PGR210 – Machine Learning and Natural Language Processing

Lecture 6: Clustering, SOM

Andrii Shalaginov

27.09.2021 (week 39)

Plan for today

- Clustering
- SOM
- Repetition

- *Please, remember the final presentation of the Assignment 1 on 29th of September during our session. Every group needs to present and deliver the assignment report by the deadline.

Overview of the course (1)*

[Machine Learning Introduction]

Lecture 1 (week 34): Machine Learning Basics
Book: Kononenko (chapter 1,2), Muller (chapter 1)
(2 hours) Introduction, AI / ML, How does ML works, qwiklabs
(2 hours) Supervised, Unsupervised, Reinforced Learning; Q&A and control questions -> self-study

Exercise 1:
(2 hours) Jupyter Notebook, Google Colab, GCP, Microsoft Azure AI; assignment
(2 hours) Statistics, ML basics, data analytics

Lecture 2 (week 35): Overview of ML software tools
Book: Kononenko (chapter 3,4,5)
(2 hours) data processing, data formats, knowledge presentation
(2 hours) the curse of dimensionality; control questions

Exercise 2:
(2 hours) Learning as a Search
(2 hours) RapidMiner, Weka, PSPP, Orange3

Lecture 3 (week 36): Overview of ML libs
Book: Kononenko (chapter 5), Muller (chapter 1)
(2 hours) Examples of ML libraries utilization
(2 hours) Practice with data and analytics on pre-defined datasets

* To be updated

Overview of the course (2)*

Exercise 3:
(2 hours) Keras, TensorFlow, dlib, scikit; example of brute-force
(2 hours) Gradient Descent, Genetic Algorithm.

Lecture 4 (week 37): Data Processing
(2 hours) Data Quality, Visualization
(2 hours) Features construction (extraction and selection), Curse of Dimensionality, PCA

Exercise 4:
Book: Kononenko (chapter 6,7), Muller (chapter 4)
(2 hours) RelieFF, InfoGain, CFS measure
(2 hours) Practice on dimensionality reduction

Lecture 5 (week 38): Symbolic and Statistical Learning; Model Evaluation
Book: Kononenko (9,10), Muller (5)
(2 hours) Symbolic Learning
(2 hours) Statistical Learning

Exercise 5:
(2 hours) Decision Trees, Decision Rules
(2 hours) Regression Trees, Regression Rules, k-NN

* To be updated

Overview of the course (3)*

Lecture 6 (week 39): Clustering, Classification
Book: Kononenko (11,12)
(2 hours) ANN, SVM, Deep Learning
(2 hours) EM, K-Means, quantization

Exercise 6:
(2 hours) Deep Neural Network, SOM, SVM
(2 hours) Multiclass problems

[Natural Language Processing]

Lecture/Exercise 7 (week 40): Introduction to NLP, popular NLP libraries, use cases and popular applications
Lecture/Exercise 8 (week 41): NLP Pipeline, Text processing
Lecture/Exercise 9 (week 42): Text visualization and basic analysis
Lecture/Exercise 10 (week 43): introduction to text representation, BoW and TF-IDF representation
Lecture/Exercise 11 (week 44): Text representations: Word2Vec and Glob2Vec representation
Lecture/Exercise 12 (week 45): Recent progress and future trends.

Overview of the topics: ML

- Lecture 1 (week 34): Machine Learning Basics
Book: Kononenko (chapter 1,2), Muller (chapter 1)

- Lecture 2 (week 35): Overview of ML software tools
Book: Kononenko (chapter 3,4,5)

- Lecture 3 (week 36): Overview of ML libs
Book: Kononenko (chapter 5), Muller (chapter 1)

- Lecture 4 (week 37): Data Processing
Book: Kononenko (chapter 6,7), Muller (chapter 4)

- Lecture 5 (week 38): Symbolic and Statistical Learning; Model Evaluation
Book: Kononenko (chapter 9,10), Muller (chapter 5)

- **Lecture 6 (week 39): Clustering, Classification**
Book: Kononenko (chapter 11,12)

Clustering

We Are Here, Because.....

- We believe our data has a structure that reflects its system of origin
- We believe that proper analysis of the data will reveal the data’s structure
- We believe that the **data structure** we discover, will give us useful information about the system of origin

Overview

- Recall mixture model of data
  - The structure of the data
  - The data’s relation to the system we are studying

- How we can extract information about the system, from the data
  - Analytical Model
    - The Math of Extracting (unmixing the mixture model)
  - Empirical Methods
    - Data analysis to extract information about the data structure
  - Unsupervised Learning
  - Self Organizing Maps (SOM)
    - Self Organizing Feature Maps

Analytical vs Empirical

- Analytical
  - The true nature of the system under study
  - Idealized model (usually mathematical)
  - Allegory of the cave: the objects, not their shadows
  - We Can Never* Have Direct Knowledge

- Empirical
  - What we can actually know about the system under study
  - Data
  - Our analysis of the data
    - Estimates of the true nature
  - Always indirect knowledge of the true nature of the system
    - Recall limitations of the senses

Supervised v Unsupervised

- Supervised Learning Vectors are Labeled
  - Explicit preconceptions about data structure
  - Costly
- Unsupervised Learning Vectors: Unlabeled
  - Are there implicit preconceptions?
    - There is at least one
  - Lower Cost

Why is labelling costly?

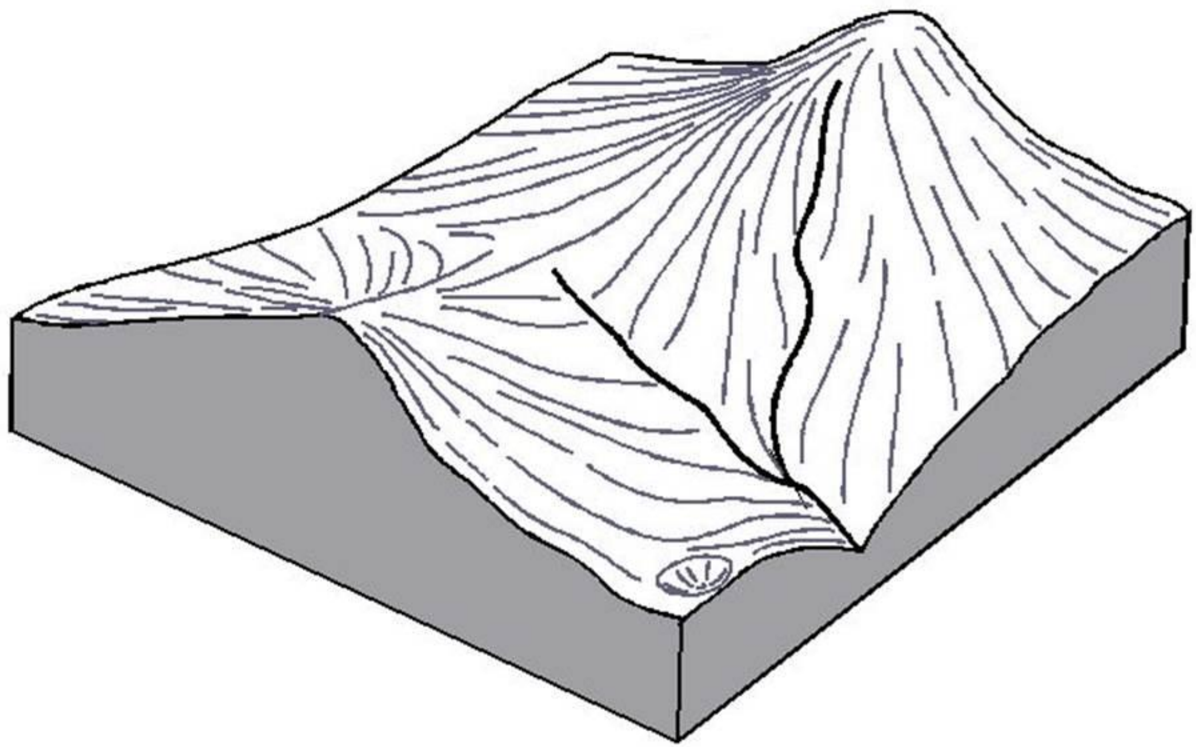
5 Reasons for Unsupervised Learning

1. Cost of Labelling
2. Data Mining
3. Dynamic Classes
4. Identify Useful Features
5. Initial Exploratory Data Analysis

Types of Unsupervised Learning

- Clustering
  - K-mean
  - GMM
- Self Organization
  - What is the organizational principal?
    - Data topology
    - Want a topology preserving projection to lower dimensional space
      - Say What?
      - Some/all of the data structure is preserved

Topology Preserving Projections I



http://www.cita.utoronto.ca/~murray/GLG130/Exercises/F2.gif

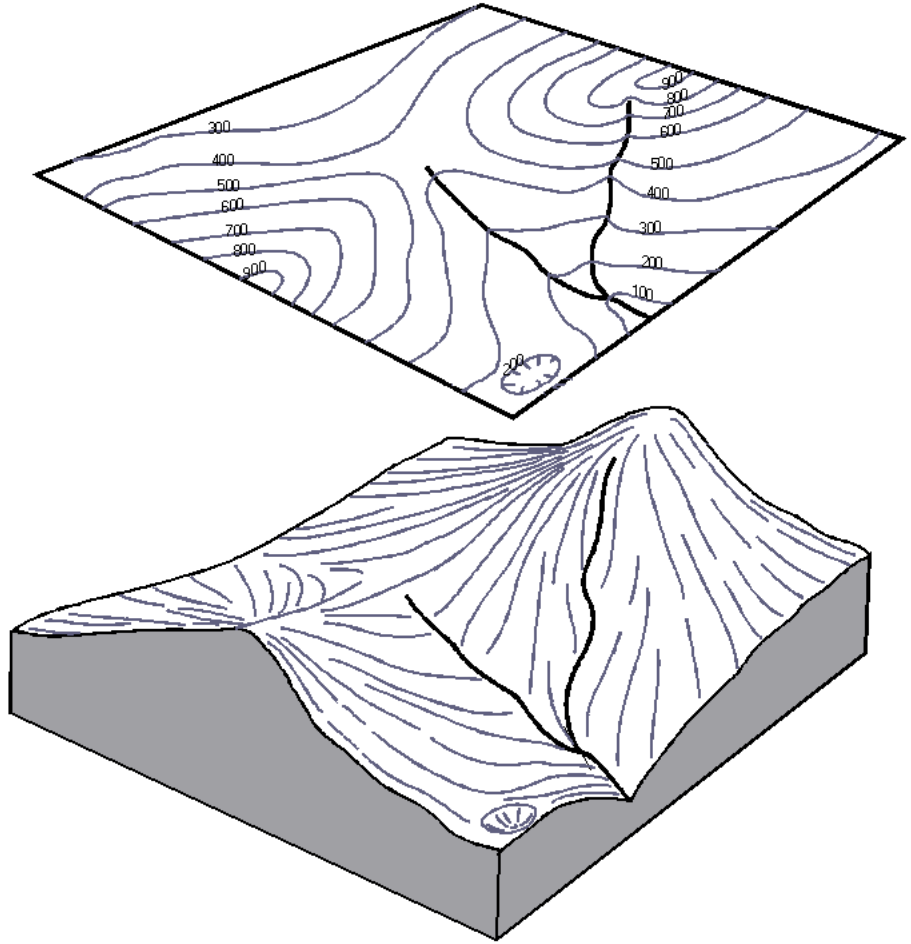
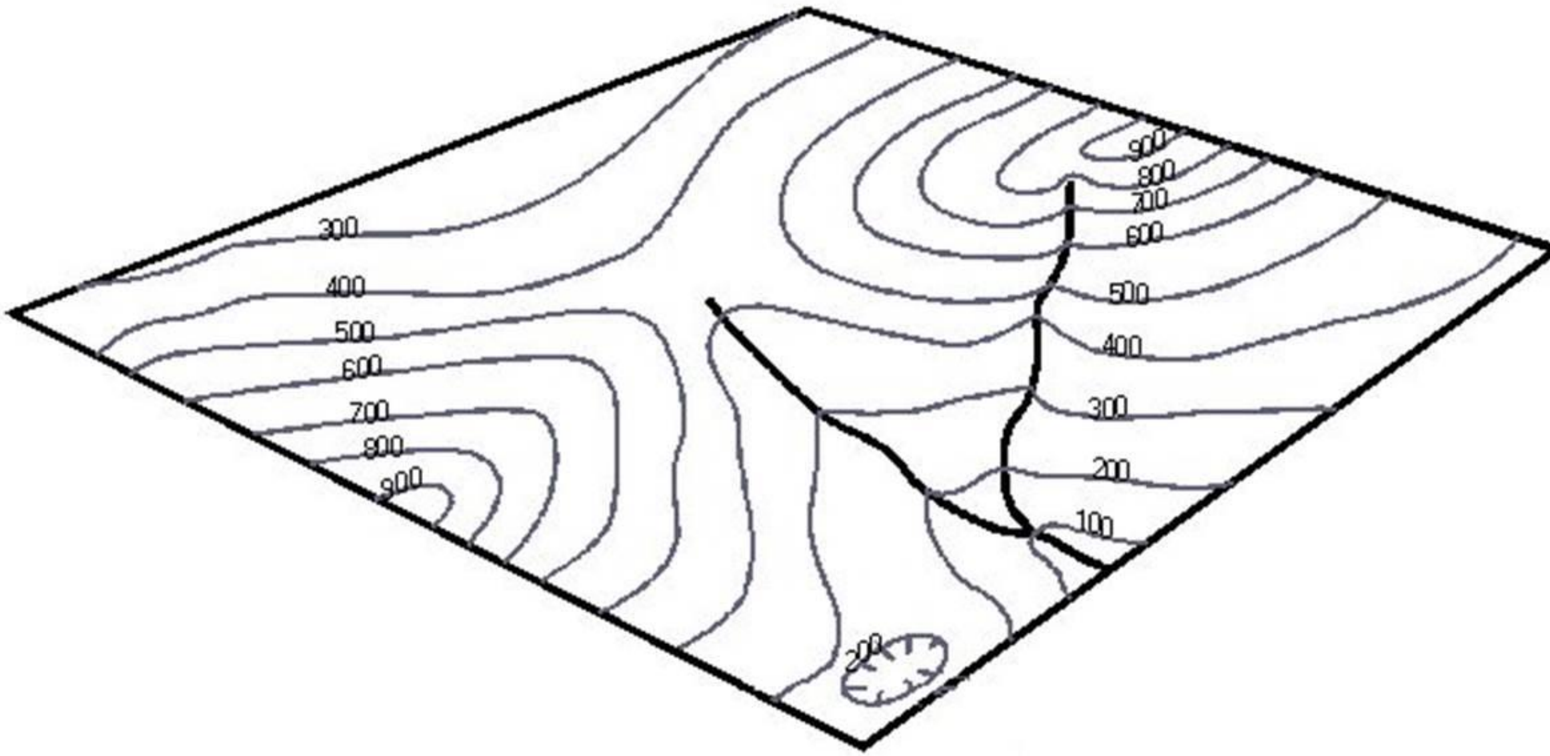


Figure 2. The relationship between a topographic map (top) and the corresponding land surface (bottom).

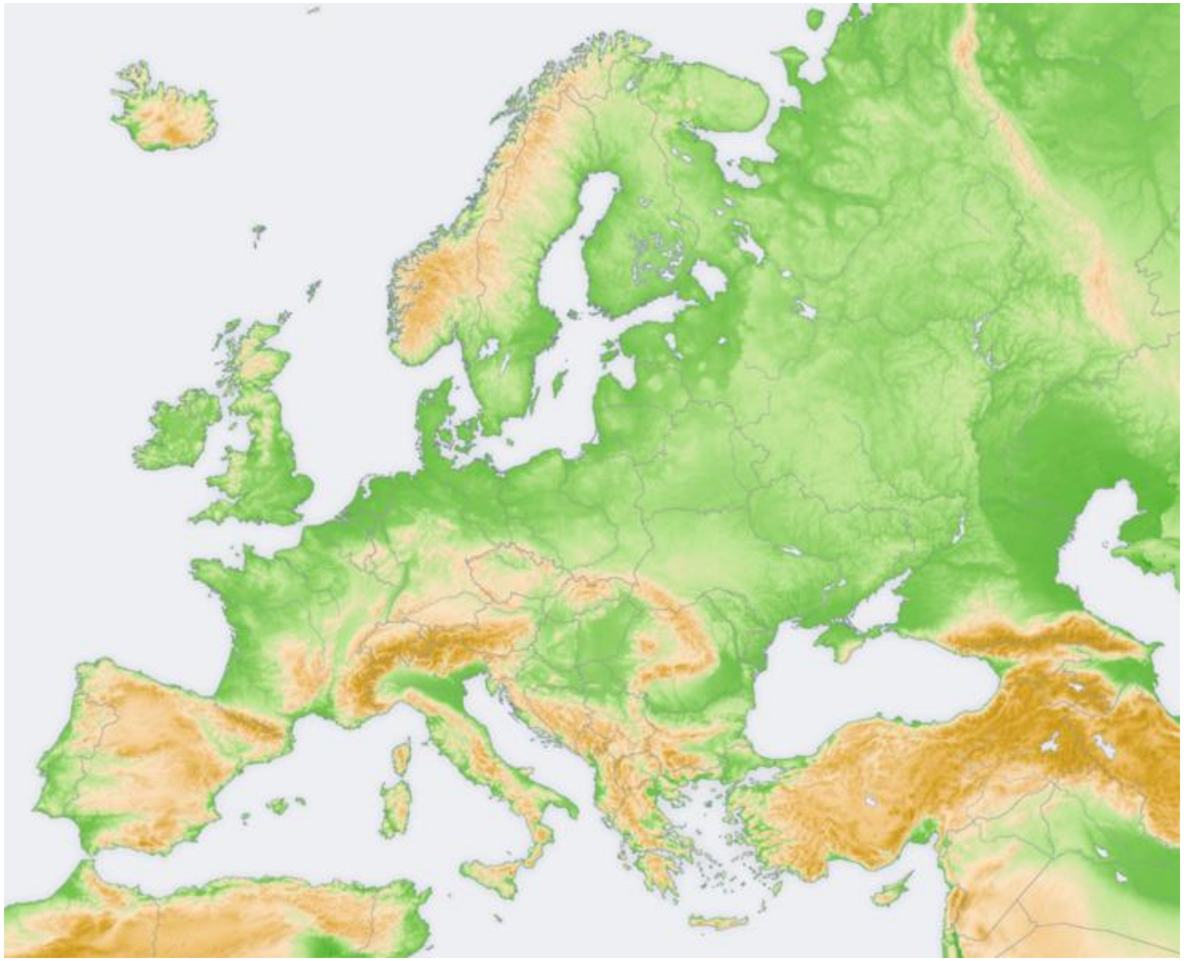
http://www.cita.utoronto.ca/~murray/GLG130/Exercises/F2.gif

Topology Preserving Projections I



http://www.cita.utoronto.ca/~murray/GLG130/Exercises/F2.gif

Topology Preserving Projections



https://commons.wikimedia.org/wiki/File:Europe_topography_map_en.png

Topology Preserving Projections

- Geographic terrain projections are limited.
  - Restricted to 3D -> 2D
    - 2D map (isomorphic projection):
      - N, S, E W -> Top, Bottom, Right, Left
    - 3D ->2D map: N, S, E, W, Higher, Lower

  - How do we visualize nD -> 2D (n>3) ???
    - n= 4, Iris Flower Data Set

  - What relationship(s) we can generalized for n dimensional spaces?

  - What do they all feature spaces have in common?
    - A distance metric!

Topology Preserving Projections

- How will the distance metric handle polymorphous data?
  - Units of time (different units of time?)
    - Sprint performance data: years of age and seconds to finish
  - Units of space
    - (meters, lightyears)
    - Surface area
    - Volumetric
  - Units of mass (grams, kilograms, tonnes)
  - Units of $$$
    - NOK
    - USD
      - Benjamins

Topology Preserving Projections

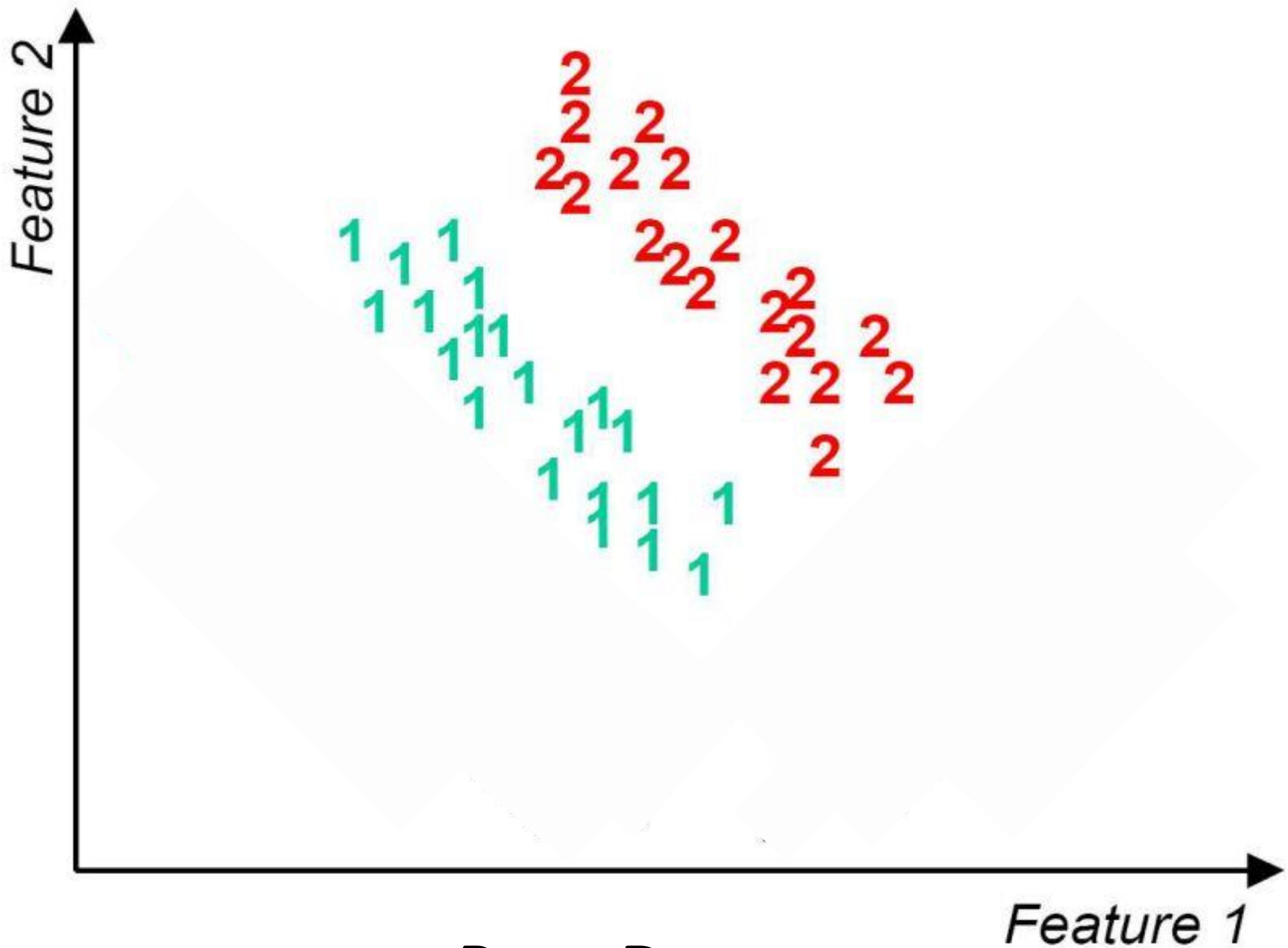
- How will the distance metric handle polymorphous data?
  - Explicit Data Standardization (z-Statistics)
  - No Data Standardization (Input raw data numbers)
    - Units are dropped, but dynamic ranges are preserved.
      - 40 years old (range: 20-65)
      - 5 years of college (0-8)
      - 50000 NOK (0-100000)
  - Fuzzification of Data input into Membership Function Values (Topic for next week)

Topology Preserving Projections

- What level of preservation is required?
  - What information can we do without?

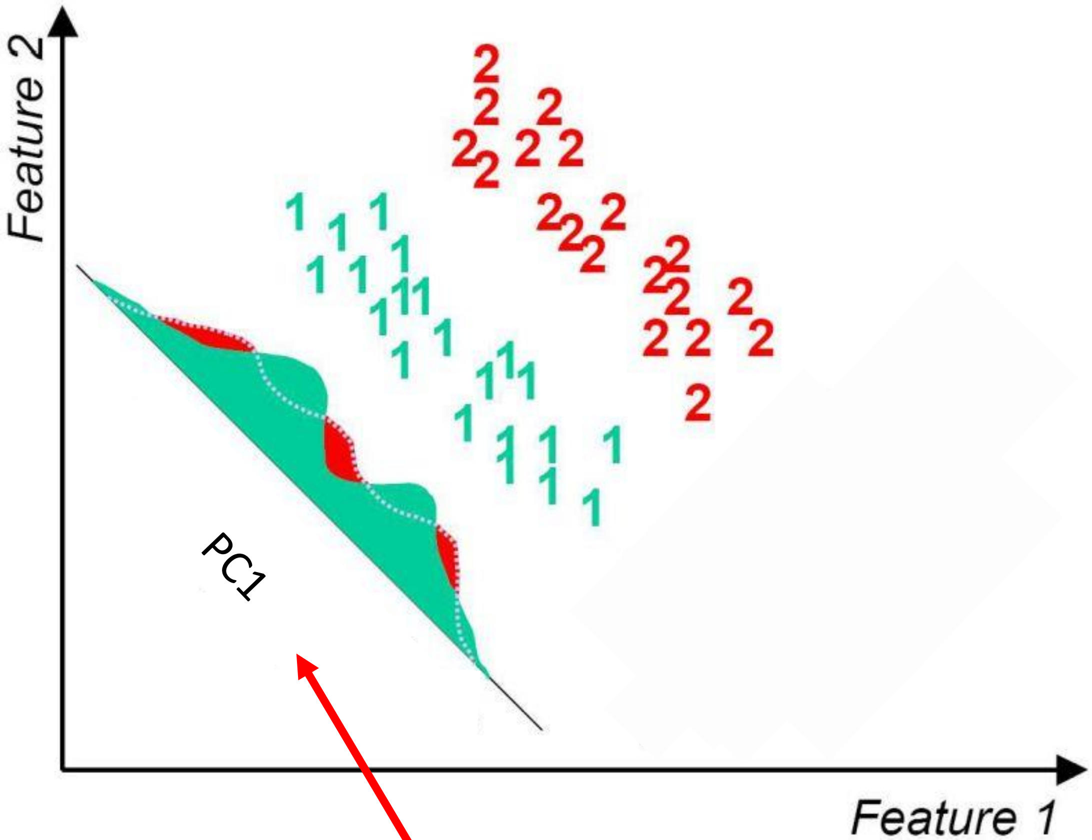
- Lossy PCA reduction for classification
  - Discarding principal components containing information
  - Do all PCs contain information?
    - Some components can be pure normal/gaussian noise
  - WARNING!
    - *DO NOT RECONSTRUCT THE DATA WITH LOSSY PCA *
      - Discuss it with me, first (cf PhD thesis: “Eigenspecters”)
    - Using Lossy PCA, without data reconstruction, is OK

Topology Preserving Projections



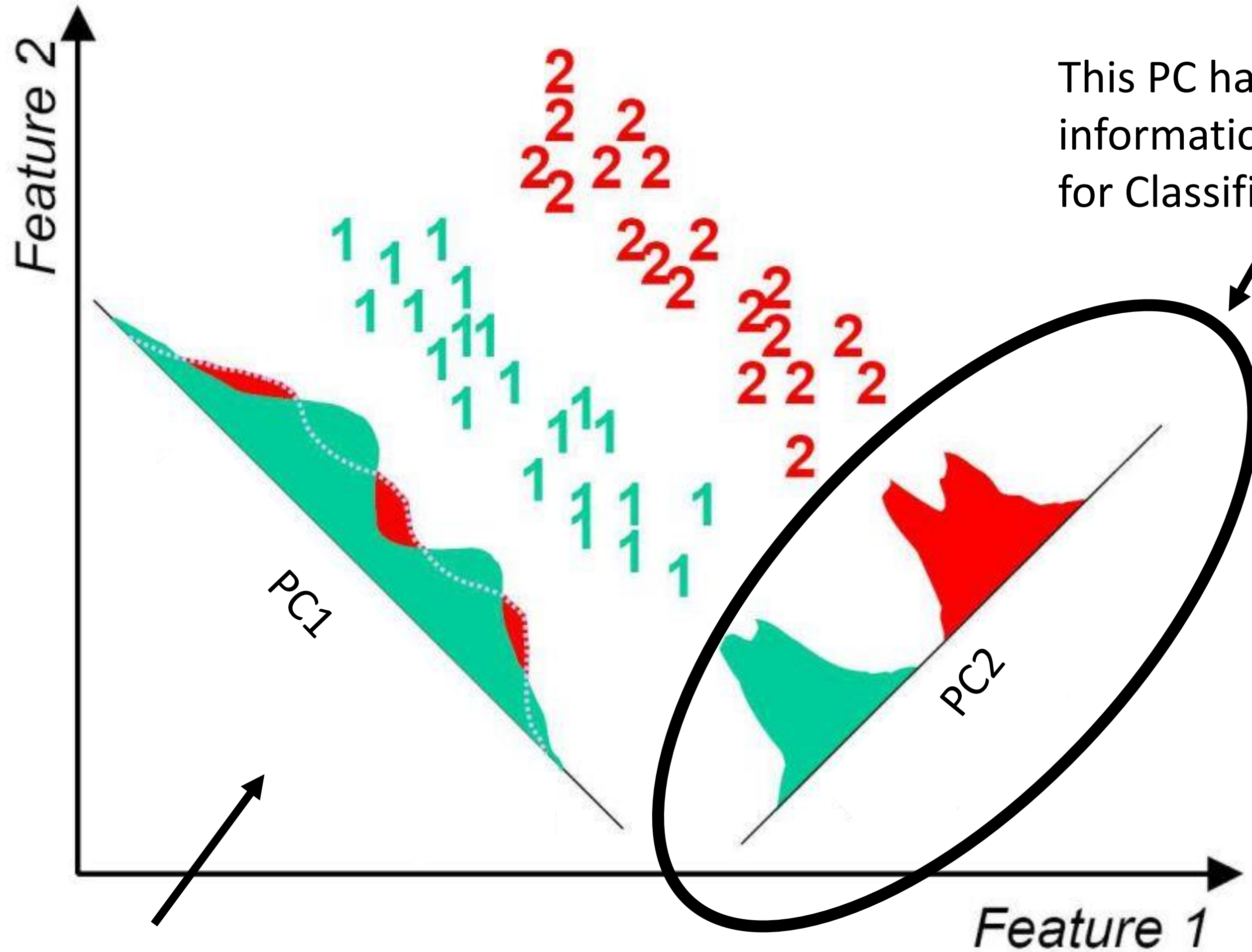
Raw Data

Lossy PCA Reduction for Classification



First PC

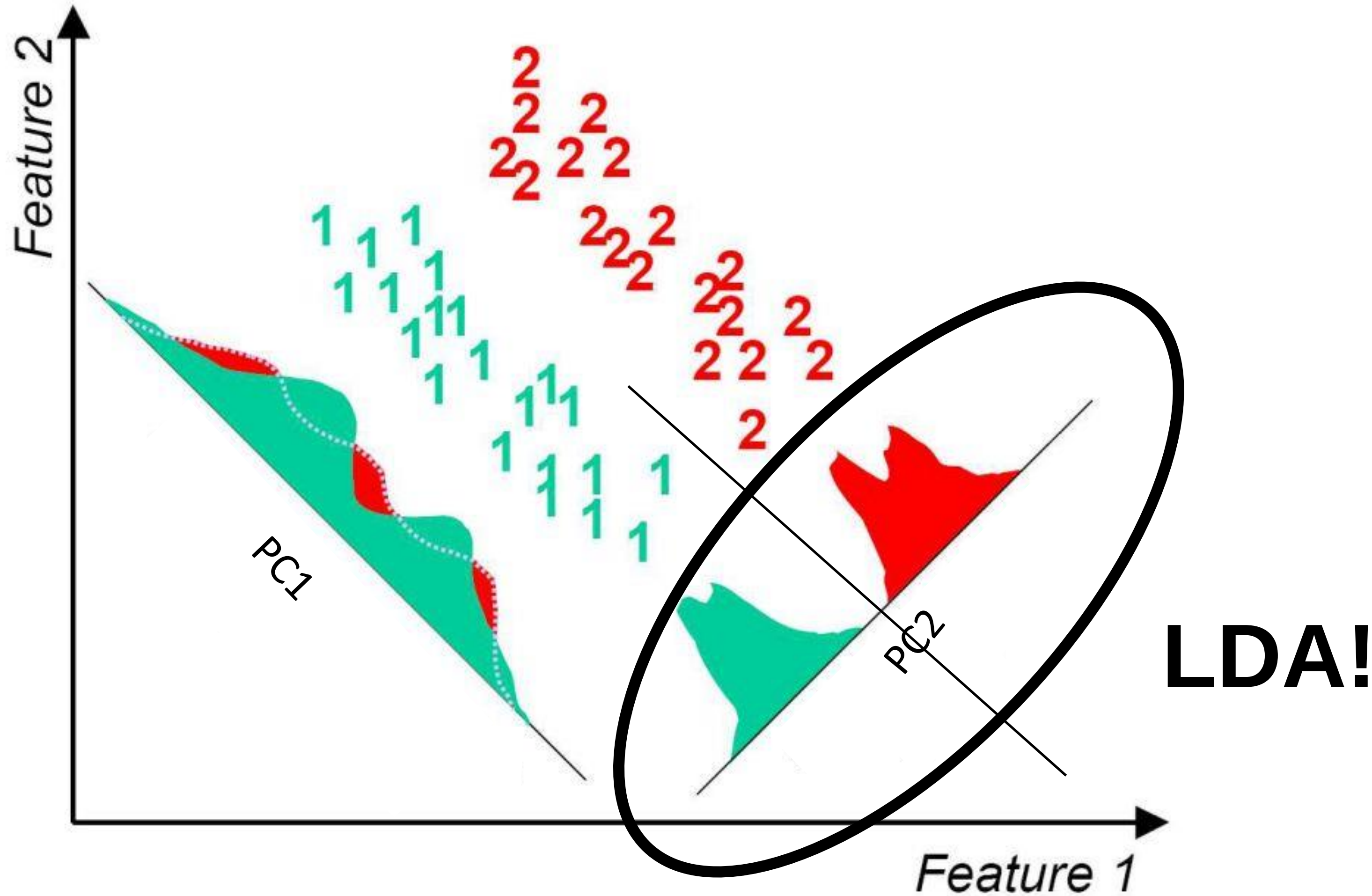
Topology Preserving Projections



This PC has all information required for Classification

So We Don't Need This PC for classification

What Type of Simple Classifier Can We Use?



Un/Supervised Clustering

- Recall k-means
  - It is semi-supervised in that we have pre-determined the number of means (number of clusters)
- Recall G-MM
  - Note how the results are affected by the initial estimate for the number of clusters

Un/Supervised Clustering

- Recall k-means and GMM-EM clustering

watch videos

www.youtube.com/watch?v=_aWzGGNrcic

www.youtube.com/watch?v=qMTuMa86NzU

www.youtube.com/watch?v=B36fzChfyGU

Un/Supervised Clustering

- Recall k-means
  - It is semi-supervised in that we have pre-determined the number of means (number of clusters)

- Recall G-MM
  - Note how the results are affected by the initial estimate for the number of clusters

- Many Artificial Neural Networks are like doing statistics with black boxes.

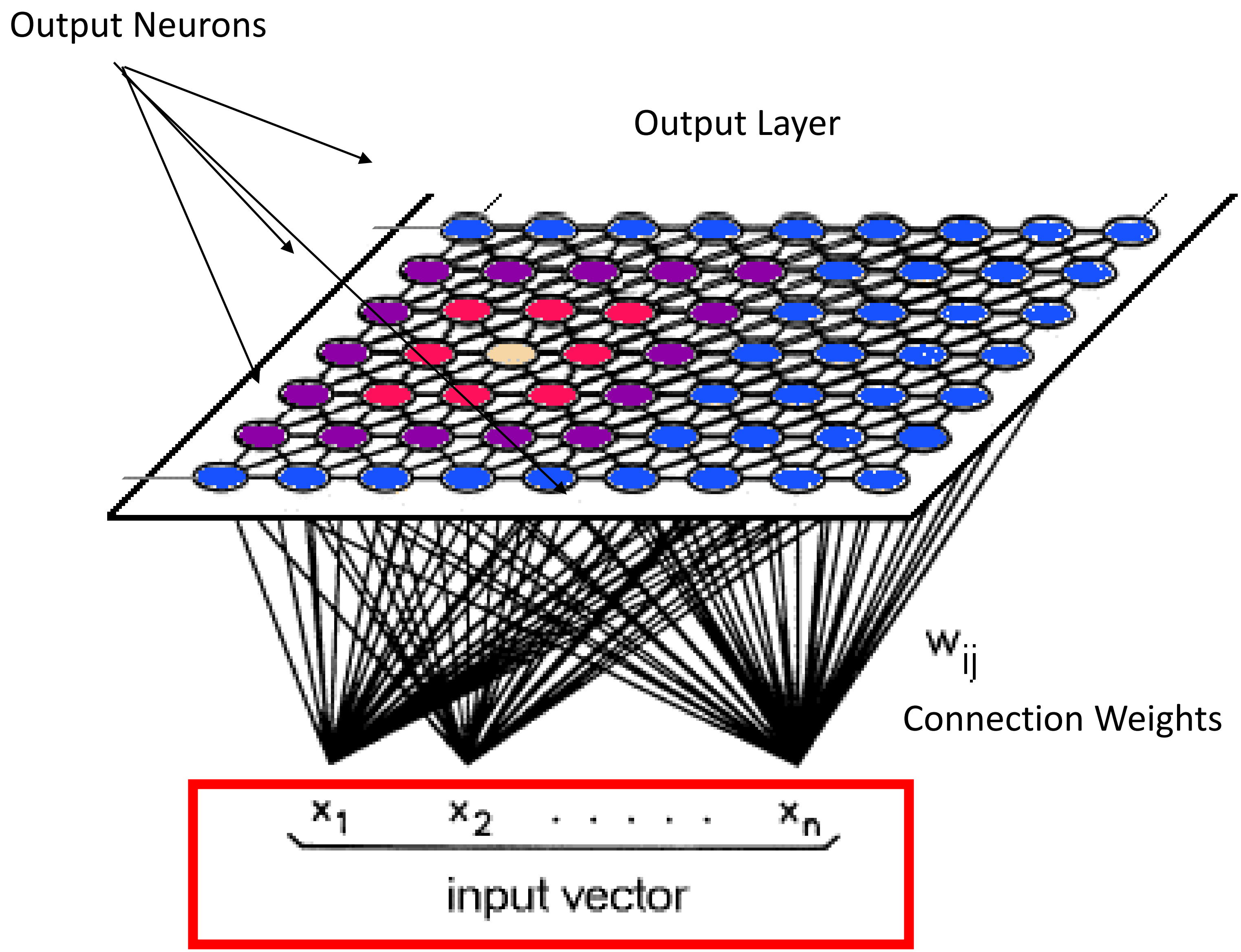
- An SOM is like doing k-means with ANN
  - We pick the number of output neurons
  - Training the SOM moves the output neurons wrt the data

SOM and Topology Preservation

- What is actually preserved?
  - Spatial Relationships
- So, we would like a way to take high-dimensional data and reduce it down to a 2-D map that preserves the spatial relationships of the higher dimensions.
- How do we do that?
  - Distance (Things nearby are similar)
  - Colour (Things with similar colors are similar)
  - Location (E, W N S –Right Left Top Bottom)*

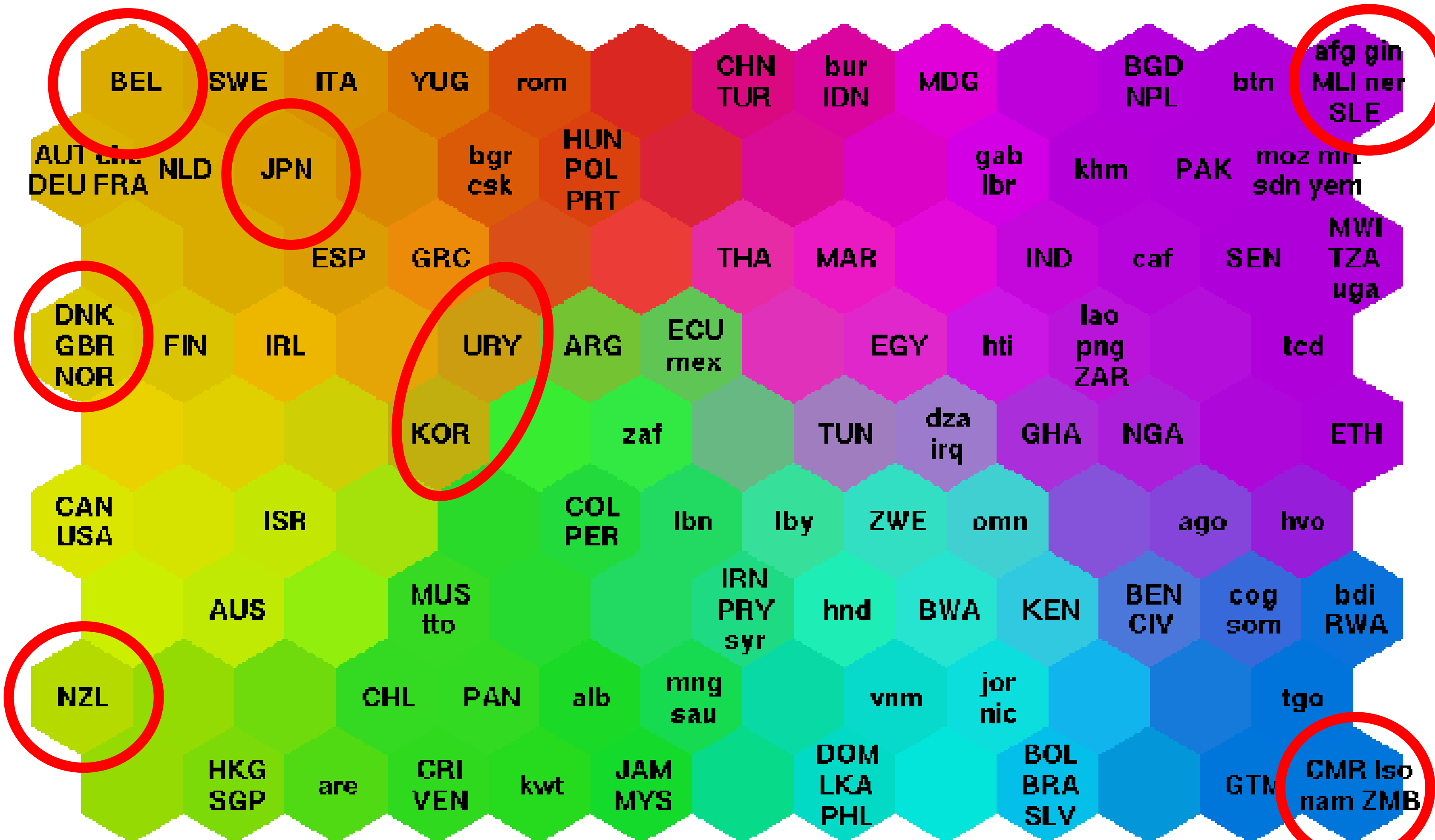
    *Might not always mean what you think

Self Organizing Maps Architecture



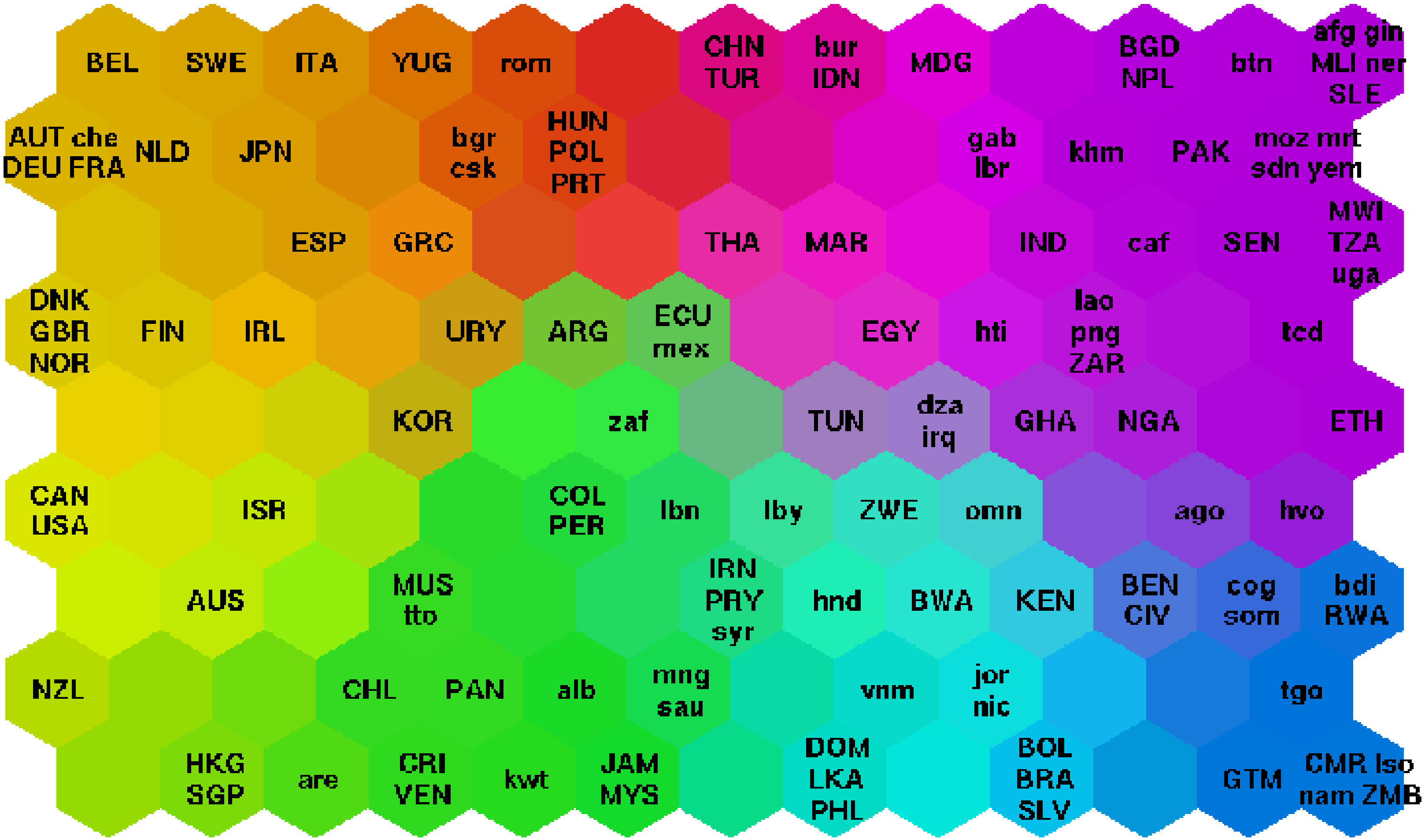
Proximity By Colour and Location

Poverty Map of the World (1997)



http://www.cis.hut.fi/research/som-research/worldmap.html

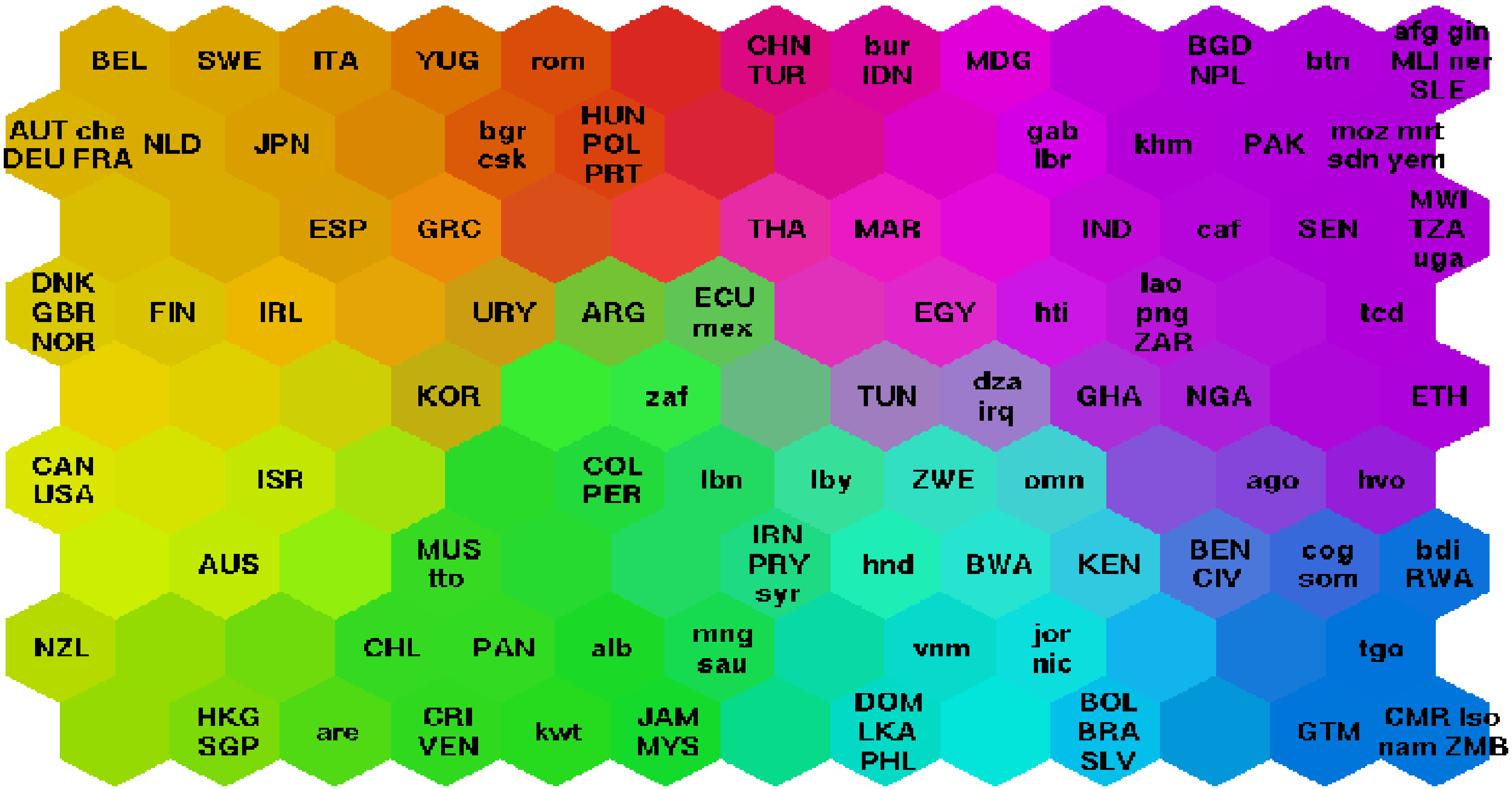
If ML Is Statistics By Other Means, Why Use ML Instead of Stats?



Is Map Orientation Important?

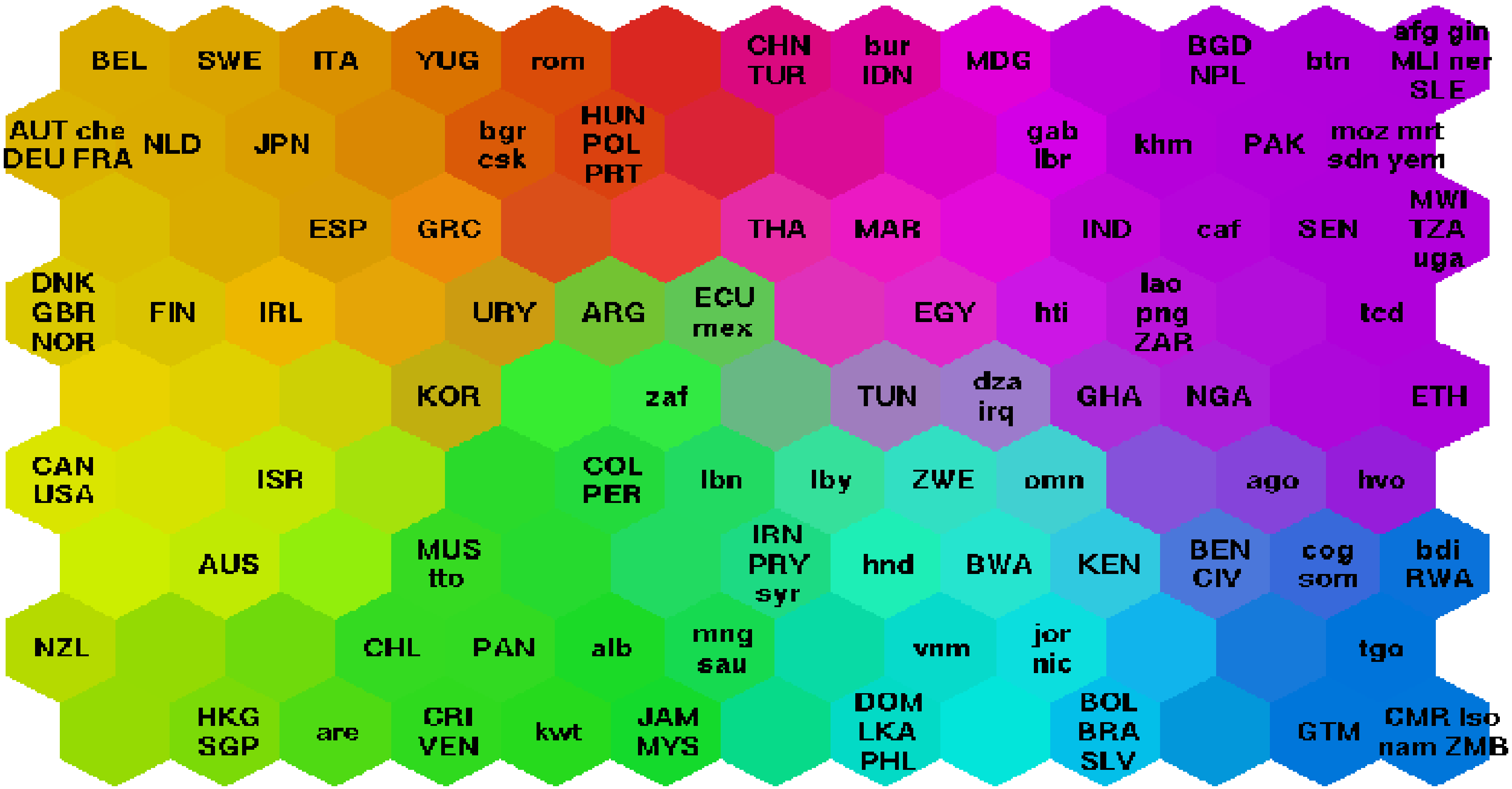
Are the Map Axes Informative?

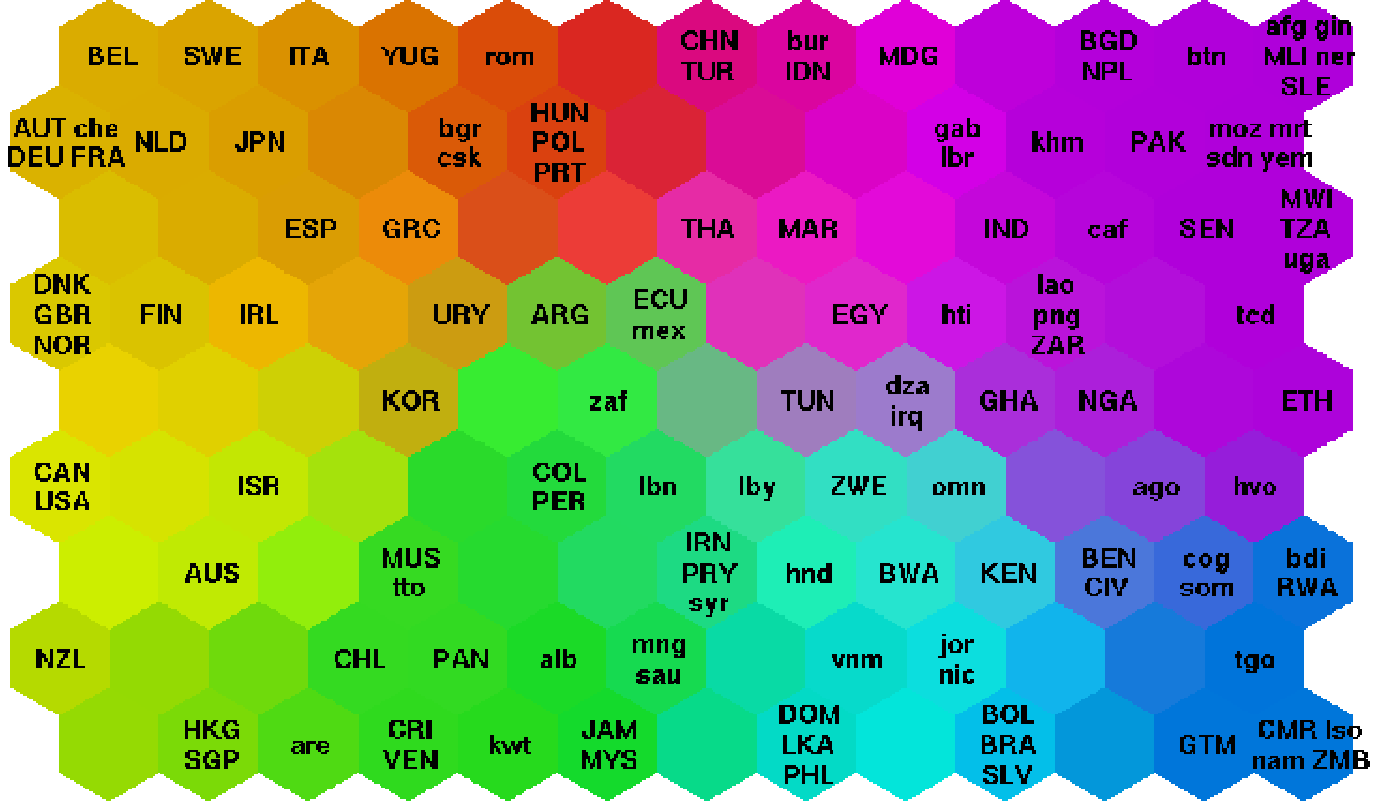
- Proximity is the most important relation
  - Data points that are in the same neighbourhood, have the closest resemblance to each other



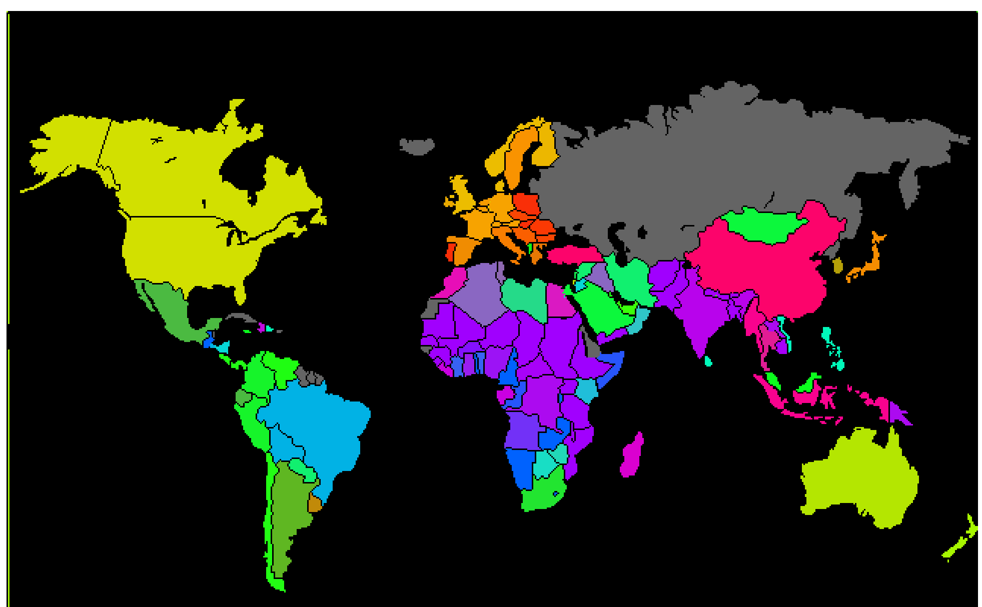
Are the Map Axes Informative?

- Data points that are to the left, right, above or below are indicating their relationship to neighbourhoods that are further away
  - Further Away = data with a less close resemblance





BEL
SWE
ITA
YUG
rom
CHN TUR
bur IDN
MDG
BGD NPL
btn
afg gin MLI ner SLE
AUT che DEU FRA
NLD
JPN
bgr csk
HUN POL PRT
gab lbr
khm
PAK
moz mrt sdn yem
ESP
GRC
THA
MAR
IND
caf
SEN
MWI TZA uga
DNK GBR NOR
FIN
IRL
URY
ARG
ECU mex
EGY
hti
lao png ZAR
tcd
KOR
zaf
TUN
dza irq
GHA
NGA
ETH
CAN USA
ISR
COL PER
lbn
lby
ZWE
omn
ago
hvo
AUS
MUS tto
IRN PRY syr
hnd
BWA
KEN
BEN CIV
cog som
bdi RWA
NZL
CHL
PAN
alb
mng sau
vnm
jor nic
tgo
HKG SGP
are
CRI VEN
kwt
JAM MYS
DOM LKA PHL
BOL BRA SLV
GTM
CMR lso nam ZMB



How Does the SOM Work?

- A competitive learning algorithm.
  - The neuron “closest to the input vector” is the winner
    - The neuron that most closely resembles a sample input.
    - Its weight vector is adjusted to move even closer to the ***current*** input vector x_i
  - The neurons that are too far away lose out completely
    - No weight adjustment for them!
- A cooperative learning algorithm
  - But the neurons in the “same neighbourhood” as the winner are partial winners
    - Their weight vectors are adjusted, based on their proximity to "winning" neurons
    - The closer the neighbour is to the winner, the more its weight vector is adjusted

How Large is the Neighbourhood?

- How big would you like it?
  - It’s a training parameter that can be set
  - A parameter that also gets smaller as training progresses
    - Like the ANN weight training step size gets smaller as training progresses

Neighbour Interconnection Topologies

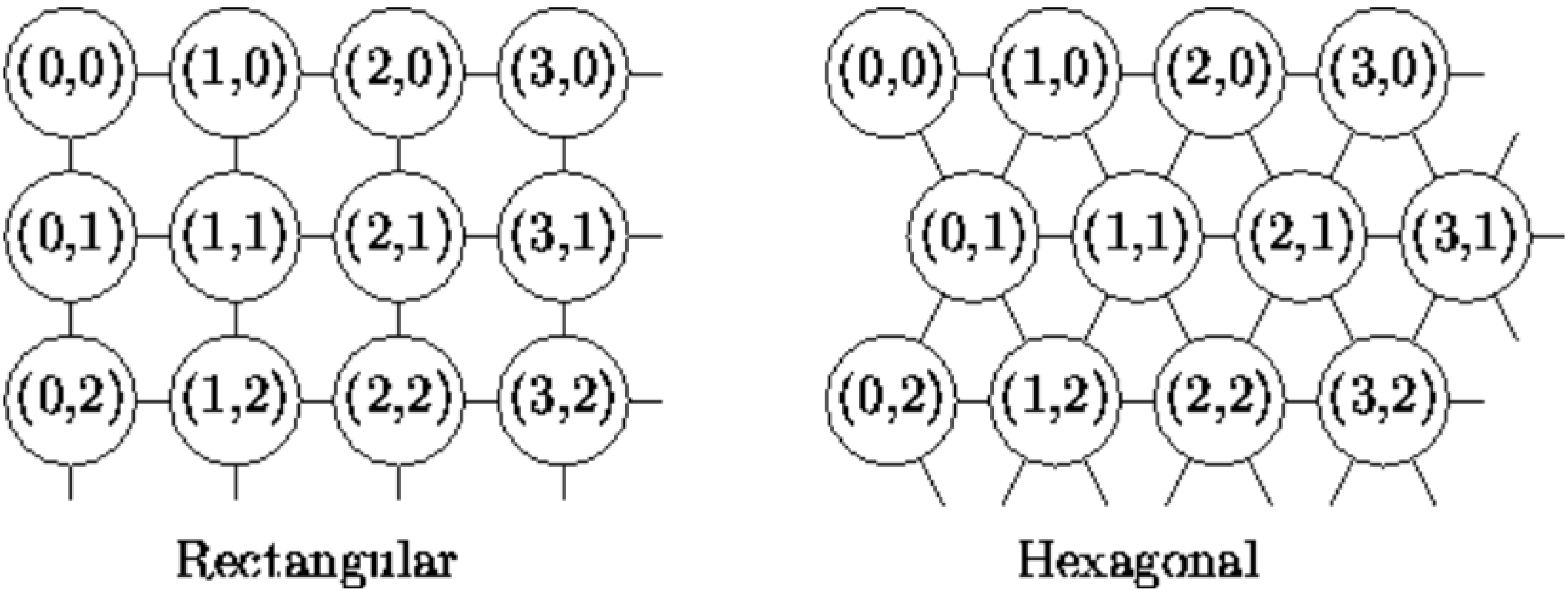
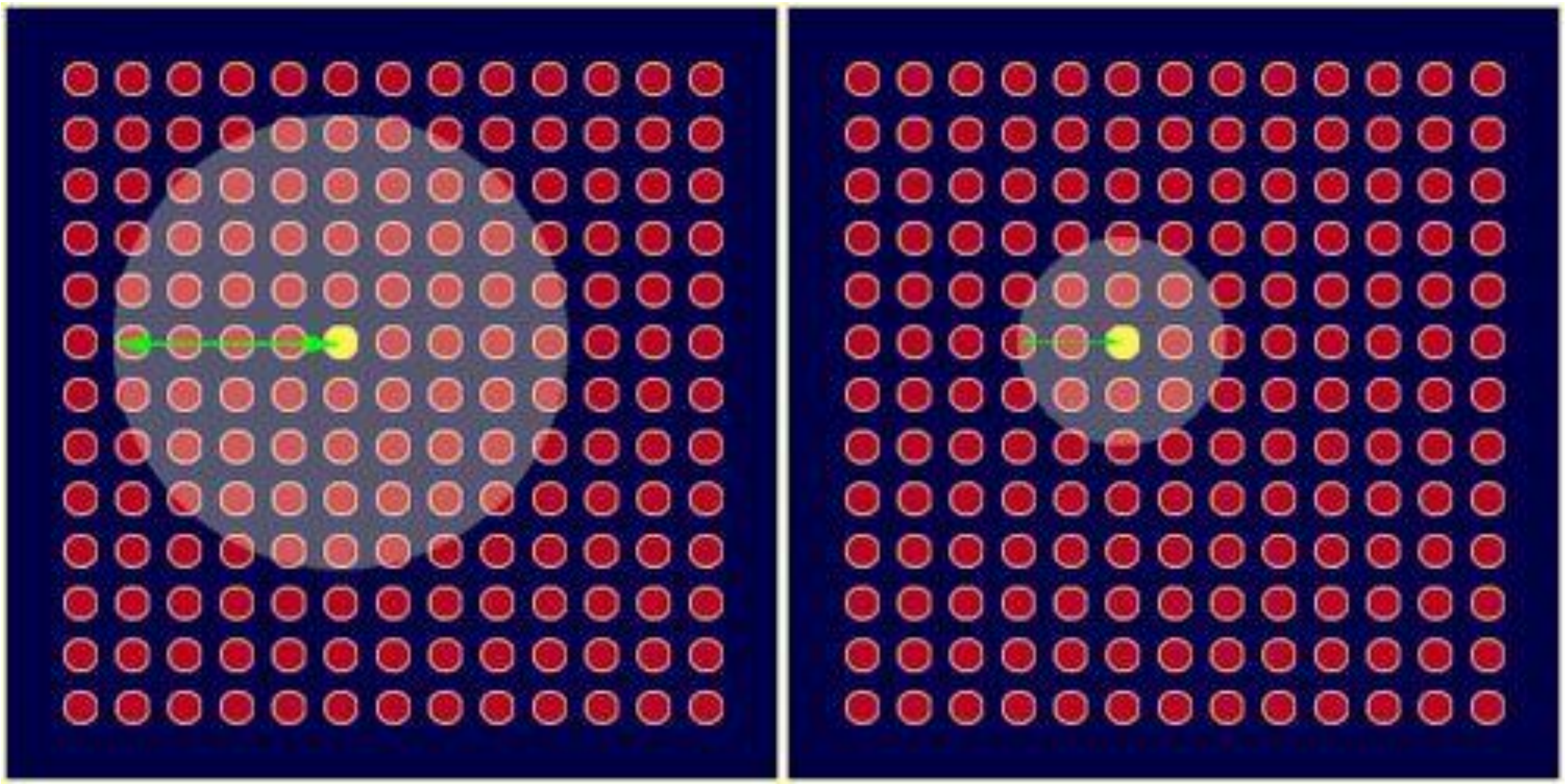


Figure 2.3: Different topologies

http://users.ics.aalto.fi/jhollmen/dippa/node9.html

Neighbourhoods in a Rectangular Map



The Hexagonal Neighbourhood

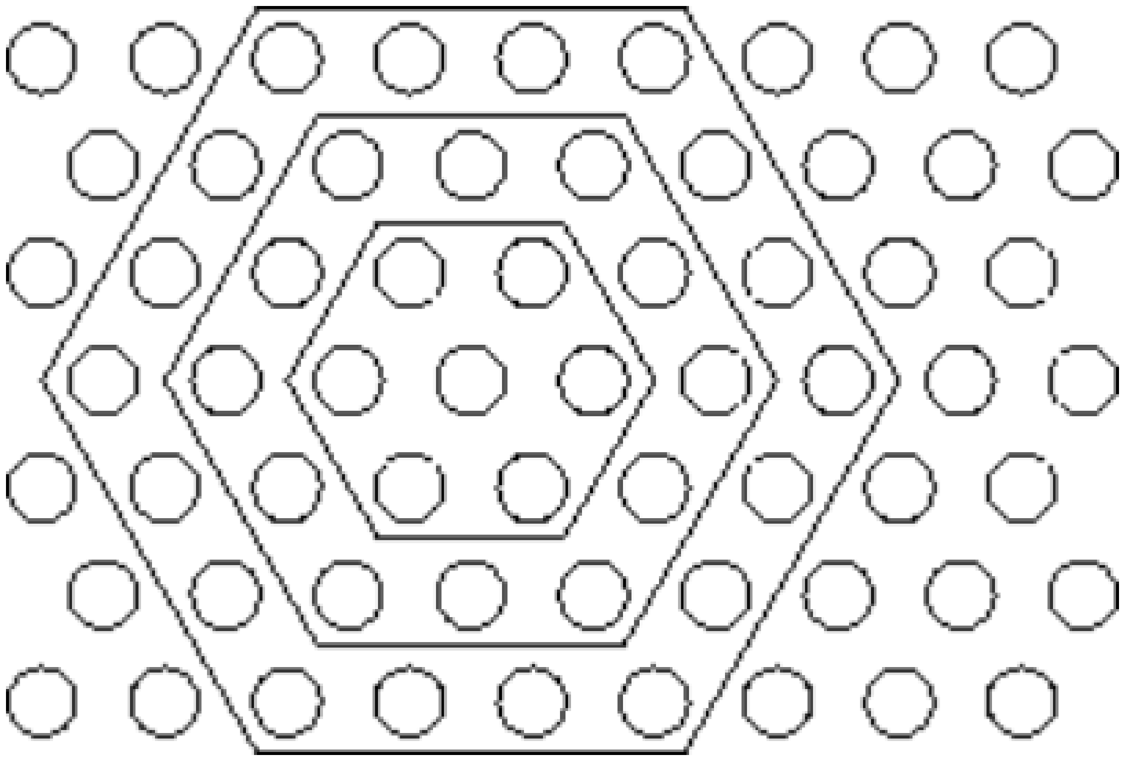


Figure 2.4: Neighborhood of a given winner unit

http://users.ics.aalto.fi/jhollmen/dippa/node9.html

Image Credits

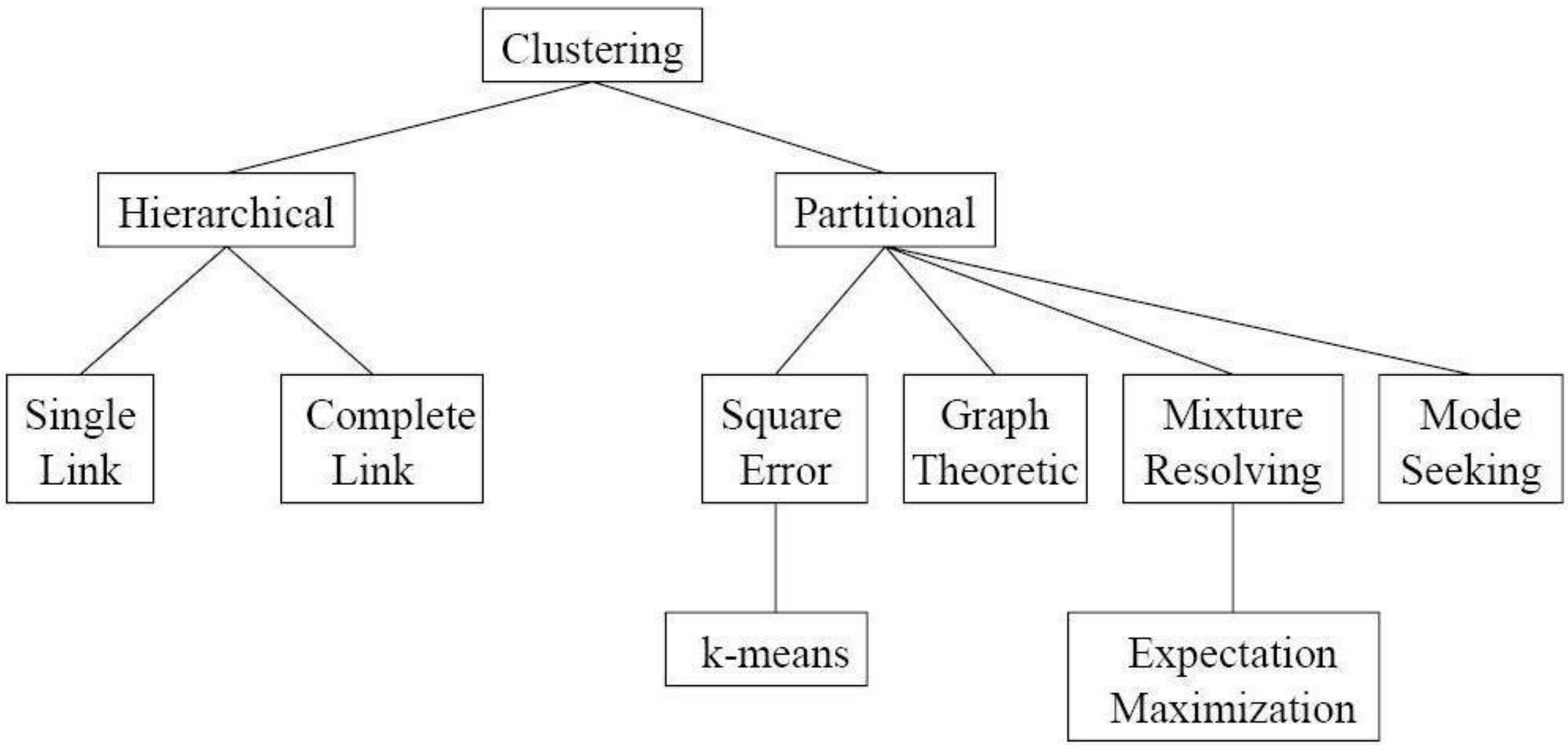
- https://12095675emilygrant3ddunitx.files.wordpress.com/2013/05/mapprojection5.gif?w=450&h=299
- https://en.wikipedia.org/wiki/Self-organizing_map#/media/File:Somtraining.svg
- https://en.wikipedia.org/wiki/File:Europe_topography_map.png
- http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html
- By User:W!B: - http://www.maps-for-free.com/, GFDL, https://commons.wikimedia.org/w/index.php?curid=5115489

Repetition

1. Sketch a possible taxonomy of clustering methods and give a short explanation for each.

Clustering (Cluster Analysis) – unsupervised learning method, which aim is to reveal *structures* in unclassified data set. These structures are composed of data instances based on *dissimilarity* measure. This means that data instance in some cluster will have bigger degree of *dissimilarity* to instances from another clusters, than from the same one.

Clustering methods

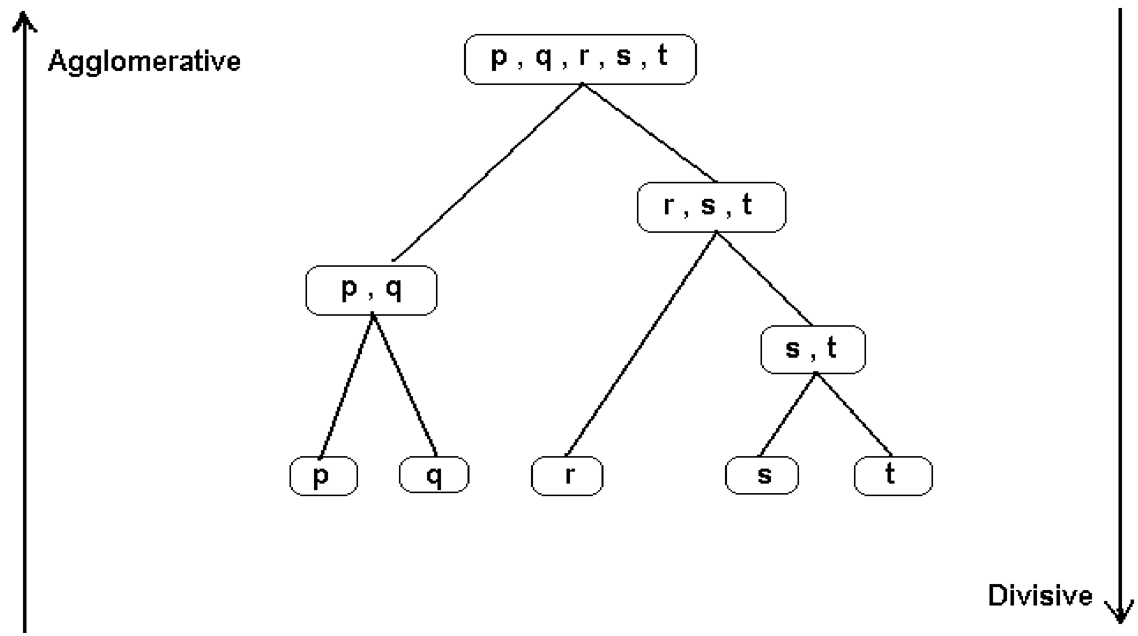


http://www.cs.gsu.edu/~wkim/index_files/SurveyParallelClustering.html

2. Present your understanding of the Hierarchical and Partitioning clustering approaches.

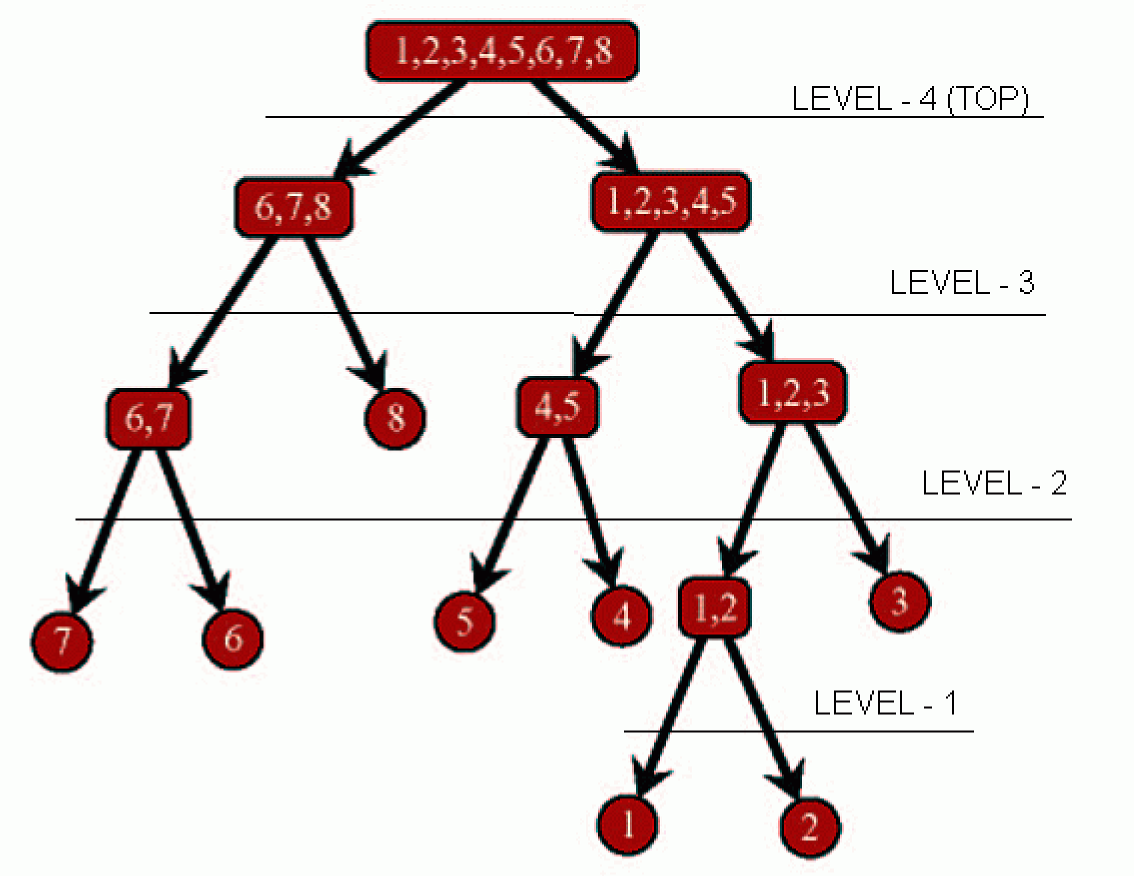
Hierarchical clustering

Hierarchical clustering – methods merge different clusters into single one based on similarity measures. Initially, each data instance is assigned to own cluster. There exist agglomerative and divisive approaches in hierarchical clustering.



http://www.solver.com/hierarchical-clustering-intro

Hierarchical clustering (2)



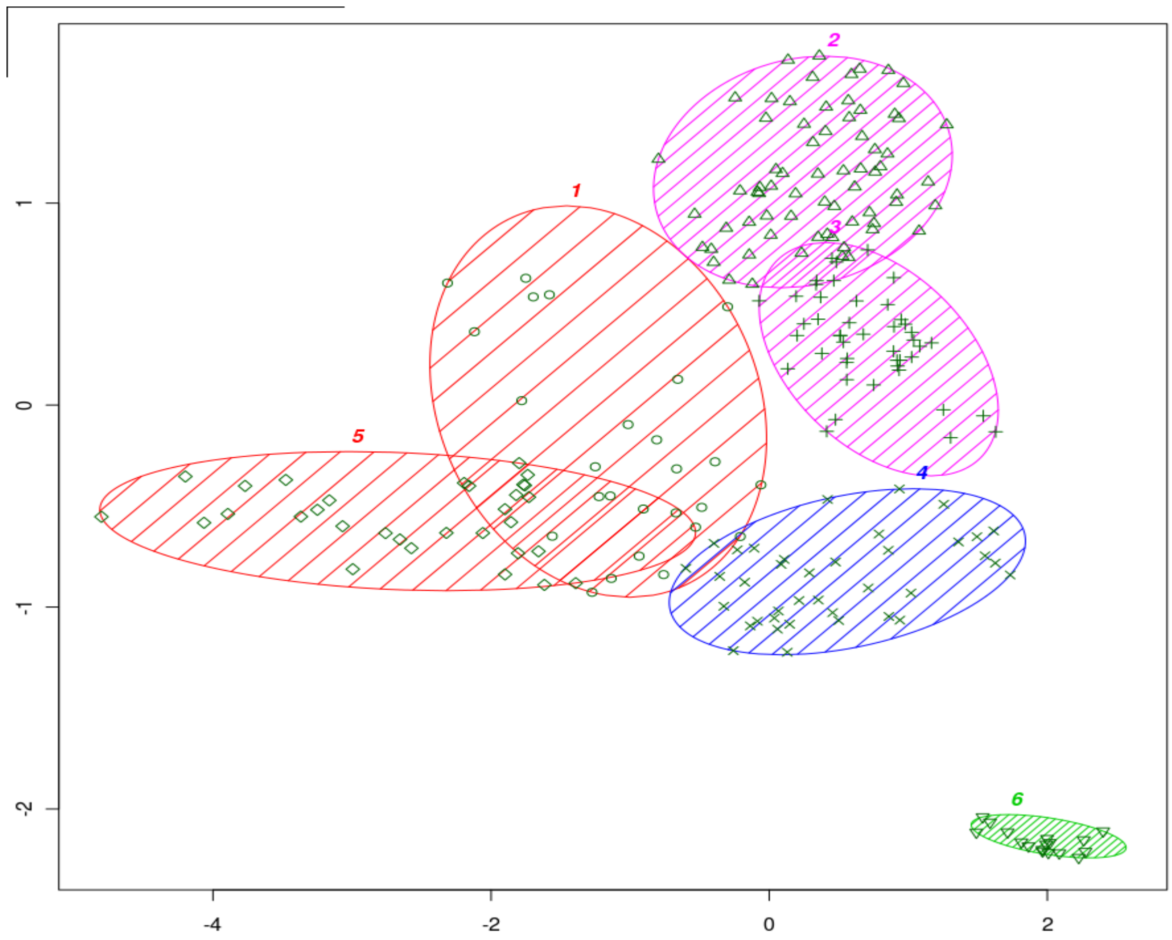
http://iv.slis.indiana.edu/sw/ward.html

Partitioning clustering

Partitioning clustering – methods reallocate given data instances into defined number of clusters. There are exist few convergence measures:

- compactness of cluster: measures dissimilarity between two instances in the same cluster,
- isolation of cluster: measures distance between a cluster and other clusters.

Partitioning clustering (2)



m.mlssoccer.com

3. What is K-means clustering. State the reasons why it can fail in some cases. What are the main parameters of this method?

K-mean configuration

- For performing optimal clustering K-means has the following requirements:
- - number of clusters K should be known before and the results of the method significantly depends on it,
- - initial centroids have to be place appropriately in order to get proper division.

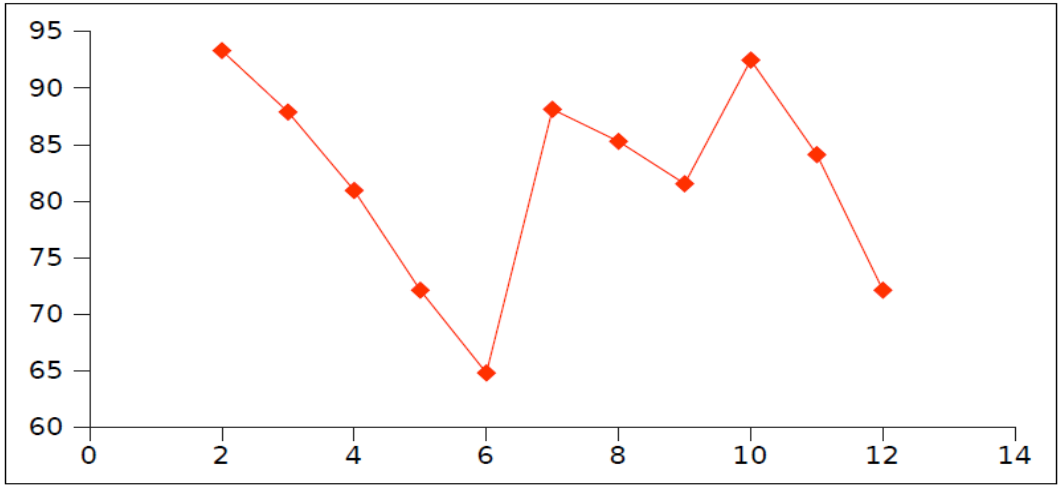
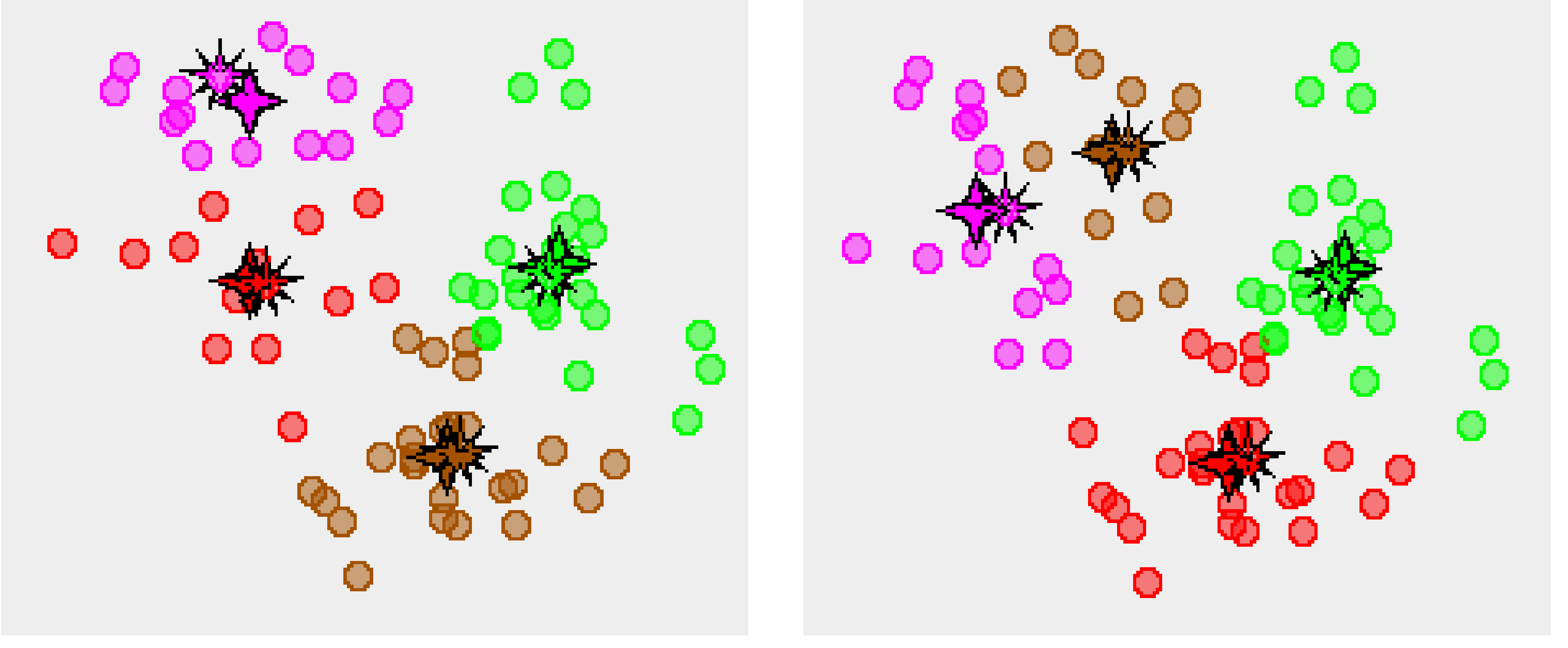


Figure 2: Dependency of RSS objective function on number of clusters

K-means clustering – initial centroids

- As you can see, we can get different results of cluster analysis for the same dataset, but different initial centroids.

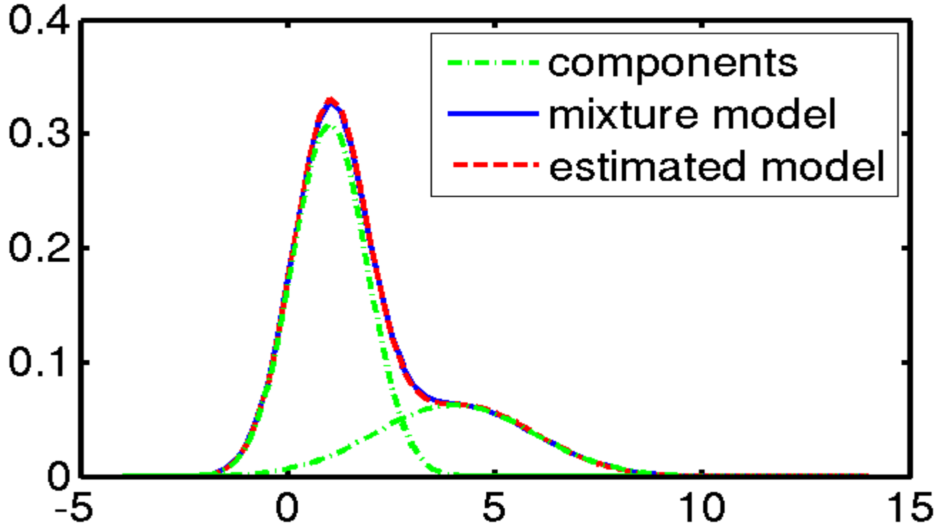


math.le.ac.uk

4. Explain the Gaussian Mixture Model.

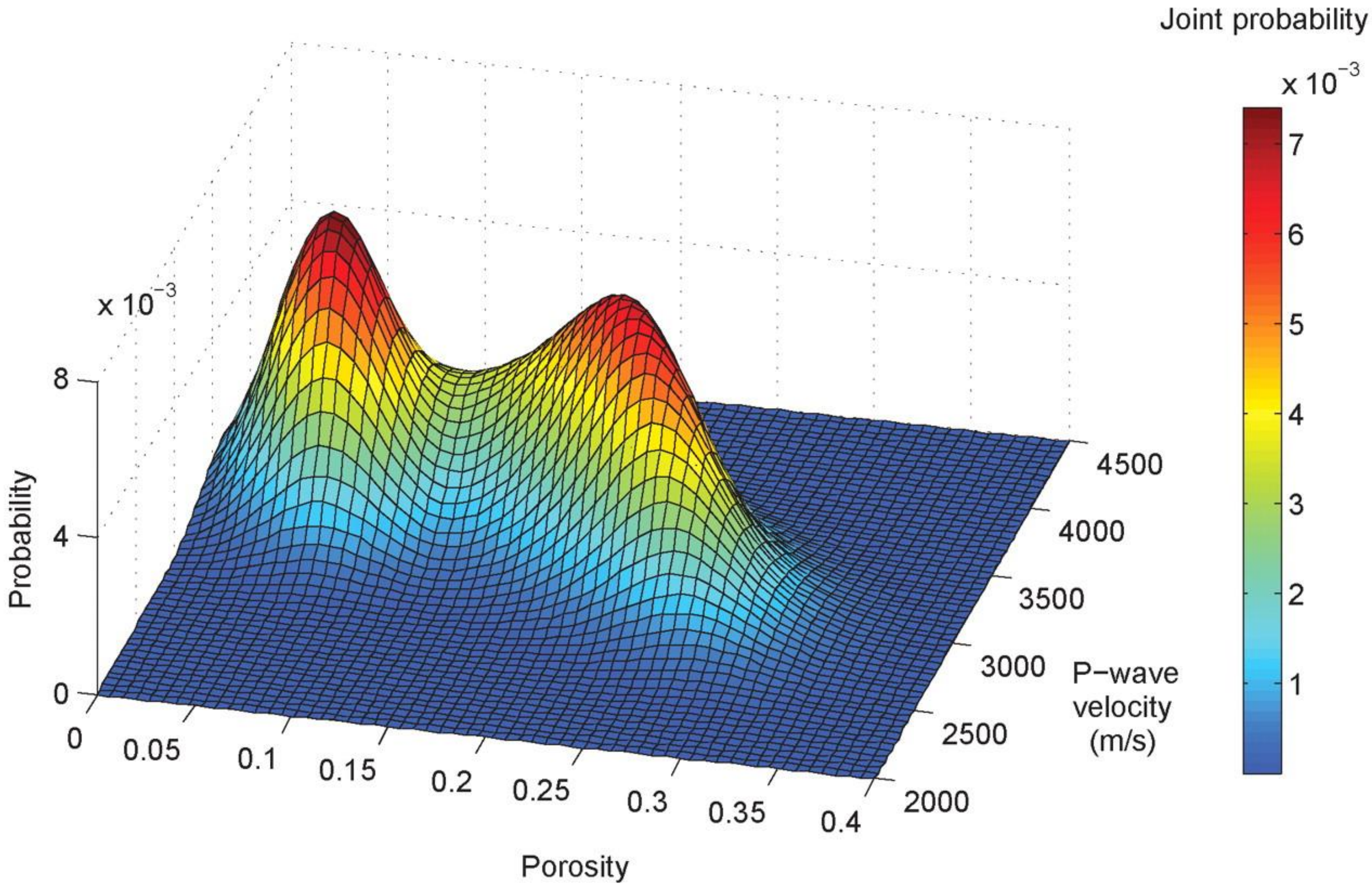
Gaussian mixture models (GMM)

Gaussian Mixture Models – clustering that is based on representation of each cluster as a convex parametric distribution **p(x)** over multiple features.



maths.adelaide.edu.au

Gaussian mixture models (2)



http://tle.geoscienceworld.org/content/30/1/54.abstract

5. What is the fuzzy clustering? Which techniques do you know?

Fuzzy clustering

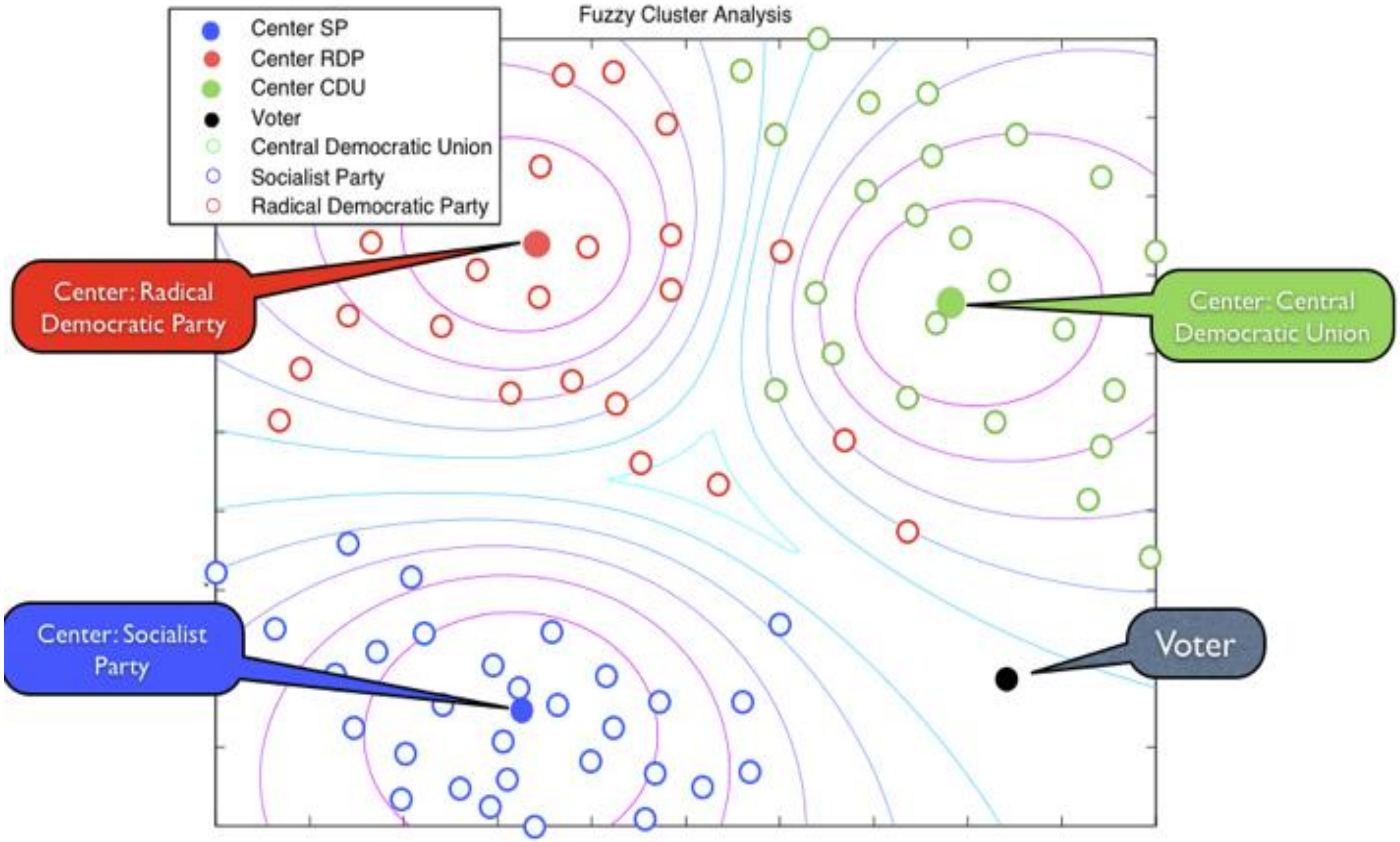
Fuzzy clustering provide a resolution for hard portioning data problem. It means that each data instance can belong to more than one cluster.

- In *c-means clustering*, the optimization problem is defined as following:

- *C* – number of clusters, *m* – degree of fuzziness, *μ* – degree of proximity (membership) of some data instance to a particular cluster.

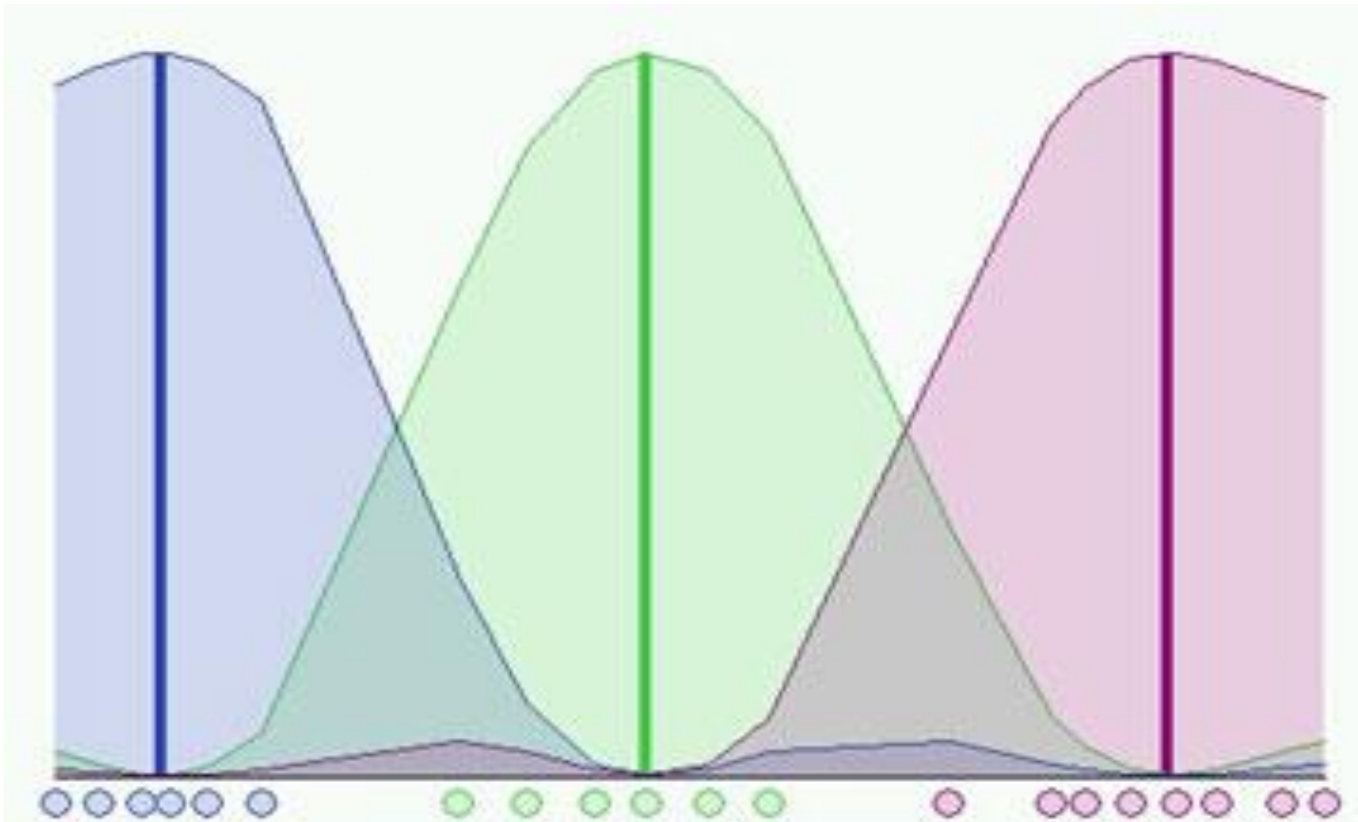
$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^{m} \left\| \mathbf{x}_i - \mathbf{v}_j \right\|^2$$

Fuzzy clustering (2)



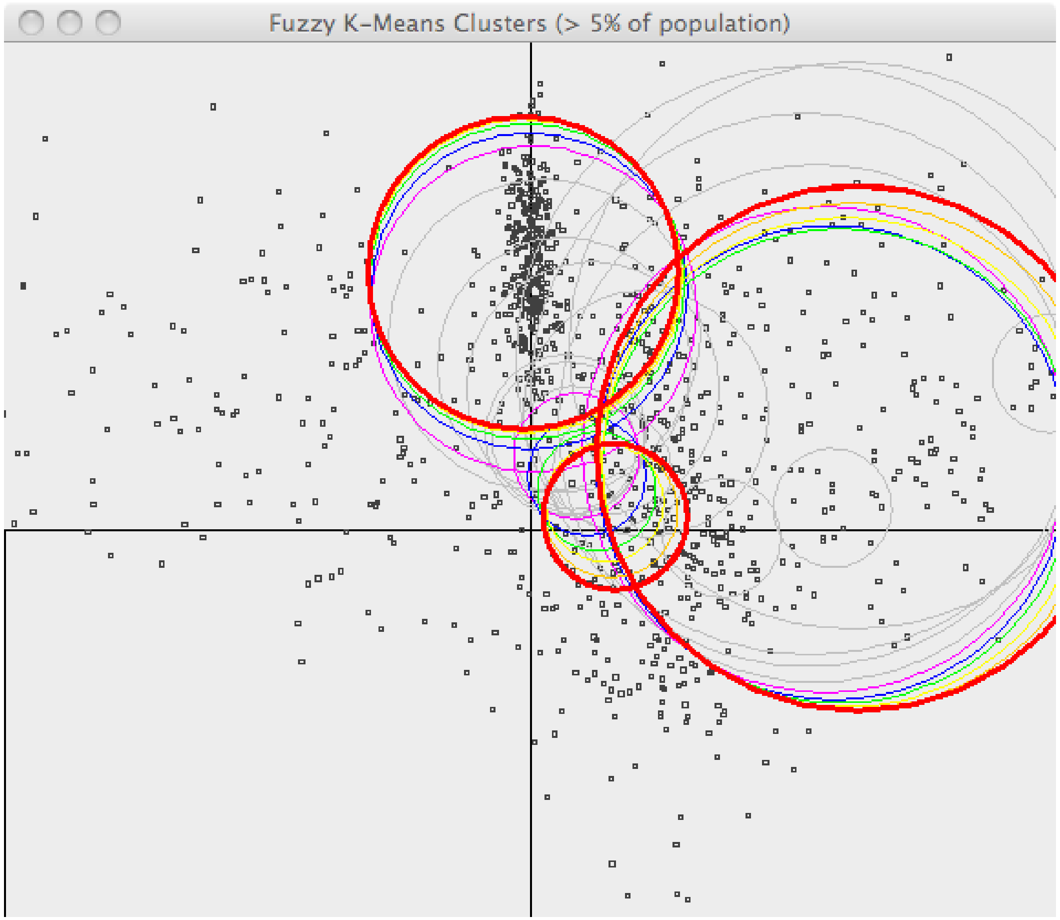
diuf.unifr.ch

Fuzzy clustering (3)



home.dei.polimi.it

Fuzzy c-means

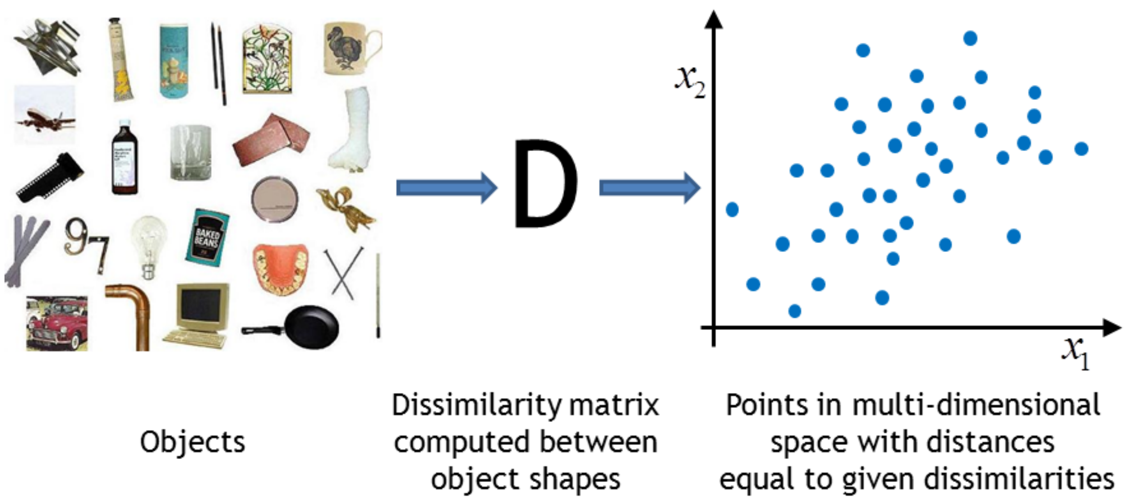


https://cwiki.apache.org/confluence/display/MAHOUT/Fuzzy+K-Means

6. What is the dissimilarity measure?

Dissimilarity measure

- In order to cluster the items in a data set, **the degree of association** between them is required.
- This may be a distance measure, or a measure of similarity or dissimilarity. Some clustering methods have a theoretical requirement for use of a specific measure (Euclidean distance, for example), but more commonly the choice of measure is at the discretion of the researcher.



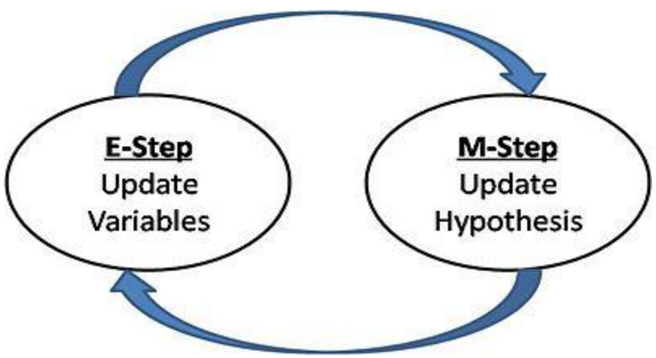
http://www.37steps.com/1264/dissimilarity-measures/

7. Give your understanding of Maximum likelihood estimation in Expectation Maximization algorithm.

Expectation maximization

The basic principle of Expectation maximization Clustering. The random initial model is fit to the observed data. Then, iteratively:

- **Expectation-Step:**
  - In the first iterations, the model changes substantially, but then converges to the given amount of modes. Replace the old estimates with the new ones.
- **Maximization-Step:**
  - Assume that the missing data values calculated in the previous step are correct, and calculate the new maximum likelihood hypothesis based on these values. Replace the old hypothesis with the new one, go to step E.



http://cse-wiki.unl.edu/wiki/index.php/Expectation_maximization

Maximum likelihood

Maximum likelihood is a method of estimation of the statistical model's parameters. First, we need to specify the joint probability density function for all given observations Xi.

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta).$$

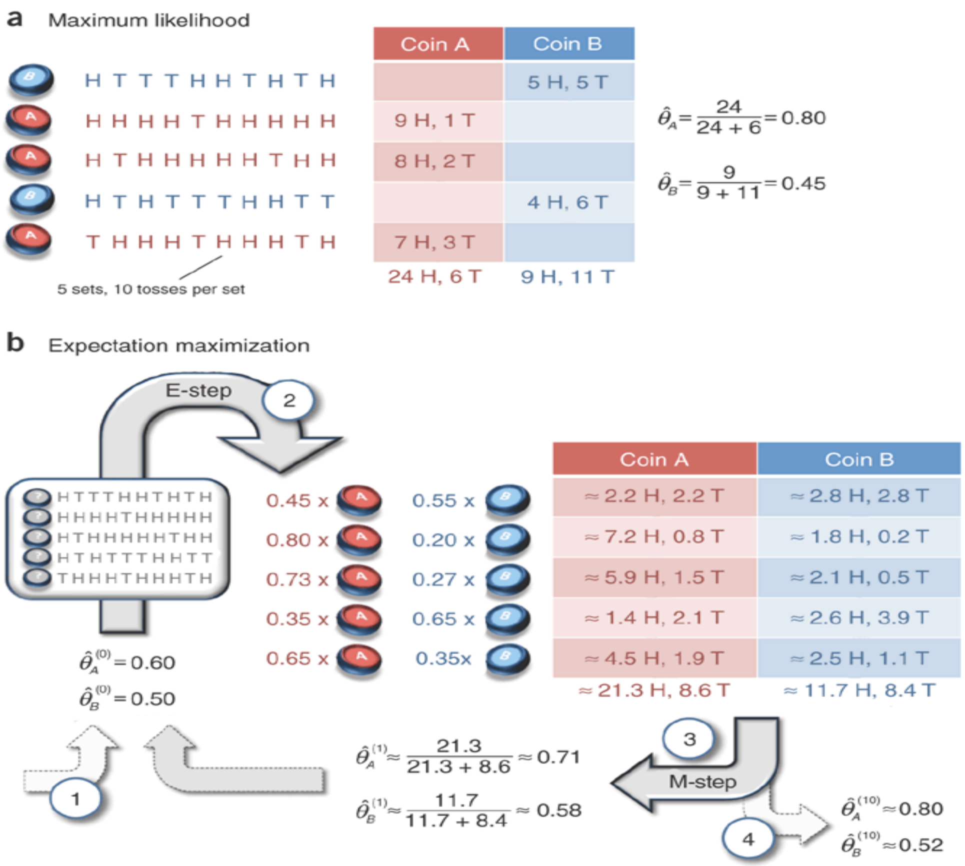
- Second, we vary the parameters 'teta' and calculate the maximum likelihood value:

$$\mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

- In practice it is more convenient to work with log-likelihood estimator. The higher value of the estimator denotes the better quality of the model:

$$\ln \mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta),$$

Example of the method



http://math.stackexchange.com/questions/25111/how-does-expectation-maximization-work

8. Perform k-means clustering for the following dataset that has 7 elements and 2 attributes. Group it in two clusters. Cross-validation obtained results using some statistical package.

Element ID	X	Y
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1 – k-means clustering

There have to be specified two important parameters for K-means success. First, we defined number of cluster as 2. Then, we should place initial centroids that will be used at the first step of K-means. For this purpose, we take two elements from the given dataset with largest distance:

	Individual	**Mean Vector (centroid)**
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2 – k-means clustering

Then, we allocate each elements from the initial cluster to one of two cluster based on the smallest measured distance:

	Cluster 1		Cluster 2	
Step	**Individual**	**Mean Vector (centroid)**	**Individual**	**Mean Vector (centroid)**
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Step 3 – k-means clustering

On the 3rd step we recalculate the coordinates of the centers (centroids) for each of two clusters:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

Step 4 – k-means clustering

Then, we repeat the step 2 and compare the distances between each elements in the given dataset and centroids of each cluster. If necessary, we repartition the clusters. For example, element 3 got the smallest distance to the Cluster 2 instead of Cluster 1:

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	**1.8**
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Step 5 – k-means clustering

On the 5th step we have obtained the following cluster partitioning. As you can see, the element 3 has migrated to the Cluster 2 from the Cluster 1. Proceeding iteratively further, we will not get any improvements. It means, that the following clusters configuration can be treated as an **optimal** one for given **number of clusters** and **initial centroids**.

	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

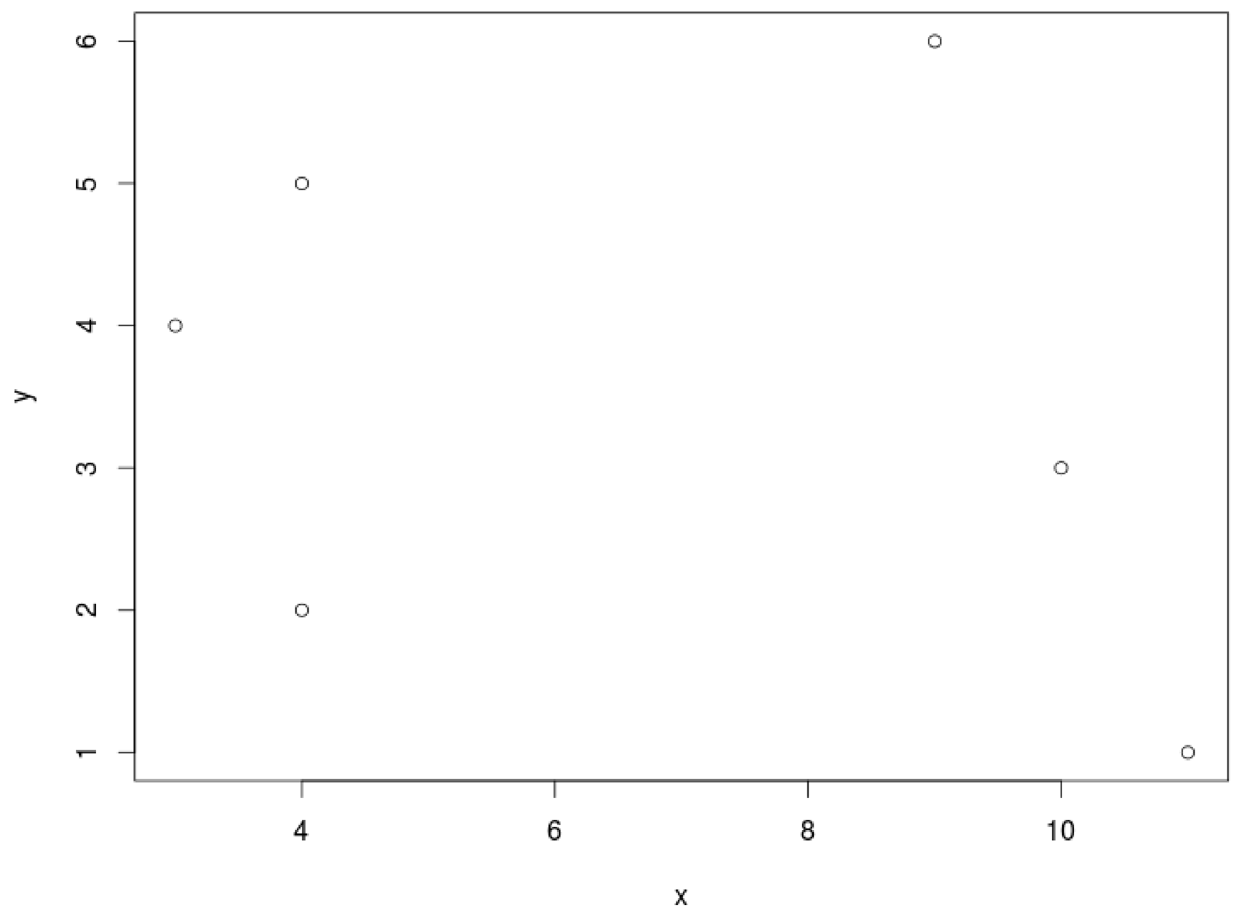
9. For this task perform hierarchical clustering until you will get two clusters in the final splitting. Reduce amount of clusters based on Euclidean distance between the objects groups on each step. Sketch provided objects in Cartesian coordinate system and estimate correctness of achieved splitting.

Objects.

A (3,4), B(4,5), C(9,6), D(10,3), E(11,1), F(4,2)

Finally, build a tree of the obtained clusters.

Plot of the given dataset



Step 1 – hierarchical clustering

On each step we should find two smallest Euclidean distances between the given points.

	A	B	C	D	E	F
A	0	1.41	6.32	7.07	8.54	2.23
B	1.41	0	5.09	6.32	8.06	3.00
C	6.32	5.09	0	3.16	5.38	6.40
D	7.07	6.32	3.16	0	2.23	6.08
E	8.54	8.06	5.38	2.23	0	7.07
F	2.23	3.00	6.40	6.08	7.07	0

Step 2 - hierarchical clustering

Then, we merge two points with the smallest distance (in our case AB and F) into a new ‘point’ ABF:

	AB	**C**	**D**	**E**	**F**
AB	0	5.09	6.32	8.06	2.23
C	5.09	0	3.16	5.38	6.40
D	6.32	3.16	0	2.23	6.08
E	8.06	5.38	2.23	0	7.07
F	2.23	6.40	6.08	7.07	0

Step 3 - hierarchical clustering

Same for the points ABF and E:

	ABF	C	D	E
ABF	0	5.09	6.08	7.07
C	5.09	0	3.16	5.38
D	6.08	3.16	0	2.23
E	7.07	5.38	2.23	0

Step 4 - hierarchical clustering

Iteratively, we merge the point C and the point DE into a new cluster CDE:

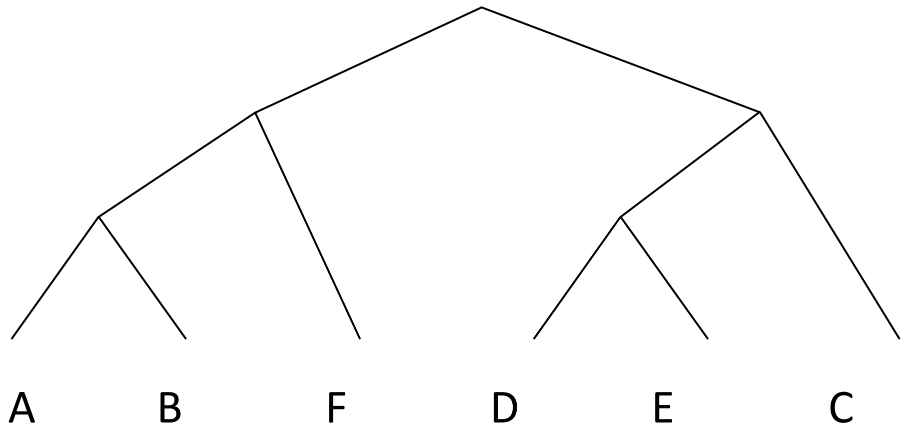
	ABF	C	DE
ABF	0	5.09	6.08
C	5.09	0	3.16
DE	6.08	3.16	0

Step 5 - hierarchical clustering

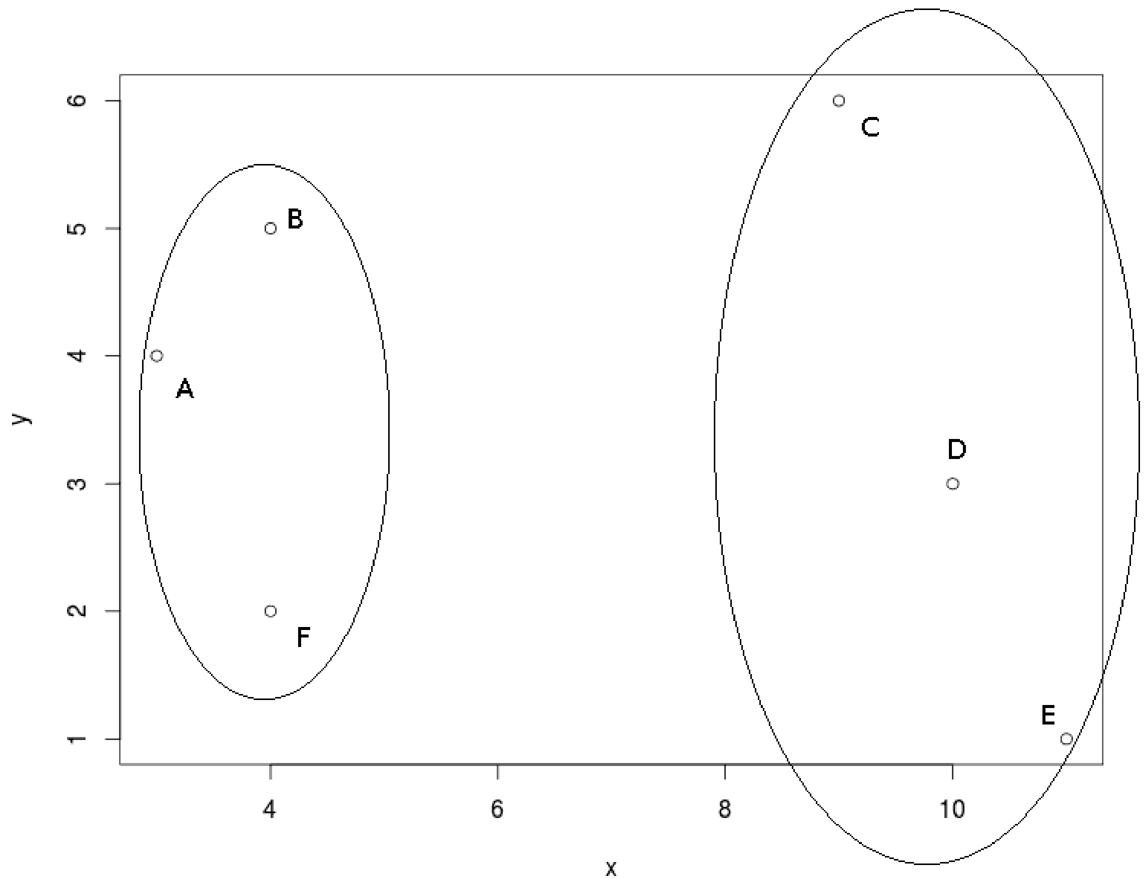
- Finally, we obtained two clusters ABF and CDE:

	ABF	CDE
ABF	0	5.09
CDE	5.09	0

- Also the clusters can be represented as a tree:

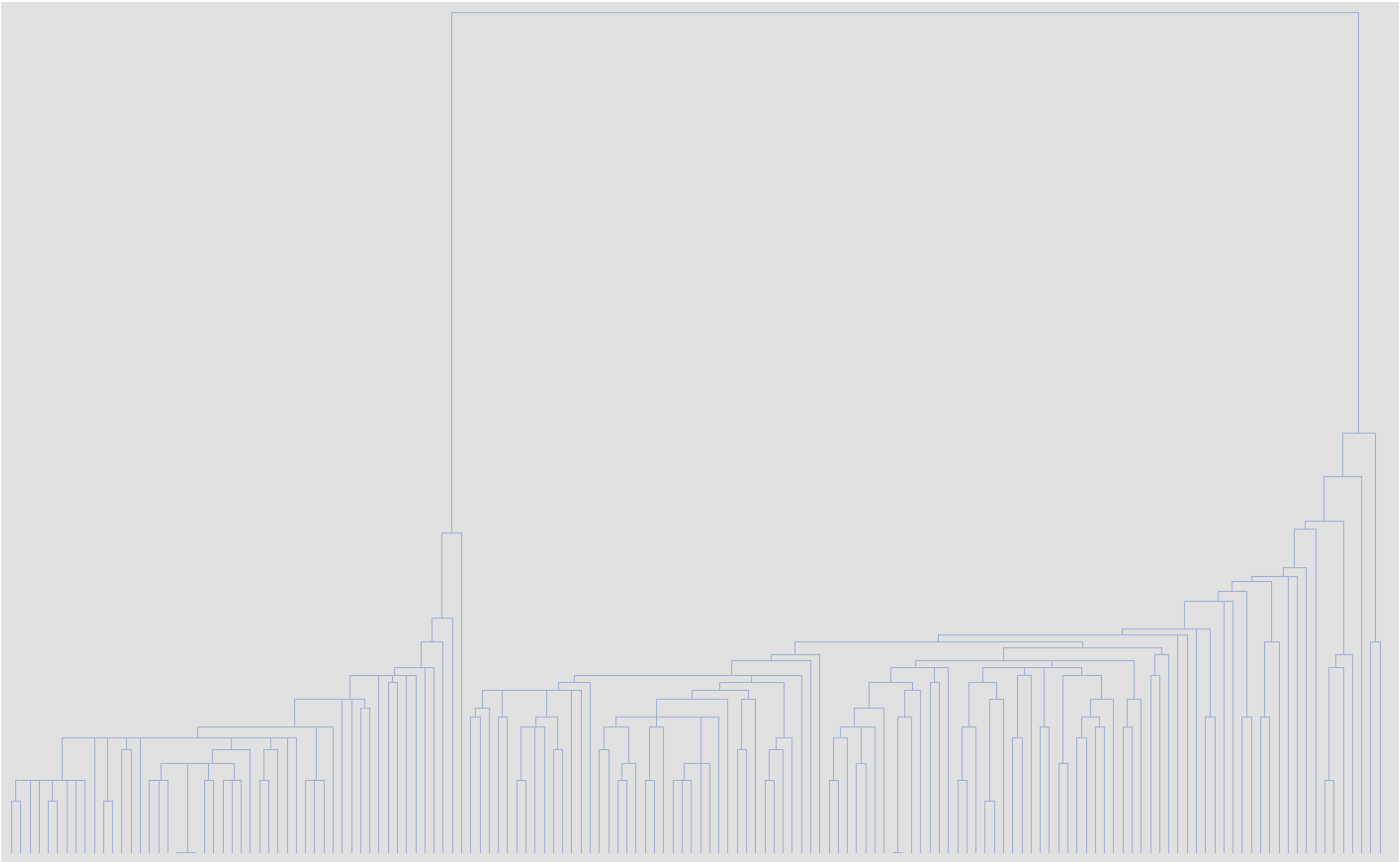


Plot of the results

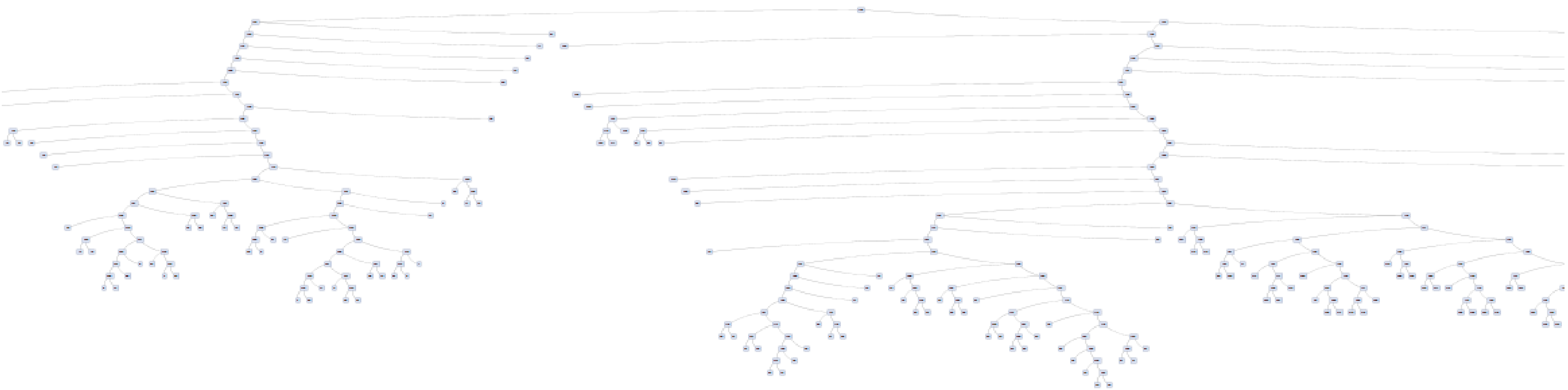


10. Use Iris dataset http://archive.ics.uci.edu/ml/datasets/Iris in order to perform the hierarchical agglomerative clustering. After building the model, present the tree or the dendrgoram of obtained results for the hierarchical clustering.

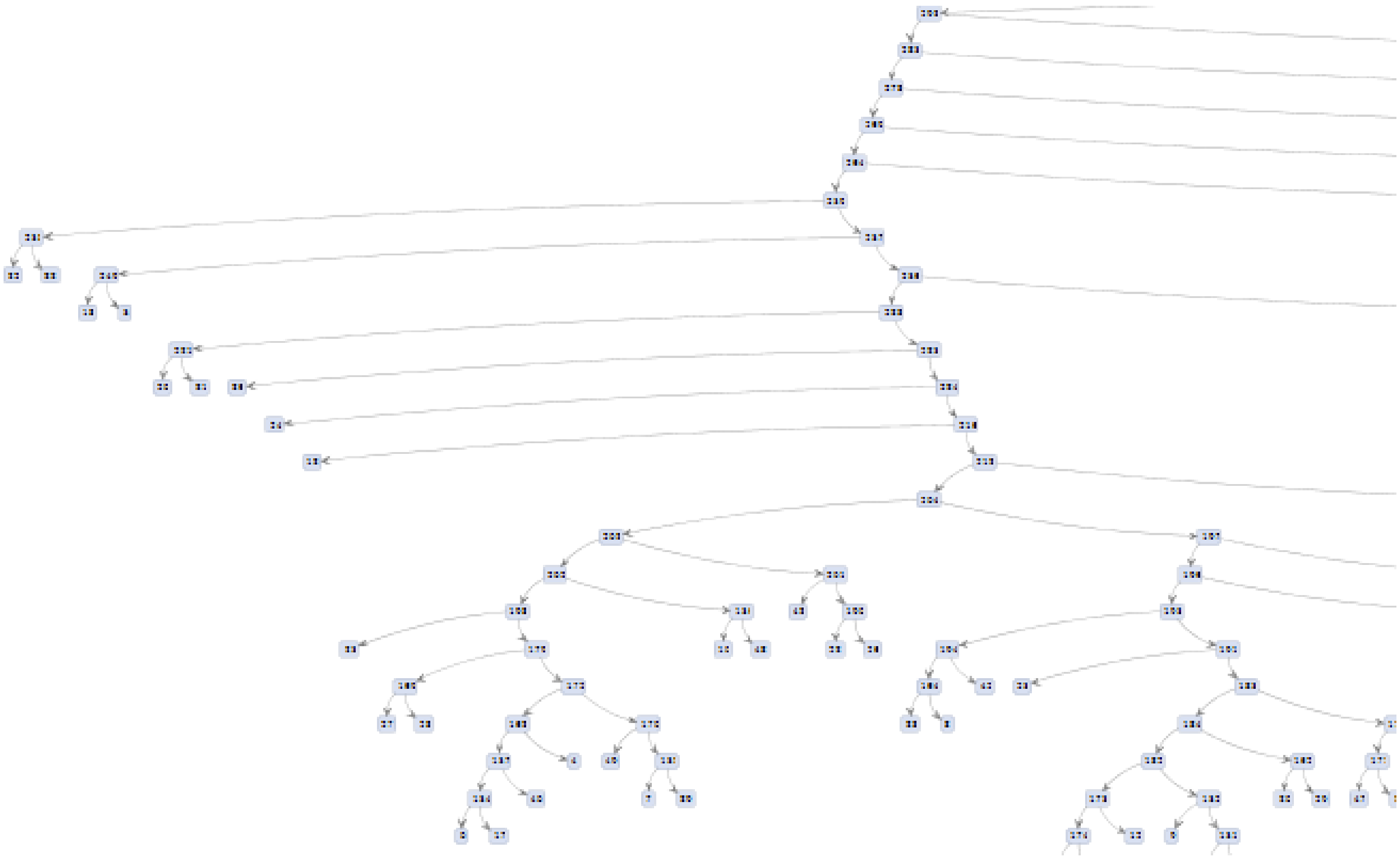
Hierarchical (agglomerative) clustering



Hierarchical (agglomerative) clustering (2)



Hierarchical (agglomerative) clustering (3)



Expectation maximization clustering

- Cross-Validation for automated detection of appropriate number of clsuters.
- Clustered Instances
- 0 28 (19%)
- 1 35 (23%)
- 2 42 (28%)
- 3 22 (15%)
- 4 23 (15%)

- Log likelihood: <u>-1.60803</u>

- 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 <u>40 %</u>

11. For this task download the Sponge data set https://archive.ics.uci.edu/ml/datasets/Sponge. You can use another dataset as well. Apply k-means and Expectation Maximization (EM) clustering methods for the dataset while changing the number of predefined clusters (just some in the range of 1-50). Compare the within cluster sum of squared errors for K-means and the log likelihood for EM. Which number of clusters give the best result? Apply EM with cross-validation in order to estimate the optimal number of clusters automatically by means of log-likelihood estimator.

K-means clustering

K=2
Number of iterations: 2
Within cluster sum of squared errors: **711.0667032163744**
Clustered Instances

0	38 (50%)
1	38 (50%)

K=10
Number of iterations: 4
Within cluster sum of squared errors: **487.05751923210806**
Clustered Instances

0	19 (25%)
1	6 (8%)
2	7 (9%)
3	5 (7%)
4	8 (11%)
5	10 (13%)
6	6 (8%)
7	5 (7%)
8	5 (7%)
9	5 (7%)

K-means clustering (2)

K=20
Number of iterations: 4
Within cluster sum of squared errors: 355.9685185185185
Clustered Instances

0	8	(11%)
1	3	(4%)
2	5	(7%)
3	3	(4%)
4	1	(1%)
5	5	(7%)
6	4	(5%)
7	5	(7%)
8	3	(4%)
9	4	(5%)
10	4	(5%)
11	3	(4%)
12	4	(5%)
13	3	(4%)
14	3	(4%)
15	8	(11%)
16	2	(3%)
17	4	(5%)
18	2	(3%)
19	2	(3%)

K-means clustering (3)

- K=50
- Number of iterations: 3
- Within cluster sum of squared errors: 133.38541666666669
- Clustered Instances
- 0 2 (3%)
- 1 1 (1%)
- 2 2 (3%)
- 3 1 (1%)
- 4 1 (1%)
- 5 2 (3%)
- 6 3 (4%)
- 7 1 (1%)
- 8 3 (4%)
- 9 1 (1%)
- 10 1 (1%)
- 11 1 (1%)
- 12 3 (4%)
- 13 2 (3%)
- 14 2 (3%)
- 15 2 (3%)
- 16 1 (1%)
- 17 2 (3%)
- 18 2 (3%)
- 19 1 (1%)

K-means clustering (4)

K=76 (the number of instances for training)
Number of iterations: 2
Within cluster sum of squared errors: <u>0.0</u>
Clustered Instances

0	1	(1%)
1	1	(1%)
2	1	(1%)
3	1	(1%)
4	1	(1%)
5	1	(1%)
6	1	(1%)
7	1	(1%)
8	1	(1%)
9	1	(1%)
10	1	(1%)
11	1	(1%)
12	1	(1%)
13	1	(1%)
14	1	(1%)
15	1	(1%)
16	1	(1%)
17	1	(1%)
18	1	(1%)
19	1	(1%)

Expectation maximization clustering

K=2
Clustered Instances
0 38 (50%)
1 38 (50%)
Log likelihood: -31.05676

K=10
Clustered Instances
0 16 (21%)
1 1 (1%)
2 12 (16%)
3 7 (9%)
4 18 (24%)
5 4 (5%)
6 3 (4%)
7 12 (16%)
9 3 (4%)
Log likelihood: **-27.62132**

Expectation maximization clustering (2)

K=20

Clustered Instances

0	2	(3%)
1	4	(5%)
2	3	(4%)
4	6	(8%)
5	5	(7%)
6	13	(17%)
7	2	(3%)
9	9	(12%)
10	4	(5%)
11	2	(3%)
12	6	(8%)
13	2	(3%)
14	6	(8%)
15	4	(5%)
16	2	(3%)
17	2	(3%)
18	1	(1%)
19	3	(4%)

Log likelihood: -30.61753

Expectation maximization clustering (3)

K=50 (empty clusters!)

Clustered Instances

0	1	(1%)
1	1	(1%)
2	2	(3%)
5	1	(1%)
11	2	(3%)
13	7	(9%)
14	1	(1%)
15	2	(3%)
20	1	(1%)
22	1	(1%)
23	3	(4%)
26	7	(9%)
29	1	(1%)
32	4	(5%)
33	18	(24%)
36	1	(1%)
39	2	(3%)
40	2	(3%)

Log likelihood: -29.05955

Expectation maximization clustering (4)

K=76 (empty clusters!)

Clustered Instances

0	1	(1%)
1	1	(1%)
4	1	(1%)
6	1	(1%)
8	3	(4%)
9	1	(1%)
10	1	(1%)
11	2	(3%)
12	7	(9%)
13	1	(1%)
14	1	(1%)
15	1	(1%)
16	1	(1%)
17	3	(4%)
18	1	(1%)
75	1	(1%)

Log likelihood: -37.90337

EM with cross-validation

According to https://archive.ics.uci.edu/ml/machine-learning-databases/sponge/sponge.info there are **12 classes**.

While applying an automated detection of cluster number, we can get the following results:

Clustered Instances
0 5 (7%)
1 13 (17%)
2 25 (33%)
3 24 (32%)
4 9 (12%)
Log likelihood: **-28.85811**

Practice - qwiklabs

Thank you for your attention!