MEDNARODNA
PODIPLOMSKA ŠOLA
JOŽEFA STEFANA

INFORMATION AND COMMUNICATION TECHNOLOGIES
PhD study programme

Data Mining and Knowledge Discovery

Petra Kralj Novak

January 21, 2020

http://kt.ijs.si/petra_kralj/dmkd3.html

So far …

- Nov. 11, 2019
    - Basic classification
    - Orange hands on data visualization and classification
- Dec. 11, 2019
    - Fitting and overfitting
    - Data leakage
    - Decision boundary
    - Evaluation methods
    - Classification evaluation metrics: confusion matrix, TP, FP, TN, FN, accuracy, precision, recall, F1, ROC
    - Imbalanced data and unequal misclassification costs
    - Probabilistic classification
    - Naïve Bayes classifier

So far …

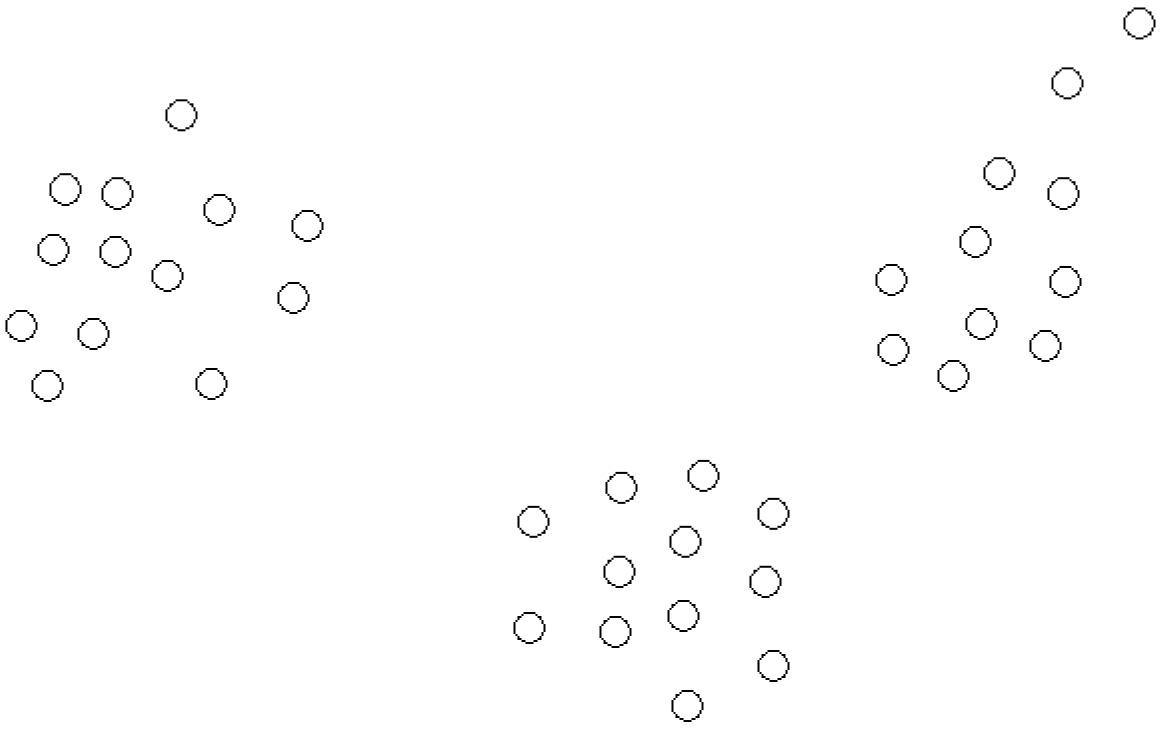
- Dec. 18 2019
  - Naive Bayes classifier
  - Laplace estimate
  - Regression (numeric prediction) and its evaluation
- Jan. 13, 2020
  - Association rules
- Jan. 15, 2020
  - Neural networks

Data mining techniques

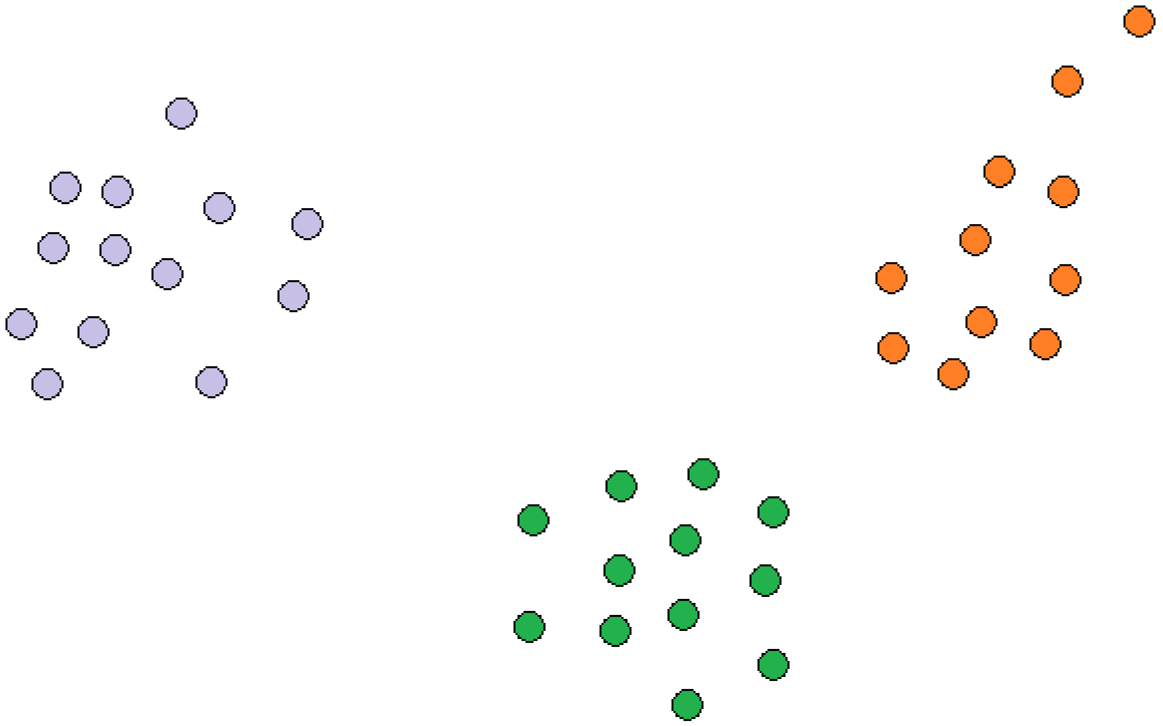
- Predictive induction
  - Classification
    - Decision trees
    - Classification rules
    - Naive Bayes classifier
    - KNN
    - SVM
    - ANN
    - …
  - Numeric prediction
    - Linear regression
    - Regression / model trees
    - KNN
    - SVM
    - ANN
    - …
- Descriptive induction
  - Association rules
    - Apriori
    - FP-growth
    - …
  - Clustering
    - Hierarchical
    - K-means
    - Dbscan
    - …

Clustering

Clustering



Clustering

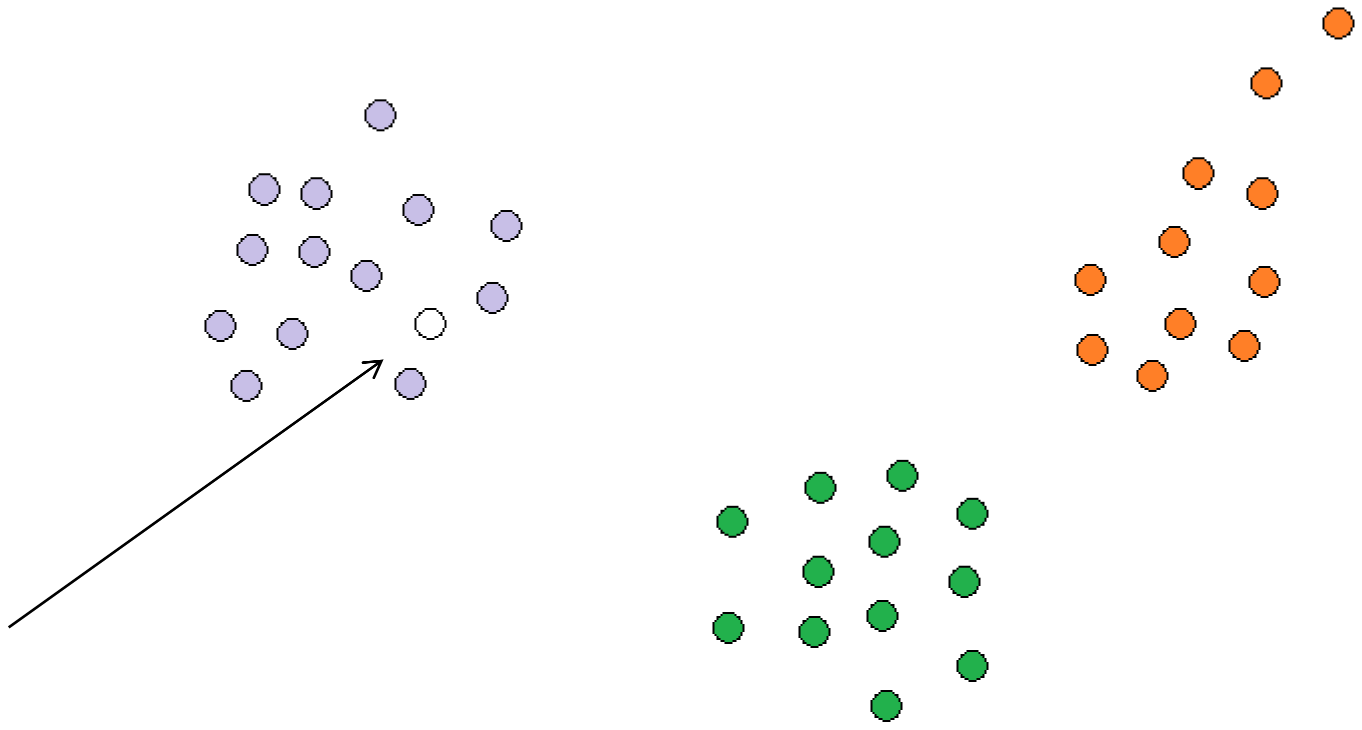


Clustering

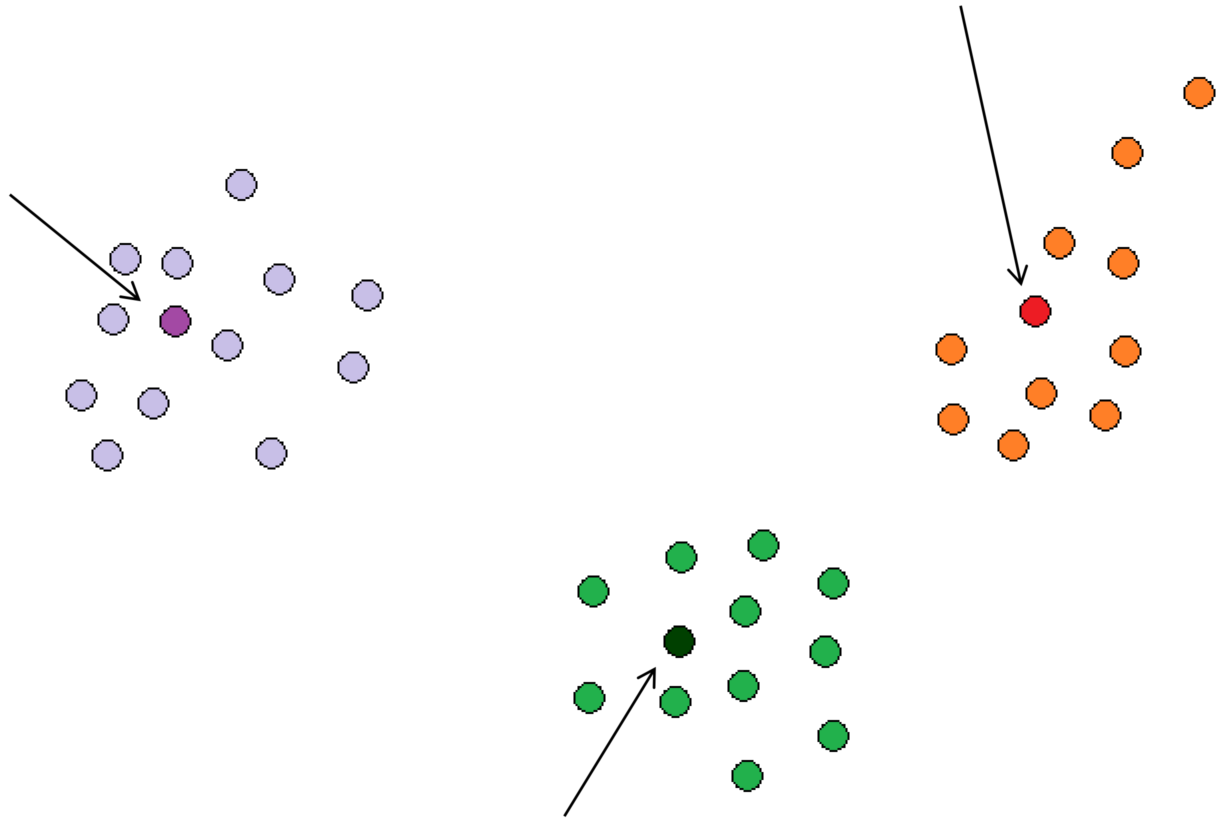
- … is the process of grouping the data instances into clusters so that objects within a cluster have high similarity but are very dissimilar to objects in other clusters.

- Wish list:
  - Identity clusters irrespective of their shapes
  - Scalability
  - Ability to deal with noisy data
  - Insensitivity to the order of input records

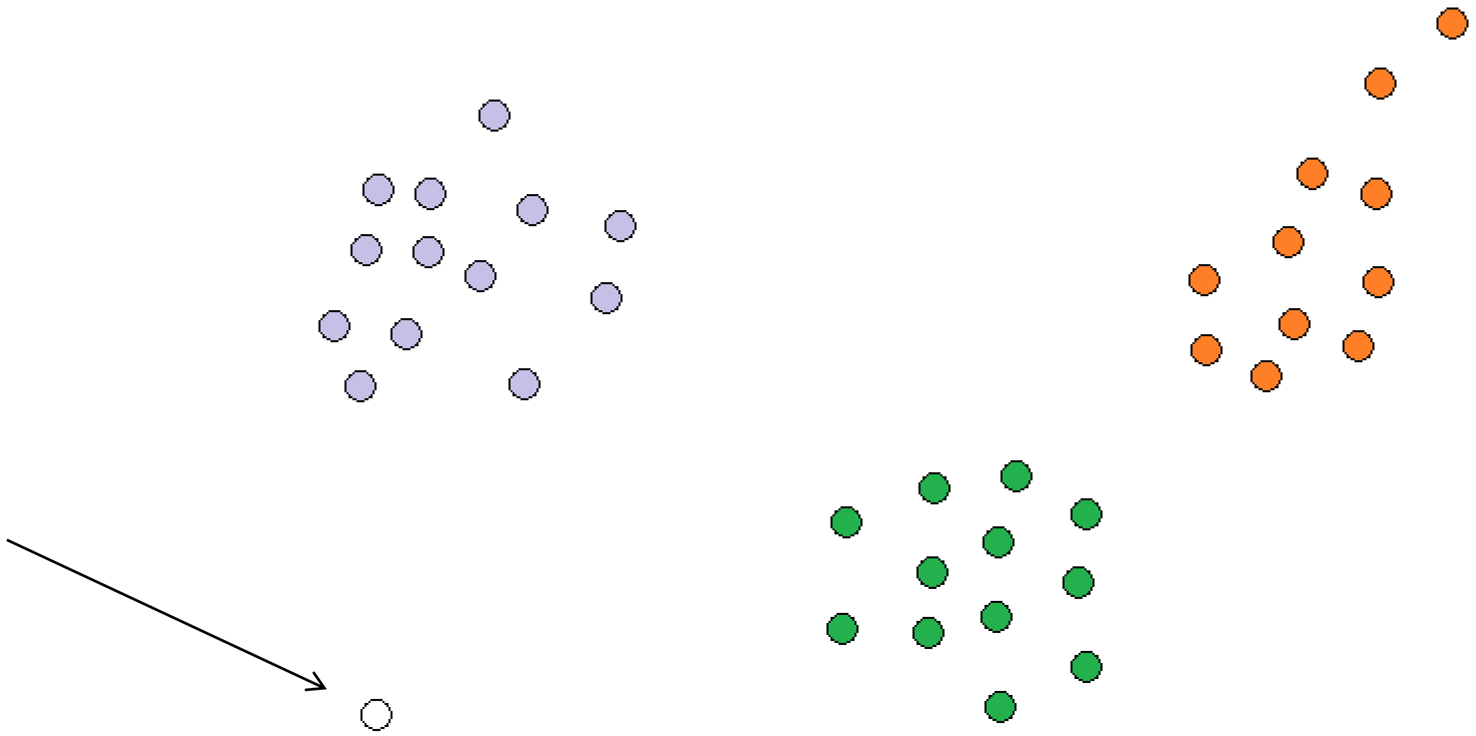
Unsupervised classification



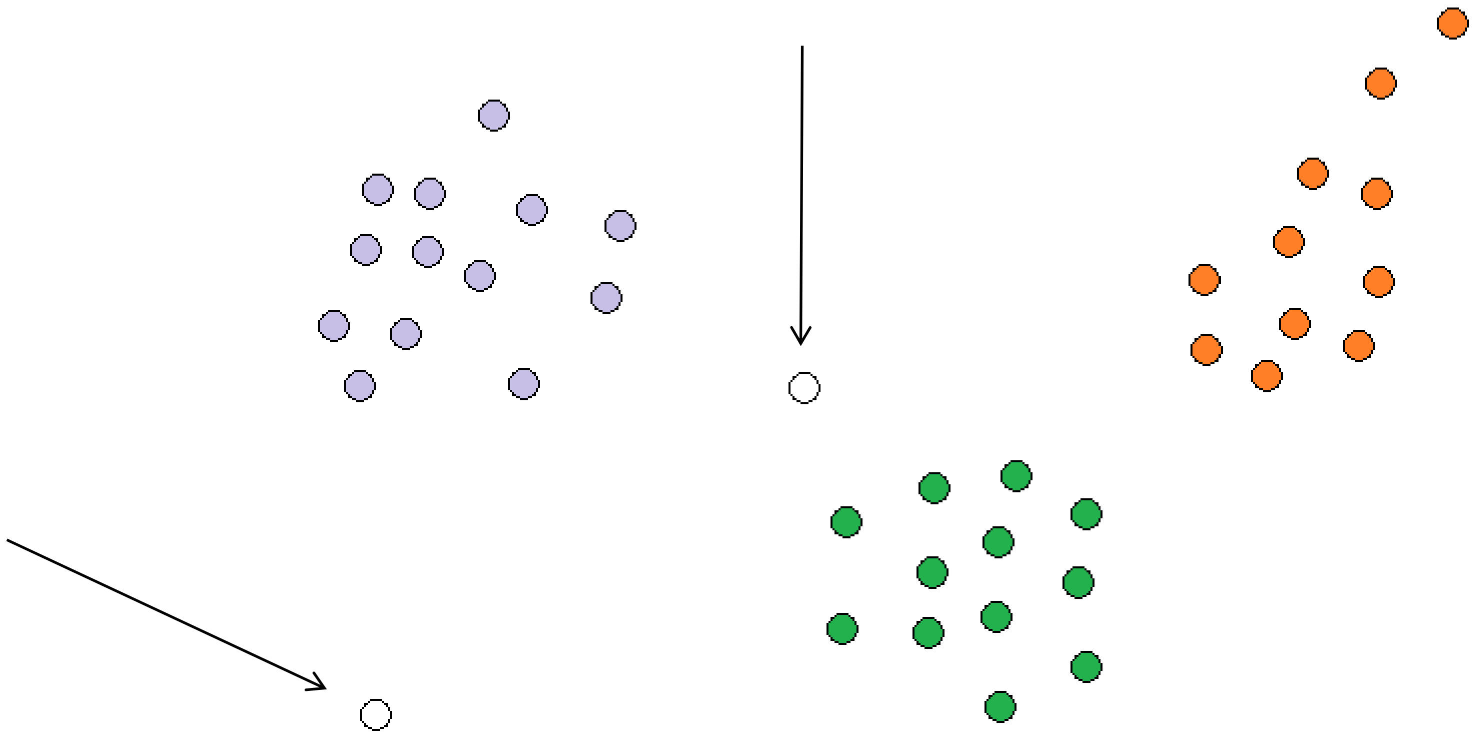
Data summarization: centroid, medoid



Outlier detection



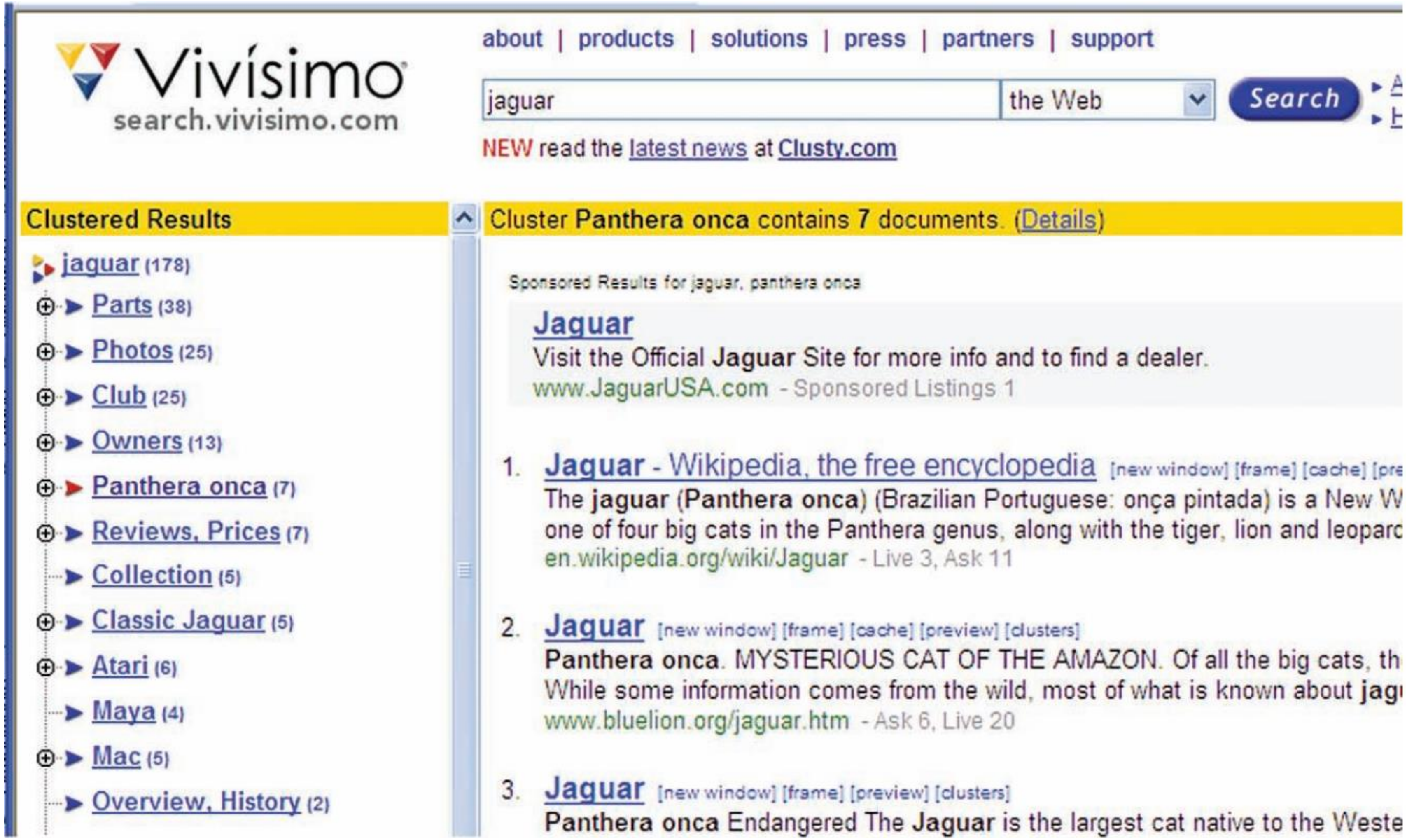
Outlier detection



Applications

- Data mining
  - Unsupervised classification
  - Data summarization
  - Outlier analysis
  - …
- Customer segmentation and collaborative filtering
- Text applications
- Social network analysis

Clustering web search results



Clustering types

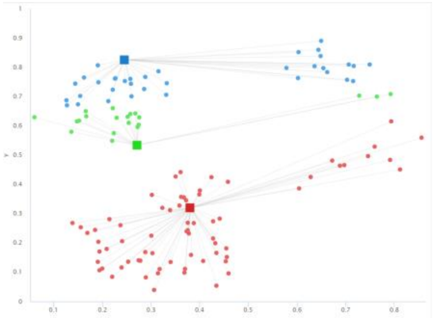
- **Partitioning**
  - k-means, k-medoids, k-modes
- **Hierarchical**
  - Agglomerative
- Grid-based
  - Multi-resolution grid structure
  - Efficient and scalable
- Density-based
  - A cluster is a dense region of points, which is separated by low density regions, from other regions of high density
  - Algorithms: DBSCAN, OPTICS, DenClue

Interactive k-Means (Educational)

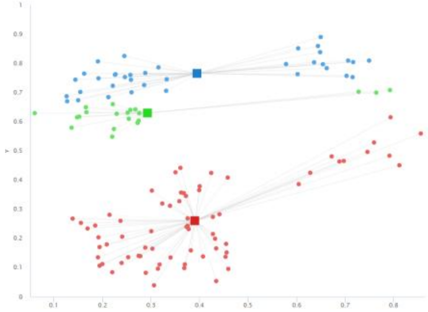
Interactive k-Means

K-Means example

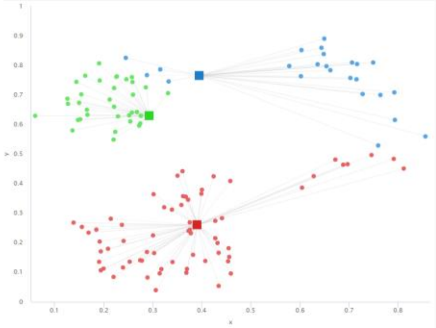
Random initialization



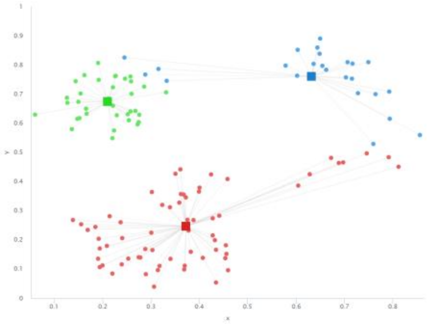
Centroid computation



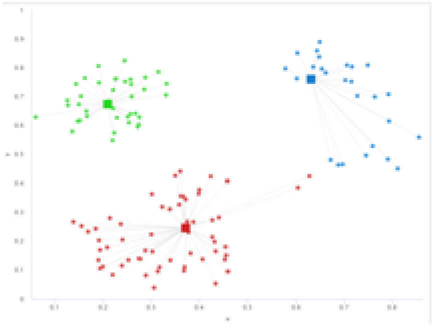
Assignment of points to the nearest centroid



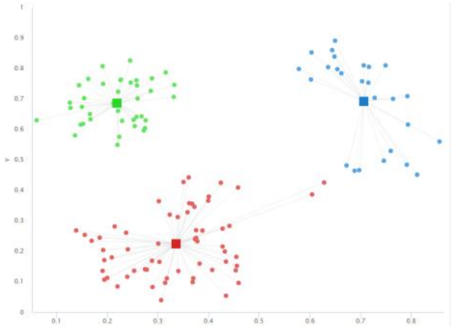
Centroid computation



Assignment of points to the nearest centroid



Centroid computation



K-means

1. Choose **k** random instances as cluster centers
2. Assign each instance to its closest cluster center
3. Re-compute cluster centers by computing the average (aka *centroid*) of the instances pertaining to each cluster
4. If cluster centers have moved, go back to Step 2

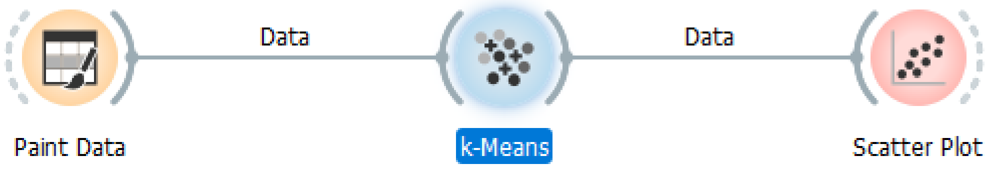
(Equivalent termination criterion: stop when assignment of instances to cluster centers has not changed)

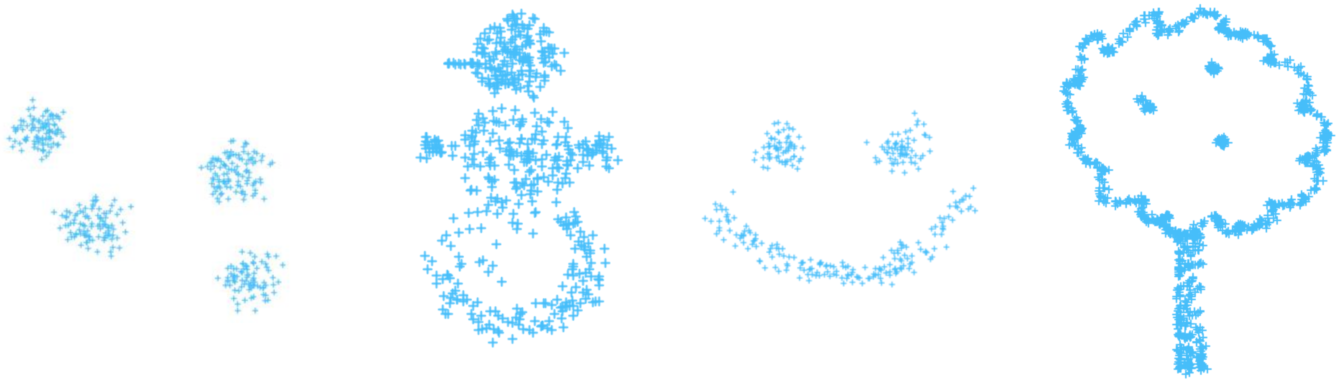
Alternatives: K-medoids, K-modes

- Might get stuck in local minima
- Silhuette for finding the optimal K

Lab exercise: clustering on drawings

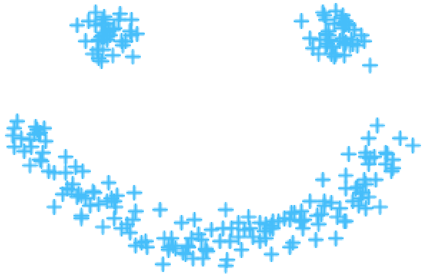
- Draw the following images in PaintData
  - Four snowballs
  - A snowman
  - A smiley face
  - An apple tree
- Compare
  - K-means
  - Hierarchical
  - DB-scan

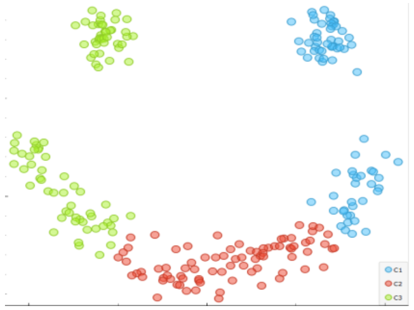


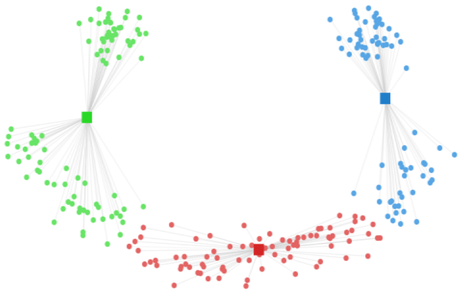


Properties of k-Means

- The number of clusters ***k*** is fixed in advance
- It is fast, it always converges
- Can converge into a local minima (bad solution because of unlucky start)
- Finds “spherical” shaped clusters
- K-Means will cluster the data even if it can’t be clustered (e.g. data that comes from uniform distributions)







Clustering evaluation

- Clustering analysis doesn't have a solid evaluation metric

- External validation criteria
  - Using the ground truth to evaluate to evaluate the clustering result
- Internal validation criteria
  - Sum of distances to centroids
  - Intracluster to intercluster distance ratio
  - Silhouette coefficient

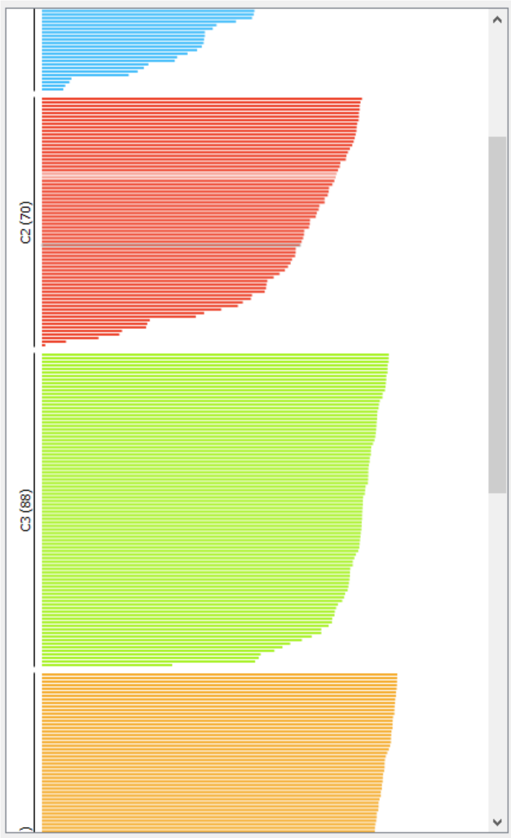
  - Parameter tuning – the "elbow" method

Aggarwal, Charu C. *Data mining: the textbook*. Springer, 2015. Chapter 6: cluster analysis, pgs 195 -201

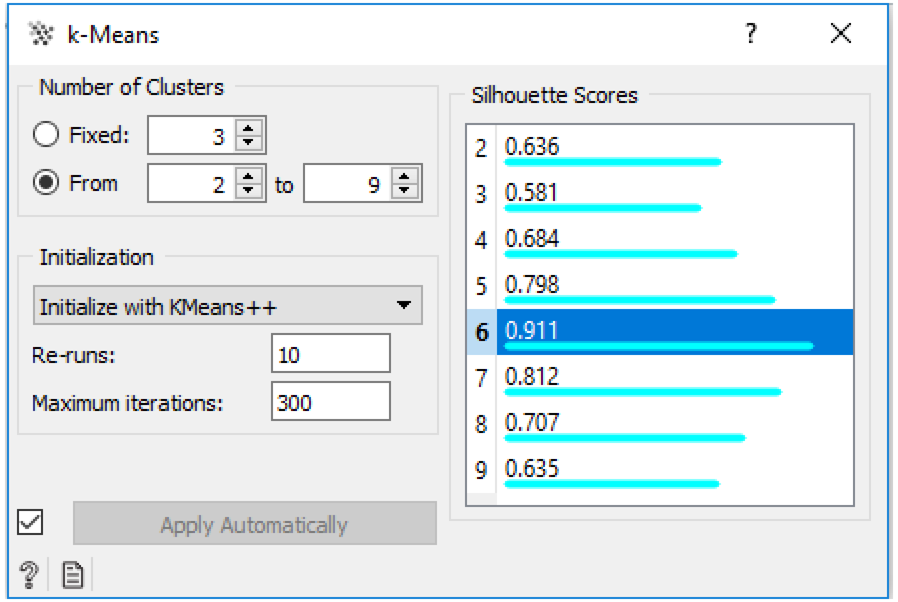
Silhouette coefficient

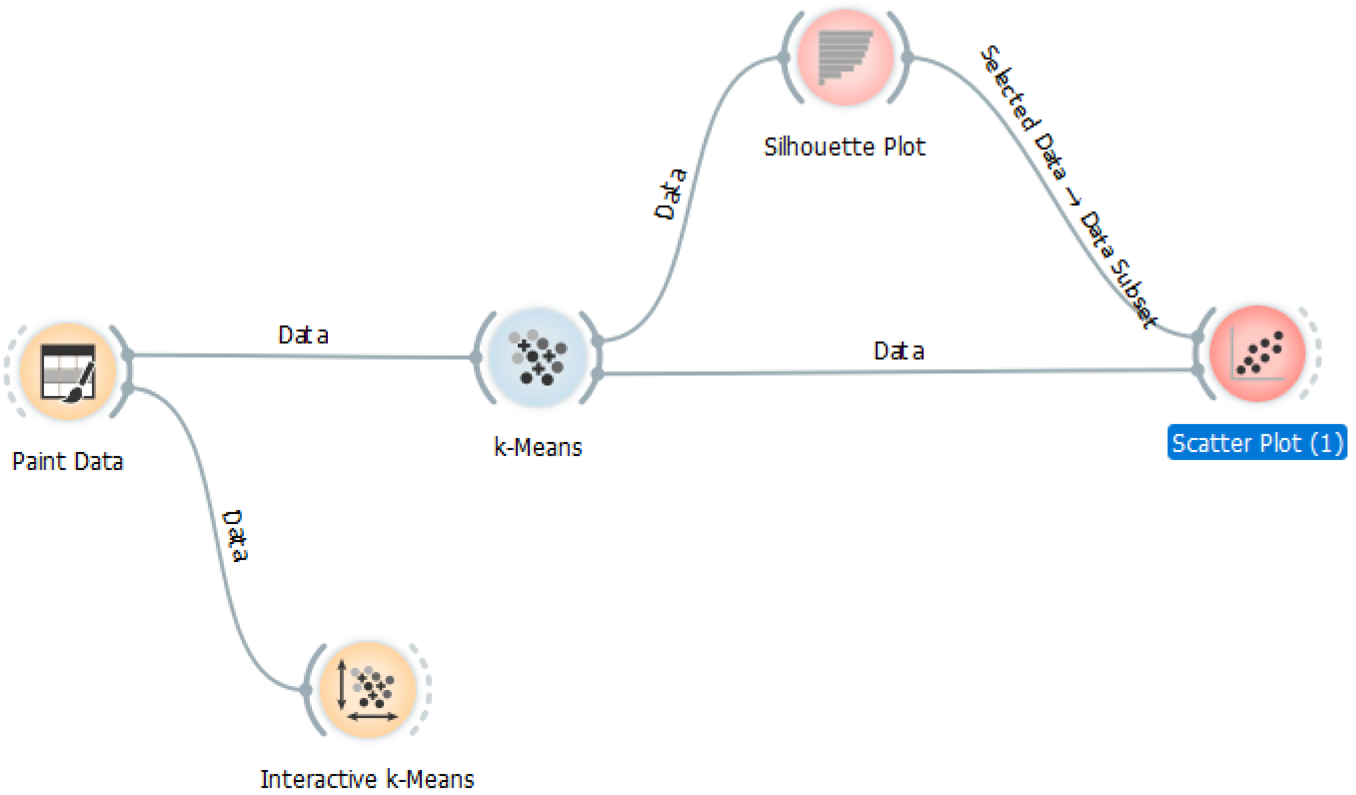
- The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

- For example x_i, its silhouette coefficient is $s_i = (b_i - a_i) / \max(a_i, b_i)$
  - a_i average distance between x_i to all other examples in its cluster.
  - b_i average distance between x_i to the examples in the “neighboring” cluster
- The overall silhouette coefficient is the average of the data point-specific coefficients.



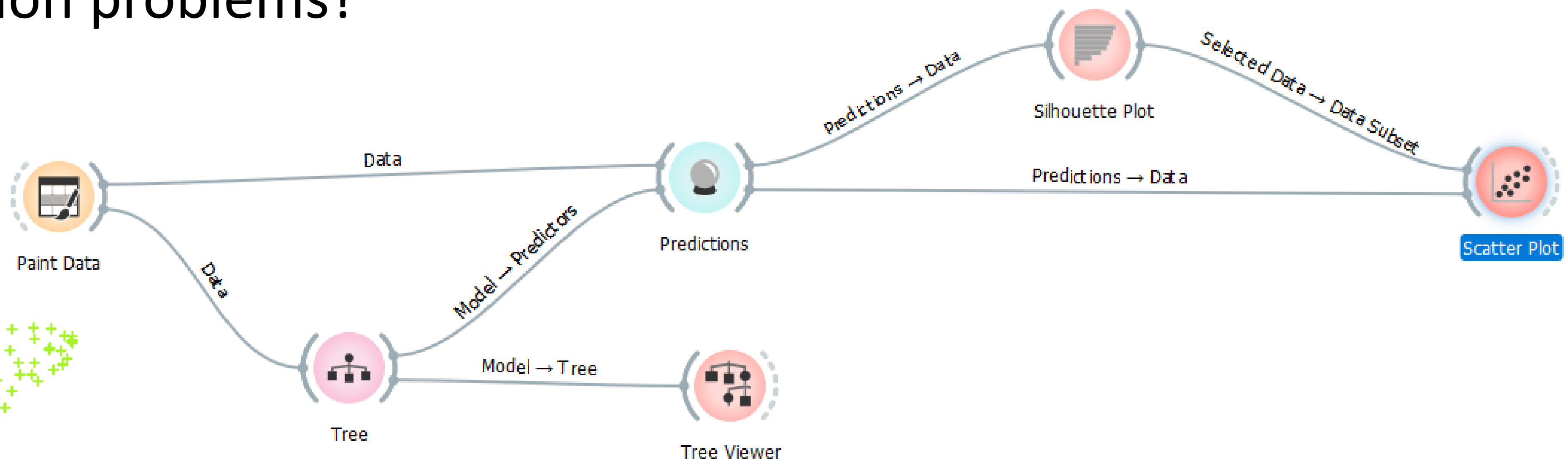
k-Means + Silhouette + „reruns“

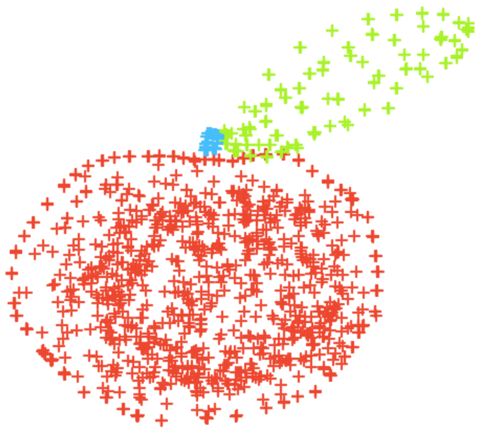




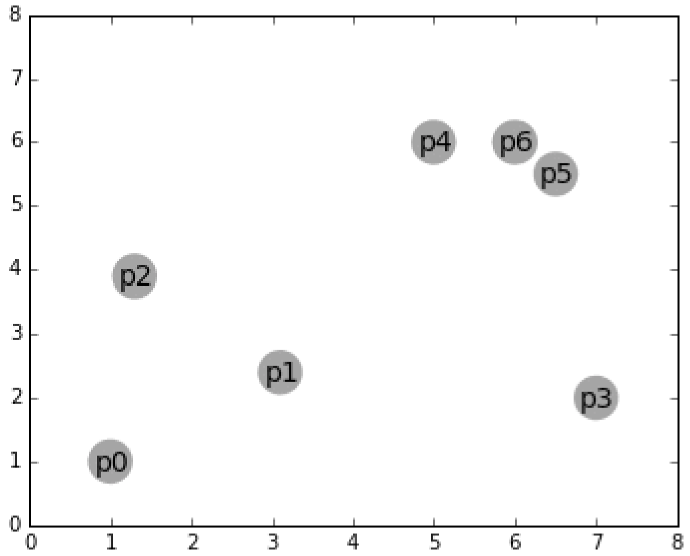
Orange workflow

- How can we use the silhouette coefficient for searching for outliers in classification problems?

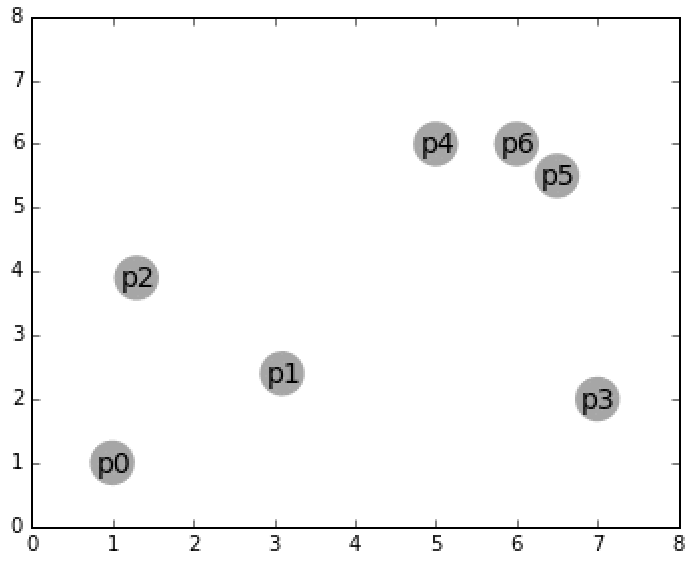


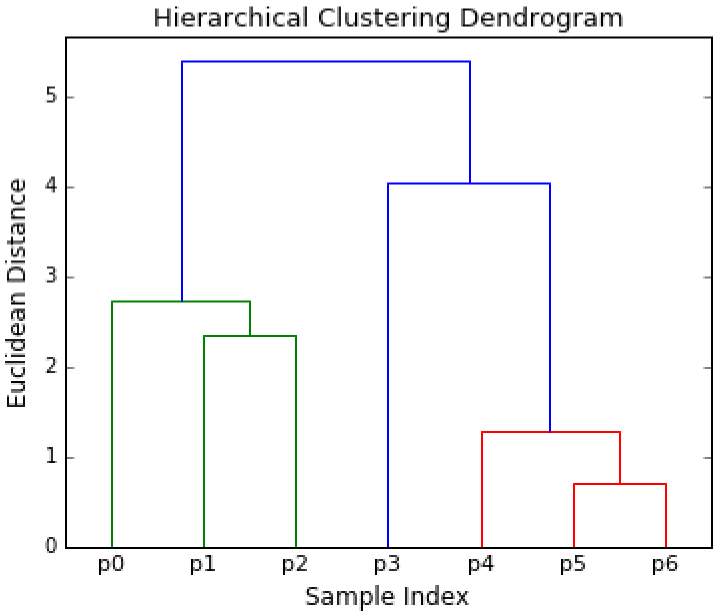


Agglomerative clustering - example



Agglomerative clustering - dendrogram





Agglomerative clustering

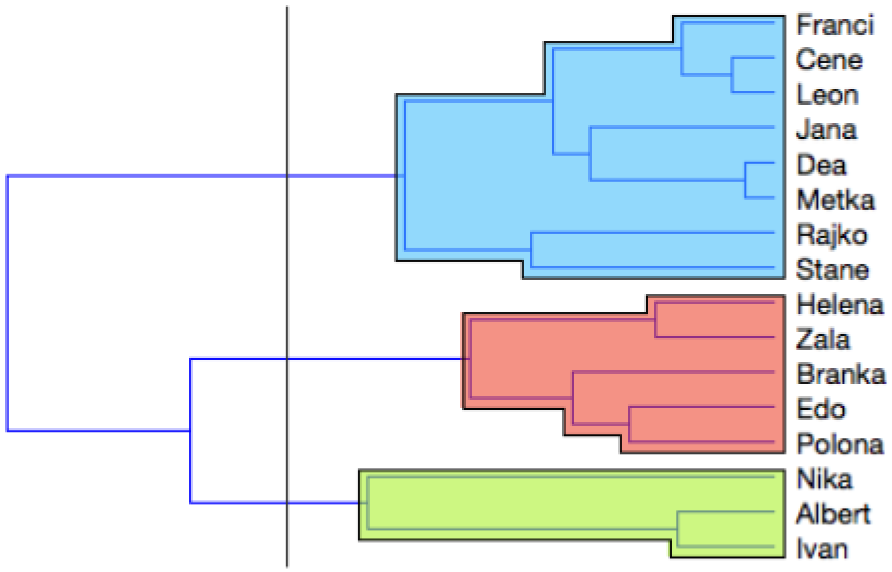
1. Start with a collection **C** of **n** singleton clusters
   - Each cluster contains one data point $\mathbf{c_i} = \{\mathbf{x_i}\}$
2. Repeat until only one cluster is left:
   1. Find a pair of clusters that is closest: min $\mathbf{D(c_i, c_j)}$
   2. Merge the clusters $\mathbf{c_i}$ and $\mathbf{c_j}$ into $\mathbf{c_{i+j}}$
   3. Remove $\mathbf{c_i}$ and $\mathbf{c_j}$ from the collection **C**, add $\mathbf{c_{i+j}}$

Some new index, not a sum

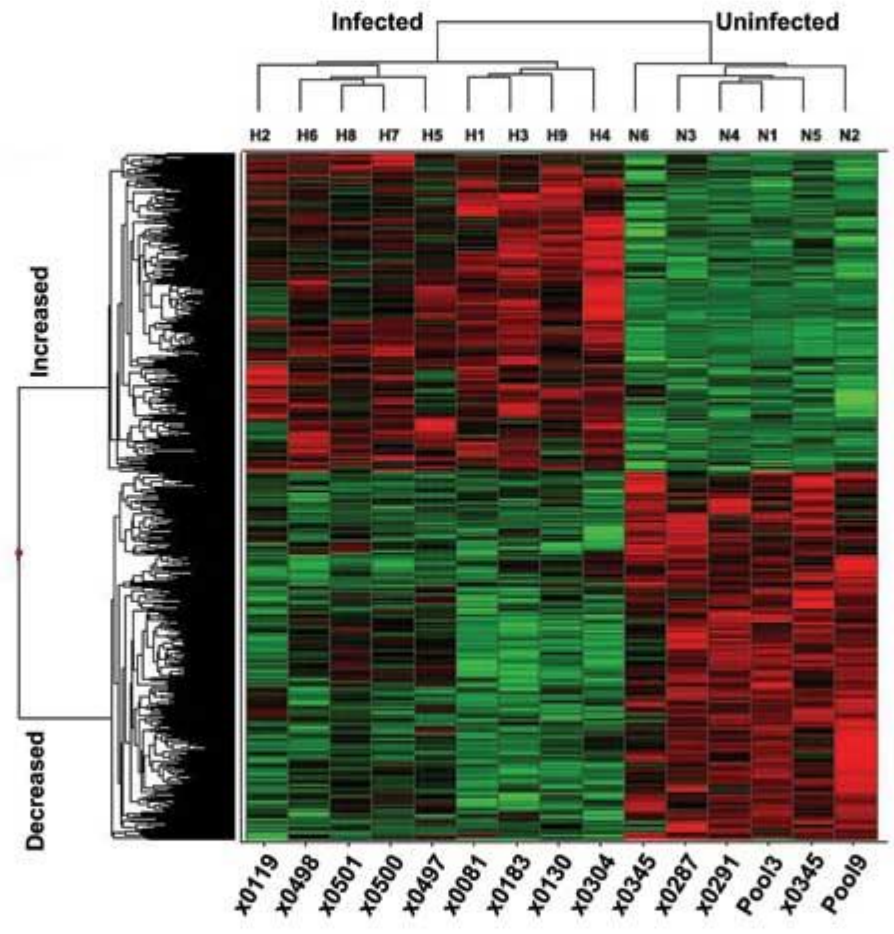
- Time and space complexity
- Sensitive to noisy data

Dendrogram

- The agglomerative hierarchical clustering algorithms result is commonly displayed as a tree diagram called a dendrogram.
- Dendrogram a tree diagram for showing taxonomic relationships.

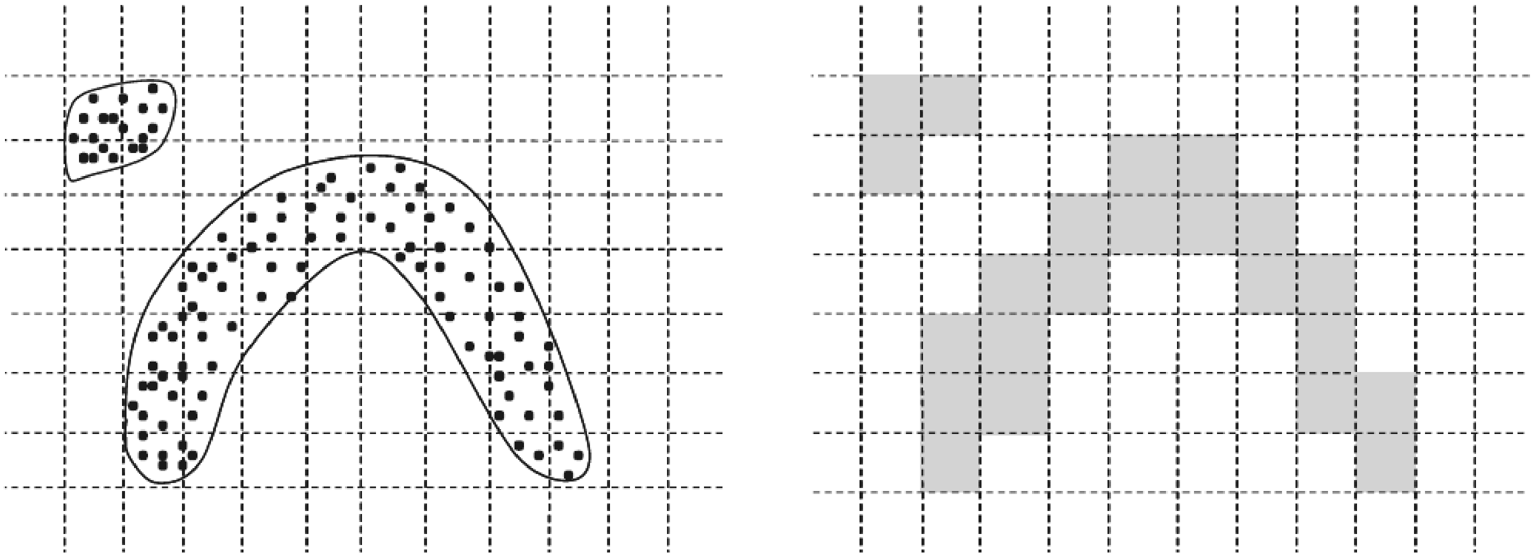


Example: Hierarchical clustering of genes



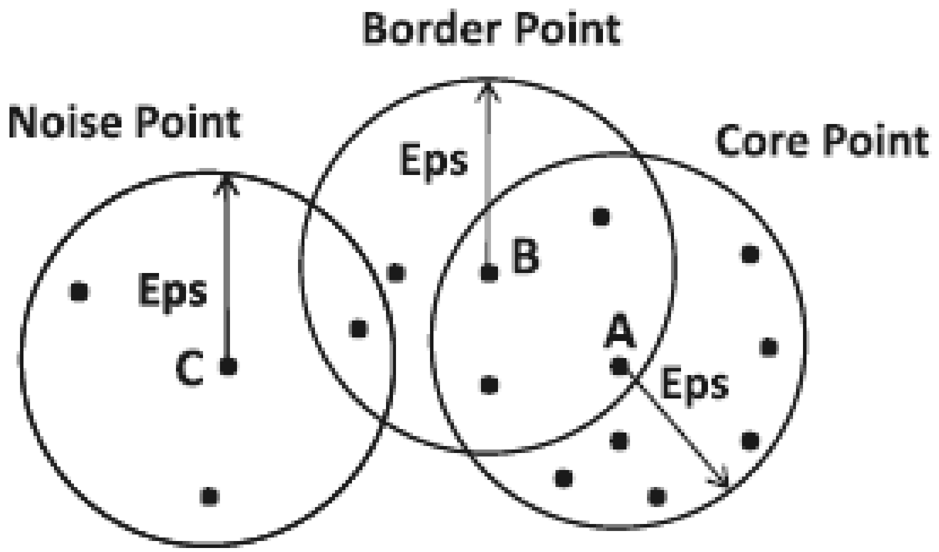
Grid-based (parameters **p** and **τ**)

1. Discretize each dimension of **D** into **p** ranges
2. Determine dense grid cells at level **τ**
3. Create graph where dense grid cells are connected if they are adjacent
4. Determine connected components of graph
5. Return: points in each connected component as a cluster



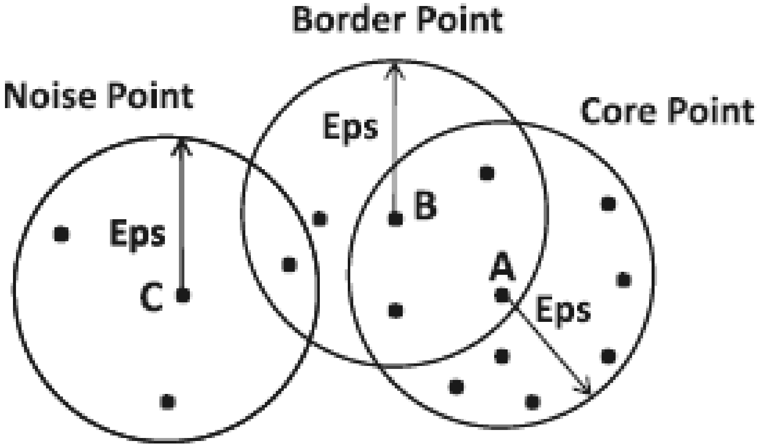
Density based clustering *DBSCAN*
(parameters: radius: *Eps*, density: τ)

- ***<u>Core point</u>:***
  - contains at least τ data points within a radius *Eps*
- ***<u>Border point</u>:***
  - not a core point
  - at least one core point within a radius *Eps*
- ***<u>Noise point</u>:***
  - neither a core point nor a border point



Density based clustering *DBSCAN* (parameters: radius: *Eps*, density: τ)

1. Determine core, border and noise points at level (*Eps, τ*);
2. Create graph in which core points are connected if they are within *Eps* of one another;
3. Determine connected components in graph;
4. Assign each border point to connected component with which it is best connected;
5. **Return** points in each connected component as a cluster;



Aggarwal, Charu C. *Data mining: the textbook*. Springer, 2015. Chapter 6: cluster analysis, pg 183

DBSCAN properties

Similar to grid-based approaches, except that it uses circular regions as building blocks.

Advantages of DBSCAN:

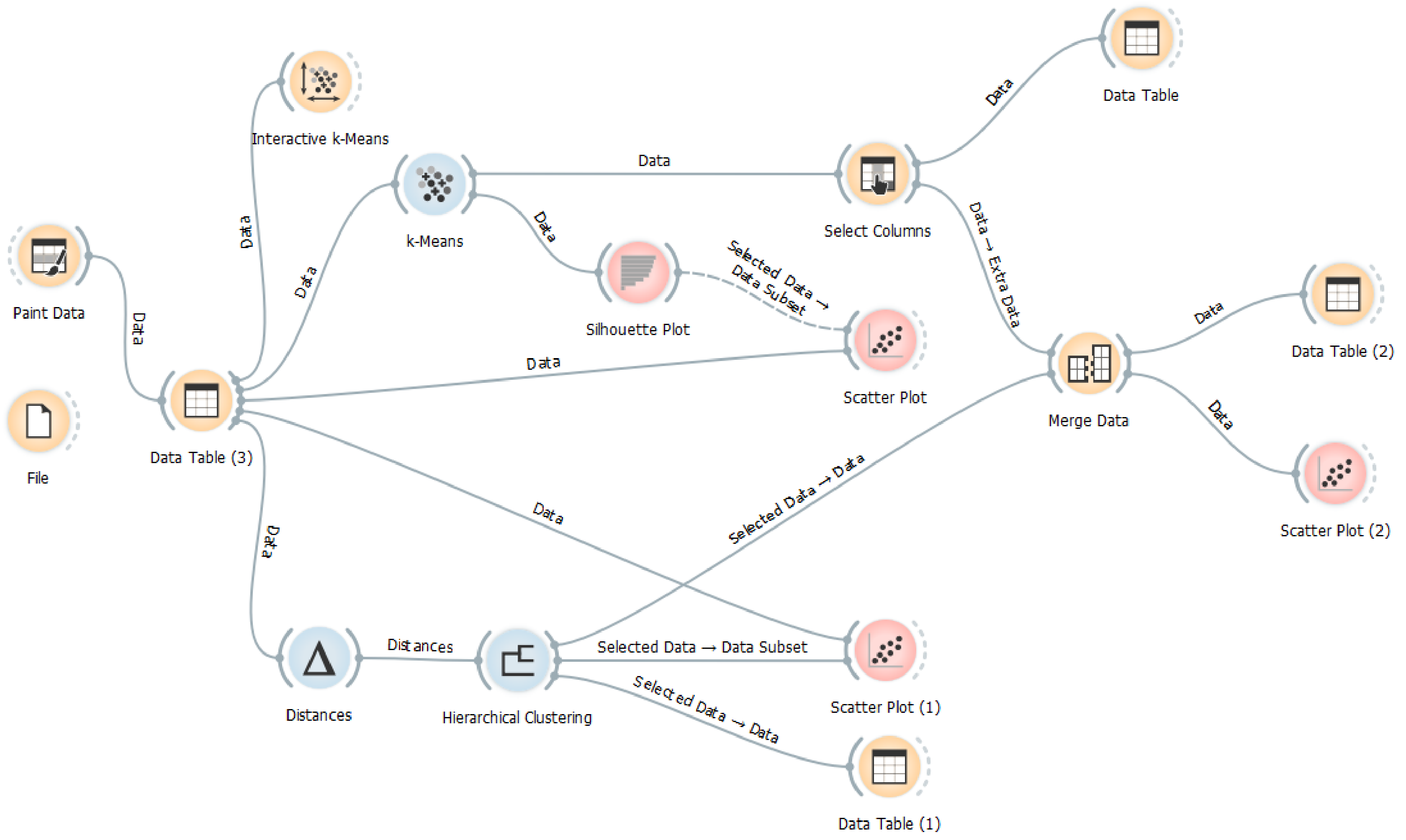
- Can detect clusters of arbitrary shape.
- Does not require the number of clusters as an input parameter.
- Not sensitive to outliers.

Disadvantages of DBSCAN:

- Computationally expensive in the first step (determining core, border and noise points)
- Susceptible to variations in the local cluster density.
- Struggles with high dimensionality data.

Lab work in Orange

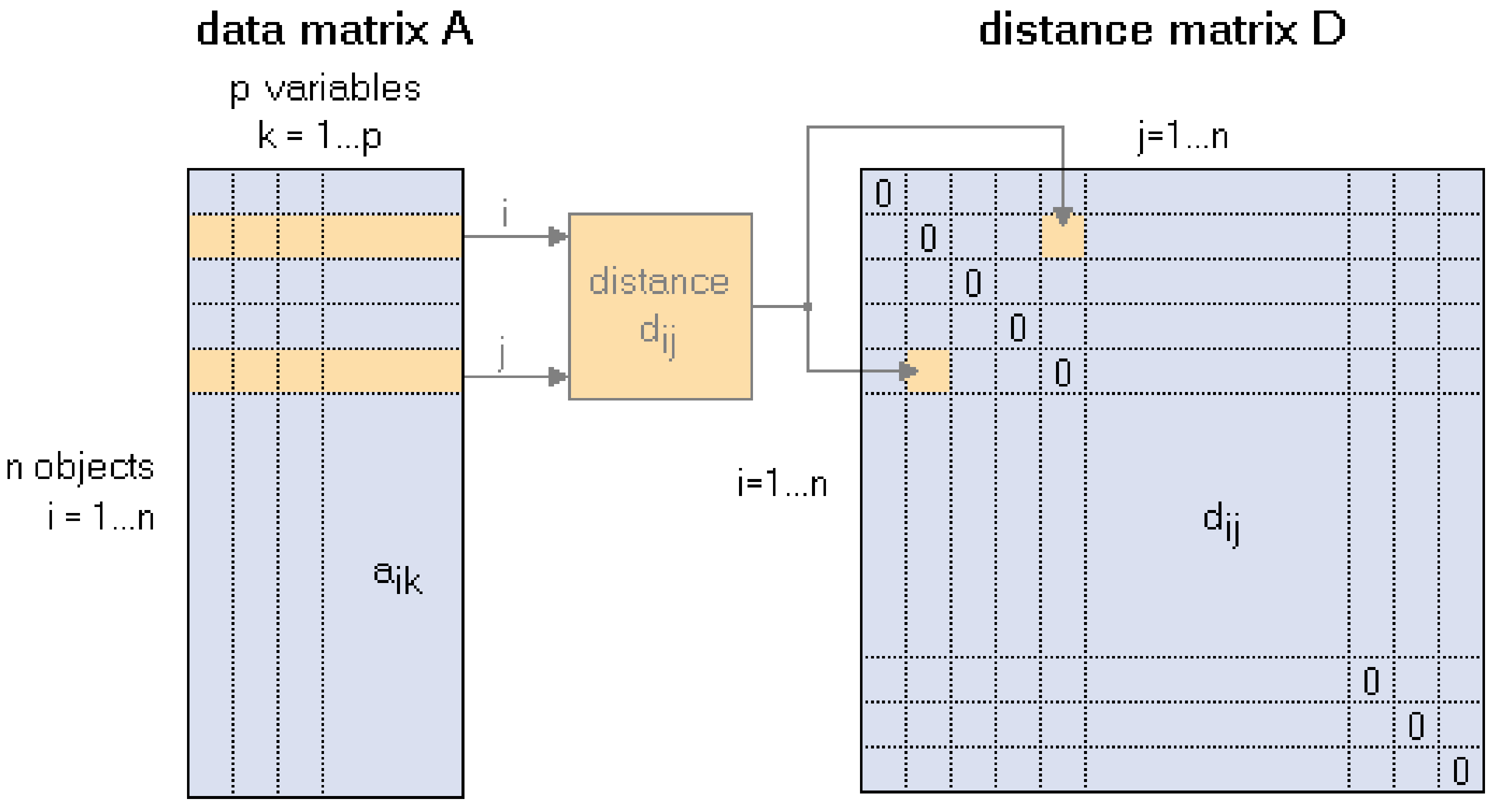
- Comparison of hierarchical and k-Means clustering on
- painted data
- „wine.tab“, where we compare also to the real classes



Similarity / distance measures

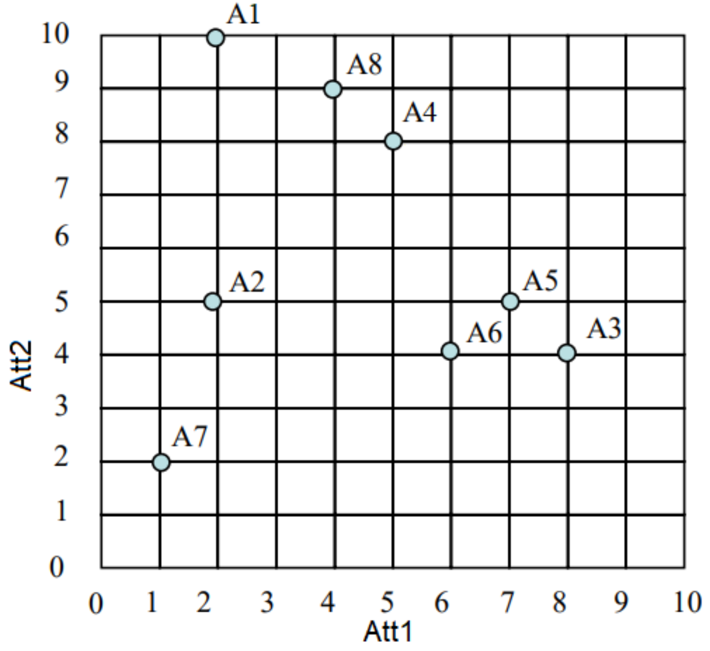
- The similarity measure depends on characteristics of the input data:
  - Attribute type: binary, categorical, continuous
  - Sparseness
  - Dimensionality
  - Type of proximity

Distance matrix



Distance matrix example

	Att1	Att2
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9



	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Euclidian → $Dist(A,B) = \sqrt[2]{(Att1(A) - Att1(B))^2 + (Att2(A) - Att2(B))^2}$

Distance measures

Euclidean	$d(x,y)=\sqrt{\sum (x_i-y_i)^2}$
Squared Euclidean	$d(x,y)=\sum (x_i-y_i)^2$
Manhattan	$d(x,y)=\sum \lvert(x_i-y_i)\rvert$
Canberra	$d(x,y)=\sum \frac{\lvert x_i-y_i\rvert}{\lvert x_i+y_i\rvert}$
Chebychev	$d(x,y)=\max(\lvert x_i-y_i\rvert)$
Bray Curtis	$d(x,y)=\frac{\sum \lvert x_i-y_i\rvert}{\sum x_i+y_i}$
Cosine Correlation	$d(x,y)=\frac{\sum (x_i y_i)}{\sqrt{\sum (x_i)^2 \sum (y_i)^2}}$
Pearson Correlation	$d(x,y)=\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (y_i-\bar{y})^2}\sqrt{\sum (y_i-\bar{y})^2}}$
Uncentered Peason Correlation	$d(x,y)=\frac{\sum x_i y_i}{\sqrt{\sum (y_i-\bar{y})^2}\sqrt{\sum (y_i-\bar{y})^2}}$
Euclidean Nullweighted	Same as Euclidean, but only the indexes where both x and y have a value (not NULL) are used, and the result is weighted by the number of values calculated. Nulls must be replaced by the missing value calculator (in dataloader).

Minkowski distance

$$D(X,Y)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p}$$

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer. (Chapter 3)

Homework

- Similarity vs. distance
- List algorithms that are based on distance/similarity

Literature

- Max Bramer: Principles of data mining (2007)
  - 14. Clustering
- Aggarwal, Charu C. *Data mining: the textbook*. Springer, 2015. Chapter 6: Cluster analysis
- Aggarwal, Charu C. *Data mining: the textbook*. Springer, 2015. Chapter 2: Similarity and distances