Clustering

Lecturer: Dr. Bo Yuan

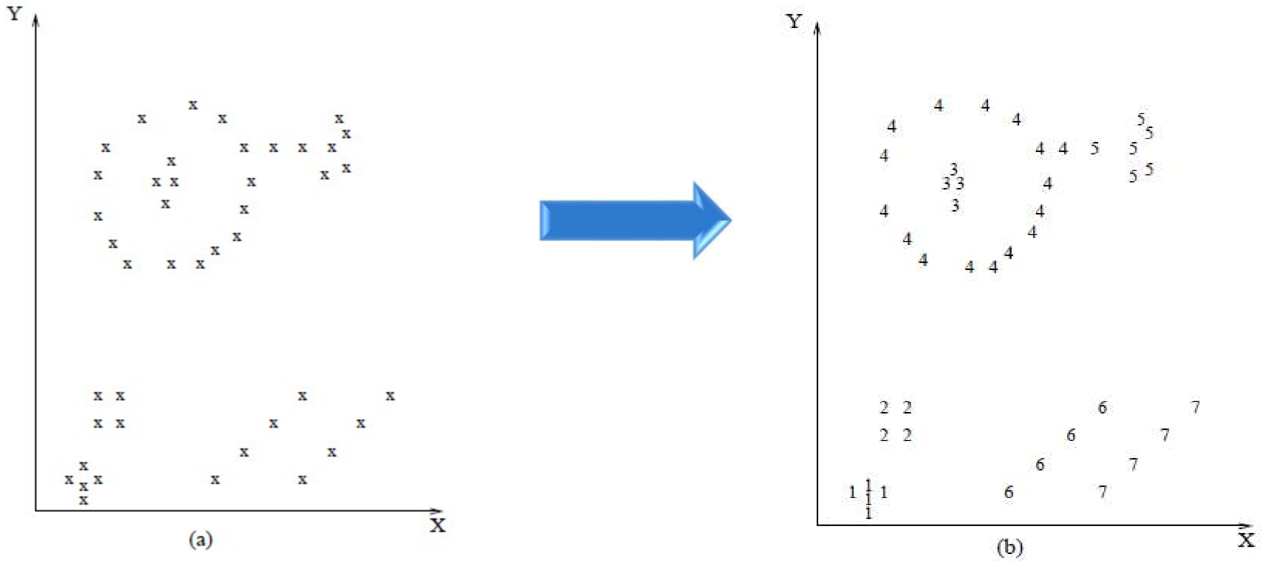
E-mail: yuanb@sz.tsinghua.edu.cn

Overview

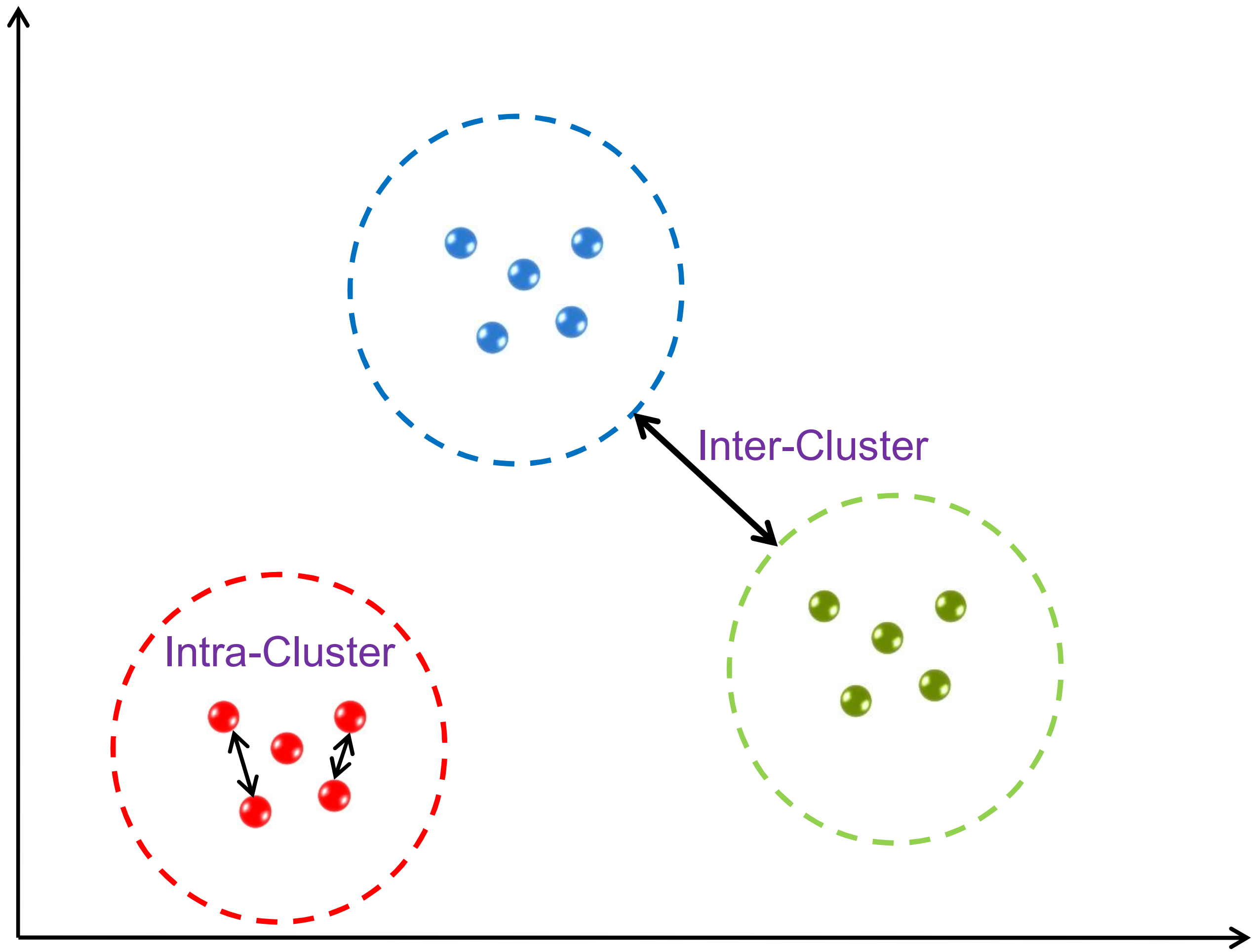
- Partitioning Methods
  - K-Means
  - Sequential Leader
  - Model Based Methods
  - Density Based Methods
- Hierarchical Methods

What is cluster analysis?

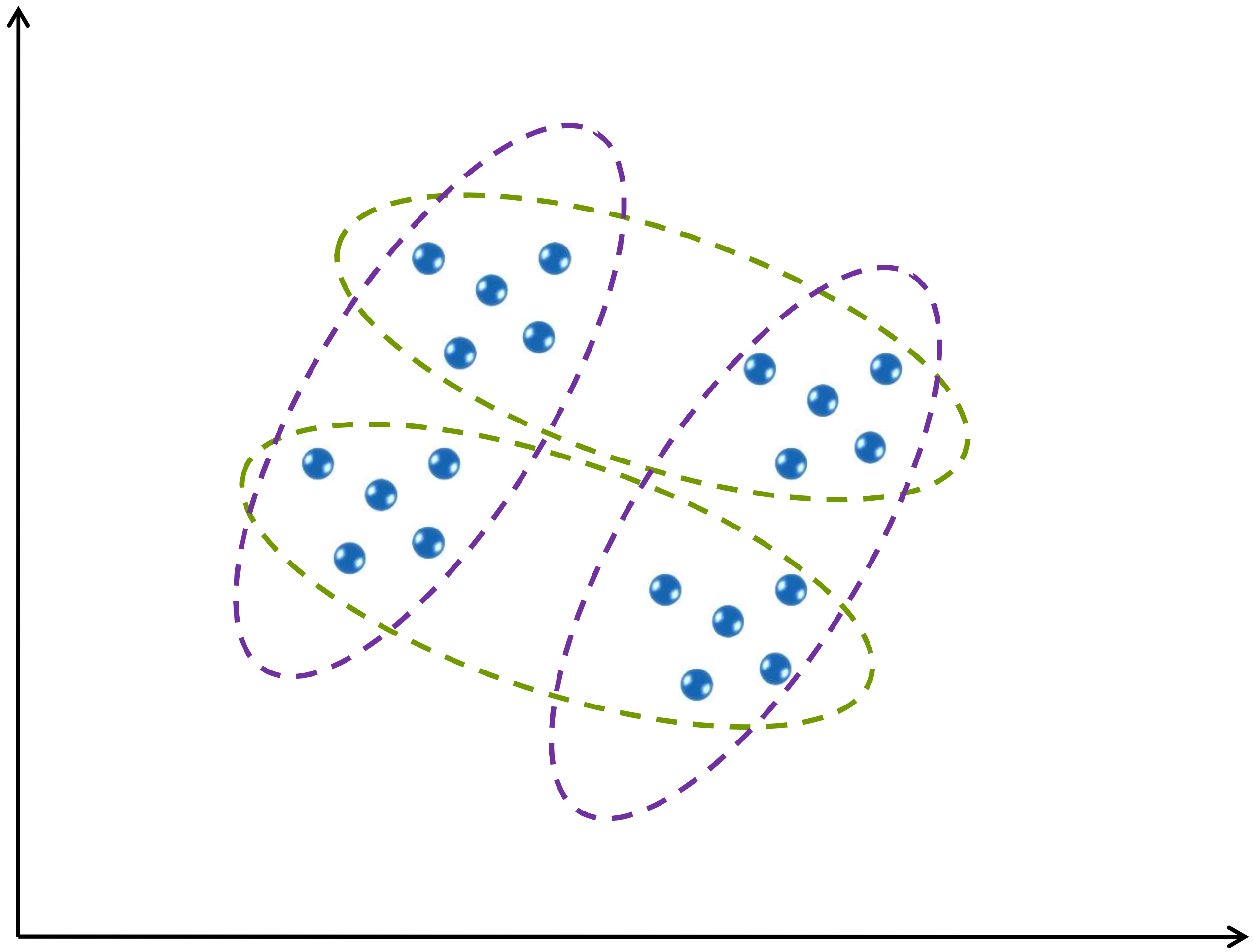
- Finding groups of objects
  - Objects similar to each other are in the same group.
  - Objects are different from those in other groups.
- Unsupervised Learning
  - No labels
  - Data driven



Clusters



Clusters



Applications of Clustering

- Marketing
  - Finding groups of customers with similar behaviours.
- Biology
  - Finding groups of animals or plants with similar features.
- Bioinformatics
  - Clustering microarray data, genes and sequences.
- Earthquake Studies
  - Clustering observed earthquake epicenters to identify dangerous zones.
- WWW
  - Clustering weblog data to discover groups of similar access patterns.
- Social Networks
  - Discovering groups of individuals with close friendships internally.

Earthquakes

GLOBAL SEISMIC HAZARD MAP

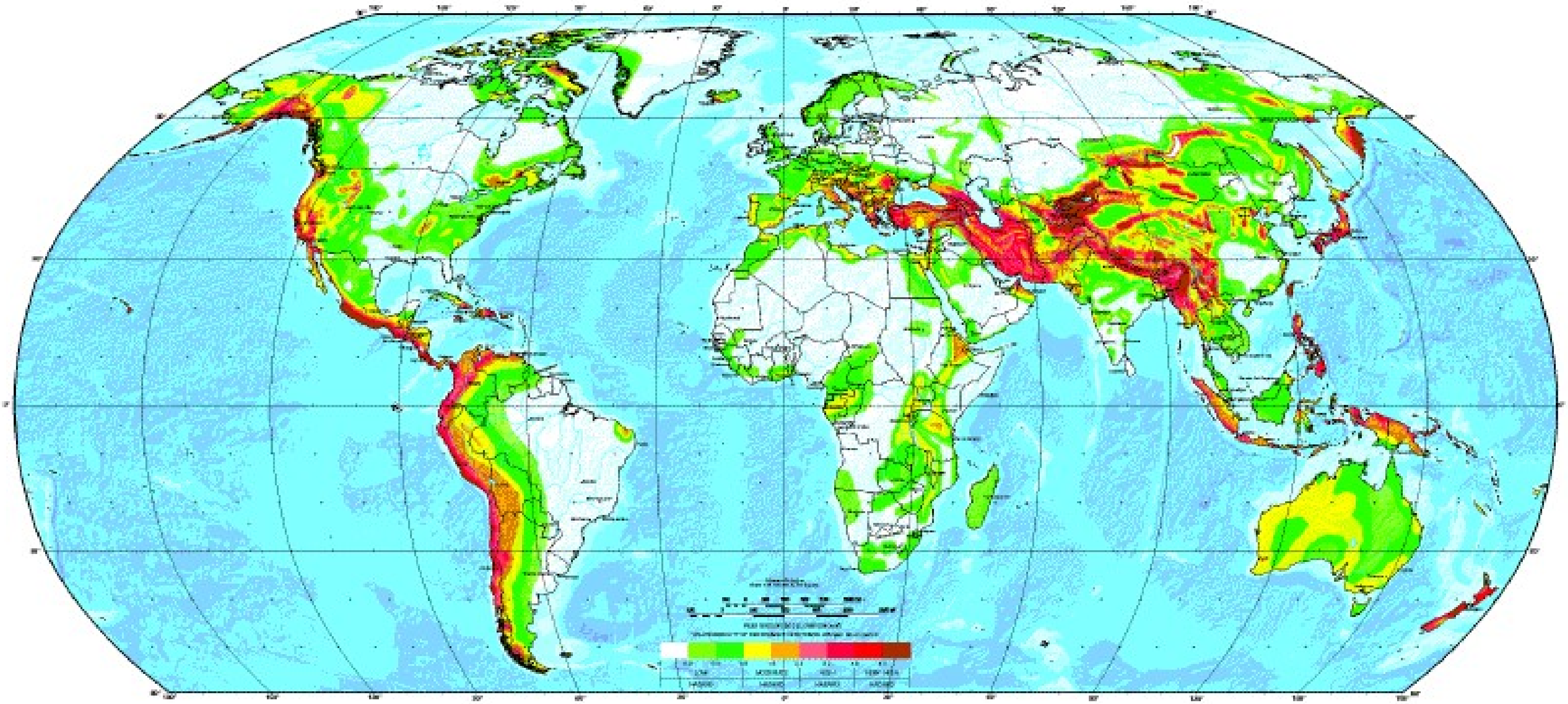
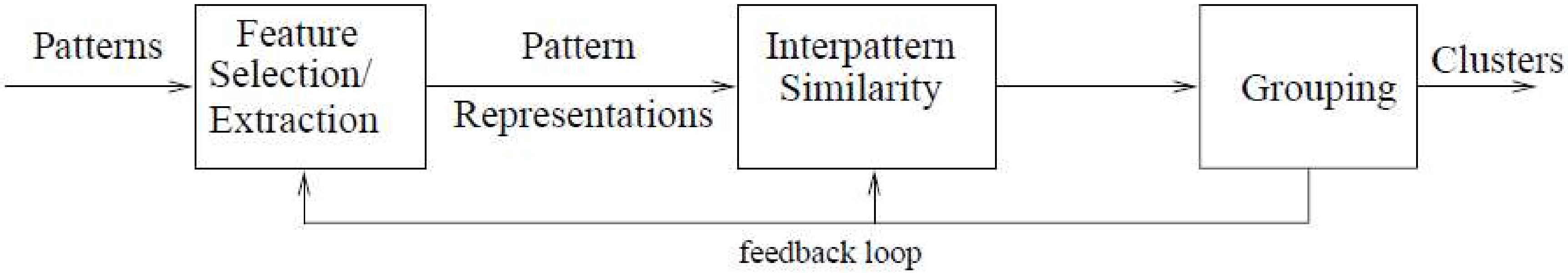


Image Segmentation

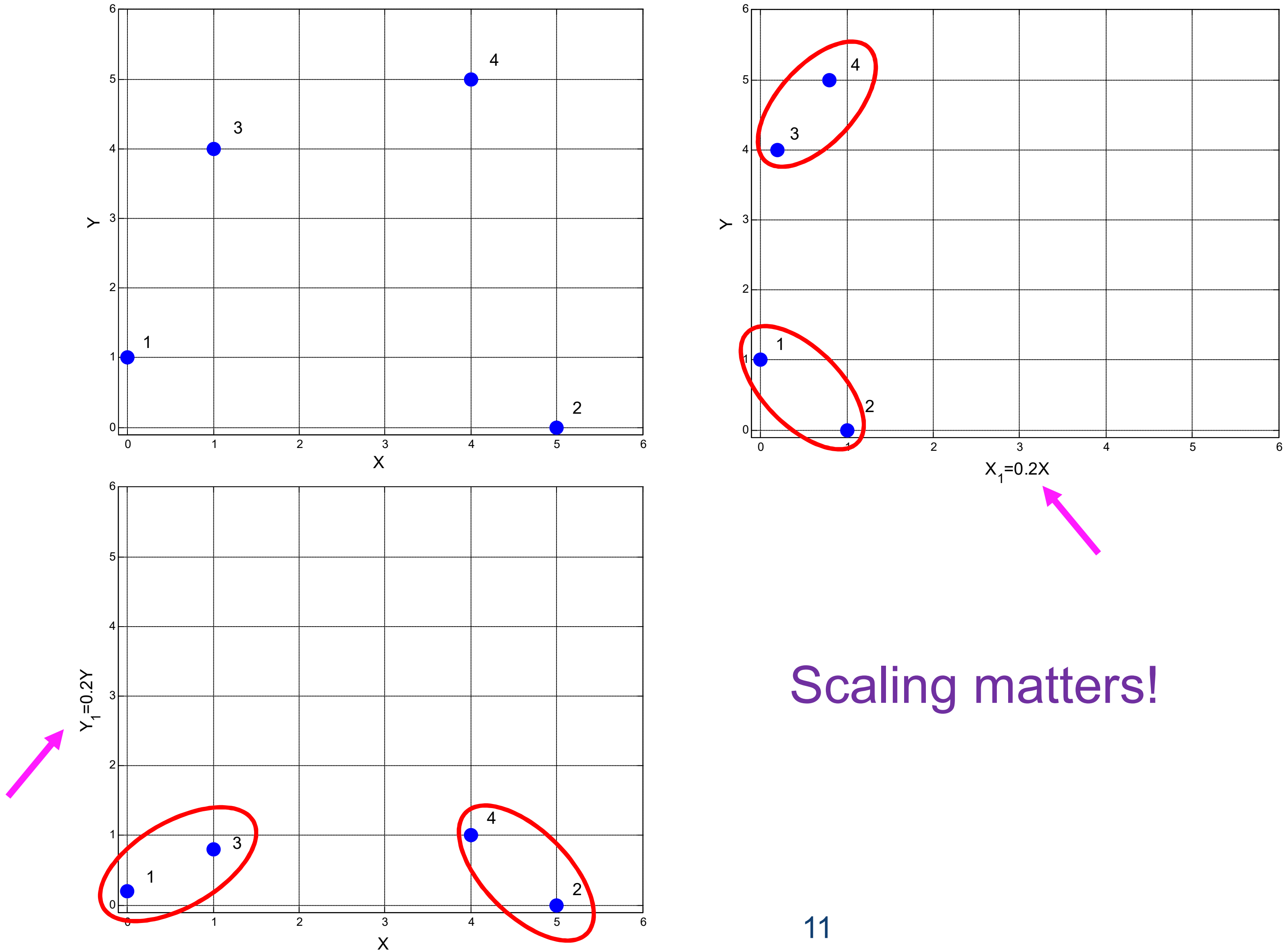
The Big Picture



Requirements

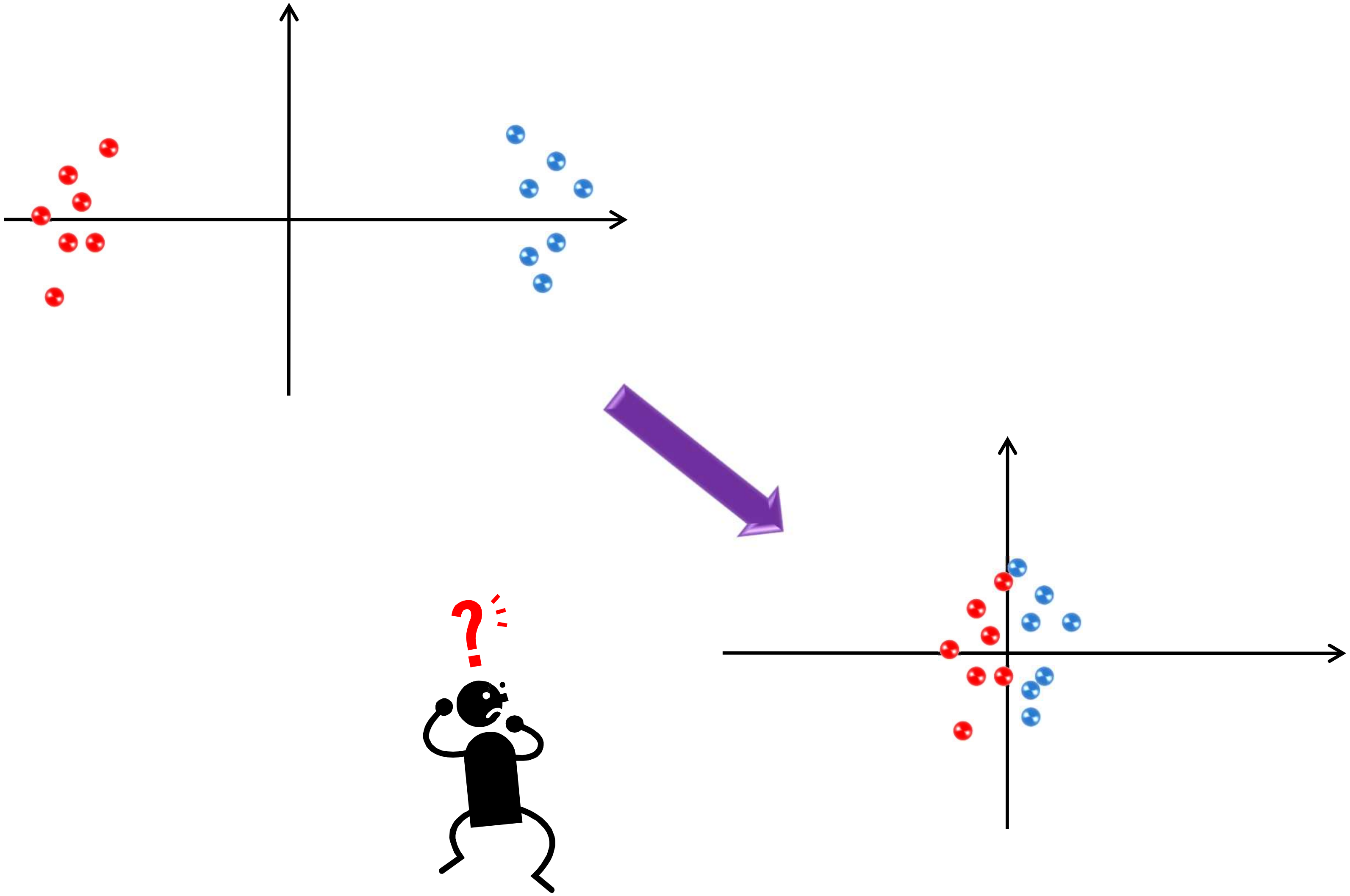
- Scalability
- Ability to deal with different types of attributes
- Ability to discover clusters with arbitrary shape
- Minimum requirements for domain knowledge
- Ability to deal with noise and outliers
- Insensitivity to order of input records
- Incorporation of user-defined constraints
- Interpretability and usability

Practical Considerations



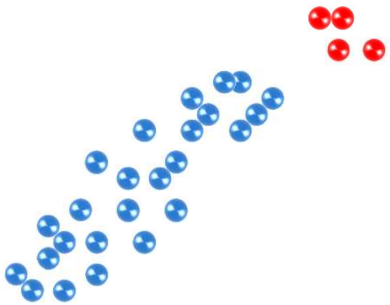
Scaling matters!

Normalization or Not?

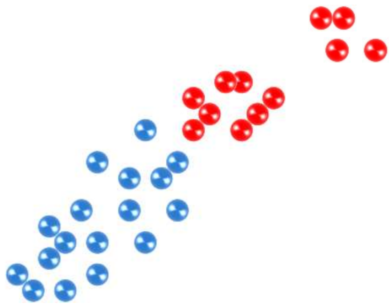


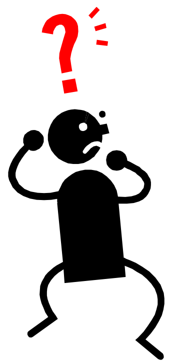
Evaluation

$$J_e = \sum_{i=1}^{c} \sum_{x \in D_i} \|x - m_i\|^2, \qquad m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

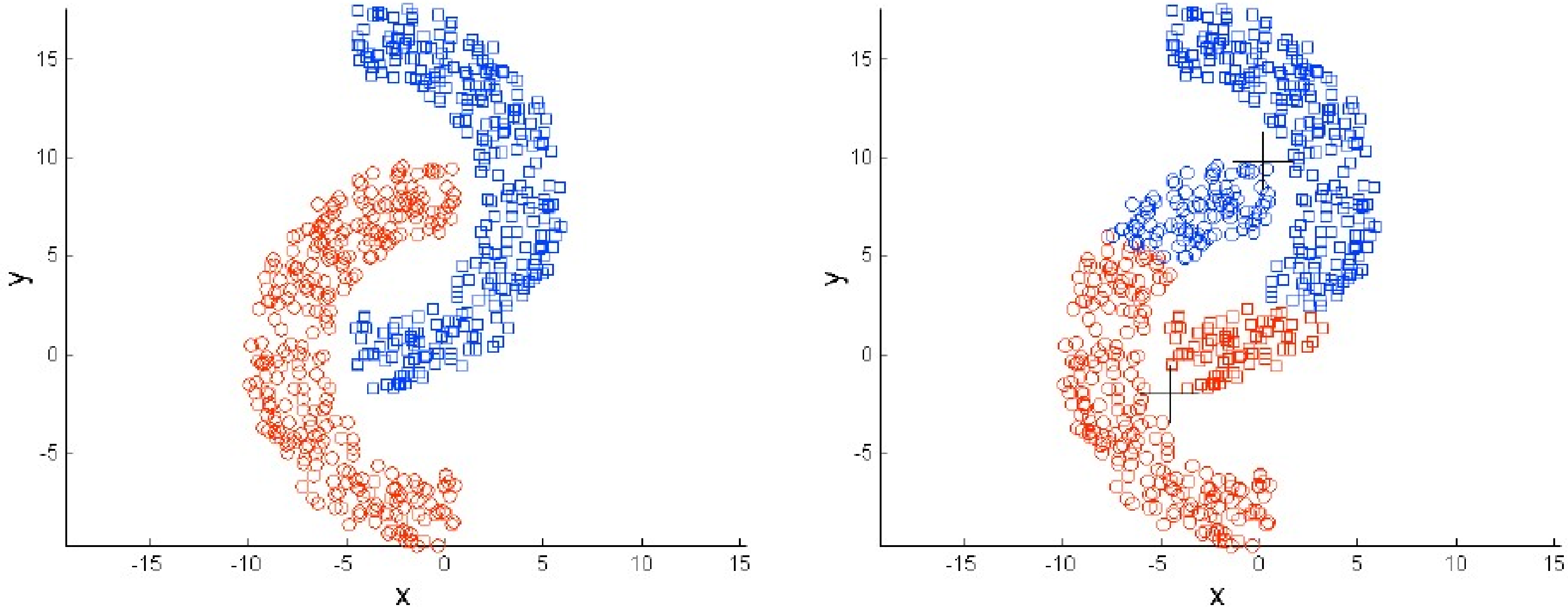


vs.





Evaluation

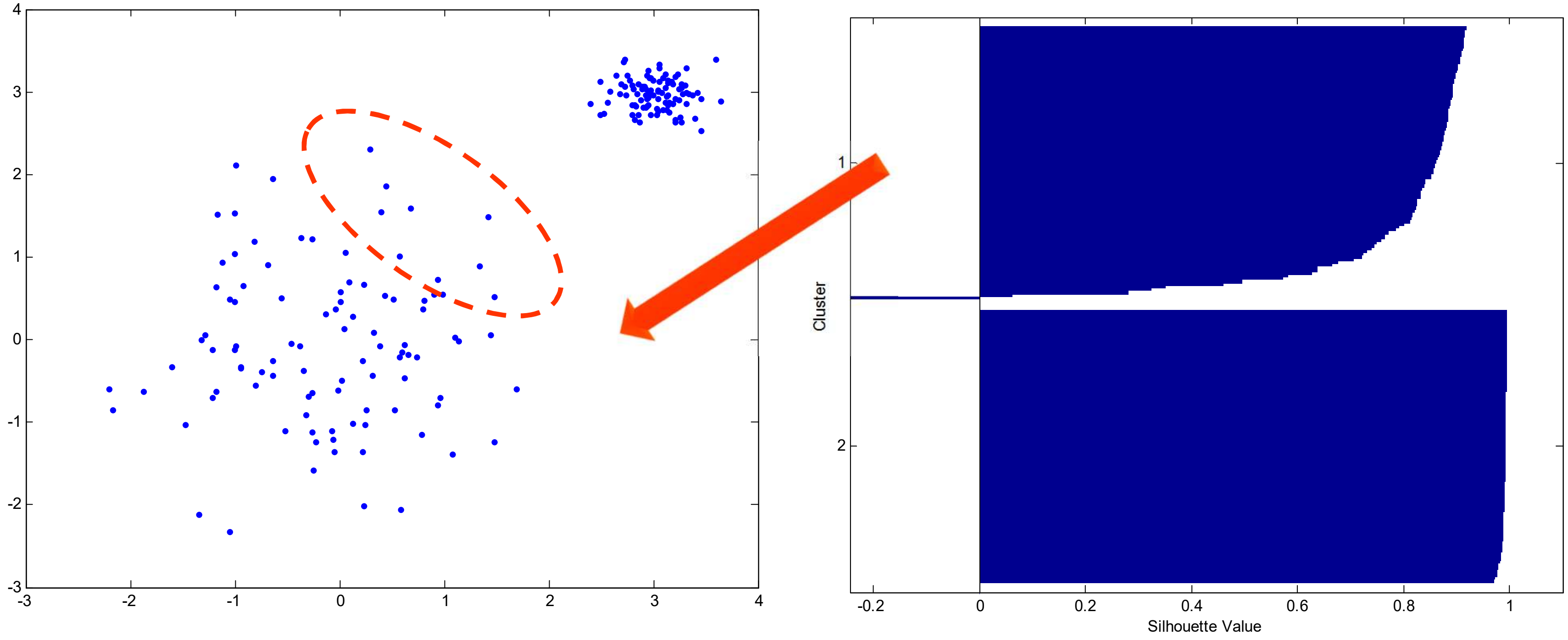


Silhouette

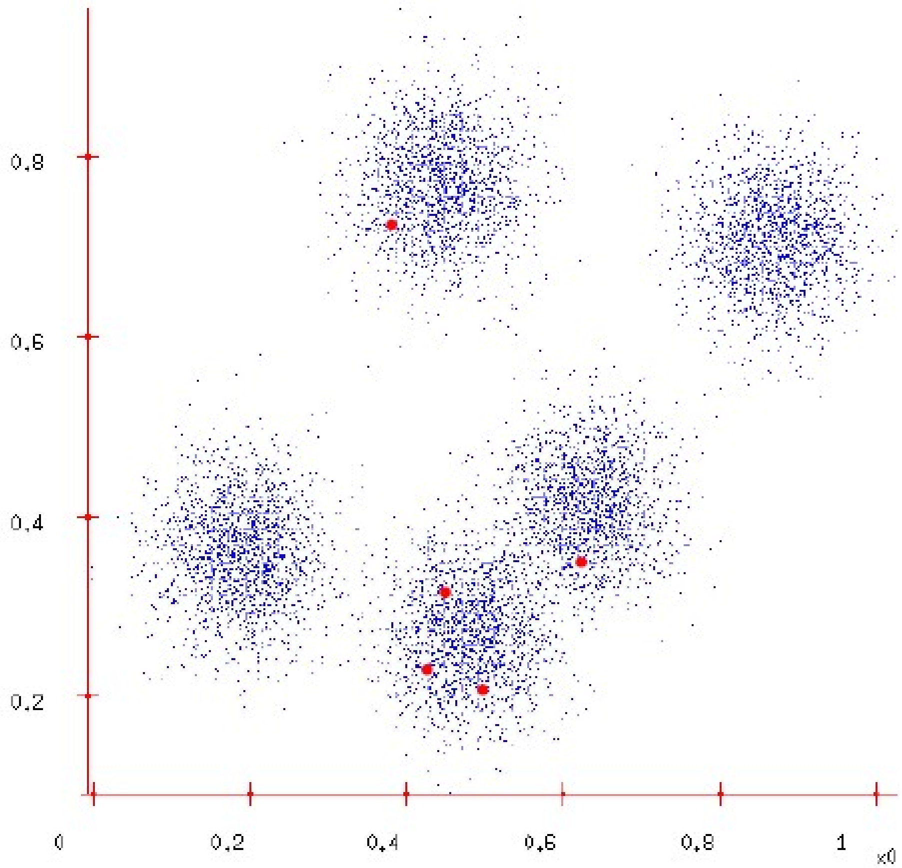
- A method of interpretation and validation of clusters of data.
- A succinct graphical representation of how well each data point lies within its cluster compared to other clusters.
- a(*i*): average dissimilarity of *i* with all other points in the same cluster
- b(*i*): the lowest average dissimilarity of *i* to other clusters

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

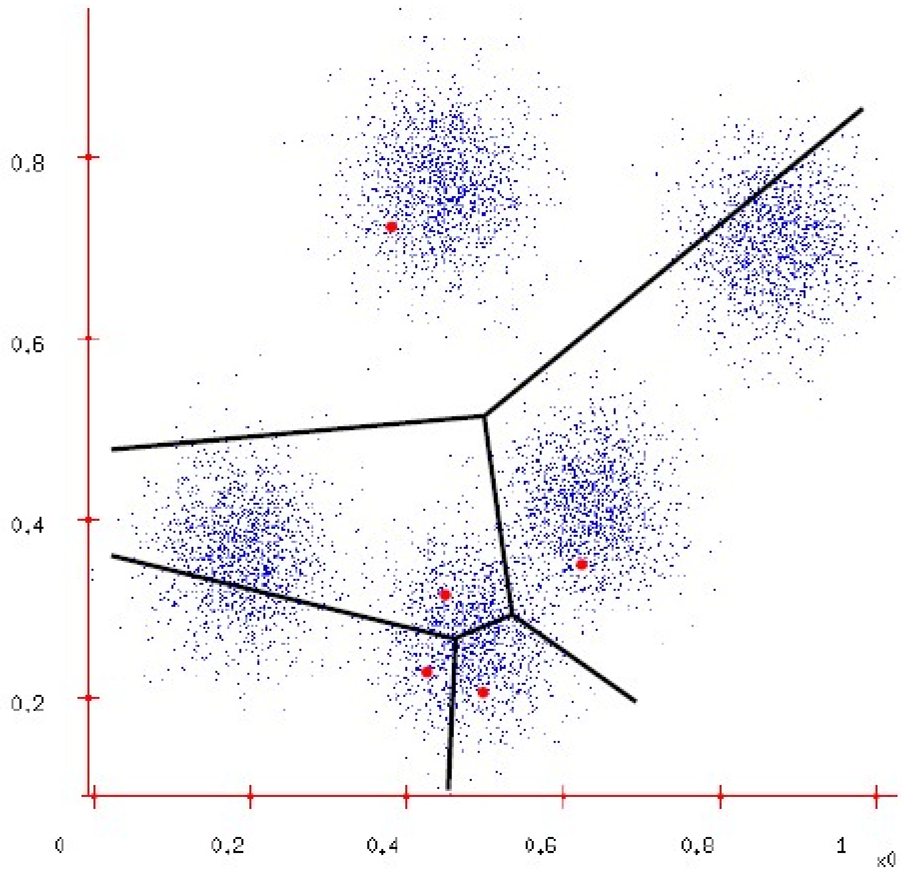
Silhouette



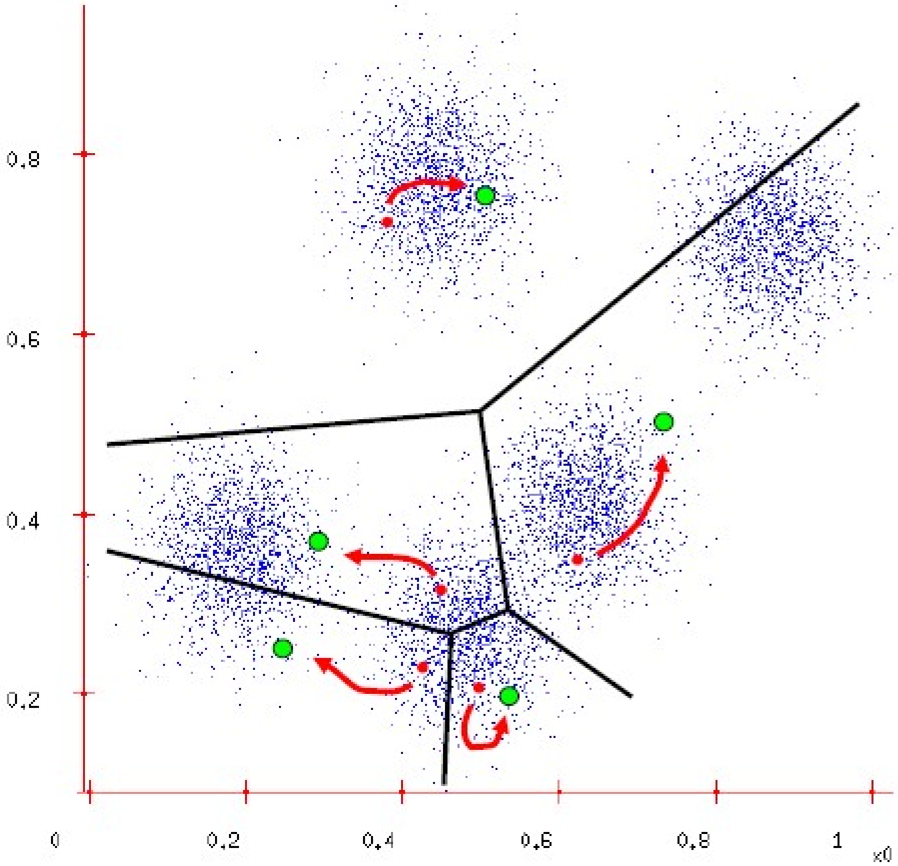
K-Means



K-Means



K-Means



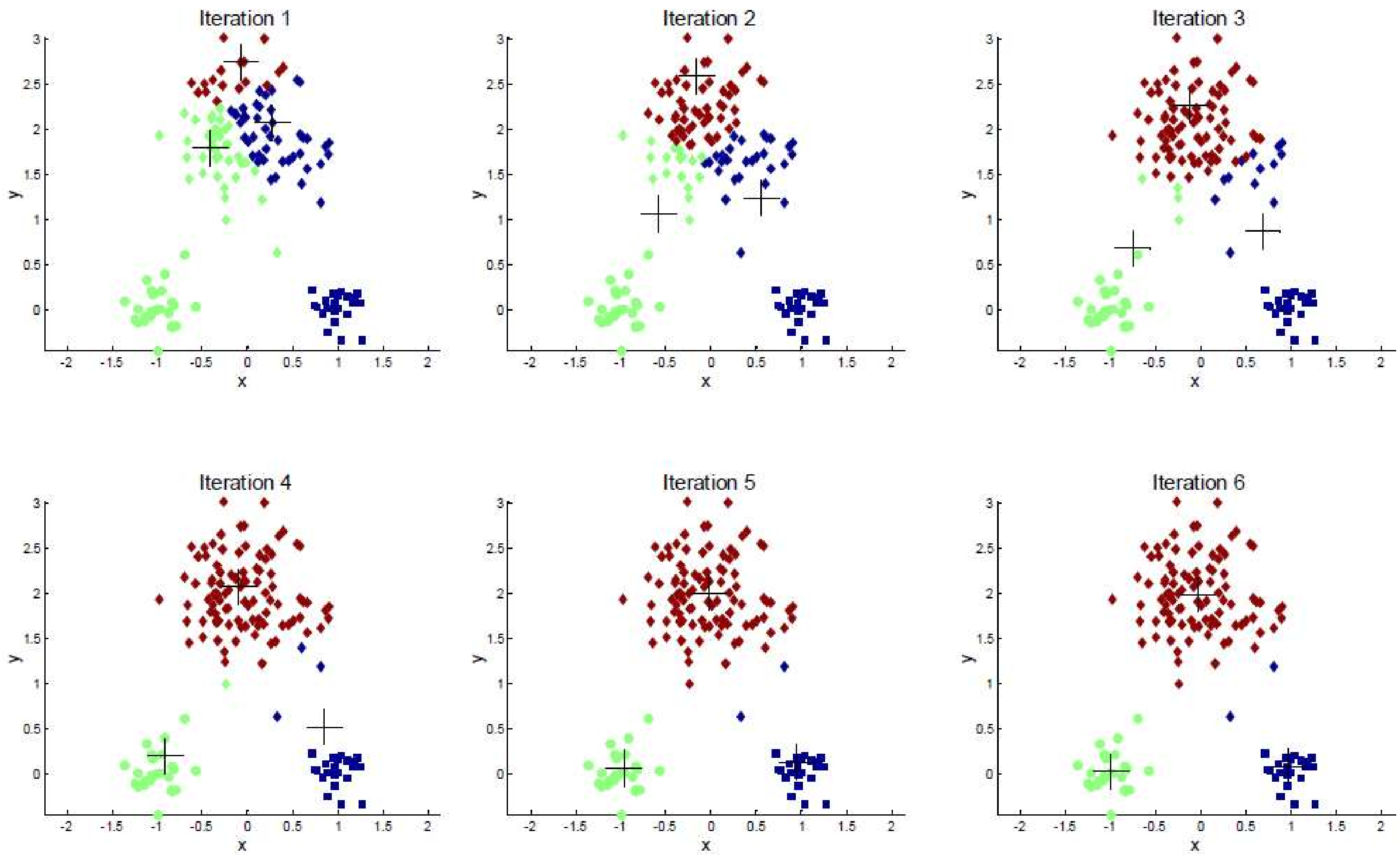
K-Means

- Determine the value of K.
- Choose K cluster centres randomly.
- Each data point is assigned to its closest centroid.
- Use the mean of each cluster to update each centroid.
- Repeat until no more new assignment.
- Return the K centroids.
- Reference
  - J. MacQueen (1967): "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol.1, pp. 281-297.

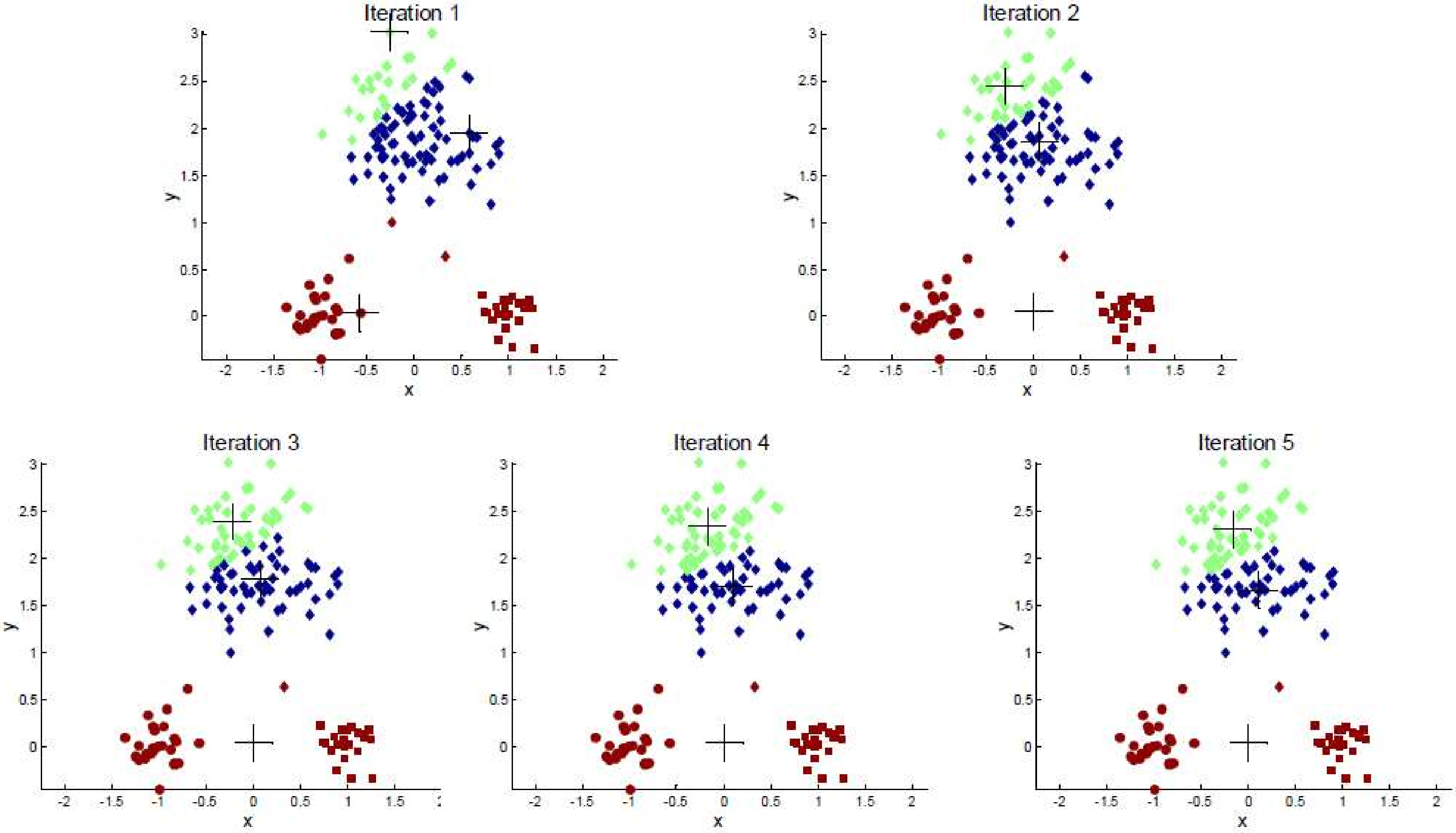
Comments on K-Means

- Pros
  - Simple and works well for regular disjoint clusters.
  - Converges relatively fast.
  - Relatively efficient and scalable $O(t \cdot k \cdot n)$
    - t: iteration; k: number of centroids; n: number of data points
- Cons
  - Need to specify the value of K in advance.
    - Difficult and domain knowledge may help.
  - May converge to local optima.
    - In practice, try different initial centroids.
  - May be sensitive to noisy data and outliers.
    - Mean of data points …
  - Not suitable for clusters of
    - Non-convex shapes

The Influence of Initial Centroids



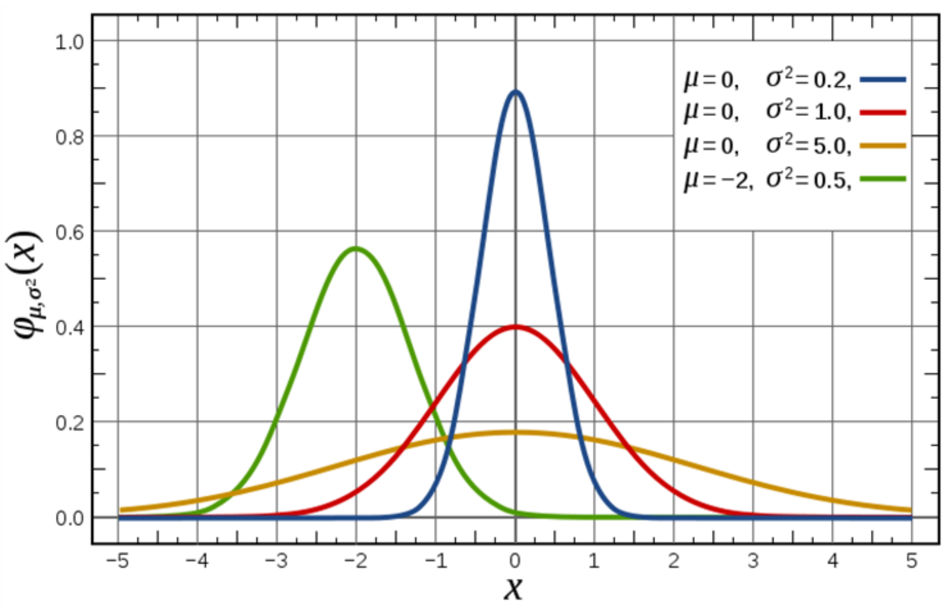
The Influence of Initial Centroids

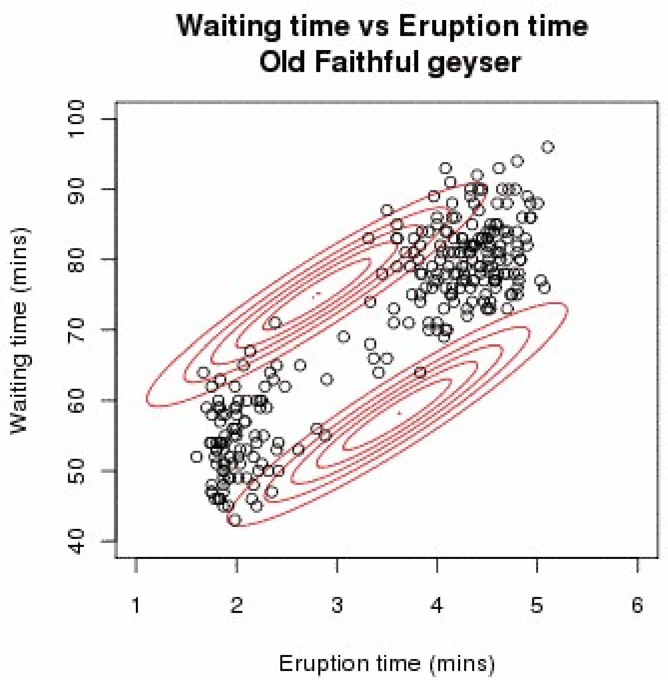


Sequential Leader Clustering

- A very efficient clustering algorithm.
  - No iteration
  - A single pass of the data
- No need to specify K in advance.
- Choose a cluster threshold value.
- For every new data point:
  - Compute the distance between the new data point and every cluster's centre.
  - If the minimum distance is smaller than the chosen threshold, assign the new data point to the corresponding cluster and re-compute cluster centre.
  - Otherwise, create a new cluster with the new data point as its centre.
- Clustering results may be influenced by the sequence of data points.

Gaussian Mixture

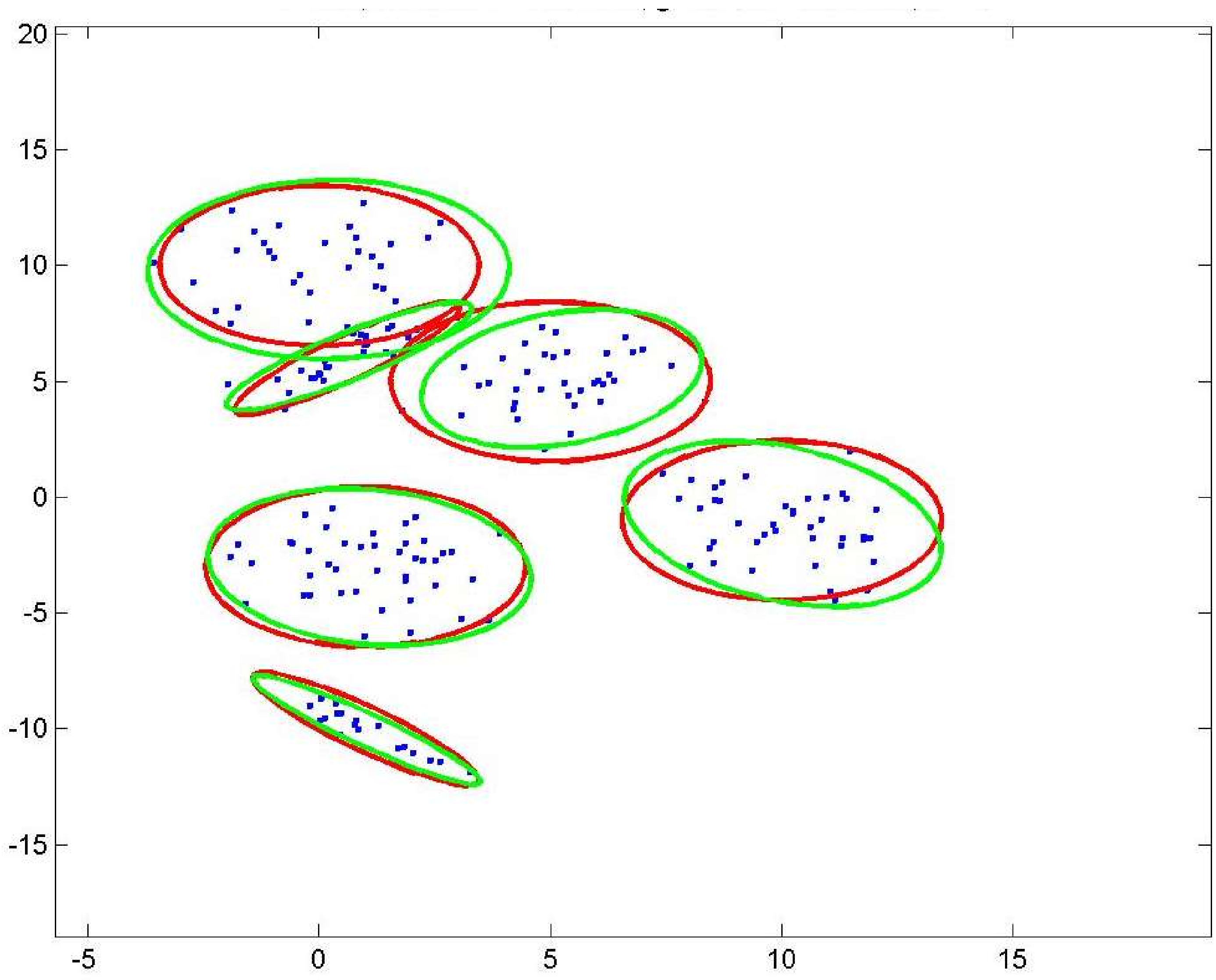




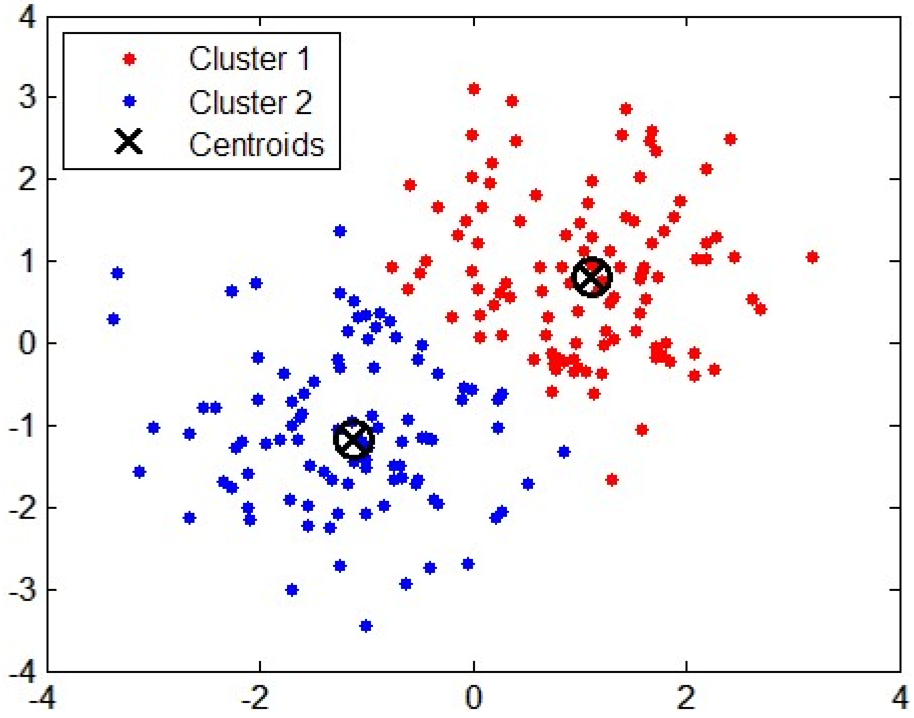
$$g(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}$$

$$f(x)=\sum_{i=1}^{n}\alpha_i g(x,\mu_i,\sigma_i),\ \alpha_i\geq 0\ \&\ \sum_i \alpha_i=1$$

Clustering by Mixture Models



K-Means Revisited



model parameters

$$\theta = \{(x_1, y_1), (x_2, y_2)\}$$

$$Z = \{Cluster_1,\ Cluster_2\}$$

latent parameters

Expectation Maximization

a Maximum likelihood

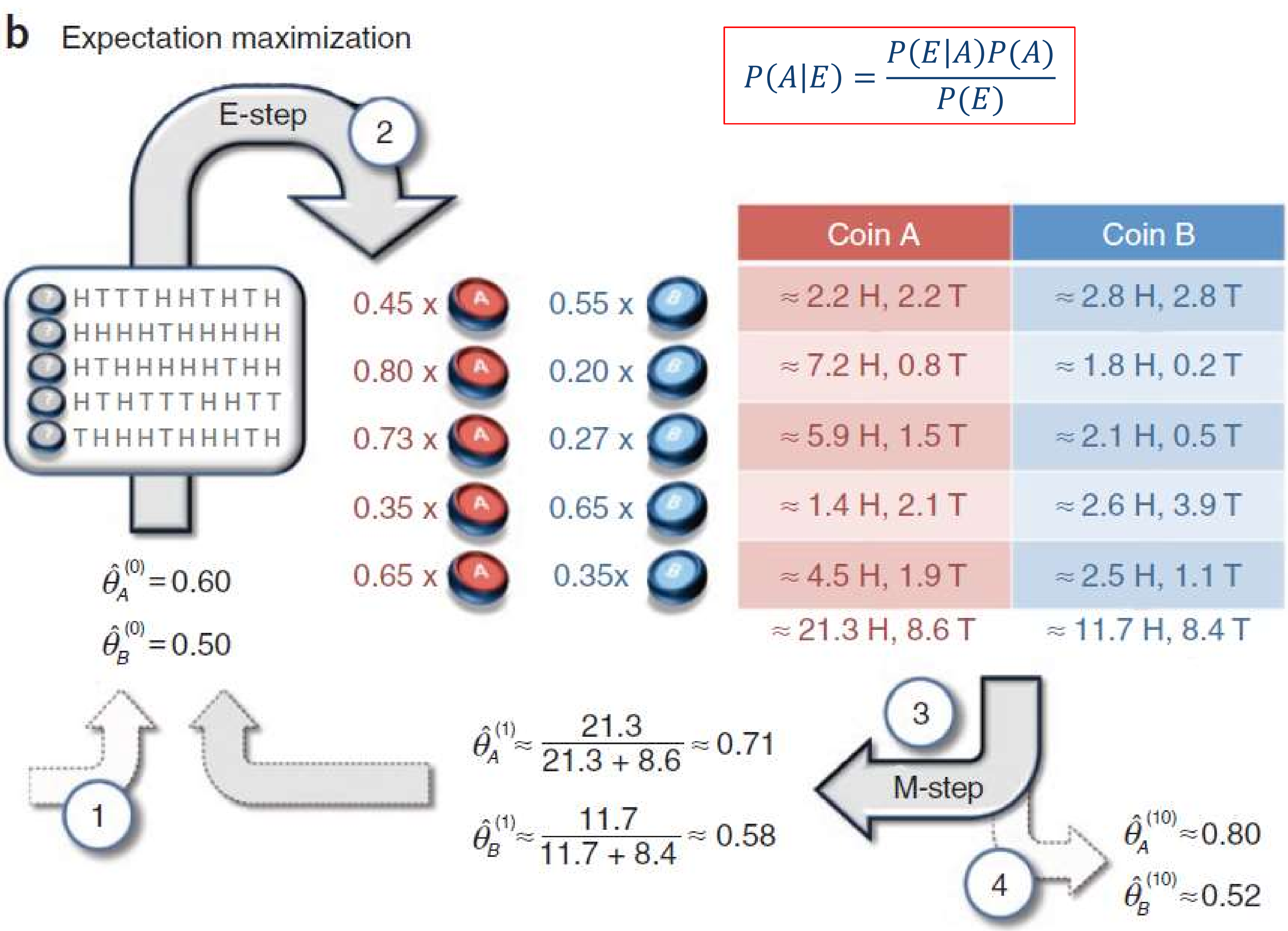
		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

5 sets, 10 tosses per set

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

$$P(A|E) = \frac{P(E|A)P(A)}{P(E)}$$



EM: Gaussian Mixture

m : the number of data points

n : the number of mixture components

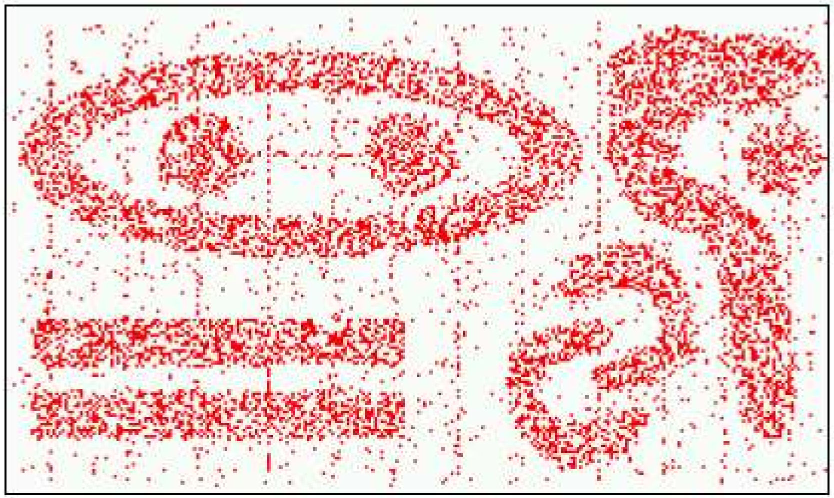
z_{ij} : whether instance i is generated by the jth Gaussian

$$E[z_{ij}] = \frac{p(x = x_i \mid \mu = \mu_j)\alpha_j}{\sum_{k=1}^{n} p(x = x_i \mid \mu = \mu_k)\alpha_k} = \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}\alpha_j}{\sum_{k=1}^{n} e^{-\frac{1}{2\sigma^2}(x_i - \mu_k)^2}\alpha_k}$$

$$\mu_j \leftarrow \frac{\sum_{i=1}^{m} E[z_{ij}]x_i}{\sum_{i=1}^{m} E[z_{ij}]} \qquad \alpha_j \leftarrow \frac{1}{m}\sum_{i=1}^{m} E[z_{ij}]$$

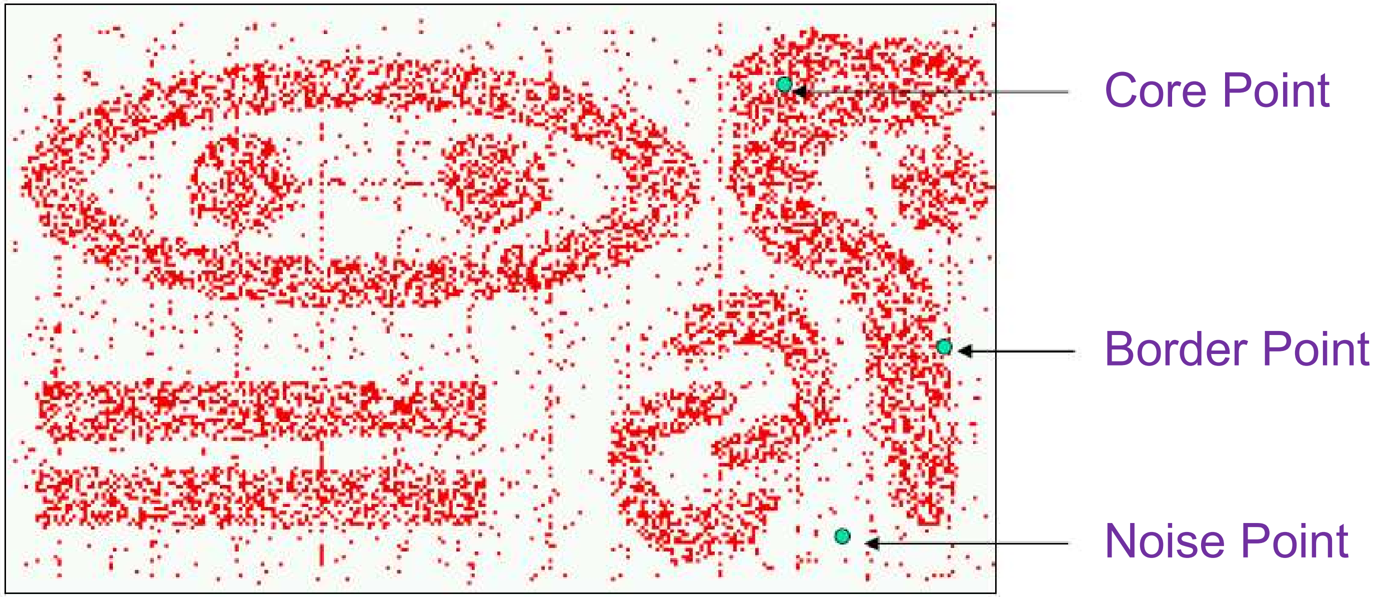
Density Based Methods

- Generate clusters of arbitrary shapes.
- Robust against noise.
- No K value required in advance.
- Somewhat similar to human vision.

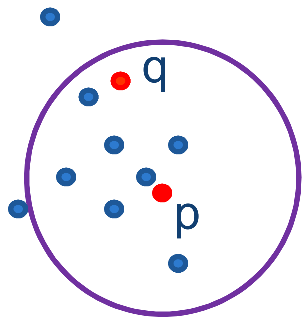


DBSCAN

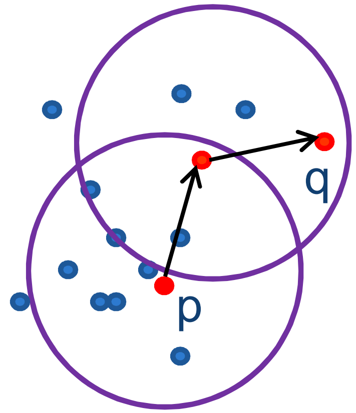
- Density-Based Spatial Clustering of Applications with Noise
- Density: number of points within a specified radius
- Core Point: points with high density
- Border Point: points with low density but in the neighbourhood of a core point
- Noise Point: neither a core point nor a border point



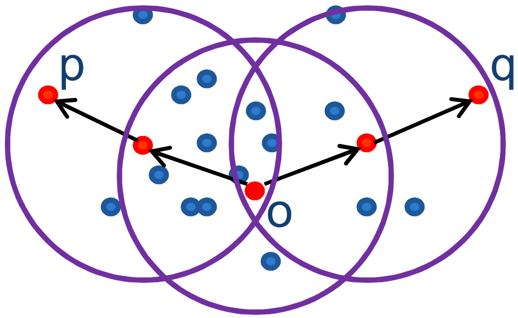
DBSCAN



directly density reachable



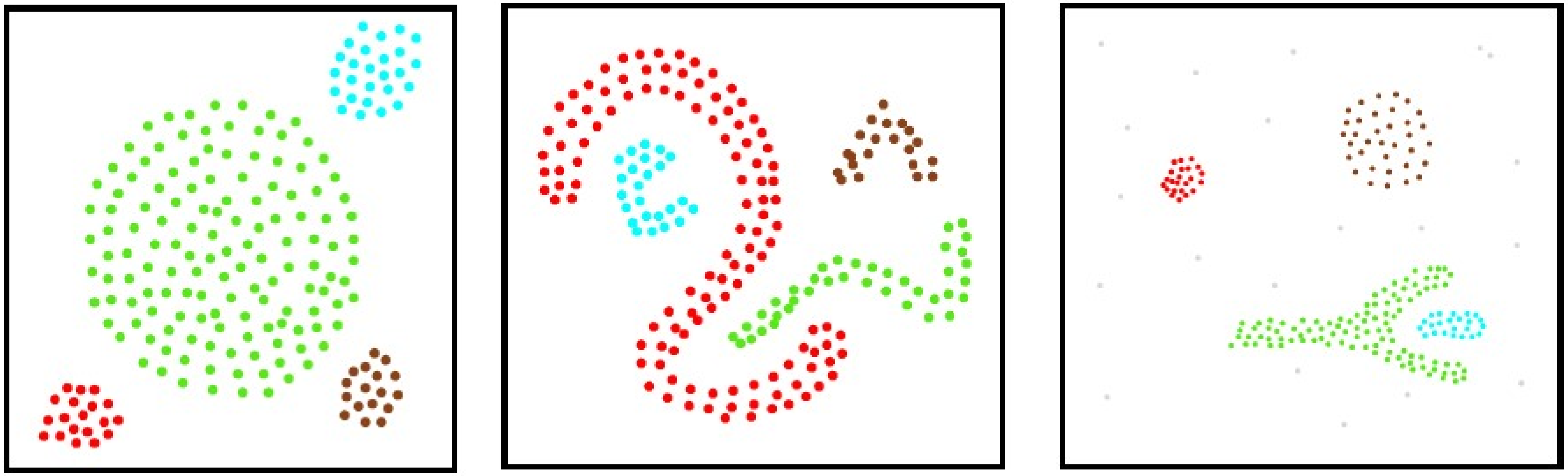
density reachable



density connected

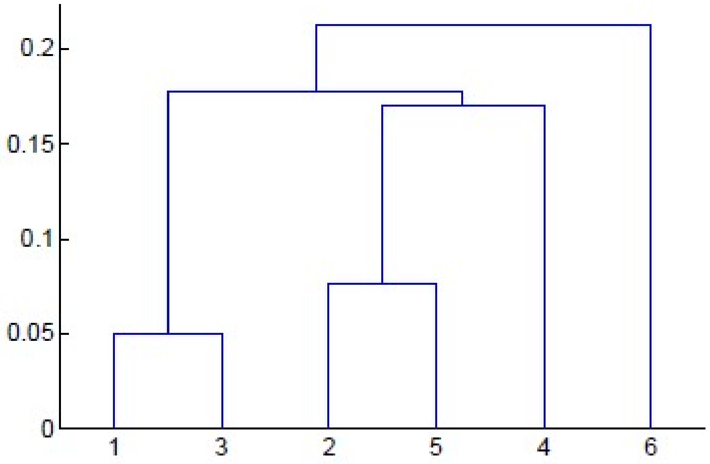
DBSCAN

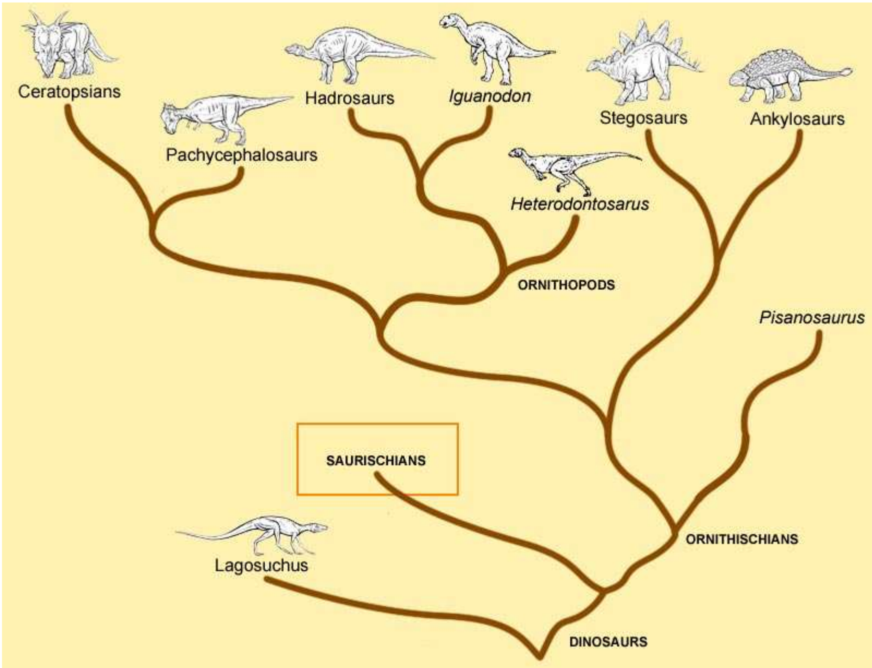
- A cluster is defined as the maximal set of density connected points.
- Start from a randomly selected unseen point P.
- If P is a core point, build a cluster by gradually adding all points that are density reachable to the current point set.
- Noise points are discarded (unlabelled).



Hierarchical Clustering

- Produce a set of nested tree-like clusters.
- Can be visualized as a dendrogram.
  - Clustering is obtained by cutting at desired level.
  - No need to specify K in advance.
  - May correspond to meaningful taxonomies.





Agglomerative Methods

- Bottom-up Method
- Assign each data point to a cluster.
- Calculate the proximity matrix.
- Merge the pair of closest clusters.
- Repeat until only a single cluster remains.
- How to calculate the distance between clusters?
- Single Link
  - Minimum distance between points
- Complete Link
  - Maximum distance between points

Example

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Single Link

Example

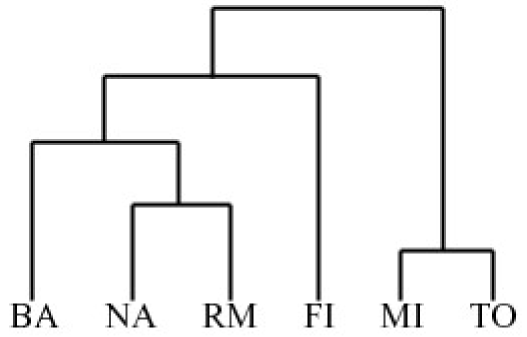
	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

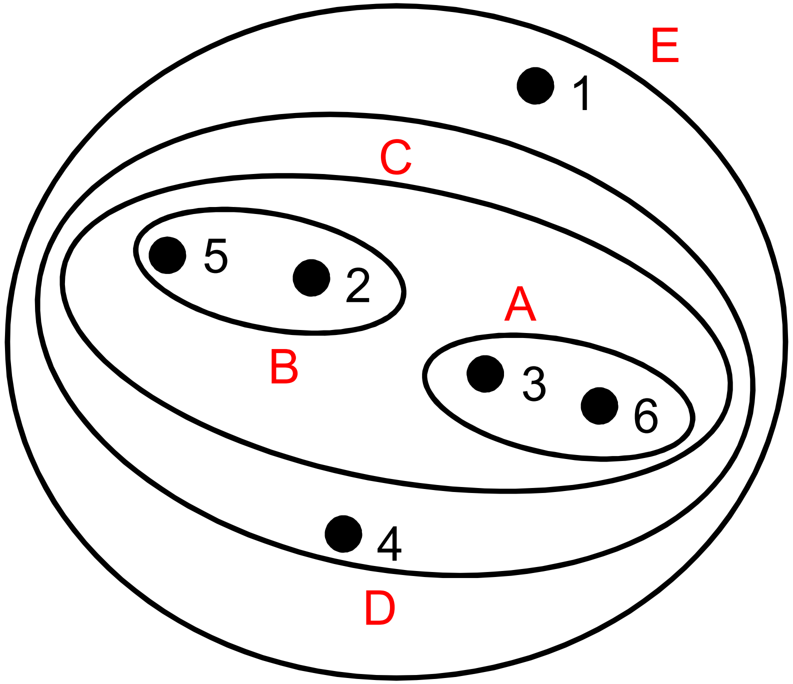
Example

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

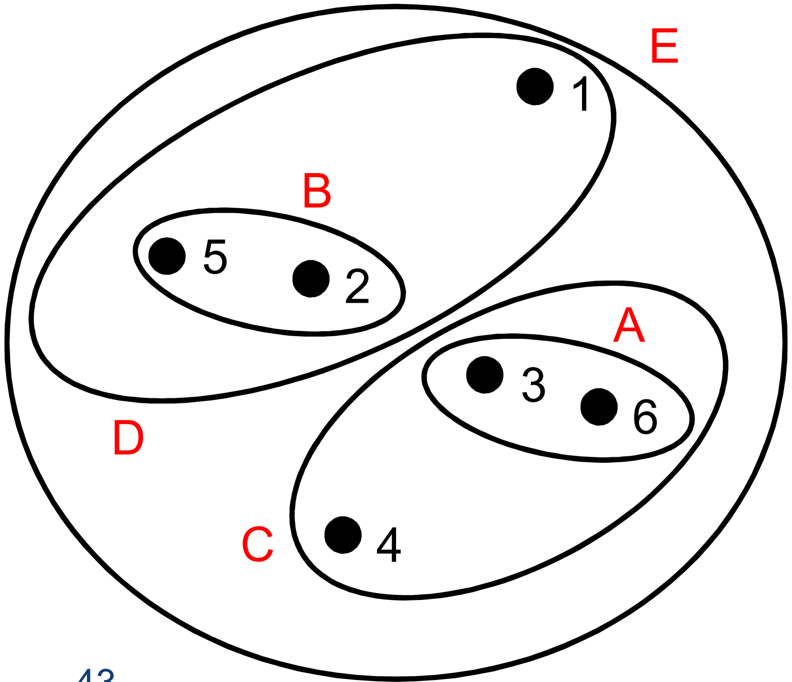
	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0



Min vs. Max



MIN



MAX

Reading Materials

- Text Books
  - R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification,* Chapter 10, 2nd Edition, John Wiley & Sons.
  - J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Chapter 8, Morgan Kaufmann.

- Survey Papers
  - A. K. Jain, M. N. Murty and P. J. Flynn (1999) “Data Clustering: A Review”, *ACM Computing Surveys*, Vol. 31(3), pp. 264-323.
  - R. Xu and D. Wunsch (2005) “Survey of Clustering Algorithms”, *IEEE Transactions on Neural Networks*, Vol. 16(3), pp. 645-678.
  - A. K. Jain (2010) “Data Clustering: 50 Years Beyond K-Means”, *Pattern Recognition Letters*, Vol. 31, pp. 651-666.

- Online Tutorials
  - http://home.dei.polimi.it/matteucc/Clustering/tutorial_html
  - http://www.autonlab.org/tutorials/kmeans.html
  - http://users.informatik.uni-halle.de/~hinnebur/ClusterTutorial