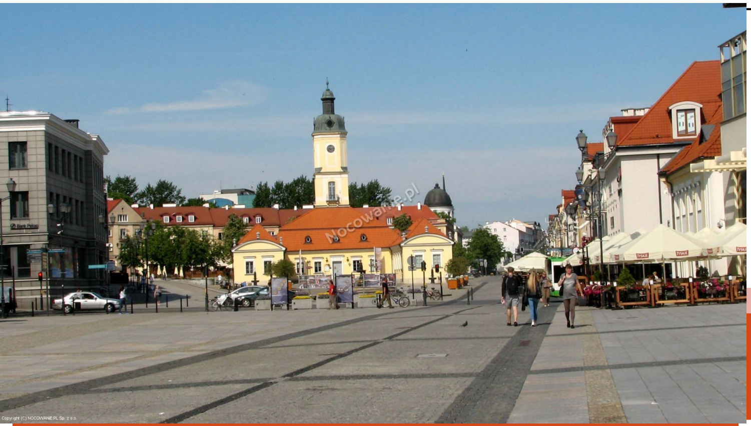
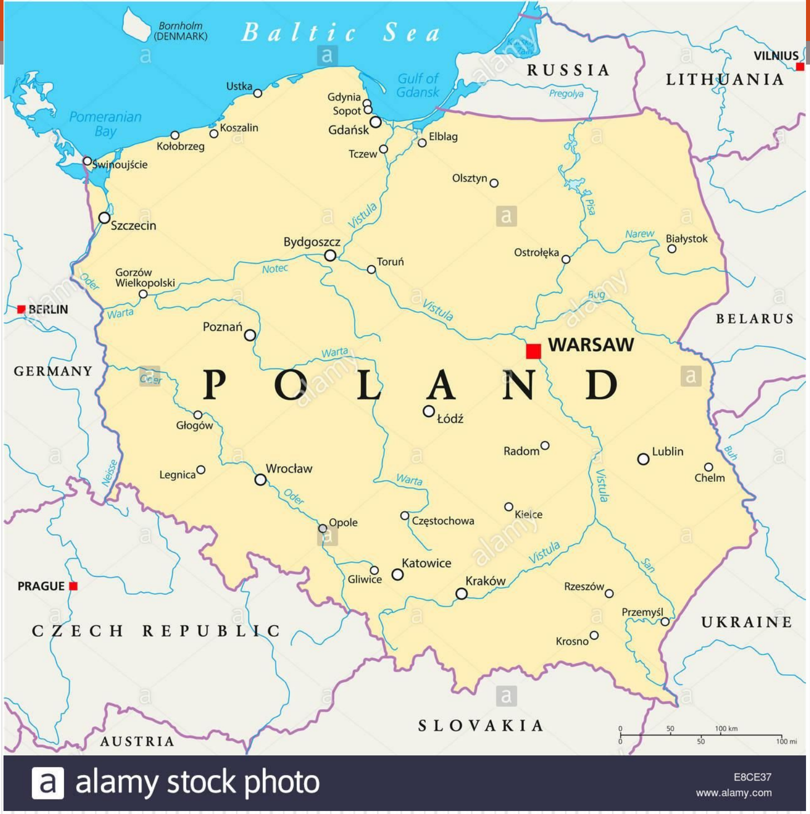
Exploration in data mining

Urszula Kużelewska, PhD
Bialystok University of Technology

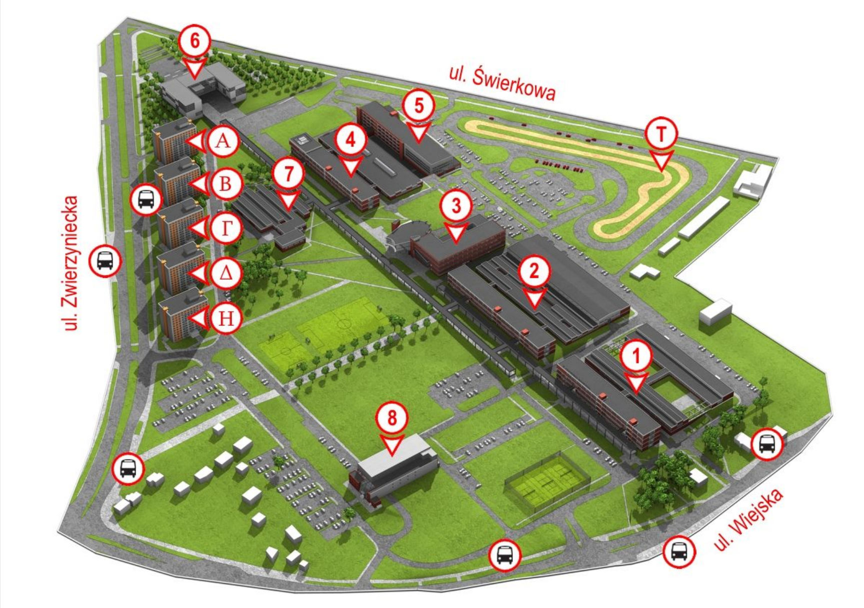
Poland



Białystok


Bornholm (DENMARK)
Baltic Sea
Gulf of Gdansk
RUSSIA
Pregolya
LITHUANIA
VILNIUS
Pomeranian Bay
Ustka
Koszalin
Kołobrzeg
Świnoujście
Gdynia
Sopot
Gdańsk
Tczew
Elblag
Olsztyn
Szczecin
Bydgoszcz
Vistula
Pisa
Narew
Białystok
Ostrołęka
Toruń
Noteć
Gorzów Wielkopolski
Oder
Warta
BERLIN
Bug
BELARUS
Poznań
Warta
WARSAW
GERMANY
POLAND
Łódź
Głogów
Radom
Lublin
Chełm
Bug
Legnica
Wrocław
Warta
Oder
Neisse
Vistula
Opole
Częstochowa
Kielce
Vistula
Katowice
Gliwice
Kraków
San
Rzeszów
Przemyśl
PRAGUE
CZECH REPUBLIC
UKRAINE
Krosno
SLOVAKIA
AUSTRIA
0 50 100 km
0 50 100 mi
alamy stock photo
E8CE37
www.alamy.com

Bialystok University of Technology



A plan of the lecture

- Introduction to clustering
- Partitioning algorithms
- Hierarchical algorithms
- Algorithms based on density
- Evaluation of clustering results
- Problems occurring in the clustering process
- Application of clustering results

Definition of clustering

Clustering - the process of **extracting knowledge** from a data set, when **no additional information** about it is available about the category assigned to individual sample points.

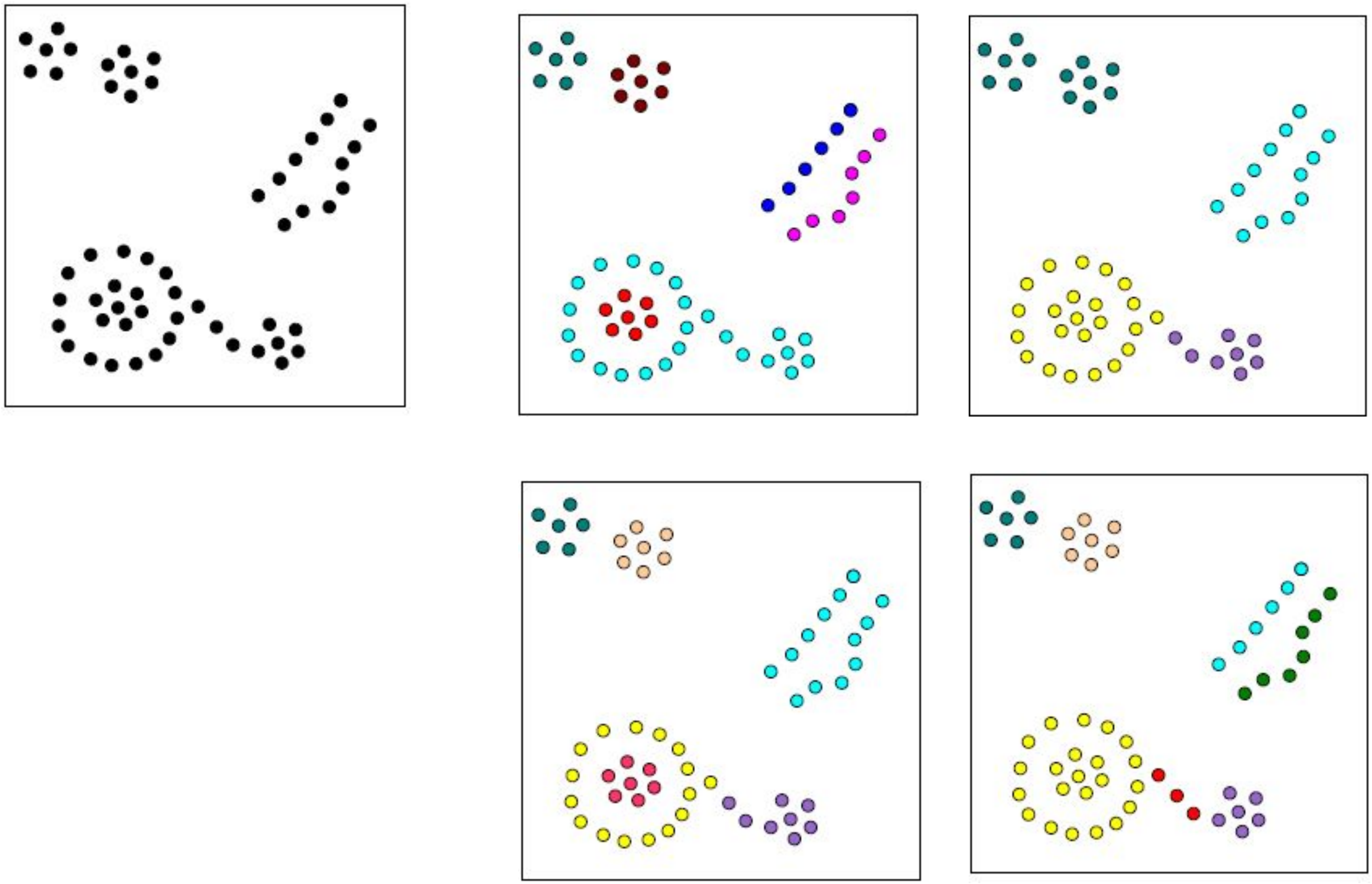
The purpose of clustering procedure is to divide a set into clusters (disjoint groups) such that each of them contains **the most similar** data in accordance with the **criterion defined *a priori***.

The main purpose of clustering algorithms is to create a **convenient and proper organization (structure) of data**, which consists of separate groups of objects.

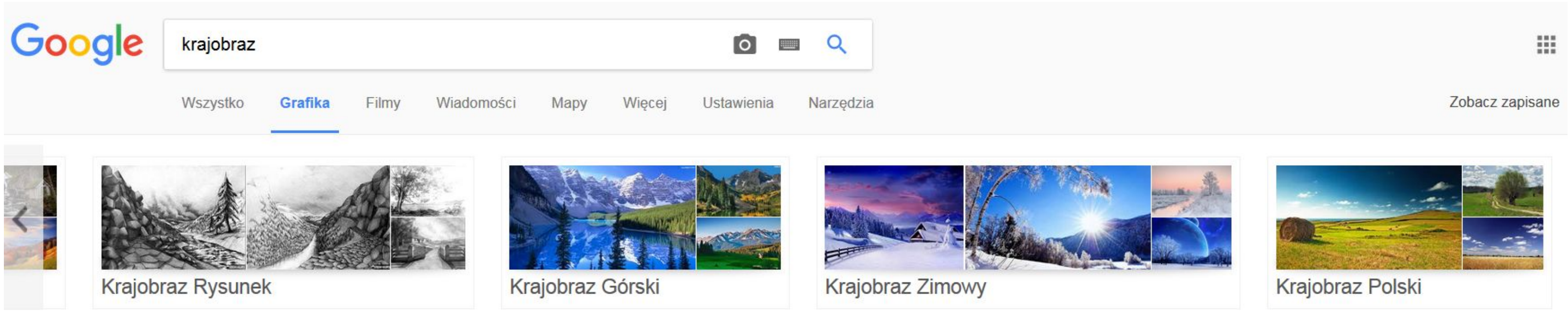
The process of identification groups from data is based on the **relations of similarity between the elements of the set**, in such a way that there is a strong similarity within the group, while among the groups is very small.

Experiment

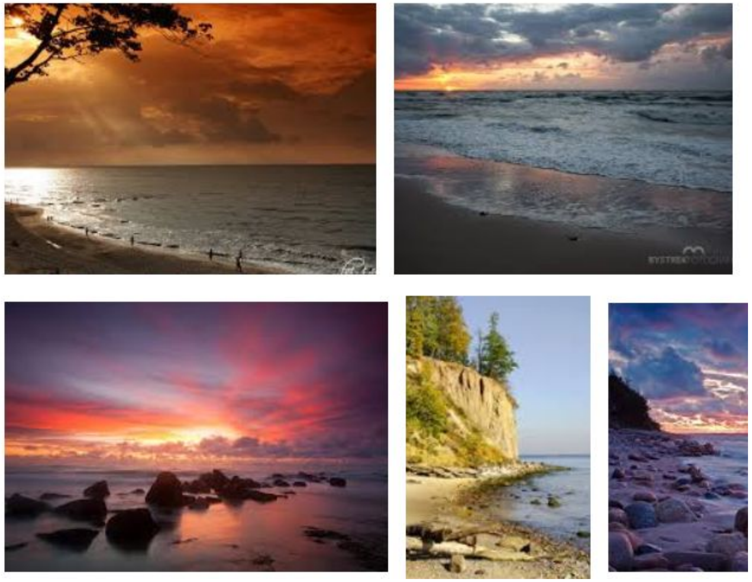
Clustering - various definitions



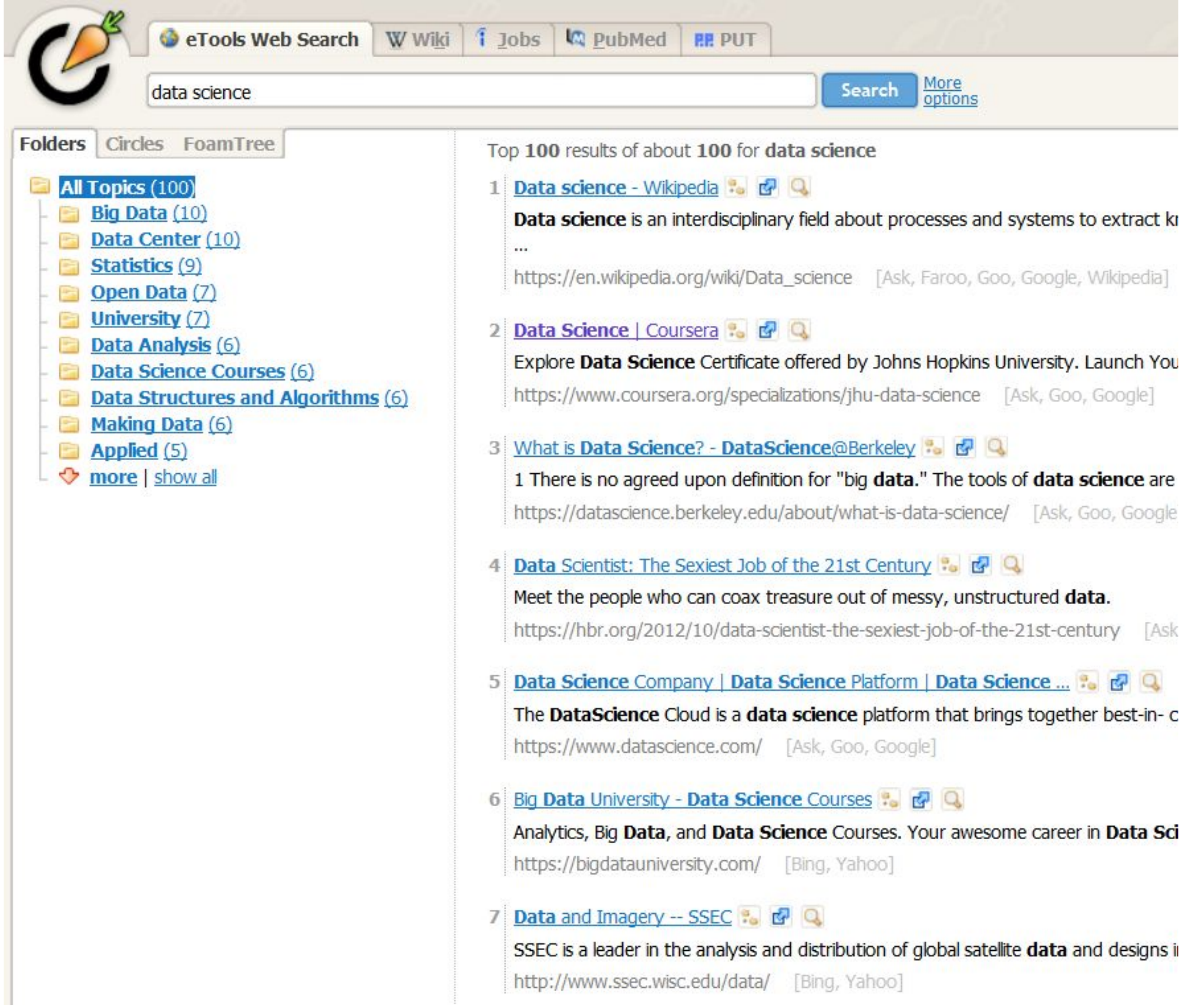
Examples of clustering applications



krajobraz



Examples of clustering applications



All Topics (100)
- Big Data (10)
- Data Center (10)
- Statistics (9)
- Open Data (7)
- University (7)
- Data Analysis (6)
- Data Science Courses (6)
- Data Structures and Algorithms (6)
- Making Data (6)
- Applied (5)
- more | show all

Top **100** results of about **100** for **data science**

1 Data science - Wikipedia
Data science is an interdisciplinary field about processes and systems to extract kı
...
https://en.wikipedia.org/wiki/Data_science [Ask, Faroo, Goo, Google, Wikipedia]

2 Data Science | Coursera
Explore **Data Science** Certificate offered by Johns Hopkins University. Launch You
https://www.coursera.org/specializations/jhu-data-science [Ask, Goo, Google]

3 What is Data Science? - DataScience@Berkeley
1 There is no agreed upon definition for "big **data**." The tools of **data science** are
https://datascience.berkeley.edu/about/what-is-data-science/ [Ask, Goo, Google

4 Data Scientist: The Sexiest Job of the 21st Century
Meet the people who can coax treasure out of messy, unstructured **data**.
https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century [Ask

5 Data Science Company | Data Science Platform | Data Science ...
The **DataScience** Cloud is a **data science** platform that brings together best-in- c
https://www.datascience.com/ [Ask, Goo, Google]

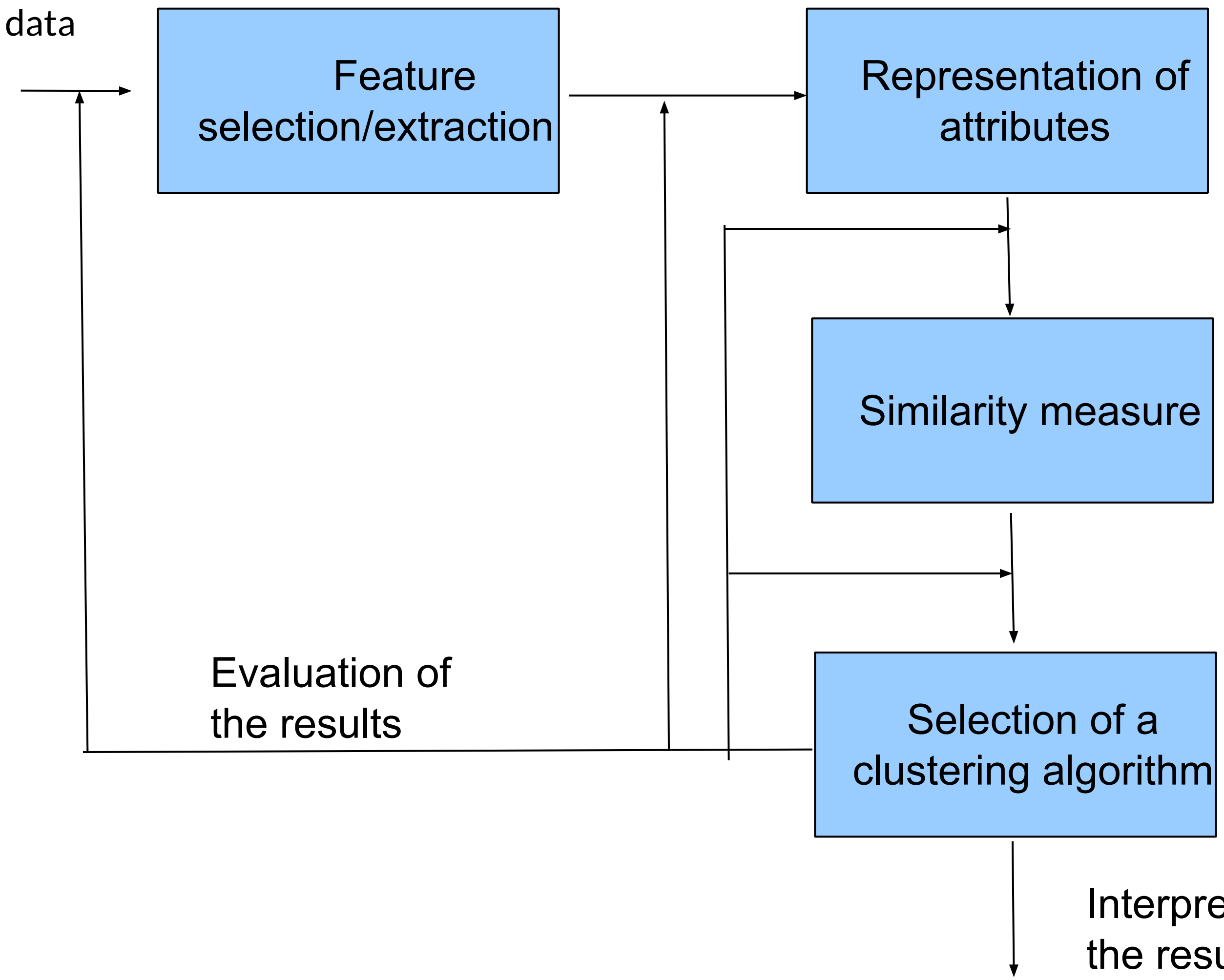
6 Big Data University - Data Science Courses
Analytics, Big **Data**, and **Data Science** Courses. Your awesome career in **Data Sci**
https://bigdatauniversity.com/ [Bing, Yahoo]

7 Data and Imagery -- SSEC
SSEC is a leader in the analysis and distribution of global satellite **data** and designs i
http://www.ssec.wisc.edu/data/ [Bing, Yahoo]

Examples of clustering applications

- Recognition of faces, letters, objects
- Compression of multimedia data
- Detection of anomalies
- Marketing - prediction of clients' preferences based on their previous behaviour
- Banking - determining the appropriate loan or type of account based on customer's earnings, type and place of work, age, etc.
- Biology - DNA matrix analysis
- Web mining - creating user profiles based on websites visited by him

Steps of clustering process



data

Feature selection/extraction

Representation of attributes

Similarity measure

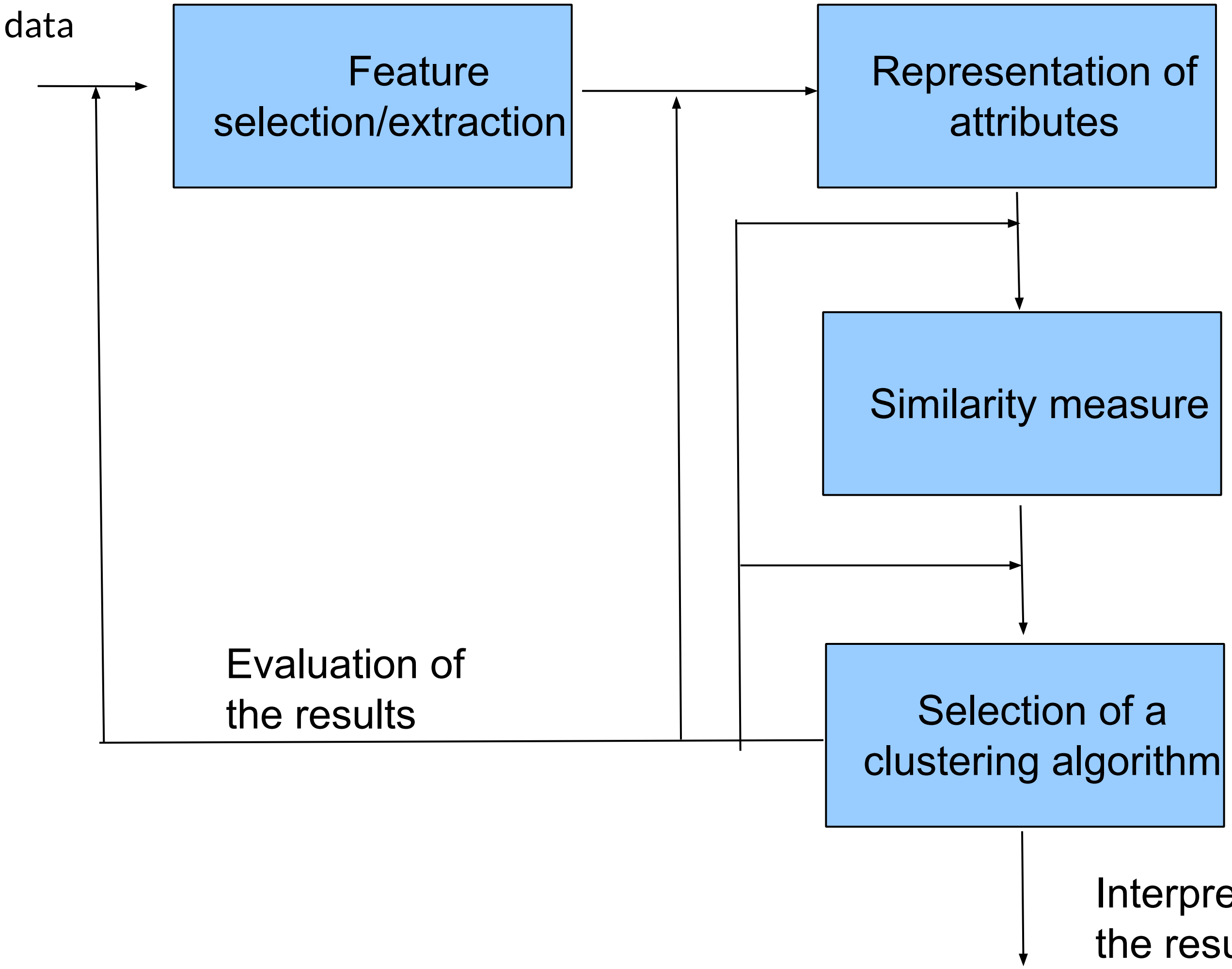
Selection of a clustering algorithm

Evaluation of the results

Interpretation of the results

The number of possible divisions of *n* objects into *k* groups: ?

Steps of clustering process


data
Feature selection/extraction
Representation of attributes
Similarity measure
Selection of a clustering algorithm
Evaluation of the results
Interpretation of the results

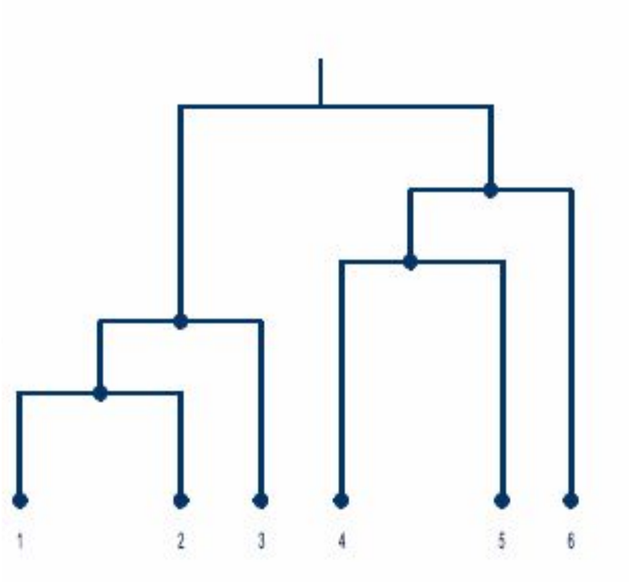
The number of possible divisions of *n* objects into *k* groups:

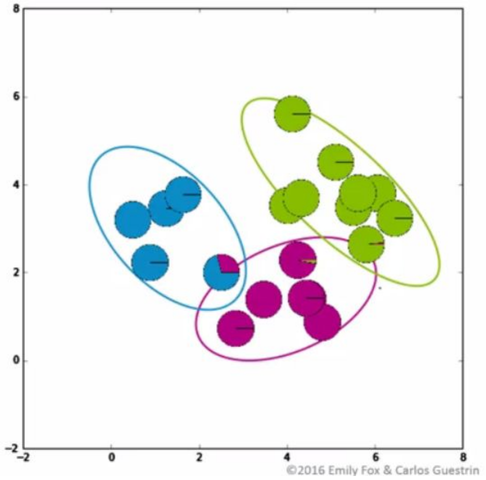
$\approx \boldsymbol{k}^{n}$

e.g. 2000 objects into 2 groups → 4 000 000

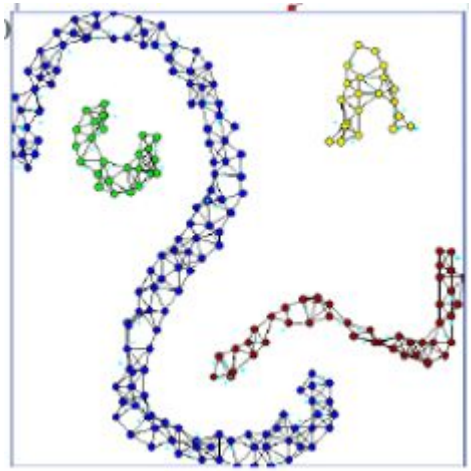
Into 20 groups → 1,048576e+66

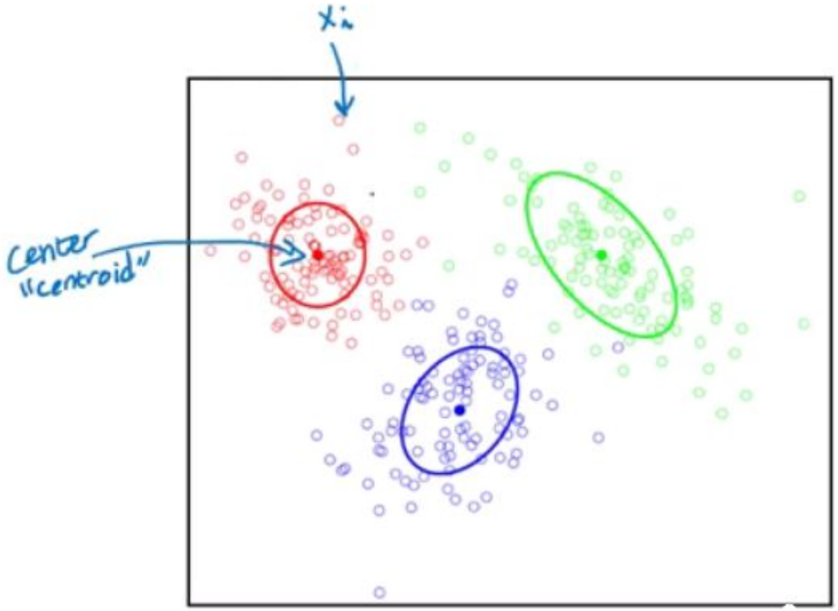
Representation of clusters



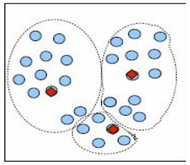


©2016 Emily Fox & Carlos Guestrin

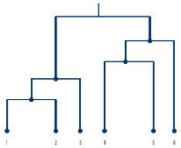




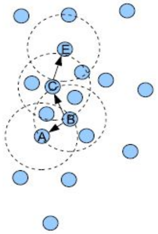
Clustering algorithms



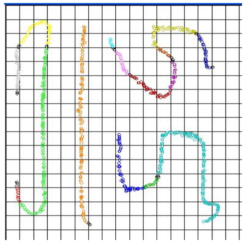
- Partitioning algorithms



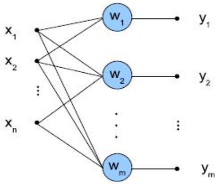
- Hierarchical algorithms



- Algorithms based on density
- Methods based on a grid generated in multidimensional space

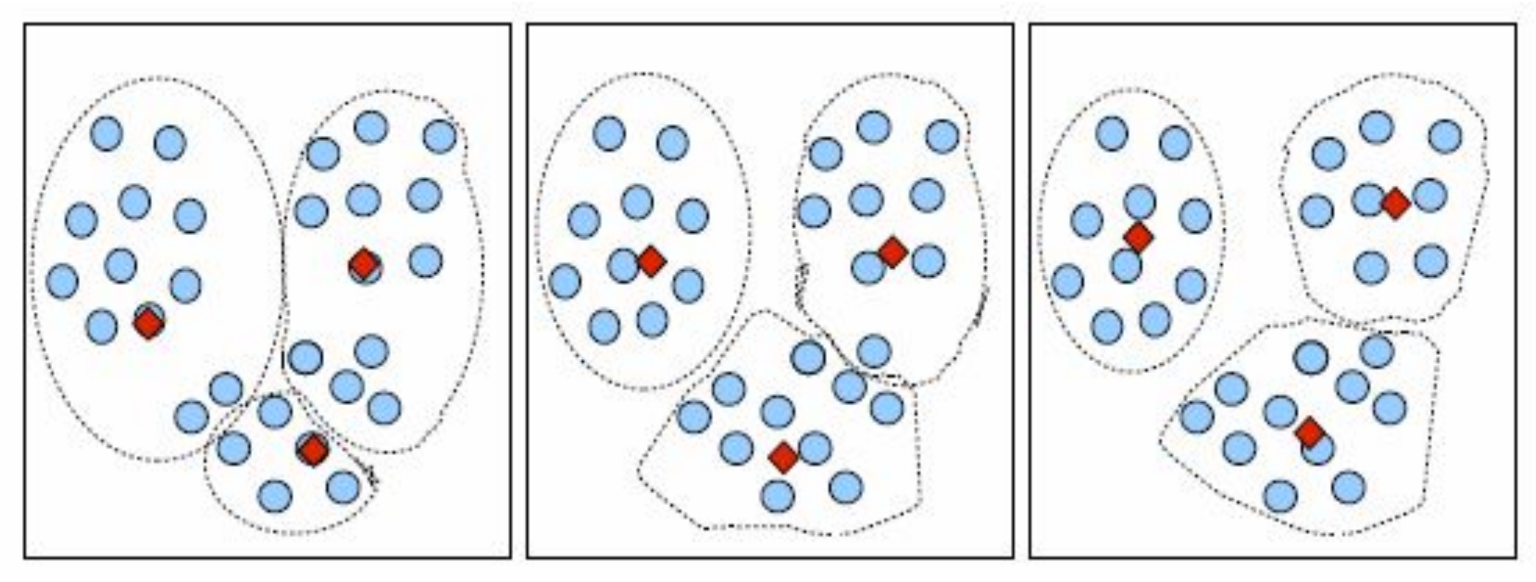


- Methods based on model and evaluation of model's parameters



Partitioning algorithms

Video on Coursera

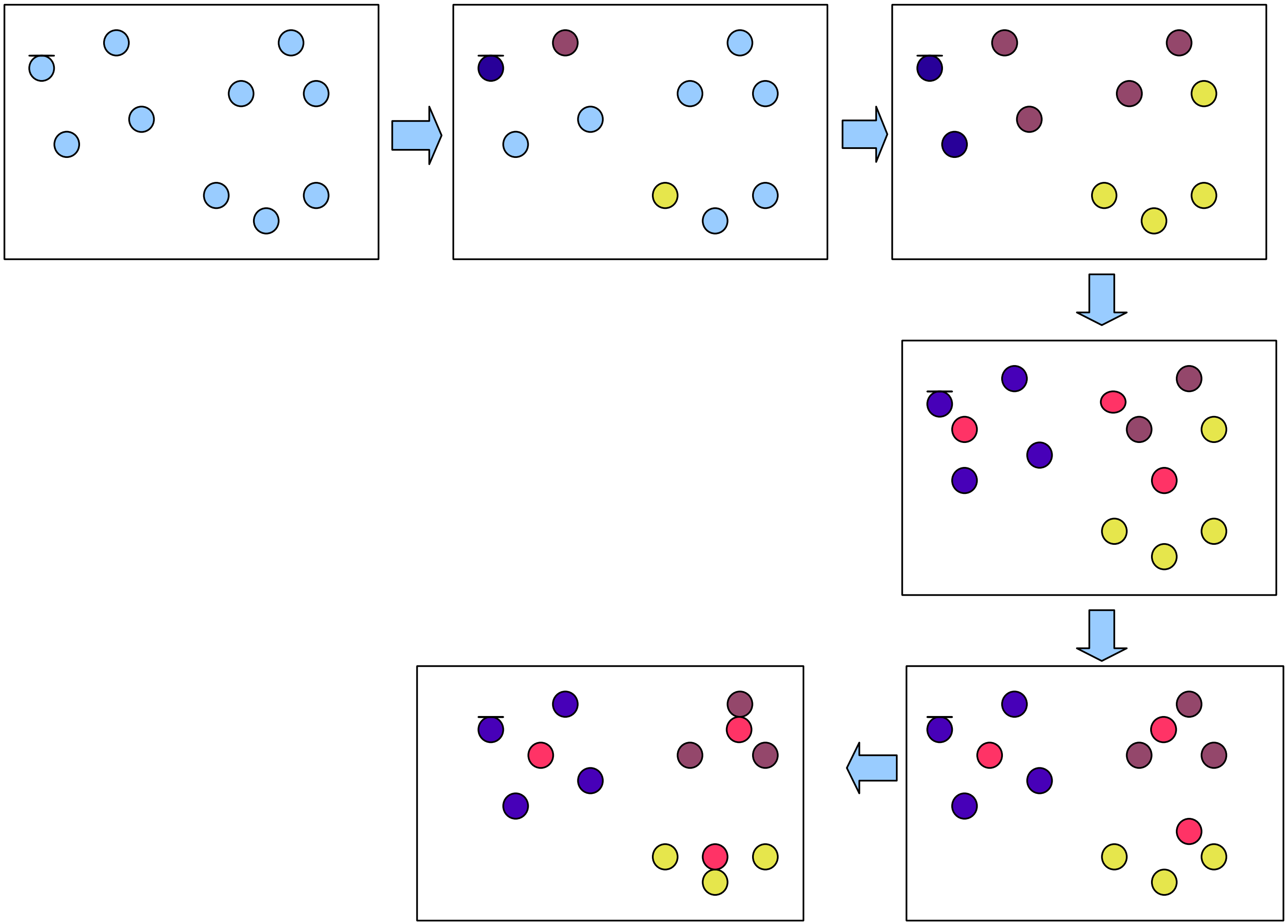


1. Optimisation of criterion function
2. Relocation of objects among groups

K-means algorithm (MacQueen'67)

1. Choose randomly *k* points from the set of data
2. Evaluate distances of all objects from dataset to every of *k* groups and denote their membership based on the closest distance.
3. If none of the points has changed their membership, stop the algorithm.
4. Calculate the mean square error of the sum of the distance of objects from the group centers.
5. If the calculated error value < determined threshold, stop the algorithm.
6. Calculate new cluster centers. Jump to p.2.

K-means algorithm

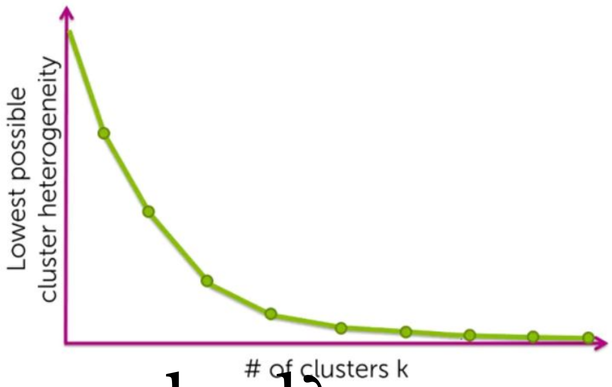


K-means advantages

- Time complexity O(tKn), where t – a number of iteration, K – a number of groups, n – a number of objects
- t,K<<n □ O(n)
- Fast?

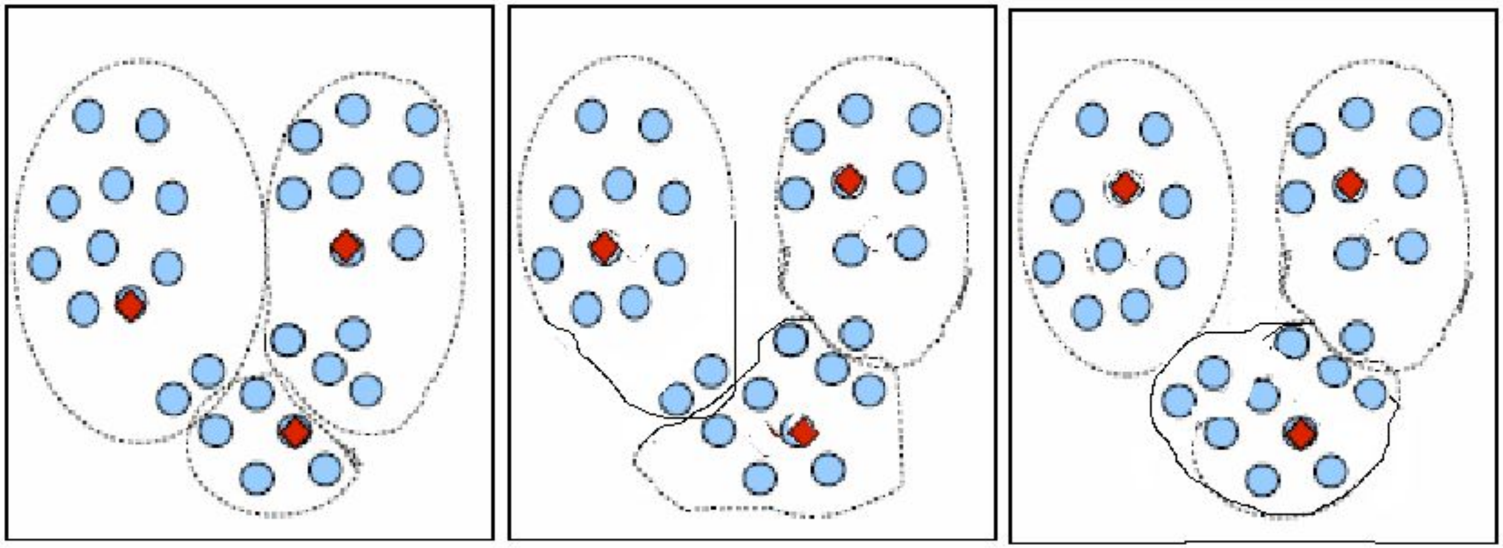
K-means disadvantages

How to choose k?



- ?
- How much is K?
- Is a global optimum always reached?
- Appropriated cluster centers initialization
- Clusters size?
- Clusters shape?
- Outliers, noise?

K-medoids algorithm (e.g. PAM - Partitioning Around Medoids, Kaufmann&Rousseeuw'87)

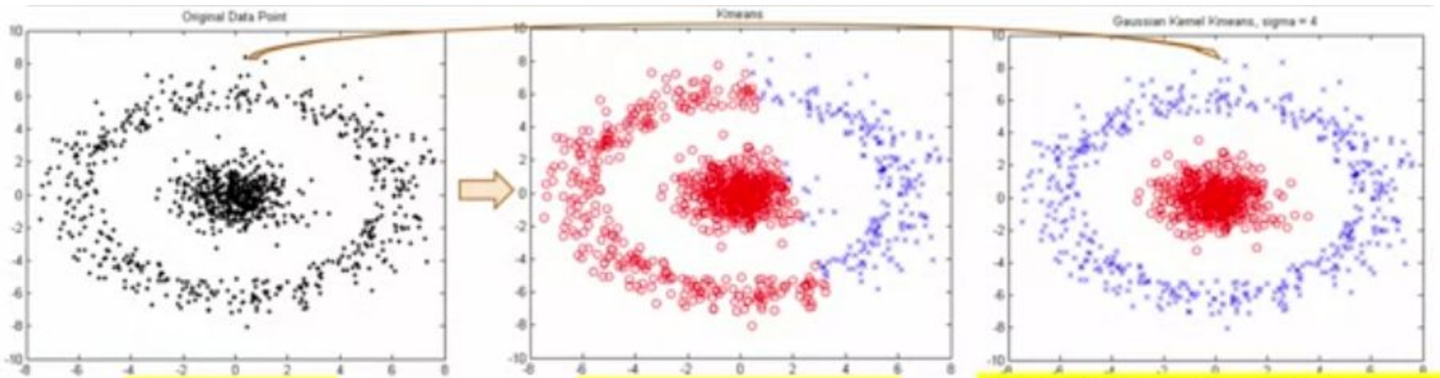


As a cluster center is always taken a point **from a dataset**

K-medoid pros&cons

- ?
- Resistant to outliers
- K?
- Fast?
- Time complexity O(n(n-K))!!!!
- Groups of spherical shape and comparable sizes
- Improved implementations:
  - CLARA (Kaufmann&Rousseeuw'90) - PAM on samples
  - CLARANS (Ng&Han'94)

Kernel k-means?



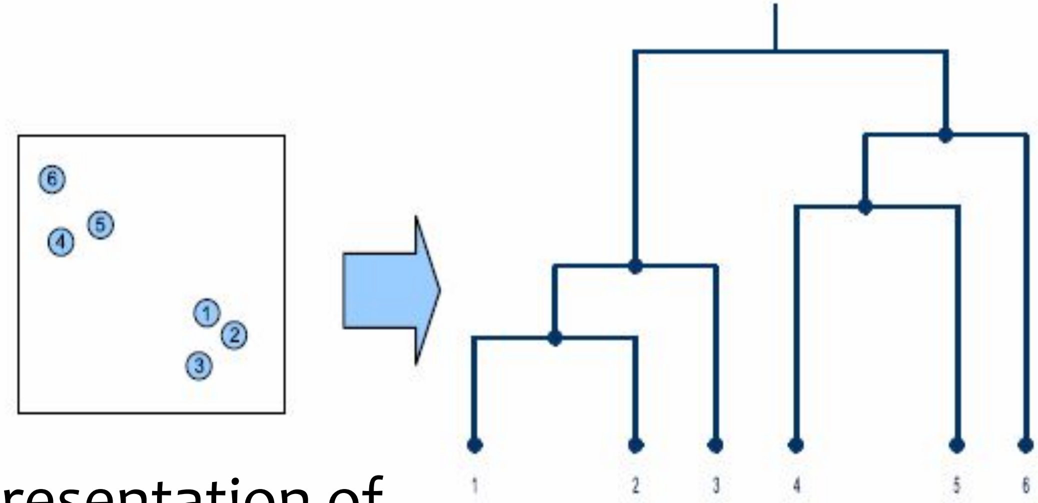
The original data set

The result of K-Means clustering

The result of Gaussian Kernel K-Means clustering

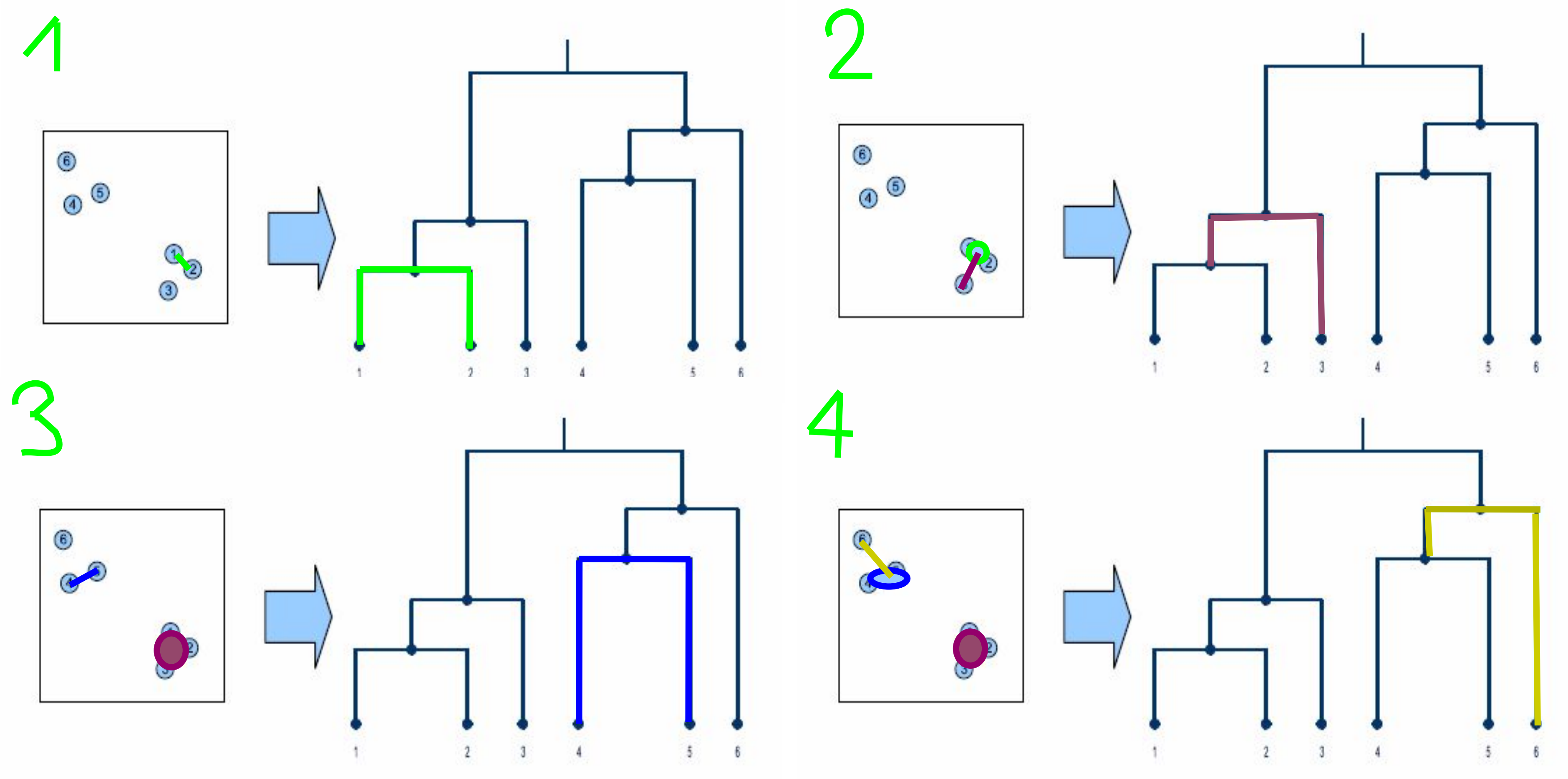
Hierarchical algorithms

- Sequential clustering
- Dendrogram formation



- Advantages: graphical presentation of relationship among the points, various methods of distance measure
- Disadvantages: time complexity $O(n^2)$, number of groups required, sensitive to outliers and noise

Hierarchical algorithms - formation of a dendrogram



Two ways of dendrogram formation

- **Divisive approach**

(top-down): starting from all objects in one group, then in every iteration large groups are split

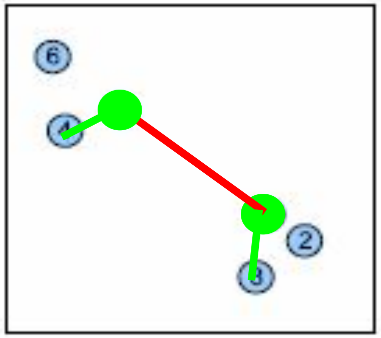
- **Agglomerative approach**

(bottom-up): starting from all objects in a separate group, then in every iteration the groups are joined

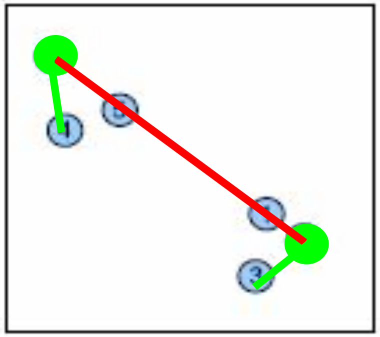
Joining of the groups

Various approaches to distance measure: e.g. single-link (hsl) or complete-link (hcl)

hsl



hcl



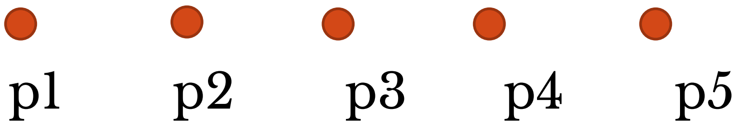
$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \left\| p - p' \right\|$$

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

Example – hierarchical bottom-up single link algorithm

	X1	x2
p1	-1,88	2,05
p2	-0,71	0,42
p3	2,41	-0,67
p4	1,85	-3,8
p5	-3,69	-1,33

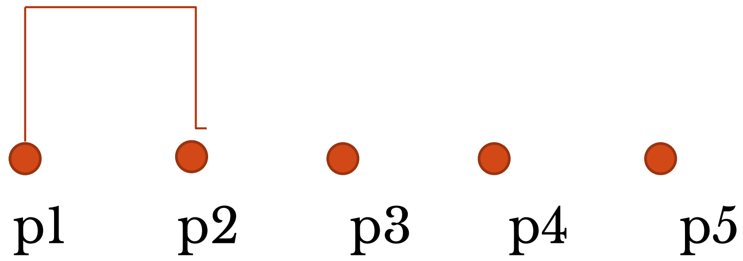
	p1	p2	p3	p4	p5
p1	0	2,0064	5,08	6,938	3,834
p2		0	3,305	4,936	3,456
p3			0	3,18	6,136
p4				0	6,066
p5					0



Example – hierarchical bottom-up single link algorithm

	X1	x2
p1	-1,88	2,05
p2	-0,71	0,42
p3	2,41	-0,67
p4	1,85	-3,8
p5	-3,69	-1,33

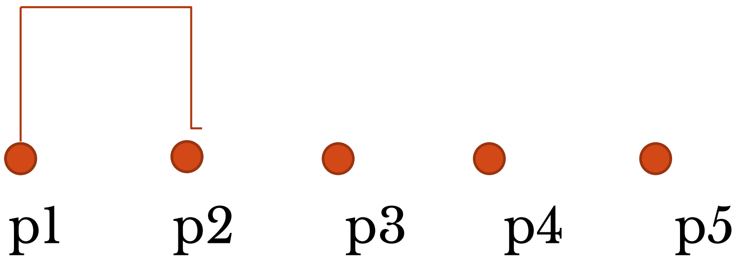
	p1	p2	p3	p4	p5
p1	0	2,0064	5,08	6,938	3,834
p2		0	3,305	4,936	3,456
p3			0	3,18	6,136
p4				0	6,066
p5					0



Example – hierarchical bottom-up single link algorithm

	p1	p2	p3	p4	p5
p1	0	2,0064	5,08	6,938	3,834
		0	3,305	4,936	3,456
			0	3,18	6,136
				0	6,066
					0

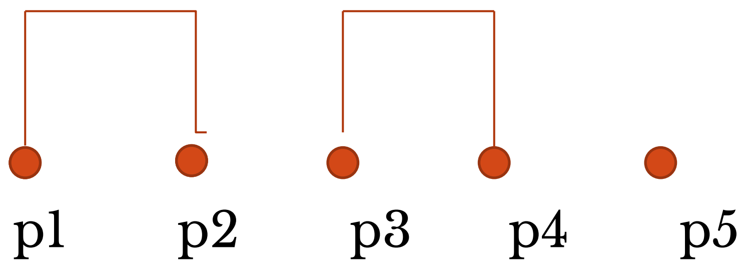
hsl	p1,p2	p3	p4	p5
p1,p2	0	3,305	4,936	3,456
p3		0	3,18	6,136
p4			0	6,066
p5				0



Example – hierarchical bottom-up single link algorithm

	X1	x2
p1	-1,88	2,05
p2	-0,71	0,42
p3	2,41	-0,67
p4	1,85	-3,8
p5	-3,69	-1,33

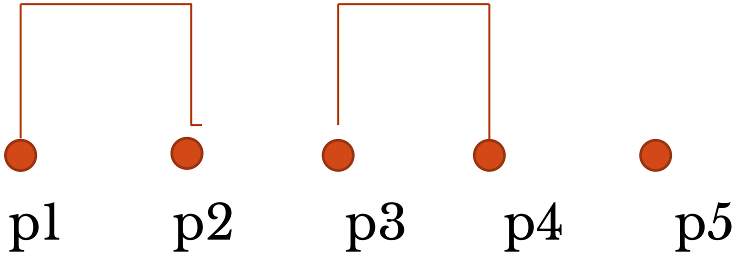
hsl	p1,p2	p3	p4	p5
p1,p2	0	3,305	4,936	3,456
p3		0	3,18	6,136
p4			0	6,066
p5				0



Example – hierarchical bottom-up single link algorithm

hsl	p1,p2	p3	p4	p5
p1,p2	0	3,305	4,936	3,456
		0	3,18	6,136
			0	6,066
				0

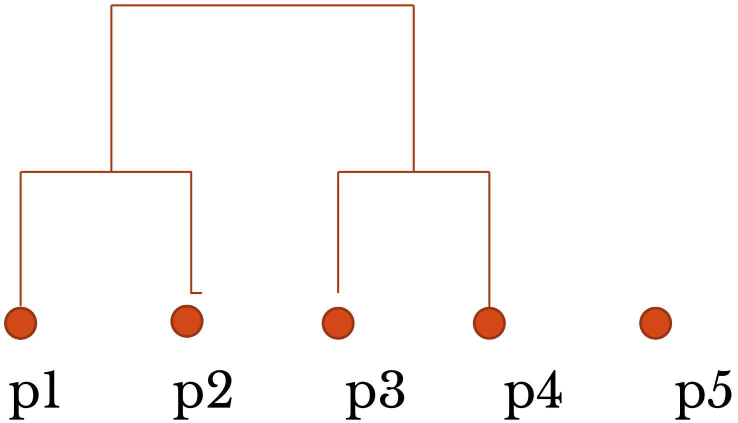
hsl	p1,p2	p3,p4	p5
p1,p2	0	3,305	3,456
p3,p4		0	6,066
p5			0



Example – hierarchical bottom-up single link algorithm

	X1	x2
p1	-1,88	2,05
p2	-0,71	0,42
p3	2,41	-0,67
p4	1,85	-3,8
p5	-3,69	-1,33

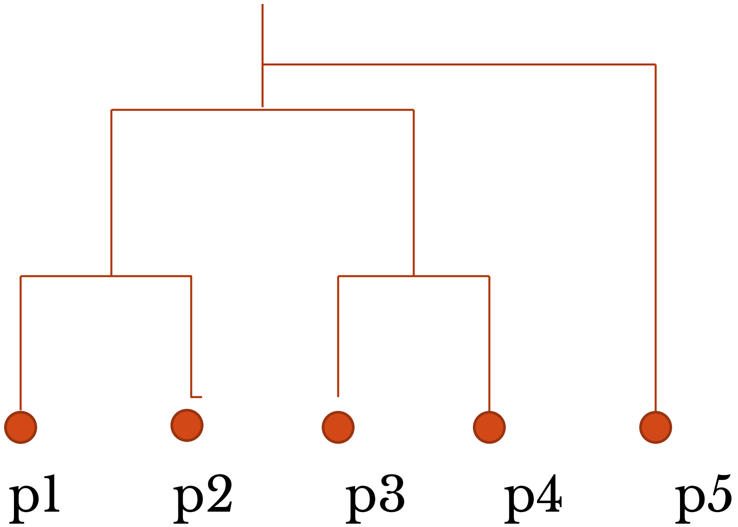
hsl	p1,p2	p3,p4	p5
p1,p2	0	3,305	3,456
p3,p4		0	6,066
p5			0



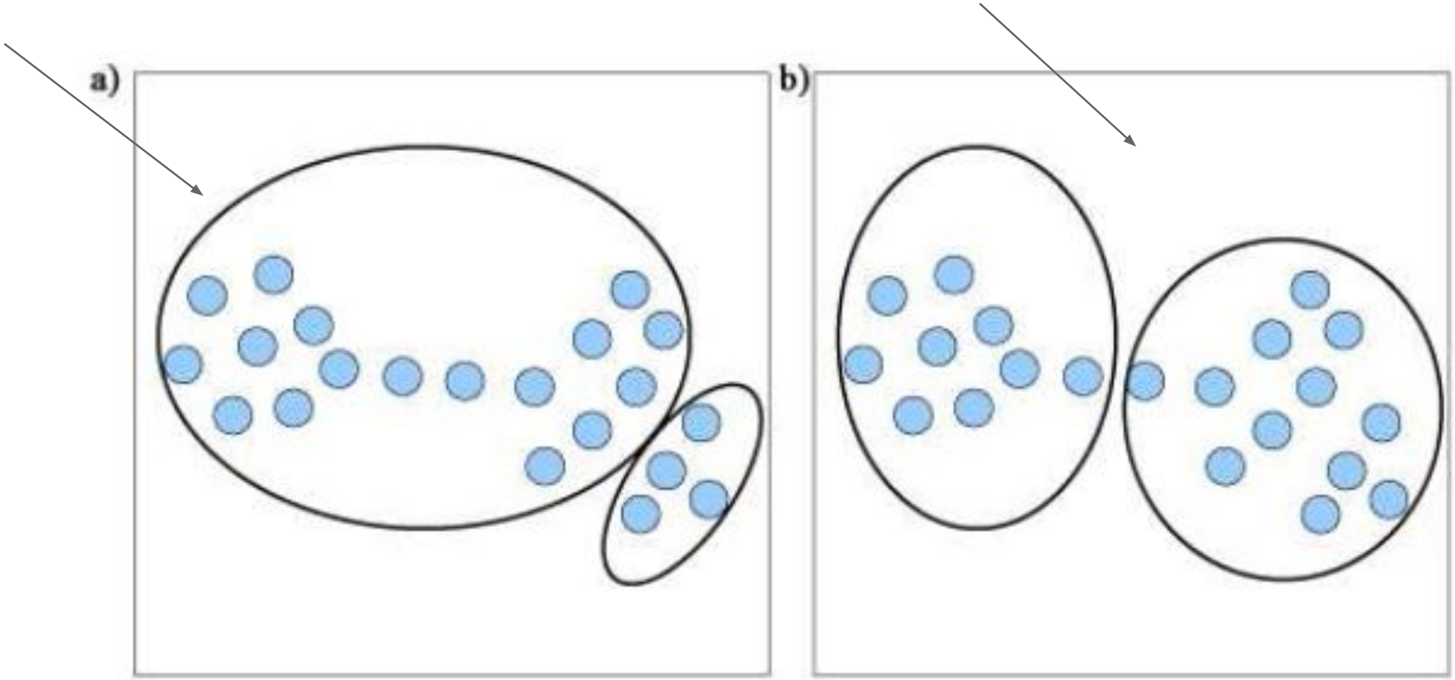
Example – hierarchical bottom-up single link algorithm

hsl	p1,p2	p3,p4	p5
p1,p2	0	3,305	3,456
		0	6,066
			0

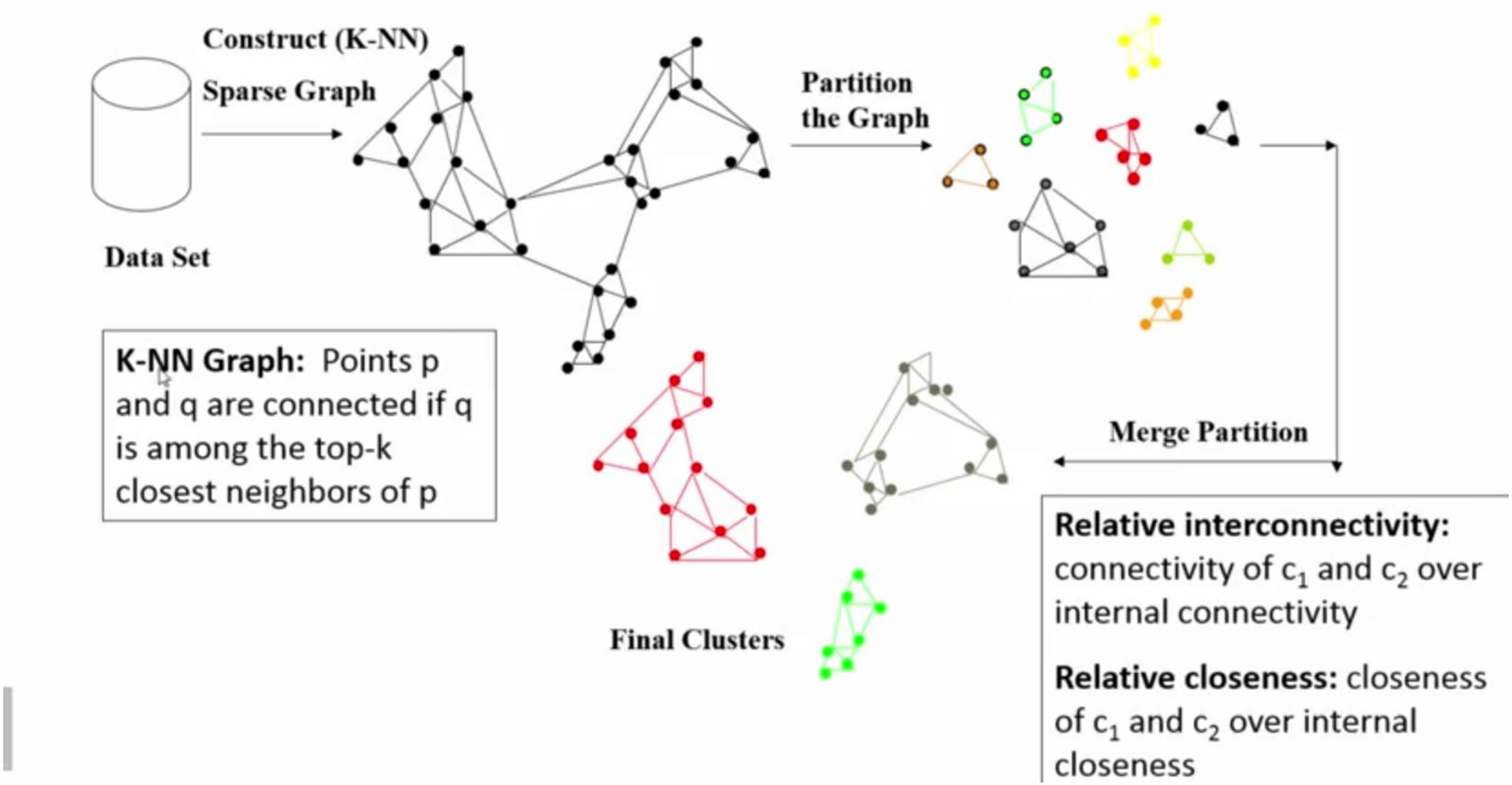
hsl	p1,p2,p3,p4	p5
p1,p2,p3,p4	0	3,456
p5		0



hsl i hcl - comparison



CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling - George Karypis

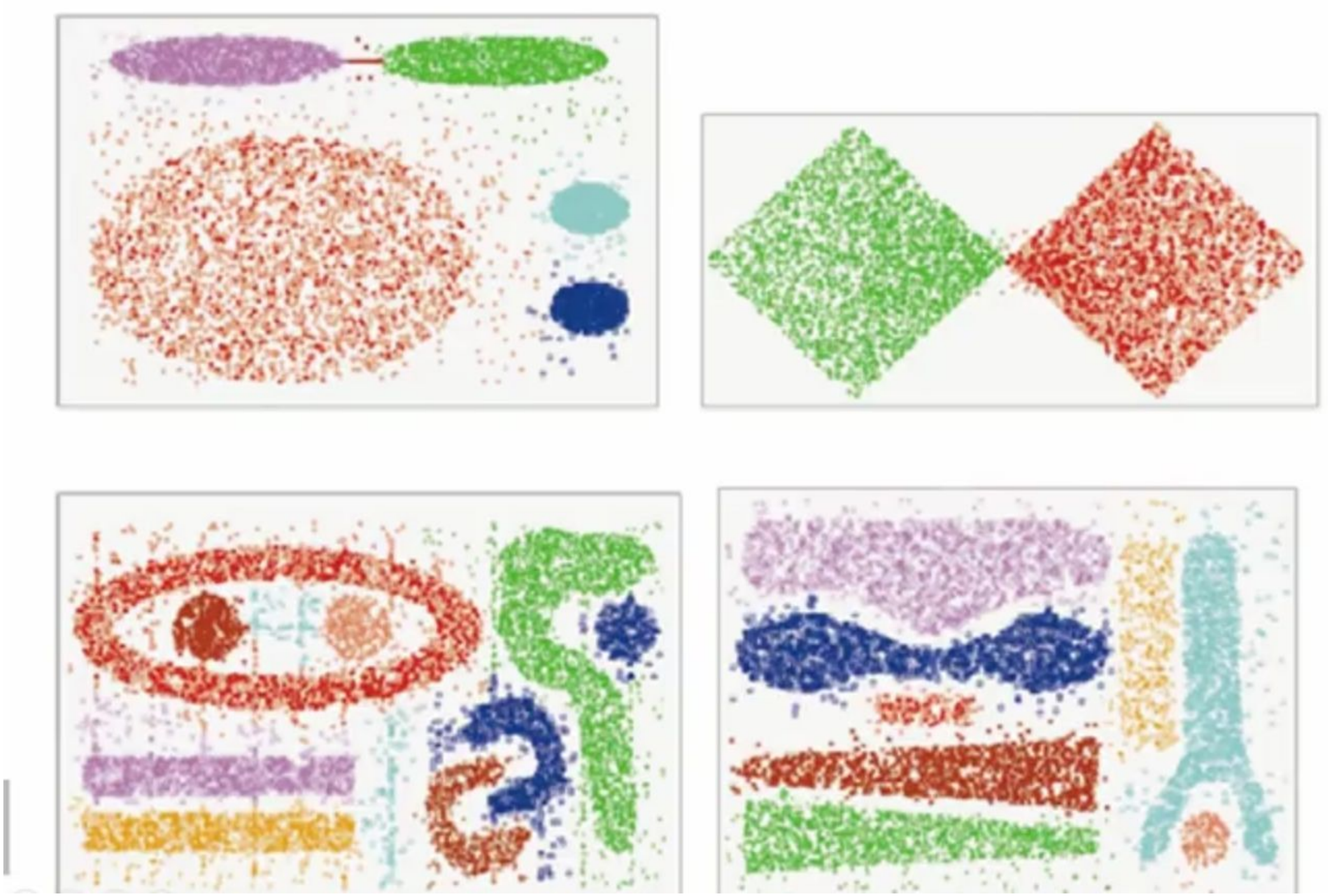


K-NN Graph: Points p and q are connected if q is among the top-k closest neighbors of p

Relative interconnectivity: connectivity of c_1 and c_2 over internal connectivity

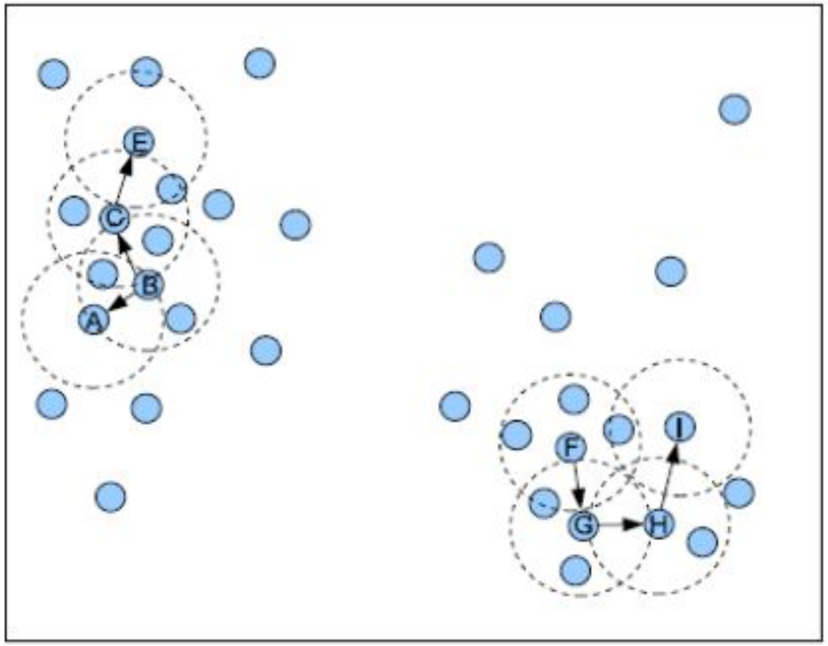
Relative closeness: closeness of c_1 and c_2 over internal closeness

CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling - George Karypis



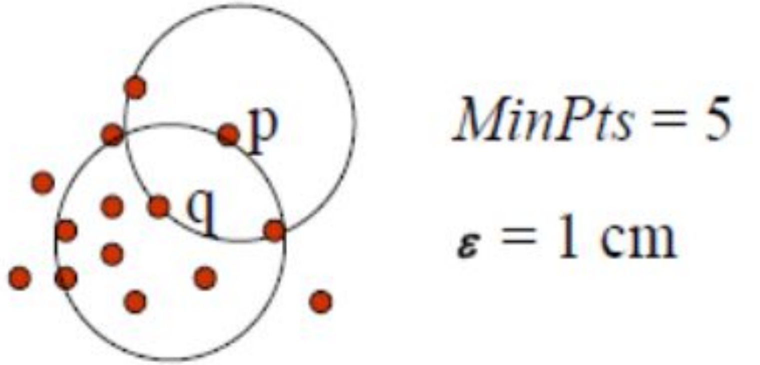
Algorithms based on density

- The areas of similar density are joined
- E.g. DBSCAN
- **Advantages:** low time complexity *O(nlogn)*, resistant to outliers and noise, no input parameter related to a number of groups, arbitrary shapes of clusters
- **Disadvantages:** other input parameters: *epsilon*, *minPts*



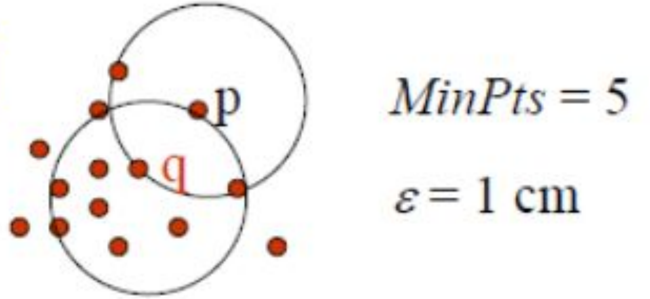
Algorithm DBSCAN

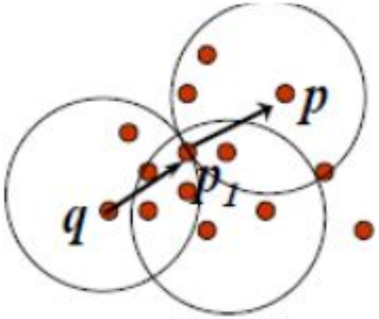
- *Epsilon* - a radius defined neighbourhood of object
- *minPts* - minimal number of points in neighbourhood
- **Core object:** an object that has at least *minPts* in its neighbourhood
- **Border object:** an object that has less than *minPts* in its neighbourhood

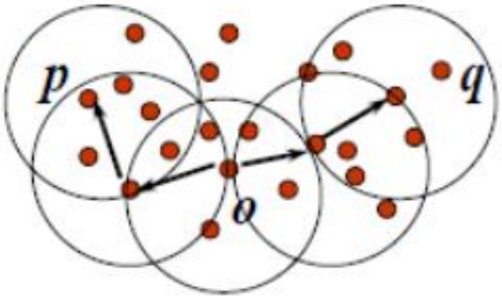


Algorithm DBSCAN

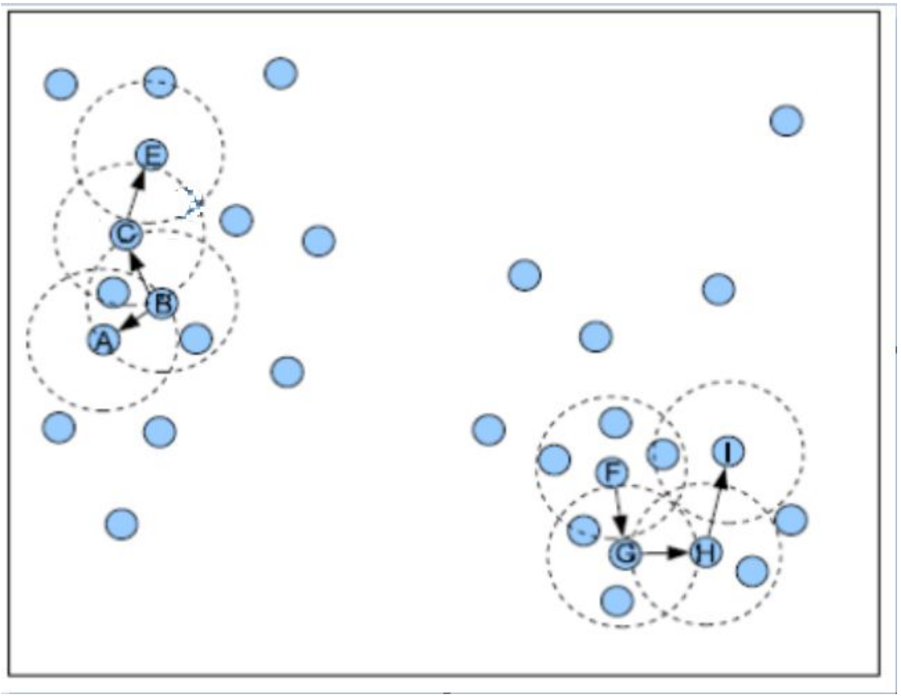
- If any object has a core object in its neighbourhood is called as **directly reachable by density**
- If any object is connected to other object through points directly reachable by density is called as **reachable by density**
- The objects that are connected with each other and between them is an object reachable by density are called **connected by density**







Example



1. Set minPts=4
2. Core points: B, F
3. Points directly reachable by density: A (from B), C (from B), G (from F)
4. Points reachable by density: A and C
5. The points E and A are connected by density through B point

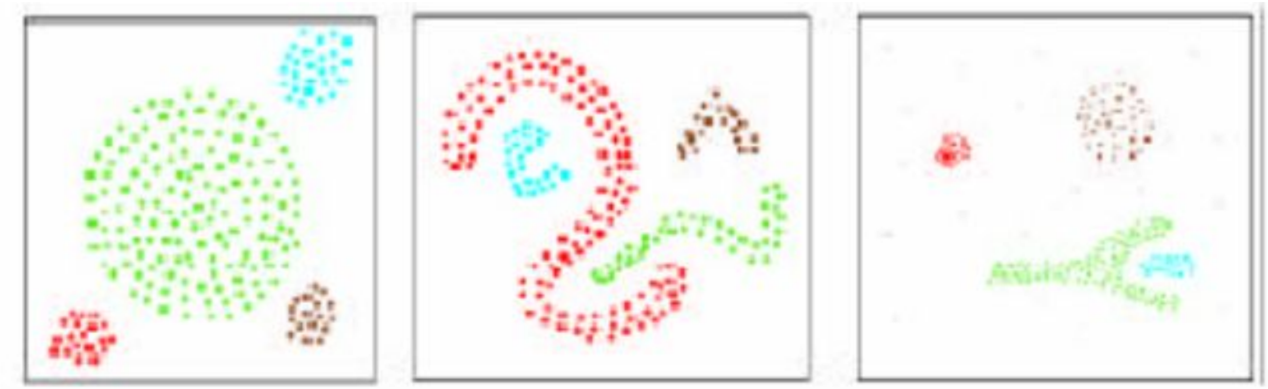
Algorithm DBSCAN

1. Set values of parameters epsilon and minPts
2. Select an arbitrary point from the dataset
3. Identify a set *G* composed of points reachable by density from the point *p*
    a. If the point *p* is a core, denote *G* as a group
    b. If *p* is a border point, go to the following point
4. If there are any unvisited points go to the step 2

Video

M. Ester, H.P. Kriegel, J. Sander, X. Xu ,96

Algorithm DBSCAN - some examples



Algorithm DBSCAN - sensitivity to input parameter's values

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

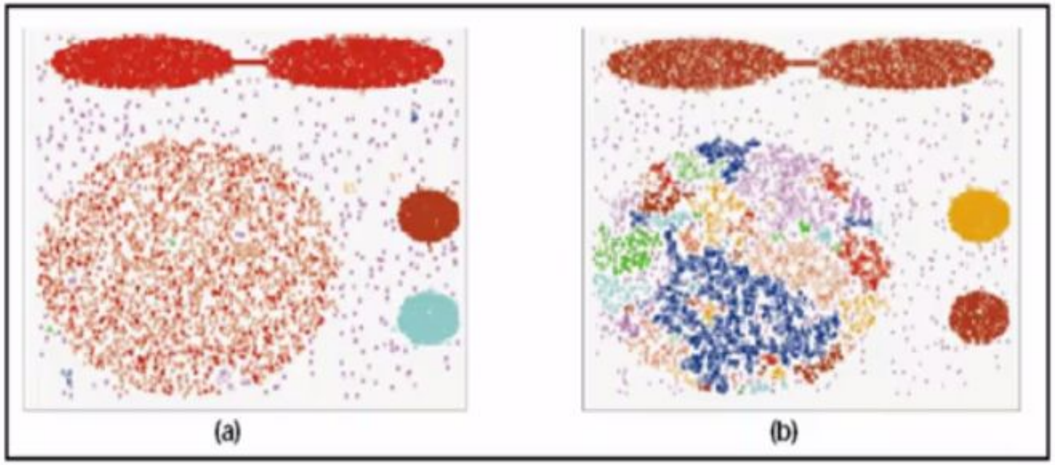
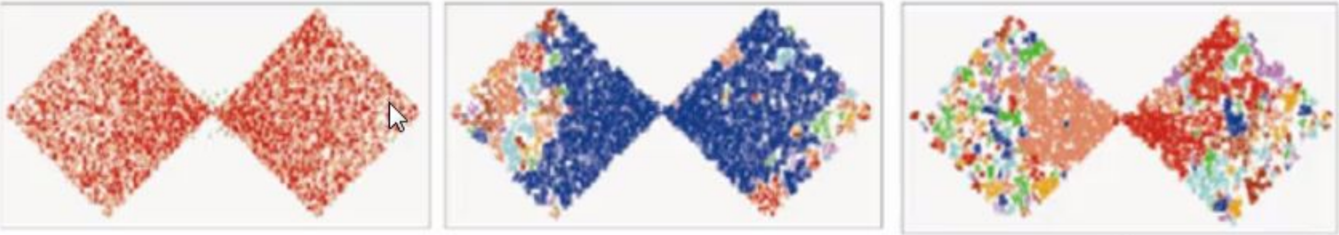
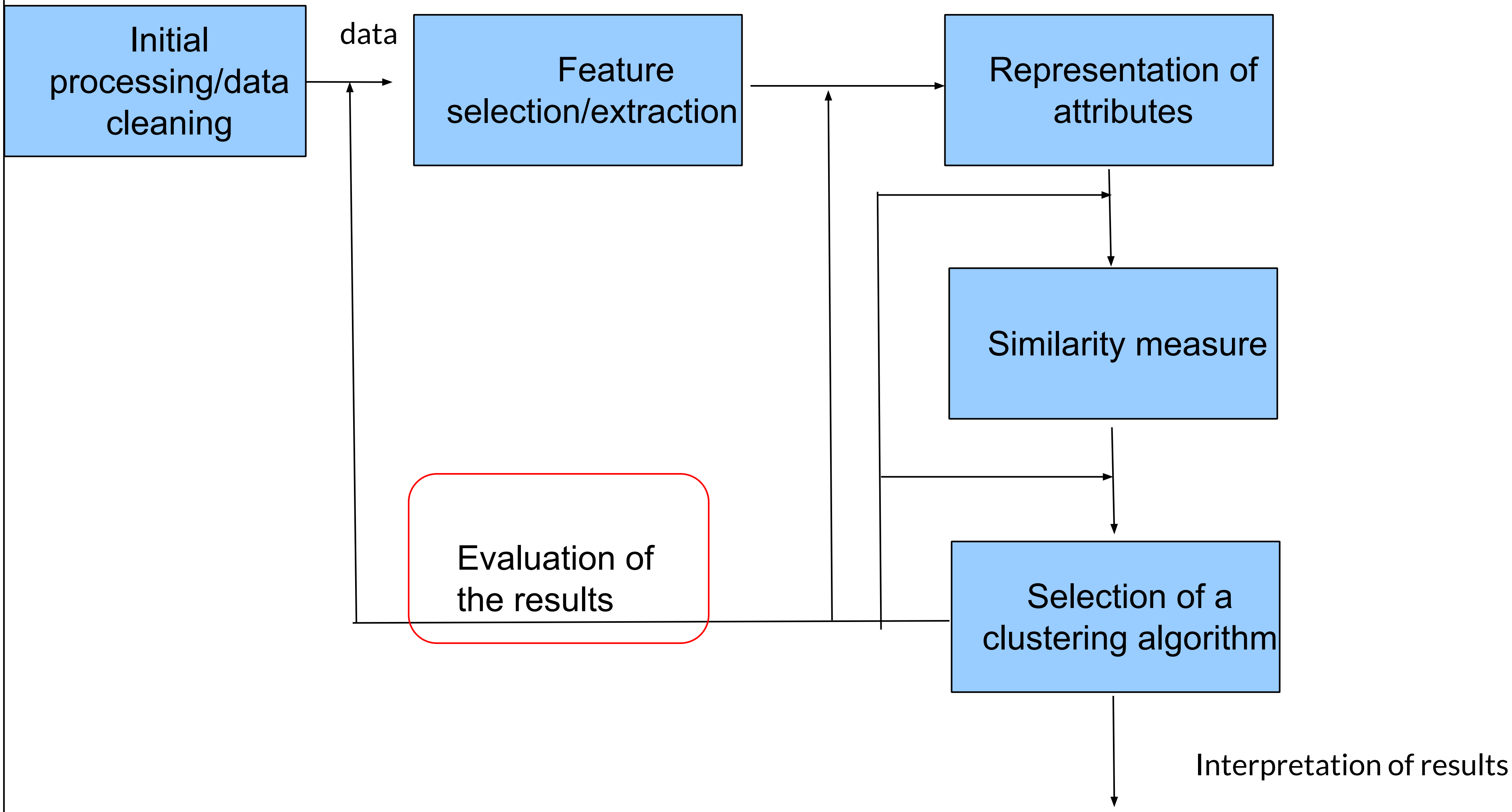


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



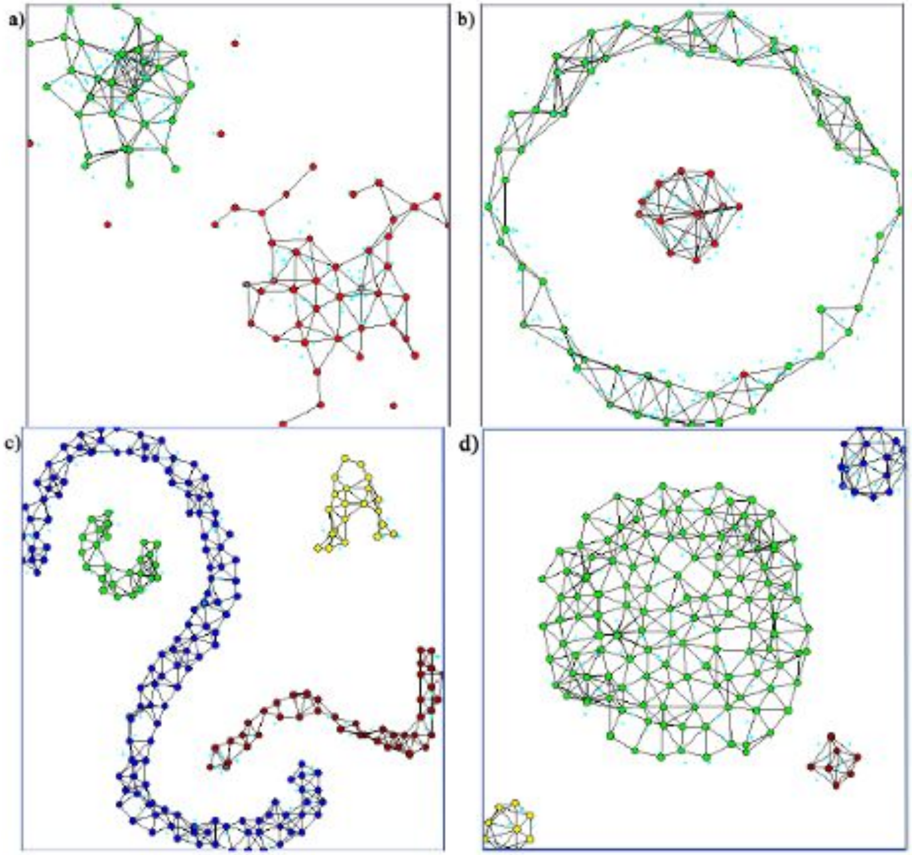
Steps of clustering process



Initial processing/data cleaning
data
Feature selection/extraction
Representation of attributes
Similarity measure
Evaluation of the results
Selection of a clustering algorithm
Interpretation of results

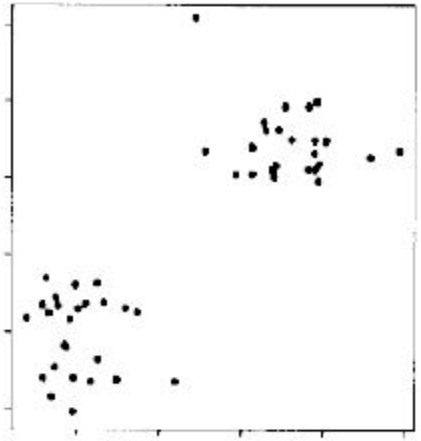
Evaluation of clustering results

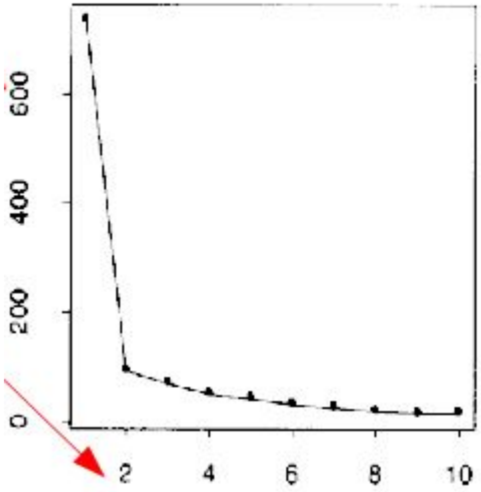
Visual evaluation



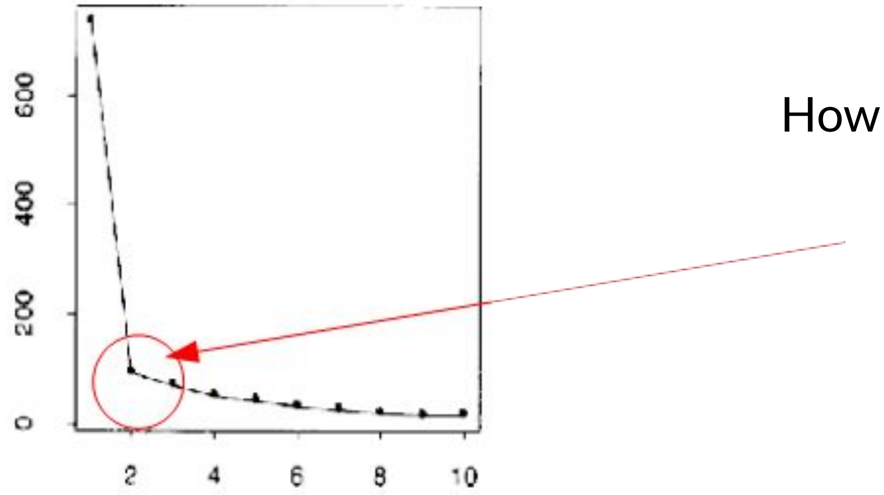
A knee-plot approach

1. Different values of input parameter + values of evaluation index
2. Examples of evaluation indices:
   a. Root Mean Square Error
   b. A scattering matrix
3. Examples of input parameters:
   a. A number of clusters
   b. Radius of neighbourhood
   c. Top k-closest neighbours

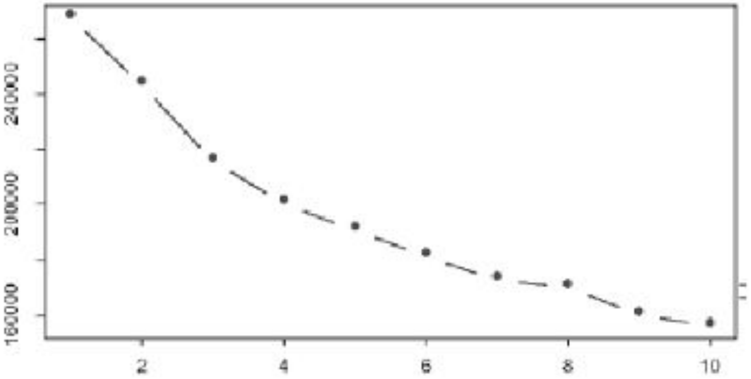




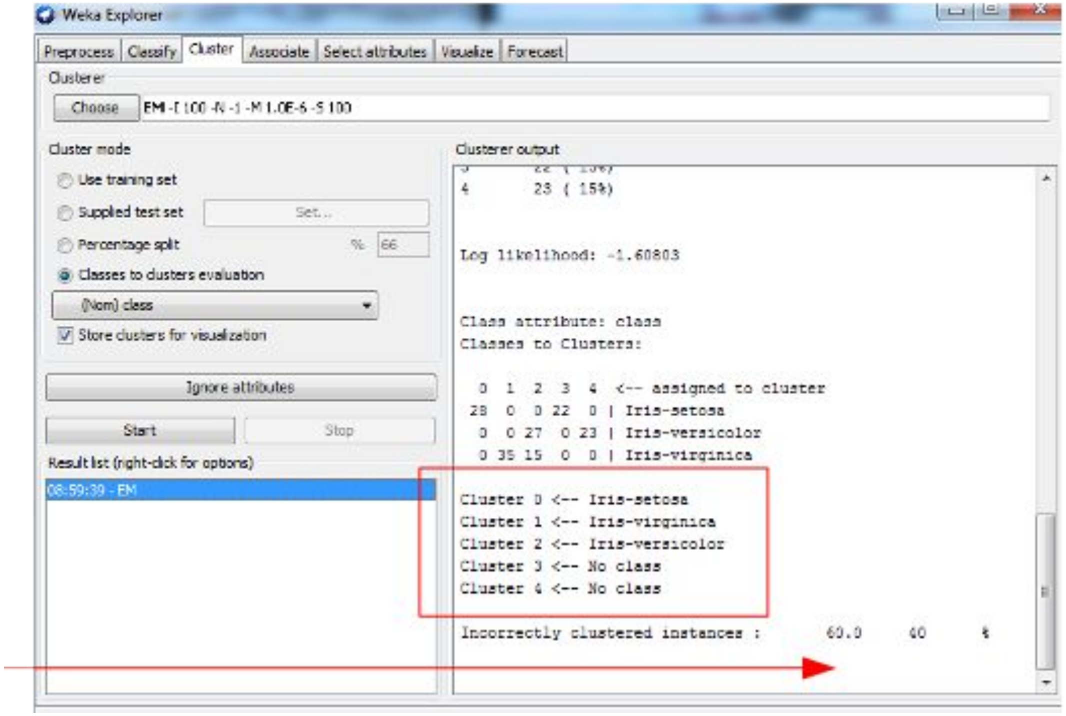
Examples of a knee-plot approach



How many groups?



External measures



Clusterer: EM -I 100 -N -1 -M 1.0E-6 -S 100

```
3       22 ( 15%)
4       23 ( 15%)


Log likelihood: -1.60803


Class attribute: class
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 28  0  0 22  0 | Iris-setosa
  0  0 27  0 23 | Iris-versicolor
  0 35 15  0  0 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-virginica
Cluster 2 <-- Iris-versicolor
Cluster 3 <-- No class
Cluster 4 <-- No class

Incorrectly clustered instances :      60.0     40      %
```

Internal measures

- A level of compactness of the groups ↑
- A level of separability of the groups ↑
- The values are calculated for various values of input parameter
- Maximal or minimal values of the measure indicates optimal value of the input parameter

Internal measures

- Dunn Index (DI) (0,+∞) ↑

$$D_{nc} = \min_{i=1,\dots,nc}\left\{\min_{j=i+1,\dots,nc}\left(\frac{d(C_i,C_j)}{\max_{k=1,\dots,nc} diam(C_k)}\right)\right\}$$

Cluster diameter

$$diam(C) = \max_{x,y\in C} d(x,y)$$

$$d(C_i,C_j) = \min_{x\in C_i, y\in C_j} d(x,y)$$

- Davies-Bouldin Measure (DB) (0,+∞) ↓

$$DB_{nc} = \frac{1}{nc}\sum_{i=1}^{nc}\left(\max_{j=1,\dots,nc, j\neq i} R_{ij}\right) \quad R_{ij} = \frac{sdev\,(C_i) + sdev\,(C_j)}{d(C_i,C_j)} \quad sdev\,(C_i) = \frac{1}{|C_i|}\sqrt{\sum_{x\in C_i}(d\,(x,\overline{x}))^2}$$

- Silhouette Index (SI) [-1,1] ↑

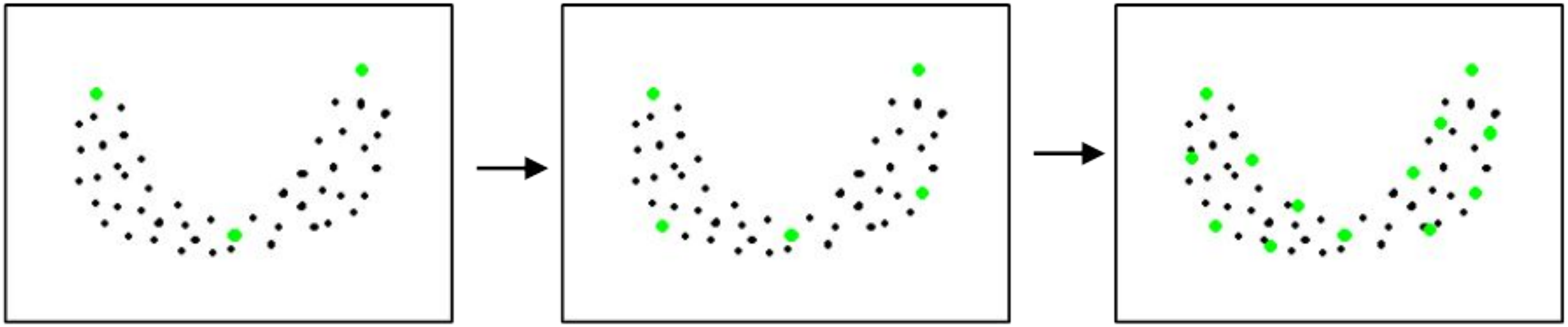
$$SI = mean_{\forall x_i\in U}\left(\frac{b_i - a_i}{\max(a_i,b_i)}\right) \quad a_i = \frac{\Sigma_{x_j\in Ck, i\neq j}\delta_{ij}}{card(C_k)-1} \quad b_i = \min_{r\neq k}\left(\frac{\Sigma_{x_j\in Cr}\delta_{ij}}{card(C_r)}\right) \quad \delta_{ij} = \frac{d_{ij}}{\max(d_{ij})}$$

- Cdbw measure (0,+∞) ↑

Average similarity of the point x_i to all the other points from the same group

Minimal average dissimilarity of the point x_i from all points from other groups

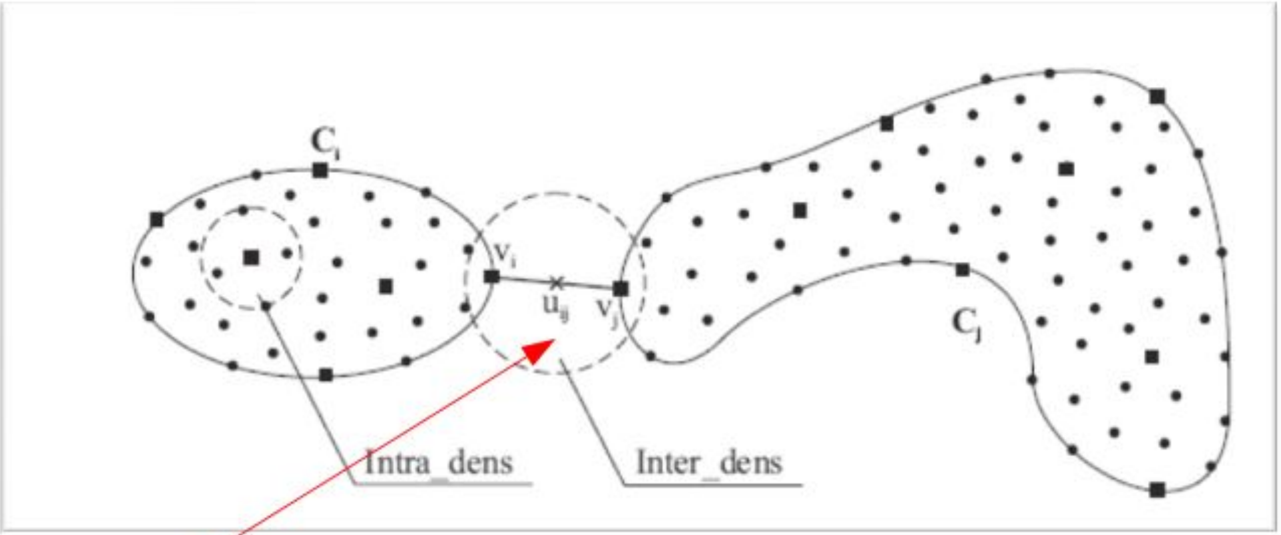
CDbw



$$CDbw(nc) = Sep(nc) \cdot Intra\_dens(nc), nc > 1$$

$$Sep(nc) = \frac{\sum_{i=1}^{nc} \sum_{j=1, i \neq j}^{nc} \min d(clos\_rep_i, clos\_rep_j)}{1 + Inter\_dens(nc)}, nc > 1$$

Separability in CDbw

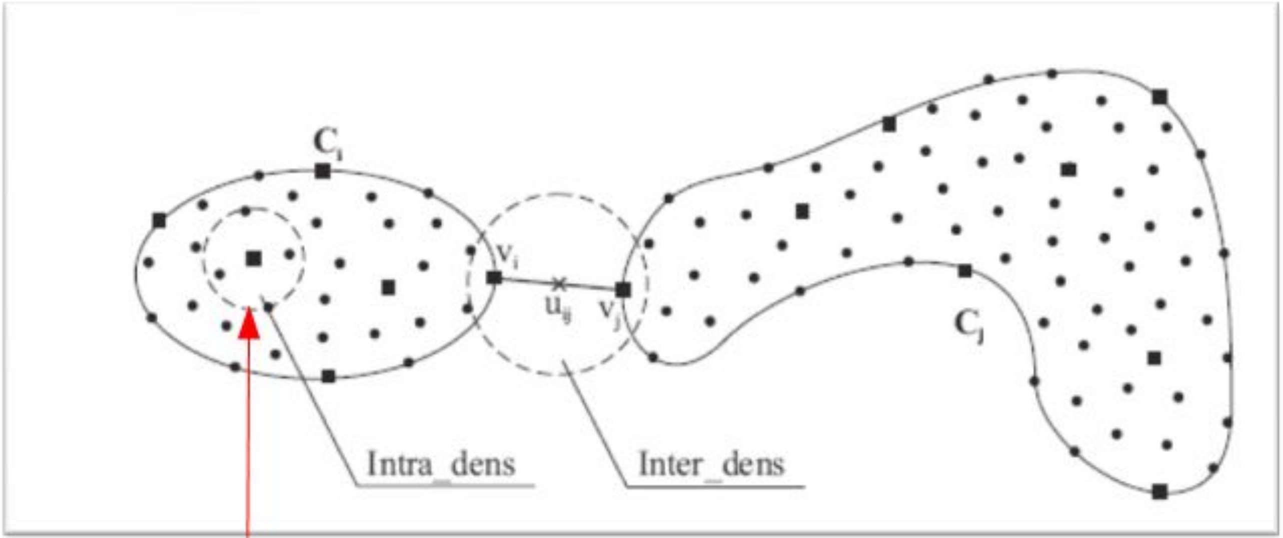


$$Inter\_dens(nc) = \sum_{i=1}^{nc} \sum_{j=1, j \neq i}^{nc} \left(\frac{d(clos\_rep_i, clos\_rep_j)}{stdev\,(C_i) + stdev\,(C_j)} \cdot density(u_{ij}) \right), nc > 1$$

$$density(u_{ij}) = \frac{\sum_{x \in C_i \cup C_j} f(x, u_{ij})}{|C_i| + |C_j|}$$

$$f(x, u_{ij}) = \begin{cases} 0 & \text{if } d(x, u_{ij}) > stdev\,(C_i) + stdev\,(C_j))/2 \\ 1 & \text{otherwise.} \end{cases}$$

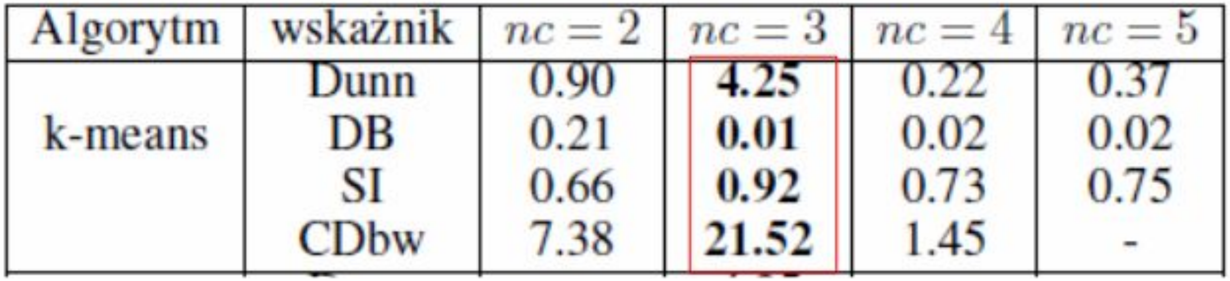
CDbw - similarity in the group



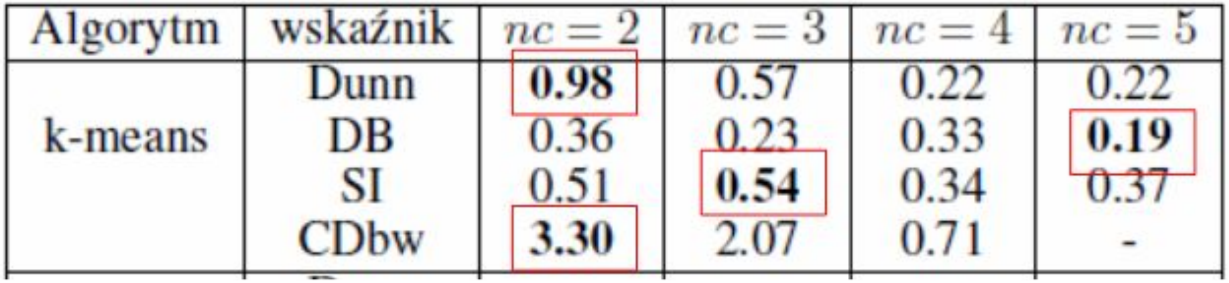
$$Intra\_dens(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\frac{1}{r}\sum_{v_{ij}\in C_i}\frac{density(v_{ij})}{stdev\,(C_i)}, nc > 1$$

$$density(v_{ij}) = \sum_{x\in C_i} g(x, v_{ij}) \qquad g(x, v_{ij}) = \begin{cases} 0 & \text{if } d(x, v_{ij}) > stdev\,(C_i) \\ 1 & \text{otherwise.} \end{cases}$$

Examples of optimal clustering evaluation

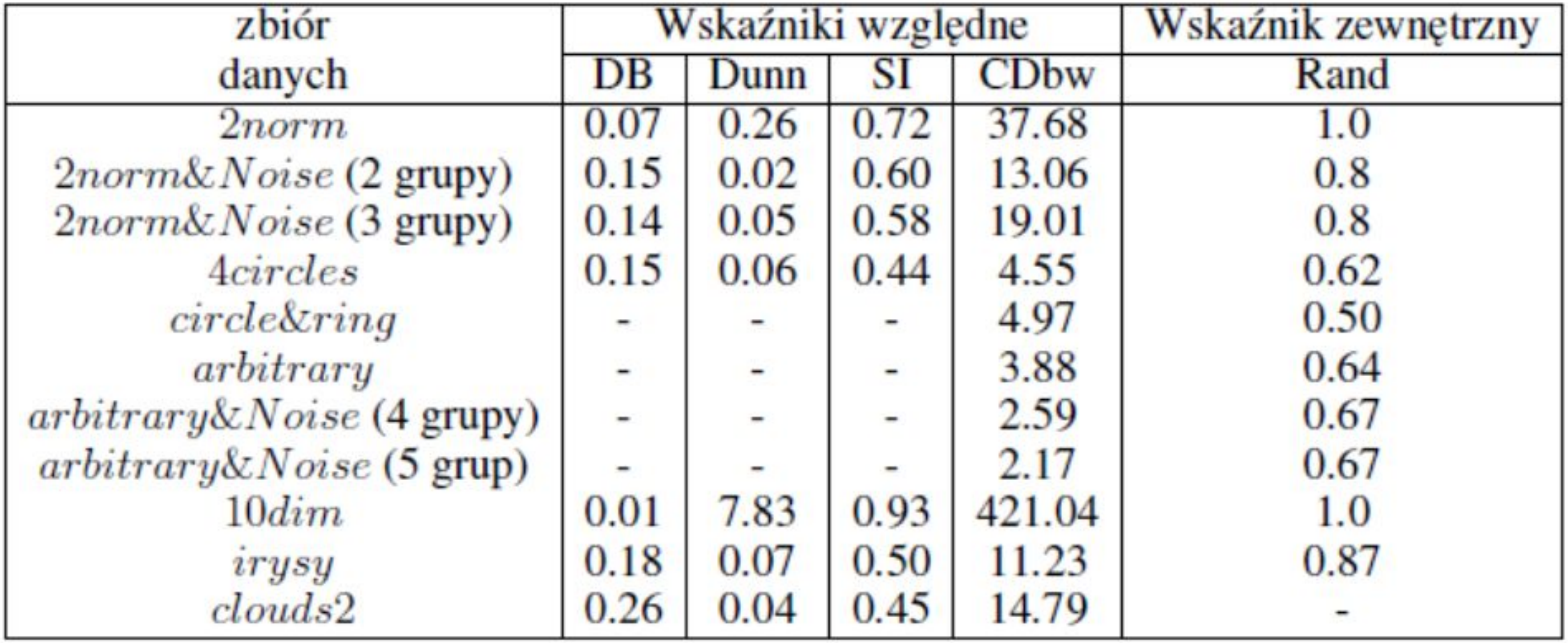


Algorytm	wskaźnik	$nc = 2$	$nc = 3$	$nc = 4$	$nc = 5$
k-means	Dunn	0.90	**4.25**	0.22	0.37
	DB	0.21	**0.01**	0.02	0.02
	SI	0.66	**0.92**	0.73	0.75
	CDbw	7.38	**21.52**	1.45	-



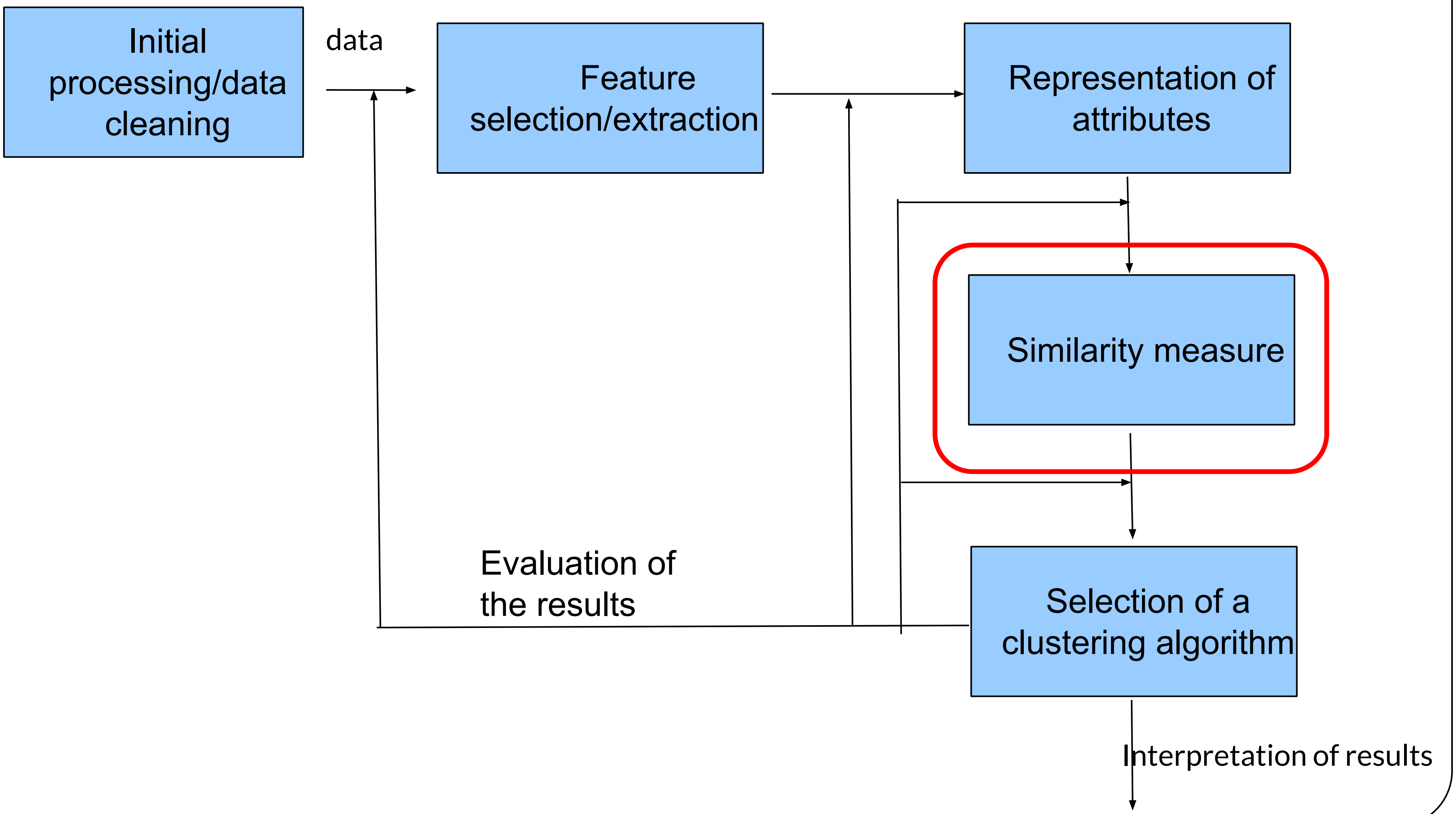
Algorytm	wskaźnik	$nc = 2$	$nc = 3$	$nc = 4$	$nc = 5$
k-means	Dunn	**0.98**	0.57	0.22	0.22
	DB	0.36	0.23	0.33	**0.19**
	SI	0.51	**0.54**	0.34	0.37
	CDbw	**3.30**	2.07	0.71	-

Examples of optimal clustering evaluation



zbiór danych	Wskaźniki względne				Wskaźnik zewnętrzny
	DB	Dunn	SI	CDbw	Rand
2norm	0.07	0.26	0.72	37.68	1.0
2norm&Noise (2 grupy)	0.15	0.02	0.60	13.06	0.8
2norm&Noise (3 grupy)	0.14	0.05	0.58	19.01	0.8
4circles	0.15	0.06	0.44	4.55	0.62
circle&ring	-	-	-	4.97	0.50
arbitrary	-	-	-	3.88	0.64
arbitrary&Noise (4 grupy)	-	-	-	2.59	0.67
arbitrary&Noise (5 grup)	-	-	-	2.17	0.67
10dim	0.01	7.83	0.93	421.04	1.0
irysy	0.18	0.07	0.50	11.23	0.87
clouds2	0.26	0.04	0.45	14.79	-

Steps of clustering process


Initial processing/data cleaning
data
Feature selection/extraction
Representation of attributes
Similarity measure
Selection of a clustering algorithm
Evaluation of the results
Interpretation of results

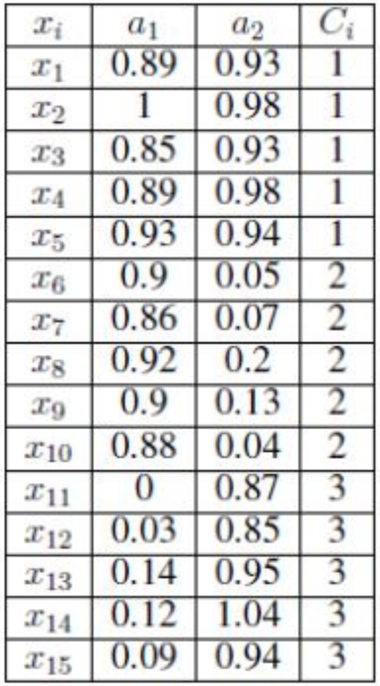
Similarity measures

- Real value attributes
- Nominal value attributes
- Binary attributes

Similarity measures

Real value attributes:

- Minkowski metrics
- Cosine distance
- Pearson correlation



x_i	a_1	a_2	C_i
x_1	0.89	0.93	1
x_2	1	0.98	1
x_3	0.85	0.93	1
x_4	0.89	0.98	1
x_5	0.93	0.94	1
x_6	0.9	0.05	2
x_7	0.86	0.07	2
x_8	0.92	0.2	2
x_9	0.9	0.13	2
x_{10}	0.88	0.04	2
x_{11}	0	0.87	3
x_{12}	0.03	0.85	3
x_{13}	0.14	0.95	3
x_{14}	0.12	1.04	3
x_{15}	0.09	0.94	3

$$d(x_1, x_2) = 0.12, \quad d(x_6, x_7) = 0.05, \quad d(x_{11}, x_{12}) = 0.03$$
$$d(x_1, x_6) = 0.89, \quad d(x_6, x_{11}) = 1.22, \quad d(x_{11}, x_1) = 0.89$$

Euclidean distance

$$d(x_1, x_2) = 0.99, \quad d(x_6, x_7) = 0.99, \quad d(x_{11}, x_{12}) = 0.99$$
$$d(x_1, x_6) = 0.73, \quad d(x_6, x_{11}) = 0.05, \quad d(x_{11}, x_1) = 0.72$$

cosine distance

Similarity measures

- Nominal attributes

$$s(x_i, x_j) = \frac{p}{k}$$

Example:
x_1: [W Yes Red]
x_2: [M Yes Blue]
x_3: [W No Red]

$s(x_1,x_2)=1/3$
$s(x_1,x_3)=2/3$
$s(x_2,x_3)=0$

Similarity measures

- Binary attributes

obiekt		x_i	
		0	1
x_j	0	a_{00}	a_{01}
	1	a_{10}	a_{11}

Simple proximity index (SPI)

$$s(x_i, x_j) = \frac{a_{00} + a_{11}}{a_{00} + a_{11} + a_{01} + a_{10}}$$

Jaccard measure

$$s(x_i, x_j) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}}$$

Example:

x_1: [1 1 0 0 1]

x_2: [1 1 1 1 1]

x_3: [0 1 0 1 0]

$s(x_1,x_2)=3/5$

$s(x_1,x_3)=2/5$

$s(x_2,x_3)=2/5$

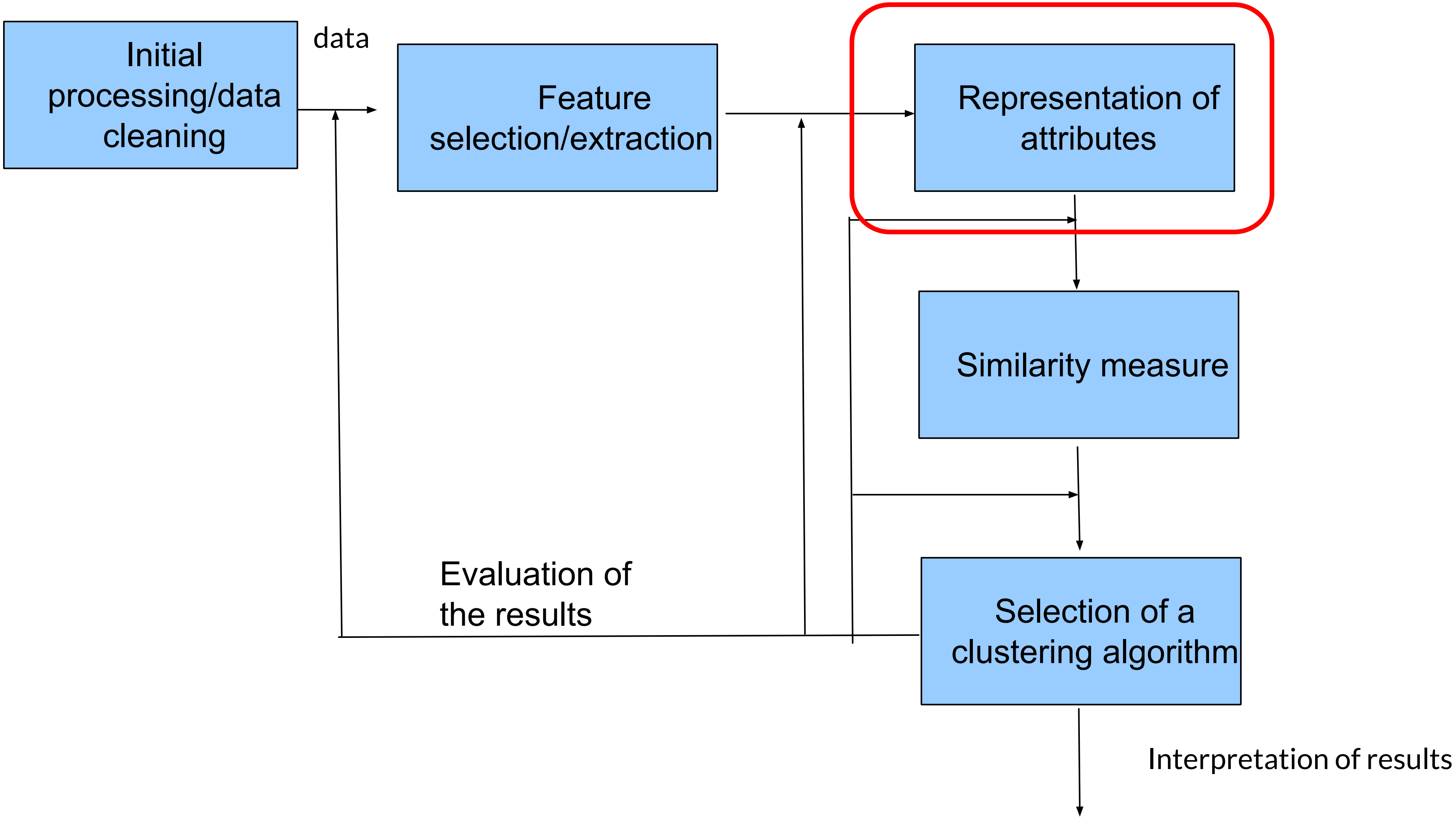
$s(x_1,x_2)=3/5$

$s(x_1,x_3)=1/4$

$s(x_2,x_3)=2/5$

- Balanced variables: SPI, Jaccard: unbalanced variables

Steps of clustering process



Initial processing/data cleaning

data

Feature selection/extraction

Representation of attributes

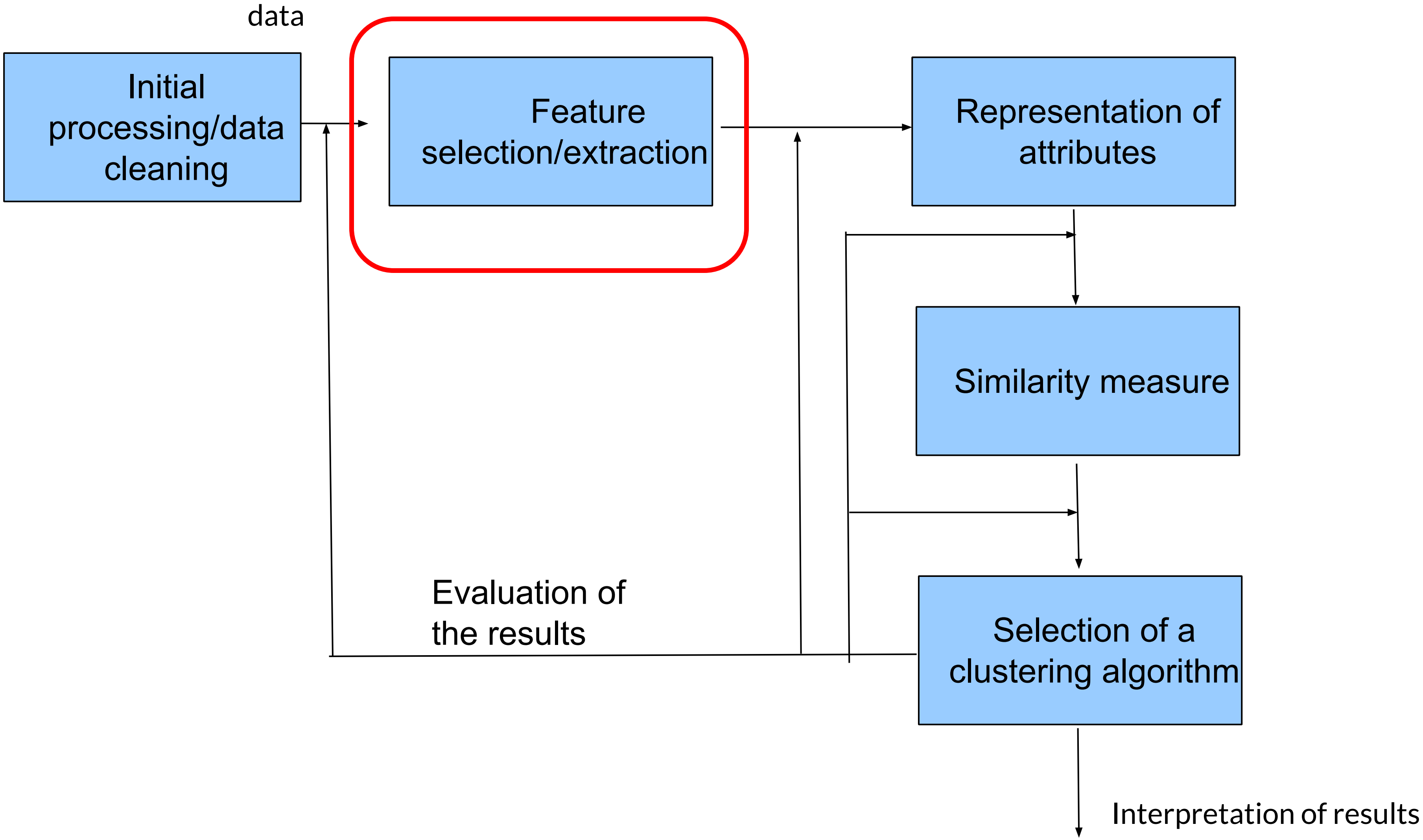
Similarity measure

Selection of a clustering algorithm

Evaluation of the results

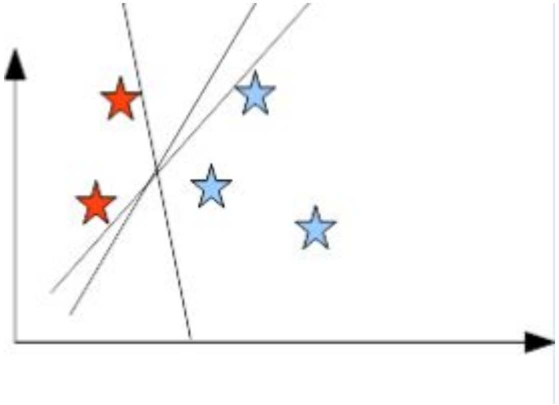
Interpretation of results

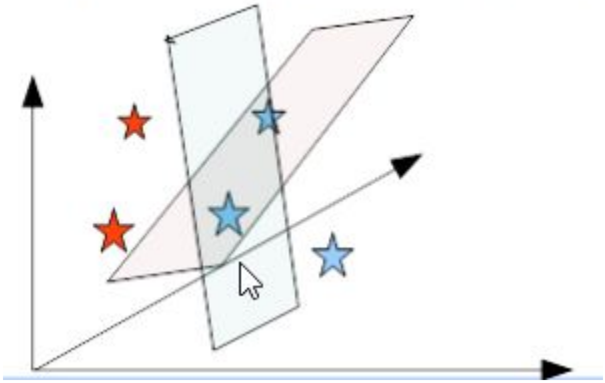
Steps of clustering process


data
Initial processing/data cleaning
Feature selection/extraction
Representation of attributes
Similarity measure
Selection of a clustering algorithm
Evaluation of the results
Interpretation of results

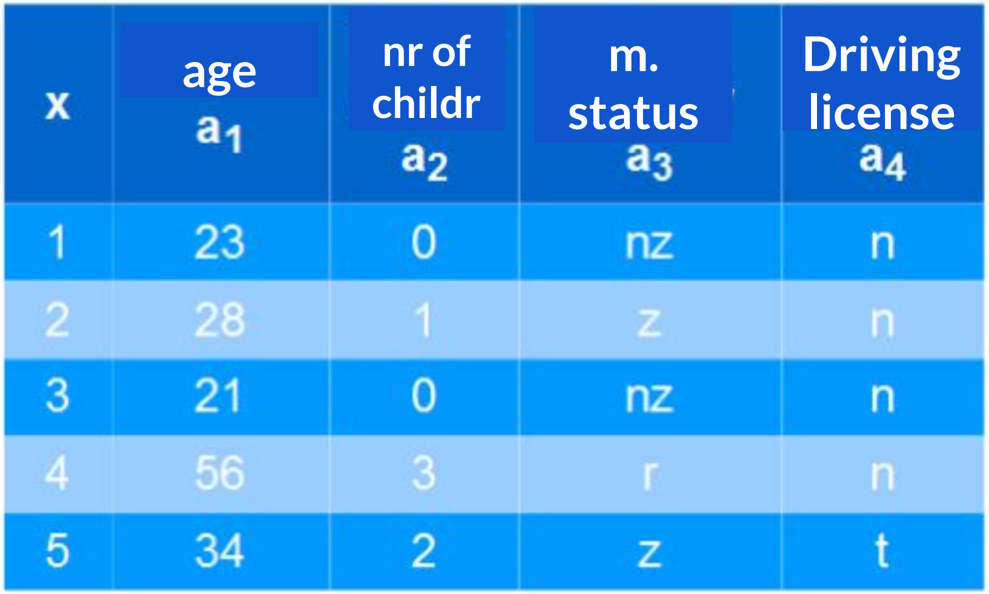
Selection of attributes

- The purpose:
  - greater separability of groups
  - easier identification of groups
- The techniques:
  - correlation measure
  - PCA

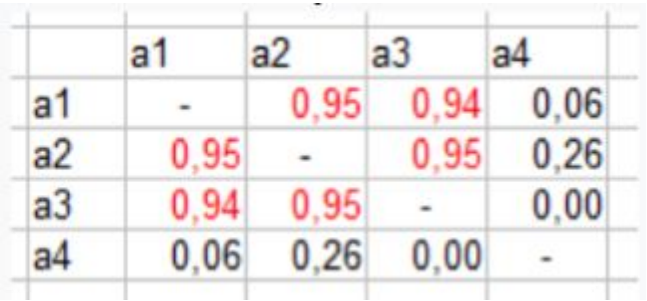




Correlation of attributes

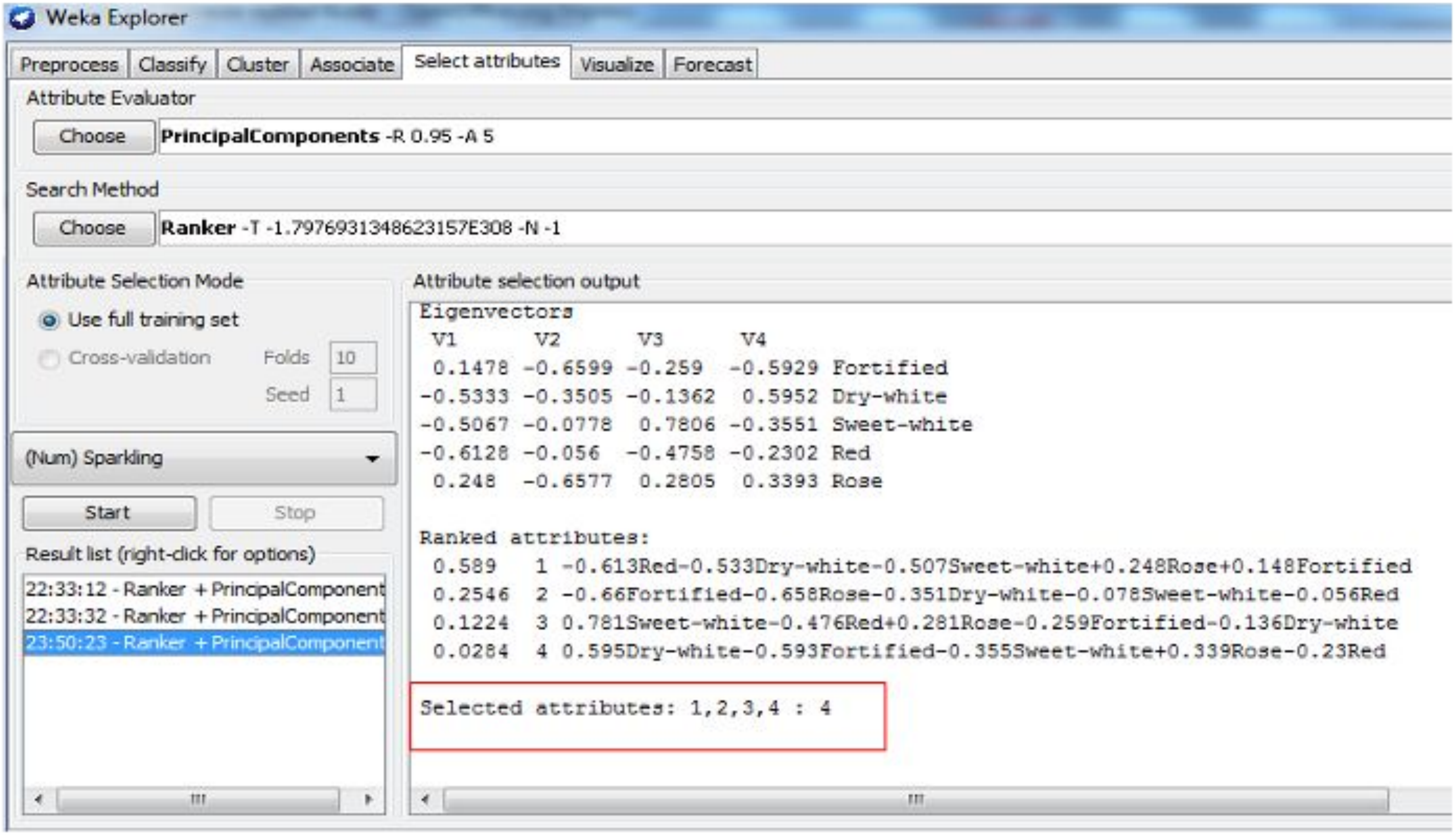


x	age a_1	nr of childr a_2	m. status a_3	Driving license a_4
1	23	0	nz	n
2	28	1	z	n
3	21	0	nz	n
4	56	3	r	n
5	34	2	z	t



	a1	a2	a3	a4
a1	-	0,95	0,94	0,06
a2	0,95	-	0,95	0,26
a3	0,94	0,95	-	0,00
a4	0,06	0,26	0,00	-

Selection based on PCA



Attribute Evaluator

PrincipalComponents -R 0.95 -A 5

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Num) Sparkling

Result list (right-click for options)

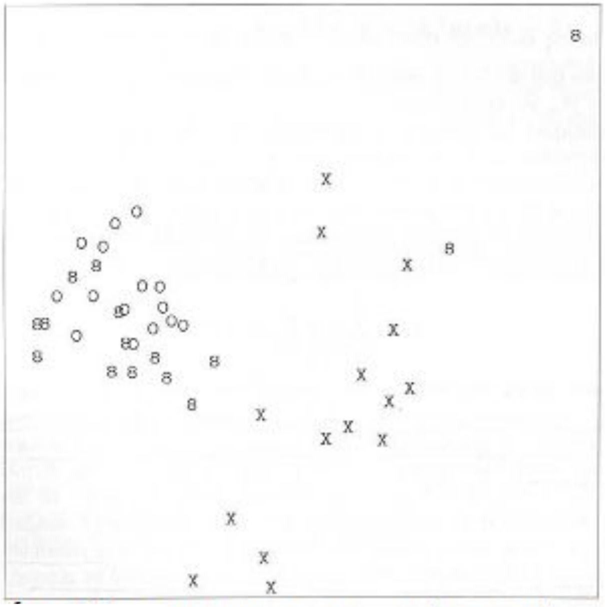
```
22:33:12 - Ranker + PrincipalComponent
22:33:32 - Ranker + PrincipalComponent
23:50:23 - Ranker + PrincipalComponent
```

Attribute selection output

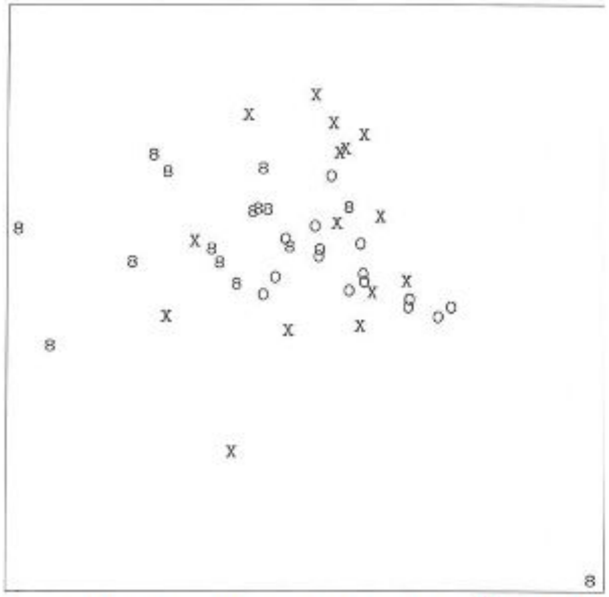
```
Eigenvectors
 V1       V2       V3       V4
 0.1478 -0.6599 -0.259   -0.5929 Fortified
-0.5333 -0.3505 -0.1362   0.5952 Dry-white
-0.5067 -0.0778  0.7806 -0.3551 Sweet-white
-0.6128 -0.056  -0.4758 -0.2302 Red
 0.248  -0.6577  0.2805  0.3393 Rose

Ranked attributes:
 0.589   1 -0.613Red-0.533Dry-white-0.507Sweet-white+0.248Rose+0.148Fortified
 0.2546  2 -0.66Fortified-0.658Rose-0.351Dry-white-0.078Sweet-white-0.056Red
 0.1224  3 0.781Sweet-white-0.476Red+0.281Rose-0.259Fortified-0.136Dry-white
 0.0284  4 0.595Dry-white-0.593Fortified-0.355Sweet-white+0.339Rose-0.23Red

Selected attributes: 1,2,3,4 : 4
```

Selection based on PCA

Principal components: 1,2

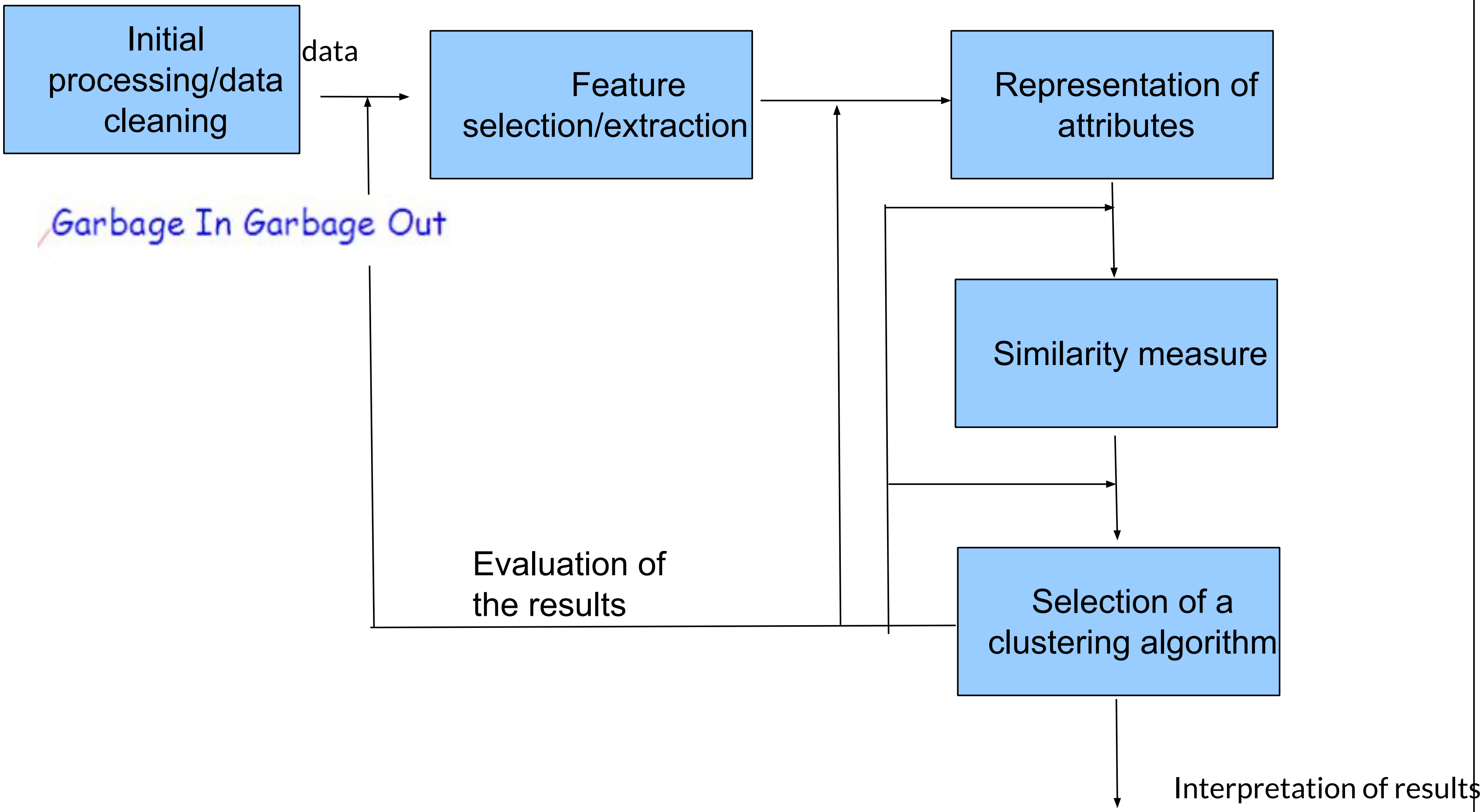


Principal components: 3,4



A.K. Jain: Algorithms for Clustering Data, Prentice Halll, 1988

Steps of clustering process



Initial processing/data cleaning

data

Feature selection/extraction

Representation of attributes

Similarity measure

Selection of a clustering algorithm

Garbage In Garbage Out

Evaluation of the results

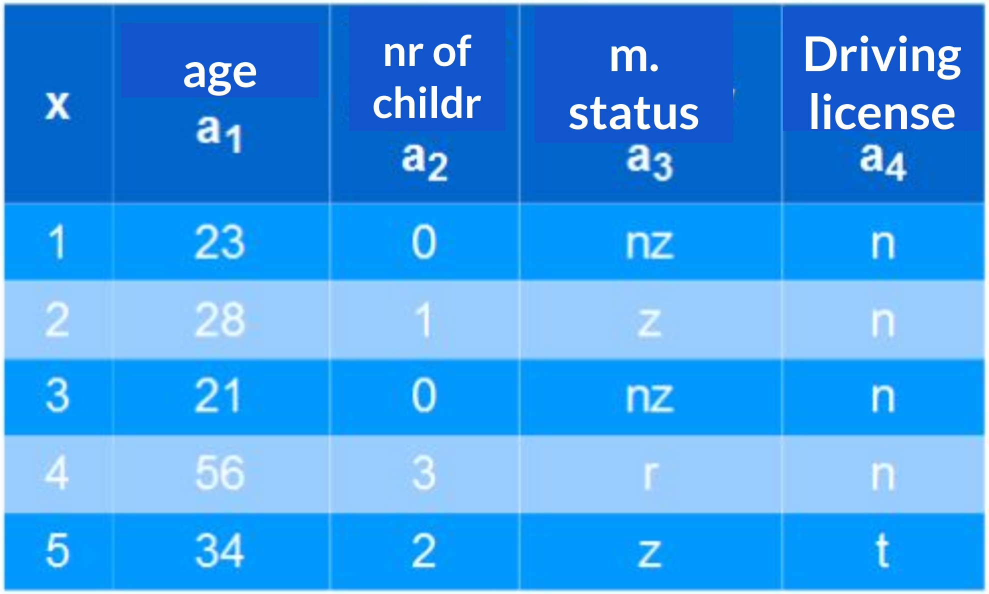
Interpretation of results

Initial processing / data cleaning

- Missing values
- Standardization
- normalization

$y=-0.133x+41.09$

[56 M 66 ? N] 1-nn:
$d(x_5,x_6)=\min$



x	age a_1	nr of childr a_2	m. status a_3	Driving license a_4
1	23	0	nz	n
2	28	1	z	n
3	21	0	nz	n
4	56	3	r	n
5	34	2	z	t

$$X^{*}=\frac{X-min(X)}{max(X)-min(X)}$$

$d(x_1,x_2)=\sqrt{(22-46)^2+(0-1)^2}=\sqrt{577}$

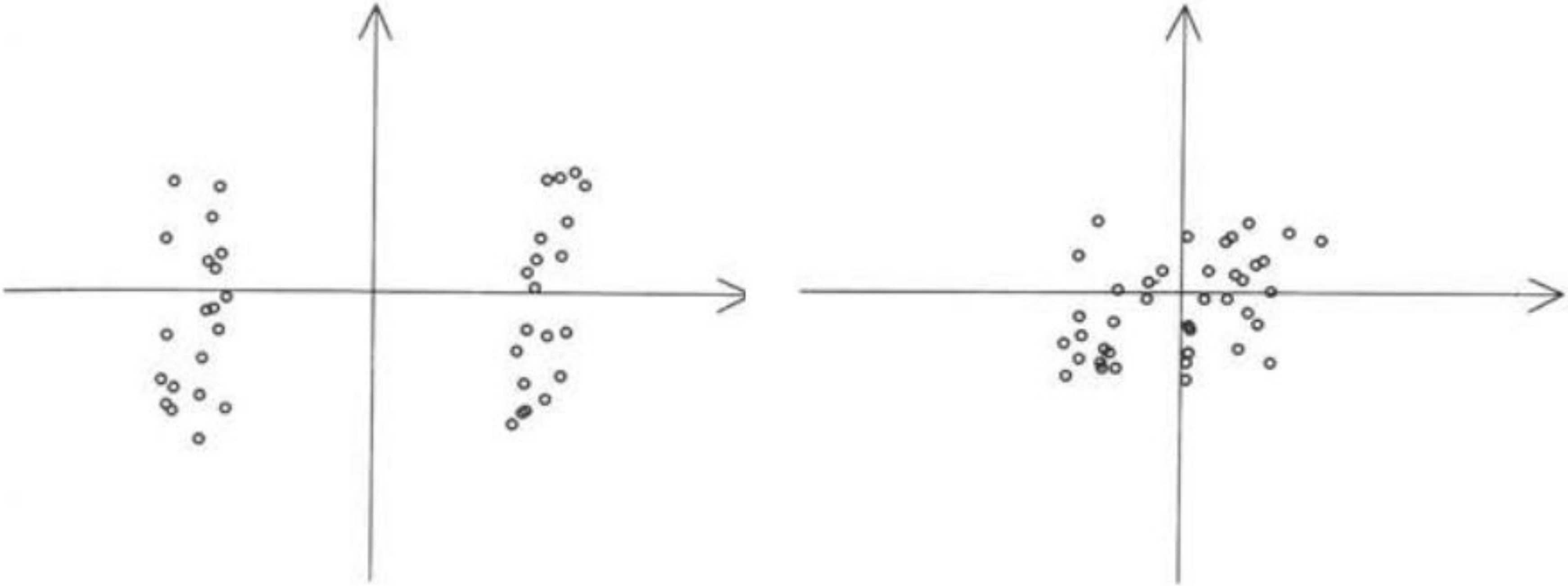
$d(x_1,x_3)=\sqrt{(22-41)^2+(0-2)^2}=\sqrt{365}$

→

$d(x_1,x_2)=\sqrt{(0-0.65)^2+(0-0.5)^2}=\sqrt{0.67}$

$d(x_1,x_3)=\sqrt{(0-0.51)^2+(0-1)^2}=\sqrt{1.26}$

Normalization vs separability



before

after

A.K. Jain: Algorithms for Clustering Data, Prentice Halll, 1988

What do you remember?

Algorithm	Category	Cluster shape	Time complexity	Input parameter

Examples of application

- Coffie marketing http://www.focus-balkans.org/res/files/upload/file/9%20Cluster_Analysis%20Schaer.pdf
- Carrot http://search.carrot2.org/stable/search
- Web resources optimisation

Thank you for your attention!

Urszula Kużelewska
email: u.kuzelewska@pb.edu.pl