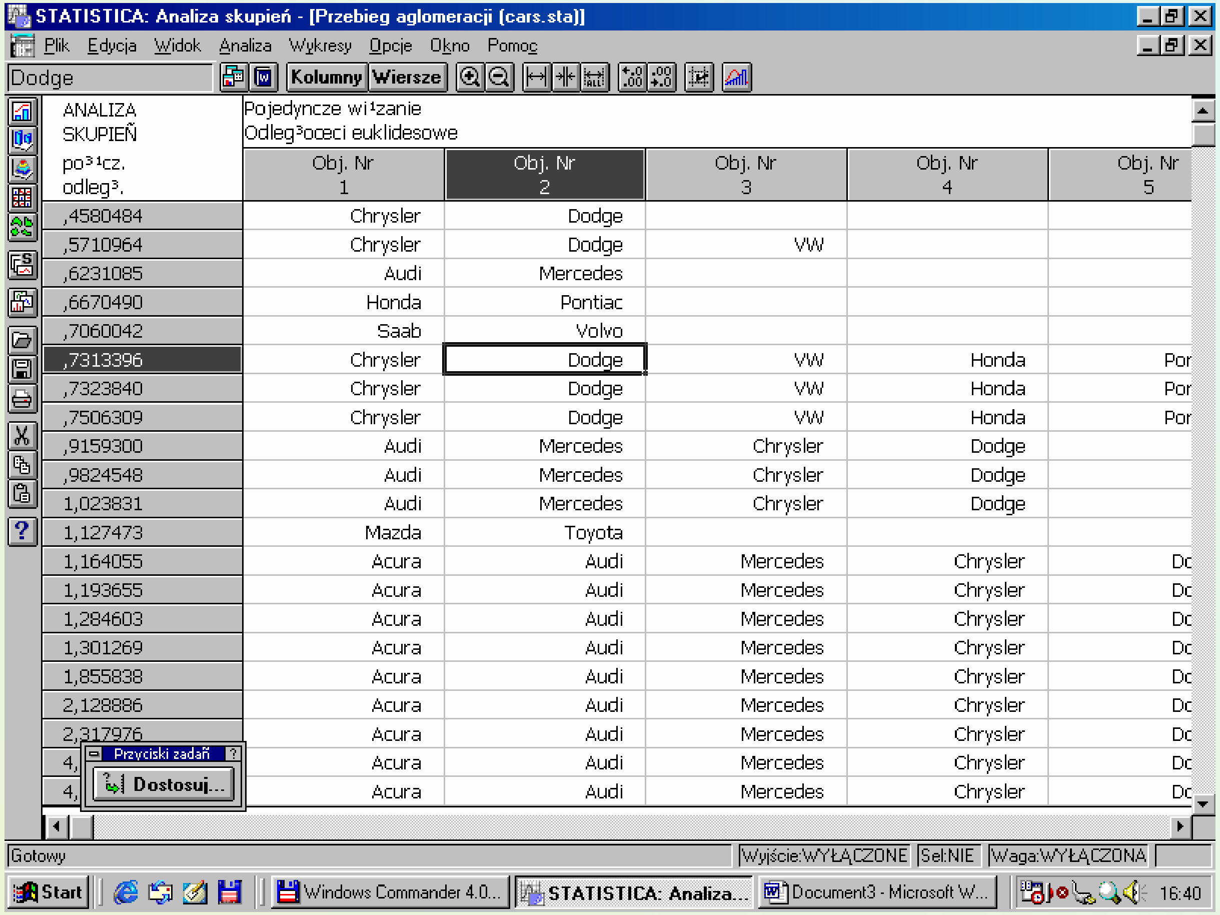
Poziomy hierarchiczny wykres drzewkowy
Pionowy wykres sopelkowy
OK
Anuluj
Macierz odległości
Przebieg aglomeracji
Statystyki opisowe
Wykres przebiegu aglomeracji
Zapisz macierz odległości

☑ Prostokątne gałęzie

□ Skaluj drzewo do odl_wiąz./odl_maks*100

Dendrogram for Single Linkage



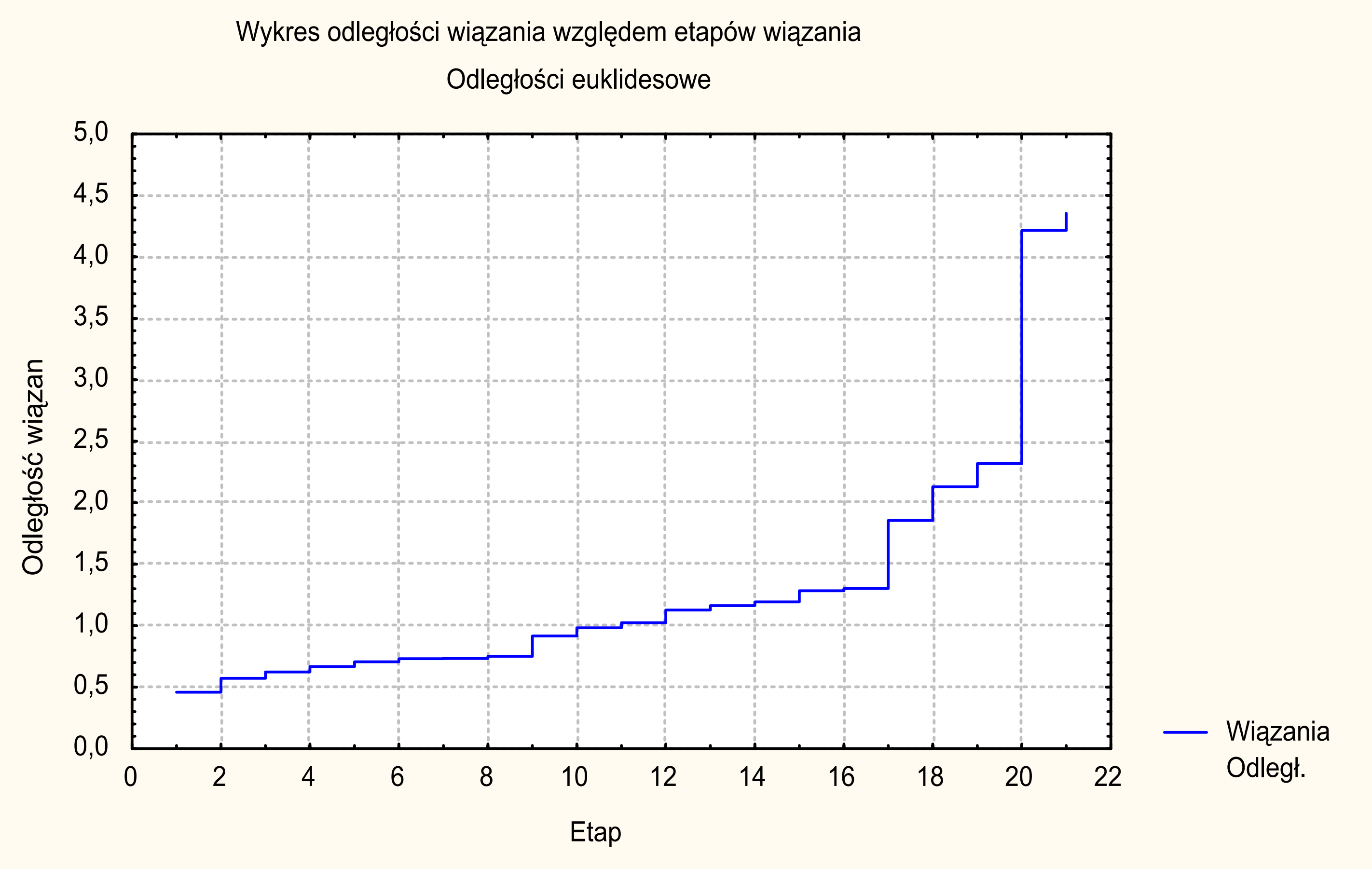


Pojedyncze wi¹zanie
Odleg³oœci euklidesowe

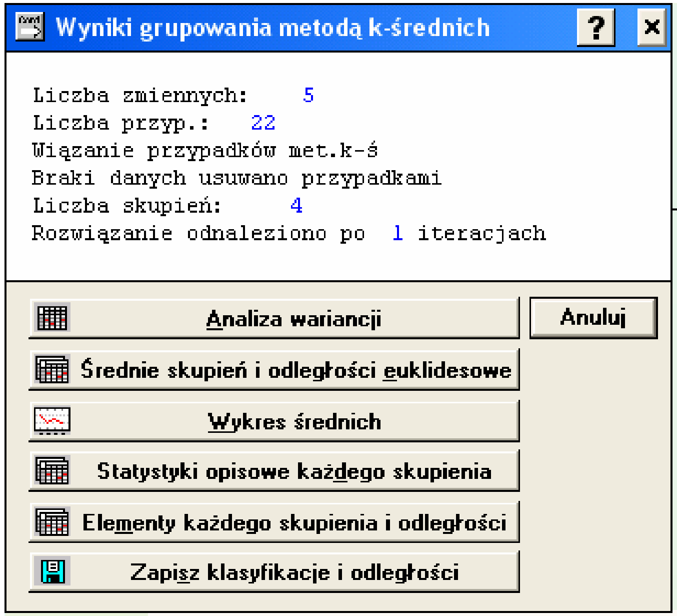
ANALIZA SKUPIEÑ po³¹cz. odleg³.	Obj. Nr 1	Obj. Nr 2	Obj. Nr 3	Obj. Nr 4	Obj. Nr 5
,4580484	Chrysler	Dodge			
,5710964	Chrysler	Dodge	VW		
,6231085	Audi	Mercedes			
,6670490	Honda	Pontiac			
,7060042	Saab	Volvo			
,7313396	Chrysler	Dodge	VW	Honda	Por
,7323840	Chrysler	Dodge	VW	Honda	Por
,7506309	Chrysler	Dodge	VW	Honda	Por
,9159300	Audi	Mercedes	Chrysler	Dodge	
,9824548	Audi	Mercedes	Chrysler	Dodge	
1,023831	Audi	Mercedes	Chrysler	Dodge	
1,127473	Mazda	Toyota			
1,164055	Acura	Audi	Mercedes	Chrysler	Dc
1,193655	Acura	Audi	Mercedes	Chrysler	Dc
1,284603	Acura	Audi	Mercedes	Chrysler	Dc
1,301269	Acura	Audi	Mercedes	Chrysler	Dc
1,855838	Acura	Audi	Mercedes	Chrysler	Dc
2,128886	Acura	Audi	Mercedes	Chrysler	Dc
2,317976	Acura	Audi	Mercedes	Chrysler	Dc
4,	Acura	Audi	Mercedes	Chrysler	Dc
4,	Acura	Audi	Mercedes	Chrysler	Dc

Analysing Agglomeration Steps

- Figure – find a cut point („kolanko” / knee)



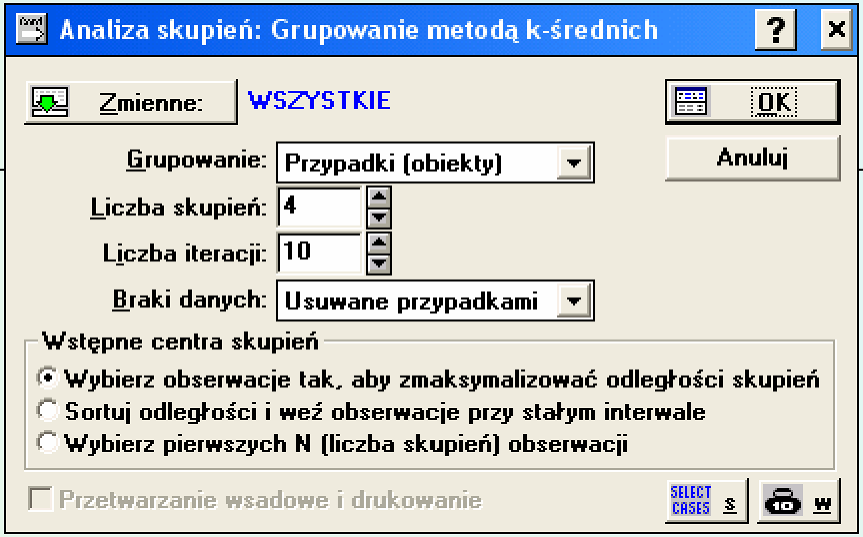
Analiza Skupień – optymalizacja k-średnich



Wyniki grupowania metodą k-średnich

```
Liczba zmiennych:    5
Liczba przyp.:   22
Wiązanie przypadków met.k-ś
Braki danych usuwano przypadkami
Liczba skupień:    4
Rozwiązanie odnaleziono po  1 iteracjach
```

Analiza wariancji
Anuluj
Średnie skupień i odległości euklidesowe
Wykres średnich
Statystyki opisowe każdego skupienia
Elementy każdego skupienia i odległości
Zapisz klasyfikacje i odległości



Analiza skupień: Grupowanie metodą k-średnich

Zmienne: WSZYSTKIE

Grupowanie: Przypadki (obiekty)

Liczba skupień: 4

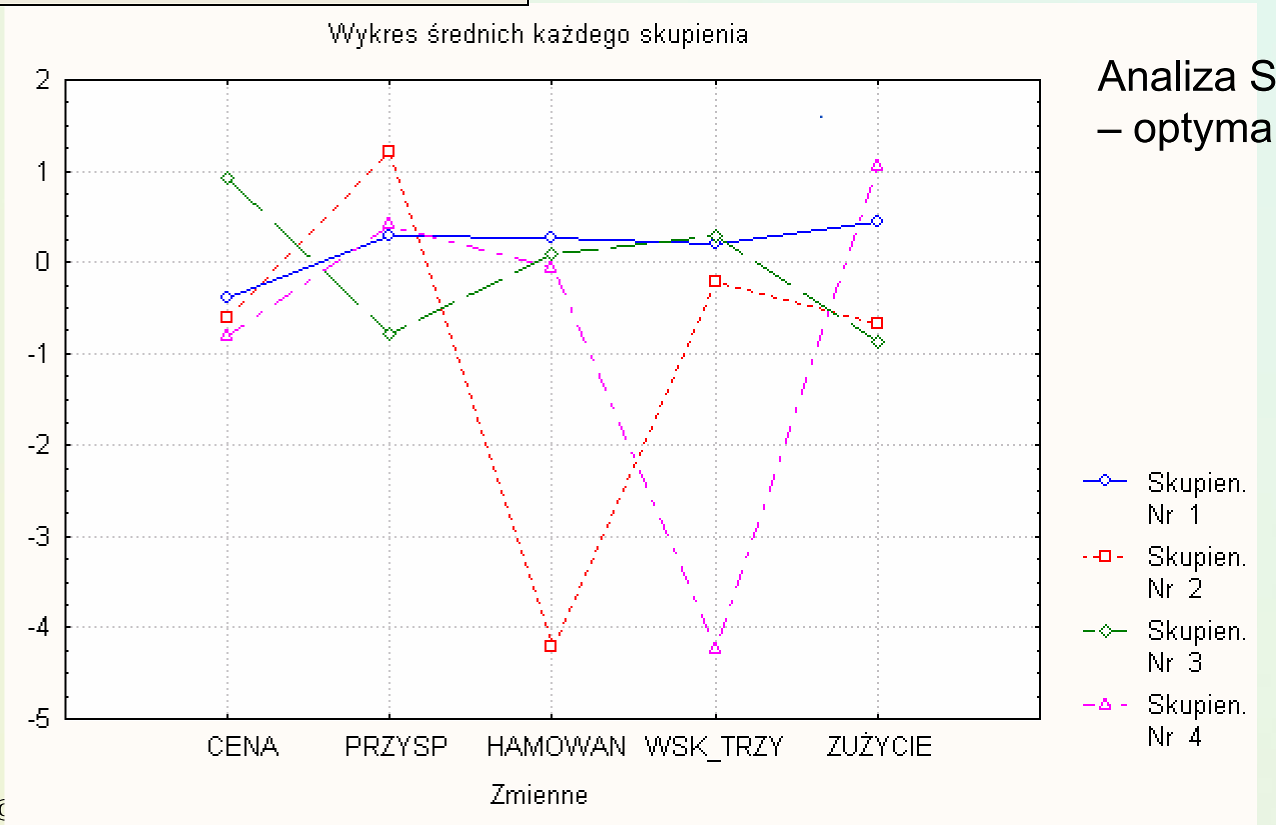
Liczba iteracji: 10

Braki danych: Usuwane przypadkami

Wstępne centra skupień

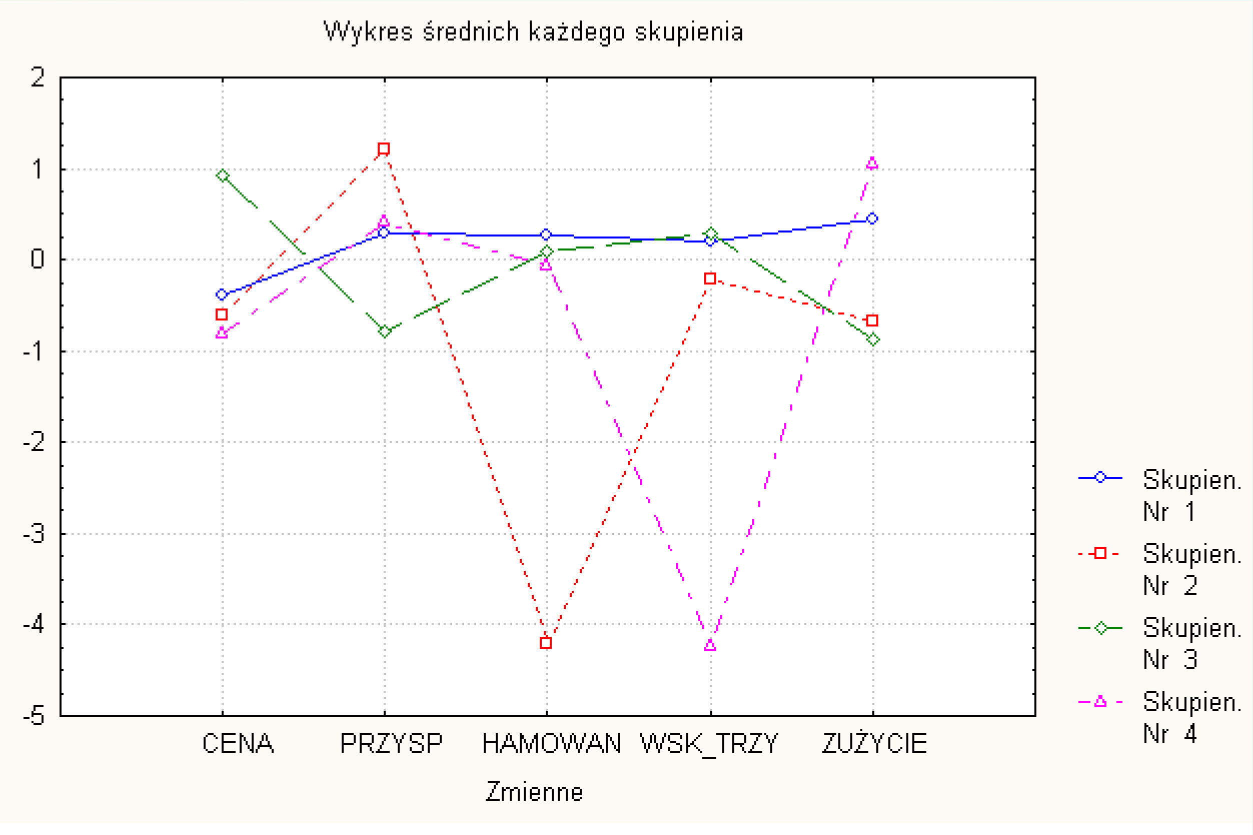
- ● Wybierz obserwacje tak, aby zmaksymalizować odległości skupień
- ○ Sortuj odległości i weź obserwacje przy stałym interwale
- ○ Wybierz pierwszych N (liczba skupień) obserwacji

□ Przetwarzanie wsadowe i drukowanie



Wykres średnich każdego skupienia

Profile - visulization



Exercise:

Clustering in WEKA

WEKA Clustering

- Implemented methods
  - *k*-Means
  - EM
  - Cobweb
  - X-means
  - FarthestFirst…
- Clusters can be visualized and compared to "true" clusters (if given)

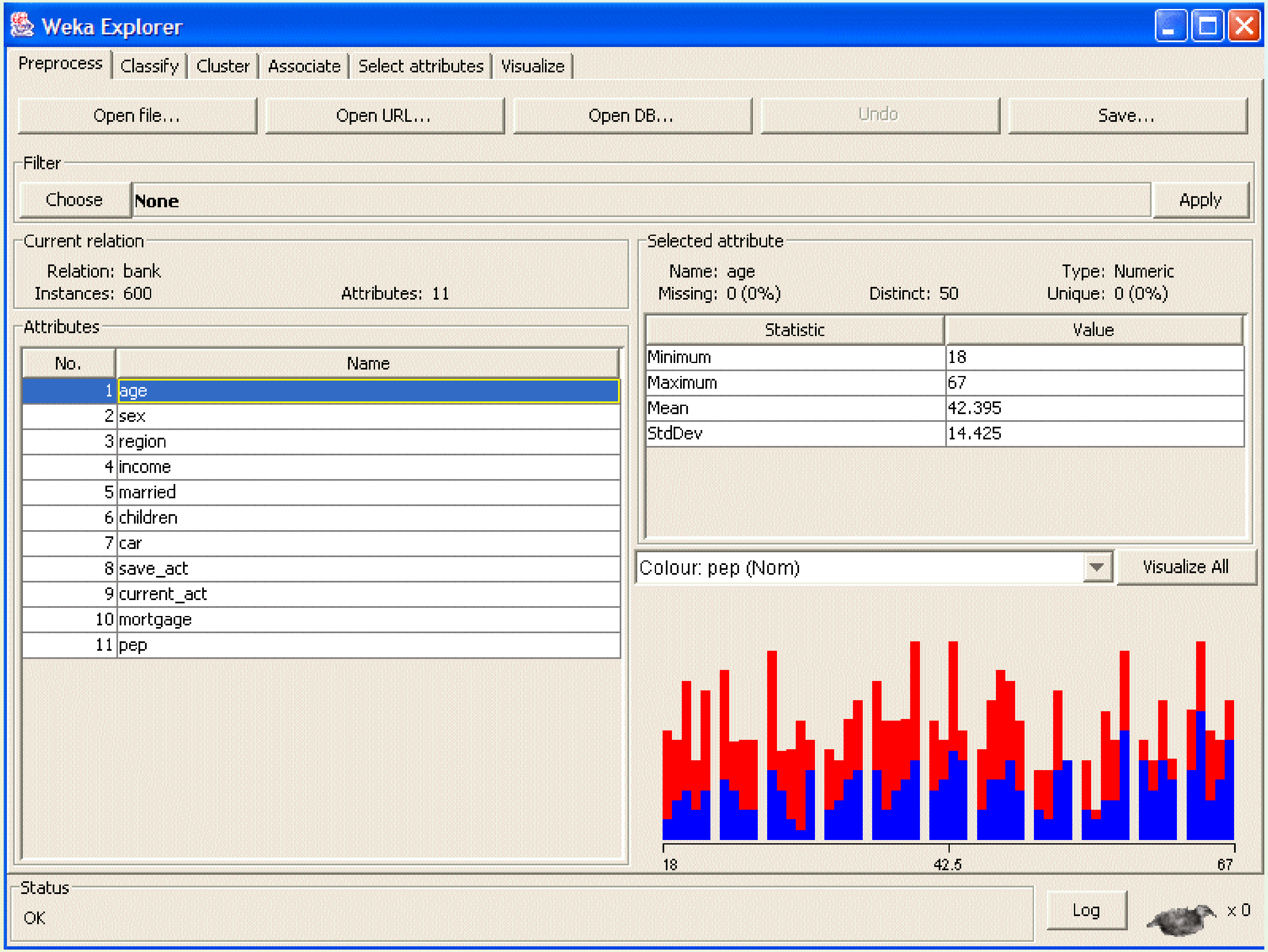
Exercise 1. K-means clustering in WEKA

- The exercise illustrates the use of the k-means algorithm.
- The example – sample of customers of the bank
  - Bank data (bank-data.cvs -> bank.arff)
  - All preprocessing has been performed on cvs
  - 600 instances described by 11 attributes

```
id,age,sex,region,income,married,children,car,save_act,current_act,mortgage,pep
ID12101,48,FEMALE,INNER_CITY,17546.0,NO,1,NO,NO,NO,NO,YES
ID12102,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
ID12103,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
ID12104,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
ID12105,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
……………………………………………………………………………..
……………………………………………………………
```

  - Cluster customers and characterize the resulting customer segments

Loading the file and analysing the data



Filter: None

Current relation

Relation: bank
Instances: 600 Attributes: 11

Attributes

No.	Name
1	age
2	sex
3	region
4	income
5	married
6	children
7	car
8	save_act
9	current_act
10	mortgage
11	pep

Selected attribute

Name: age Type: Numeric
Missing: 0 (0%) Distinct: 50 Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	67
Mean	42.395
StdDev	14.425

Colour: pep (Nom)

Status

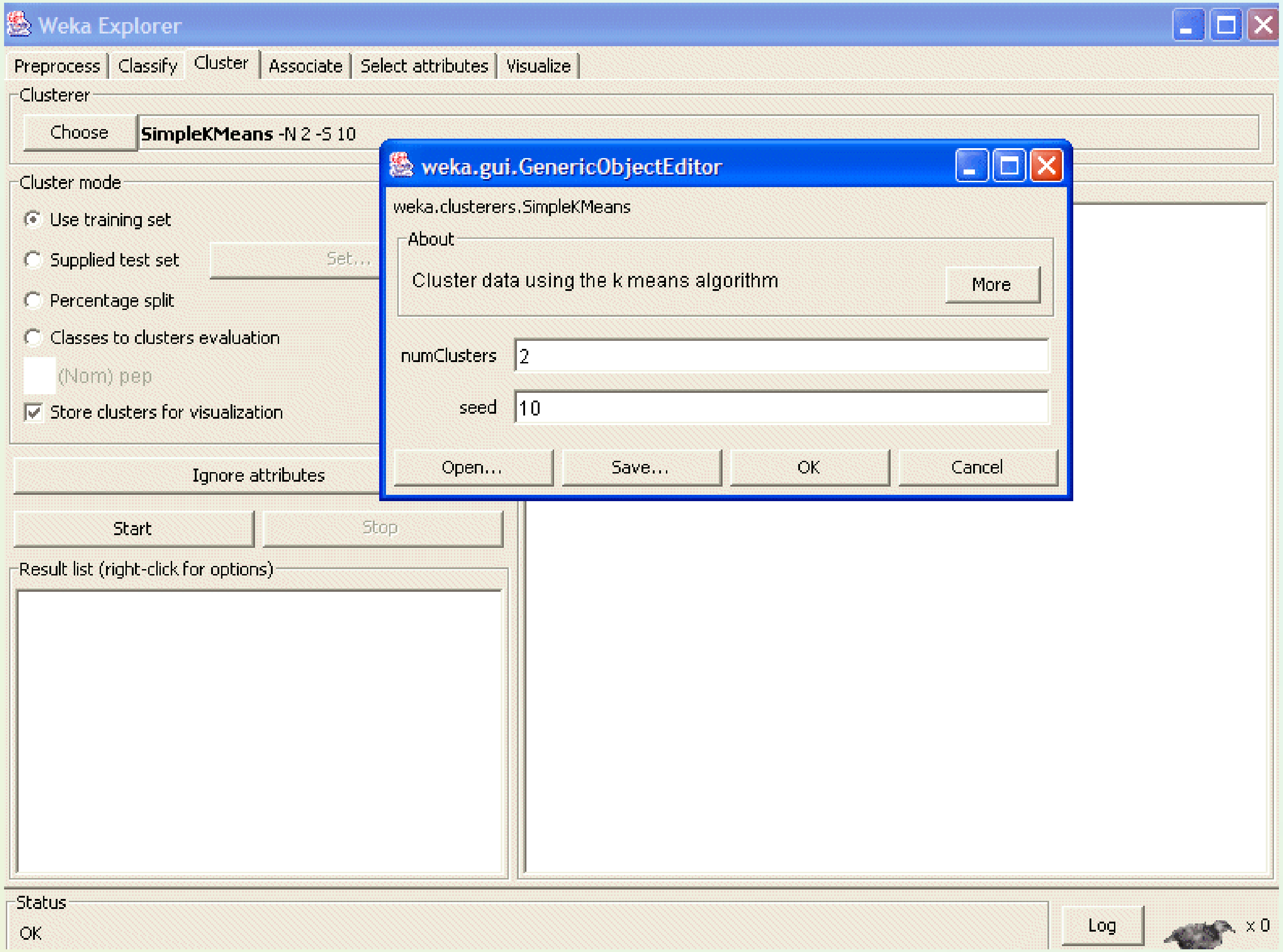
OK

Preprocessing for clustering

- What about non-numerical attributes?
  - Remember about Filters
- Should we normalize or standarize attributes?
- How it is handled in WEKA k-means?

Choosing Simple k-means

- Tune proper parameters



Weka Explorer

Clusterer: SimpleKMeans -N 2 -S 10

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split

○ Classes to clusters evaluation

(Nom) pep

☑ Store clusters for visualization

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

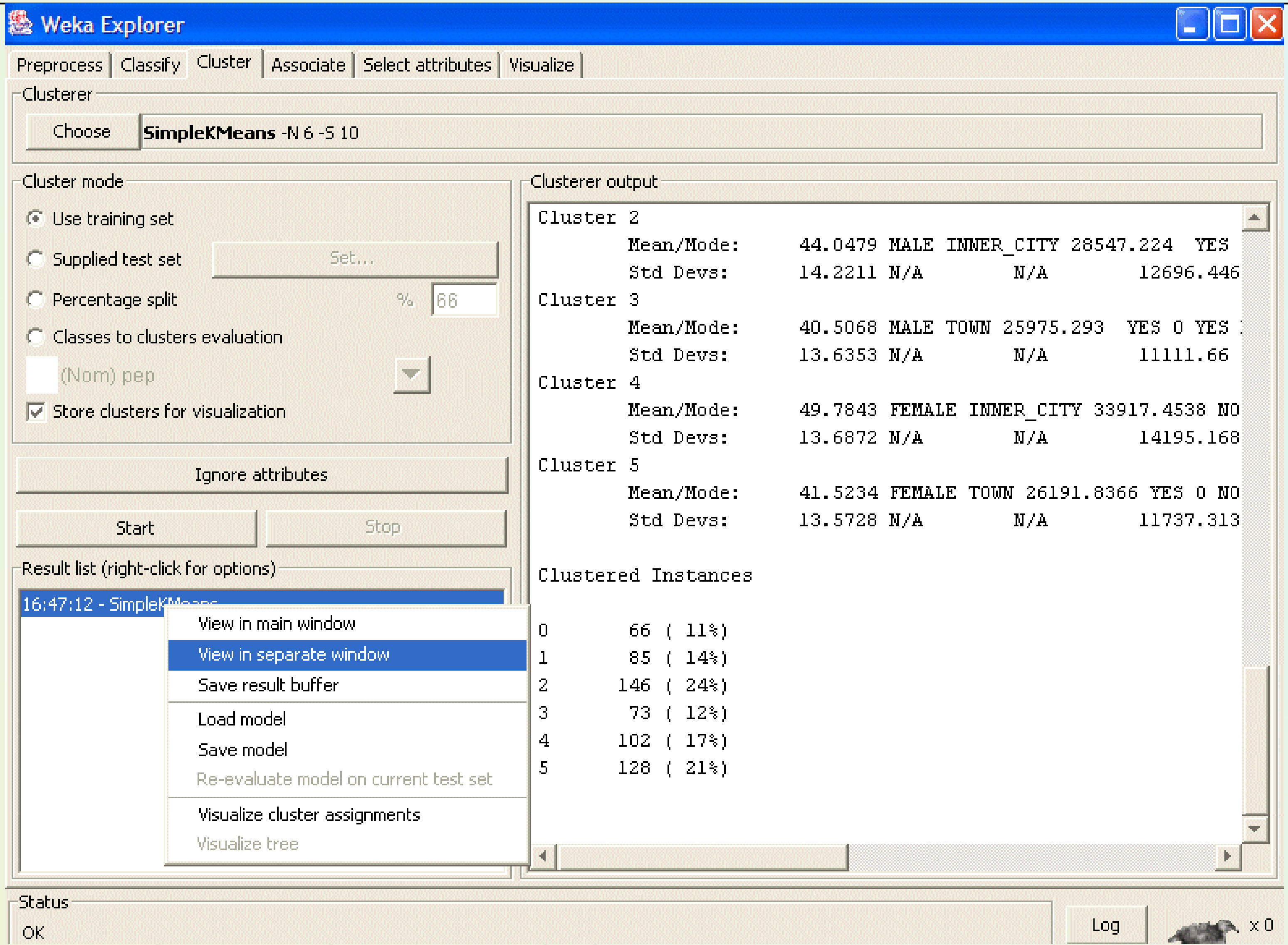
Cluster data using the k means algorithm

numClusters: 2

seed: 10

Status: OK

Clustering results



Clusterer: SimpleKMeans -N 6 -S 10

Cluster mode

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Nom) pep
- ☑ Store clusters for visualization

Result list (right-click for options)

16:47:12 - SimpleKMeans

Clusterer output

```
Cluster 2
        Mean/Mode:     44.0479 MALE INNER_CITY 28547.224  YES
        Std Devs:      14.2211 N/A        N/A        12696.446
Cluster 3
        Mean/Mode:     40.5068 MALE TOWN 25975.293  YES 0 YES
        Std Devs:      13.6353 N/A        N/A        11111.66
Cluster 4
        Mean/Mode:     49.7843 FEMALE INNER_CITY 33917.4538 NO
        Std Devs:      13.6872 N/A        N/A        14195.168
Cluster 5
        Mean/Mode:     41.5234 FEMALE TOWN 26191.8366 YES 0 NO
        Std Devs:      13.5728 N/A        N/A        11737.313

Clustered Instances

0        66 ( 11%)
1        85 ( 14%)
2       146 ( 24%)
3        73 ( 12%)
4       102 ( 17%)
5       128 ( 21%)
```

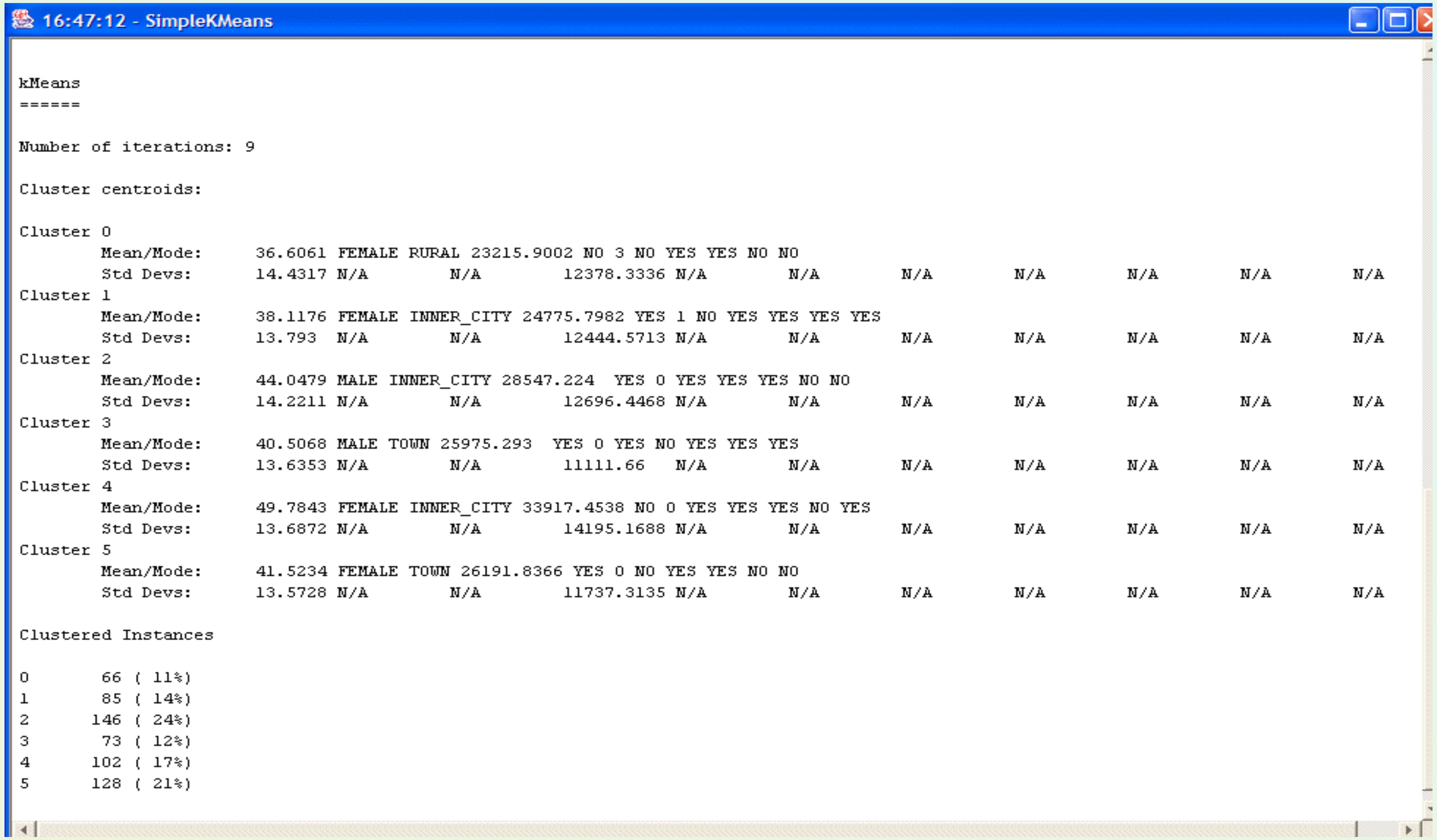
Status

OK

- Analyse the result window

Characterizing cluster

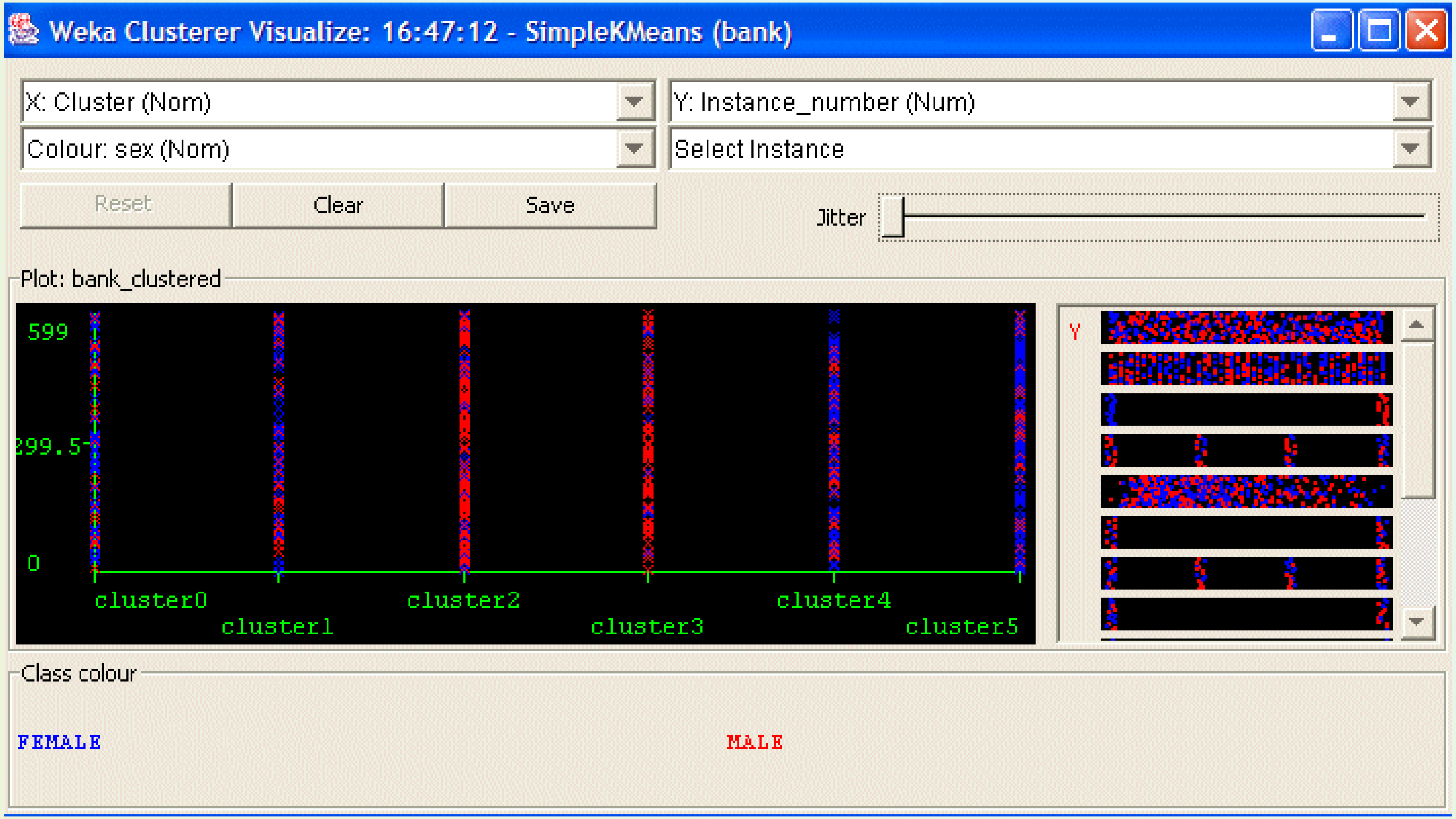
- How to describe clusters?
- What about descriptive statistics for centroids?



```
kMeans
======

Number of iterations: 9

Cluster centroids:

Cluster 0
        Mean/Mode:     36.6061 FEMALE RURAL 23215.9002 NO 3 NO YES YES NO NO
        Std Devs:      14.4317 N/A        N/A        12378.3336 N/A        N/A        N/A        N/A        N/A        N/A        N/A
Cluster 1
        Mean/Mode:     38.1176 FEMALE INNER_CITY 24775.7982 YES 1 NO YES YES YES YES
        Std Devs:      13.793  N/A        N/A        12444.5713 N/A        N/A        N/A        N/A        N/A        N/A        N/A
Cluster 2
        Mean/Mode:     44.0479 MALE INNER_CITY 28547.224  YES 0 YES YES YES NO NO
        Std Devs:      14.2211 N/A        N/A        12696.4468 N/A        N/A        N/A        N/A        N/A        N/A        N/A
Cluster 3
        Mean/Mode:     40.5068 MALE TOWN 25975.293  YES 0 YES NO YES YES YES
        Std Devs:      13.6353 N/A        N/A        11111.66   N/A        N/A        N/A        N/A        N/A        N/A        N/A
Cluster 4
        Mean/Mode:     49.7843 FEMALE INNER_CITY 33917.4538 NO 0 YES YES YES NO YES
        Std Devs:      13.6872 N/A        N/A        14195.1688 N/A        N/A        N/A        N/A        N/A        N/A        N/A
Cluster 5
        Mean/Mode:     41.5234 FEMALE TOWN 26191.8366 YES 0 NO YES YES NO NO
        Std Devs:      13.5728 N/A        N/A        11737.3135 N/A        N/A        N/A        N/A        N/A        N/A        N/A

Clustered Instances

0        66 ( 11%)
1        85 ( 14%)
2       146 ( 24%)
3        73 ( 12%)
4       102 ( 17%)
5       128 ( 21%)
```

© Stefanowski 2008

Understanding the cluster characterization through visualization



Weka Clusterer Visualize: 16:47:12 - SimpleKMeans (bank)

X: Cluster (Nom)

Y: Instance_number (Num)

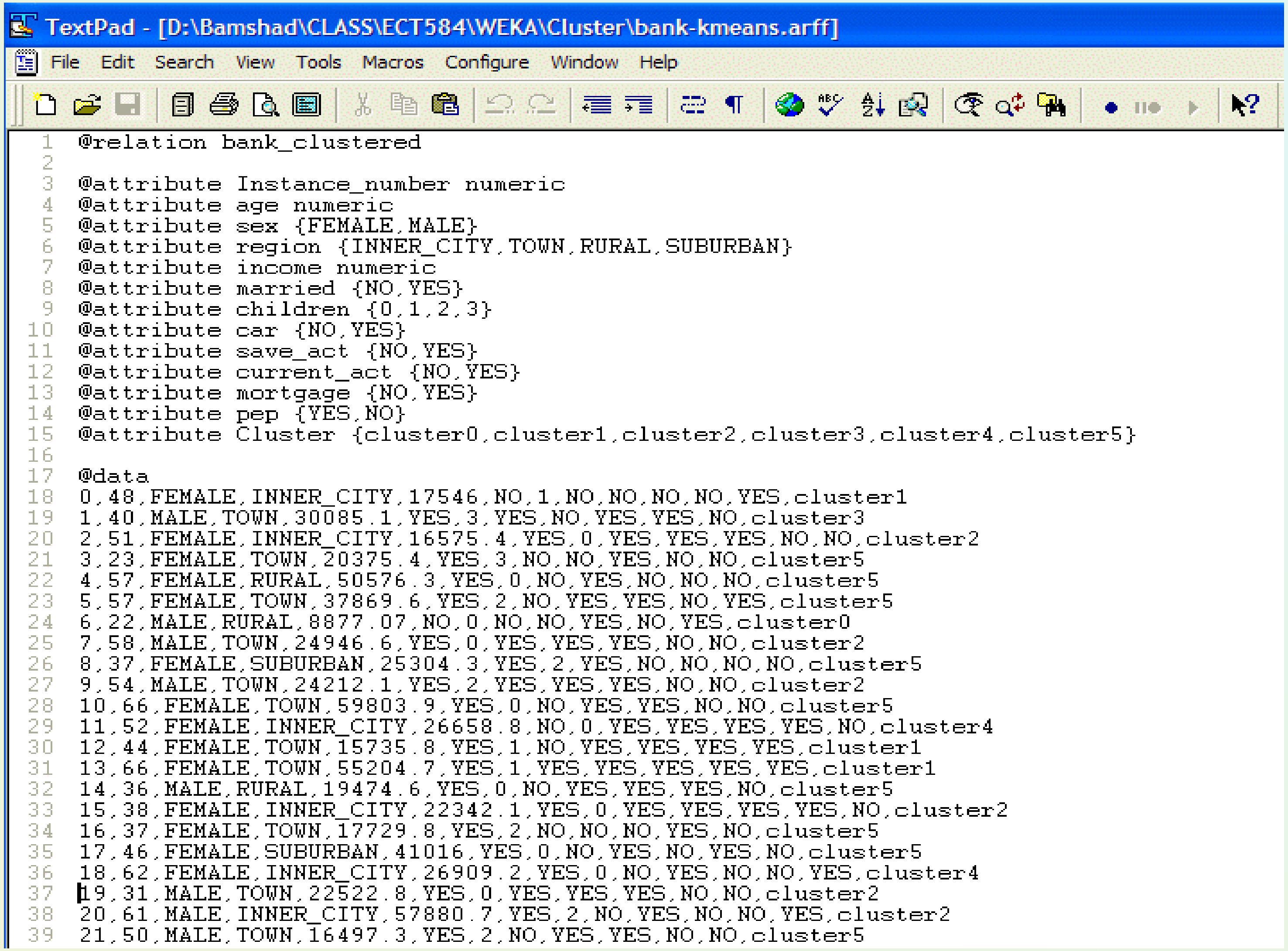
Colour: sex (Nom)

Select Instance

Class colour

FEMALE MALE

Finally, cluster assignments


TextPad - [D:\Bamshad\CLASS\ECT584\WEKA\Cluster\bank-kmeans.arff]
File Edit Search View Tools Macros Configure Window Help

```
@relation bank_clustered

@attribute Instance_number numeric
@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
@attribute Cluster {cluster0,cluster1,cluster2,cluster3,cluster4,cluster5}

@data
0,48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES,cluster1
1,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO,cluster3
2,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO,cluster2
3,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO,cluster5
4,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO,cluster5
5,57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES,cluster5
6,22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES,cluster0
7,58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO,cluster2
8,37,FEMALE,SUBURBAN,25304.3,YES,2,YES,NO,NO,NO,NO,cluster5
9,54,MALE,TOWN,24212.1,YES,2,YES,YES,YES,NO,NO,cluster2
10,66,FEMALE,TOWN,59803.9,YES,0,NO,YES,YES,NO,NO,cluster5
11,52,FEMALE,INNER_CITY,26658.8,NO,0,YES,YES,YES,YES,NO,cluster4
12,44,FEMALE,TOWN,15735.8,YES,1,NO,YES,YES,YES,YES,cluster1
13,66,FEMALE,TOWN,55204.7,YES,1,YES,YES,YES,YES,YES,cluster1
14,36,MALE,RURAL,19474.6,YES,0,NO,YES,YES,YES,NO,cluster5
15,38,FEMALE,INNER_CITY,22342.1,YES,0,YES,YES,YES,YES,NO,cluster2
16,37,FEMALE,TOWN,17729.8,YES,2,NO,NO,NO,YES,NO,cluster5
17,46,FEMALE,SUBURBAN,41016,YES,0,NO,YES,NO,YES,NO,cluster5
18,62,FEMALE,INNER_CITY,26909.2,YES,0,NO,YES,NO,NO,YES,cluster4
19,31,MALE,TOWN,22522.8,YES,0,YES,YES,YES,NO,NO,cluster2
20,61,MALE,INNER_CITY,57880.7,YES,2,NO,YES,NO,NO,YES,cluster2
21,50,MALE,TOWN,16497.3,YES,2,NO,YES,YES,NO,NO,cluster5
```