APPLIED MACHINE LEARNING

Clustering

Part I - Principle

Clustering Principle

Clustering aims at grouping data that resemble one another into ***clusters.***

Applications

- Grouping documents (e.g. webpages) into similar topics
  - *Speed up search of documents based on matches for similar queries*
- Group customers (e.g. on amazon) into types / profiles for marketing
  - Offer related products that may interest the customer
- Find groups of genes with similar phylogenetic characteristics
  - Relate some diseases and determine risks of developing a disease

Clustering Principle

Clustering can be used as:

- *Feature extraction method*: for identifying underlying structure in data and salient features, best visualized through **cluster prototype**.
- *Compression method*: for organizing the data and summarizing it through **cluster prototypes.**

A cluster prototype can be:

- A typical datapoint, best representative of the cluster
- The average (centroid) of the datapoints in the cluster

Clustering Principle

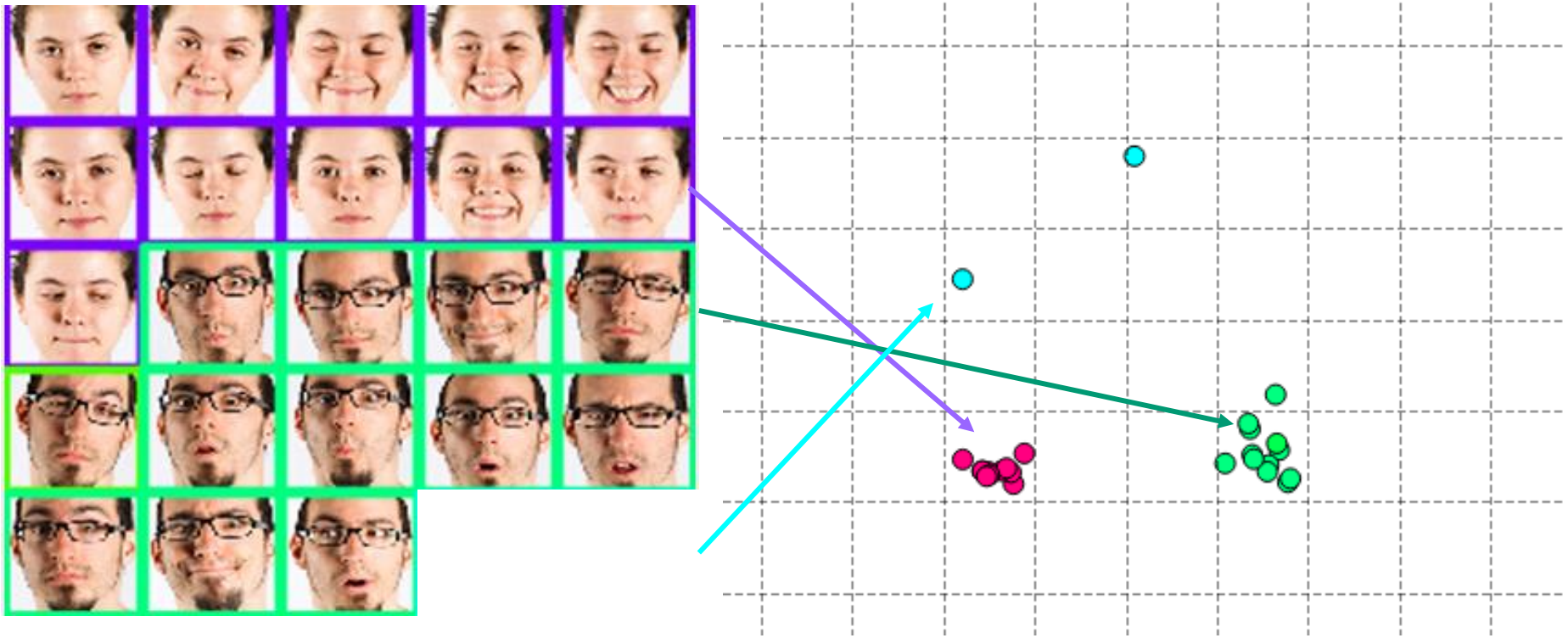
Groups of points are said to belong to the same cluster
if they are similar enough.

→ measure of similarity.

Which clusters?

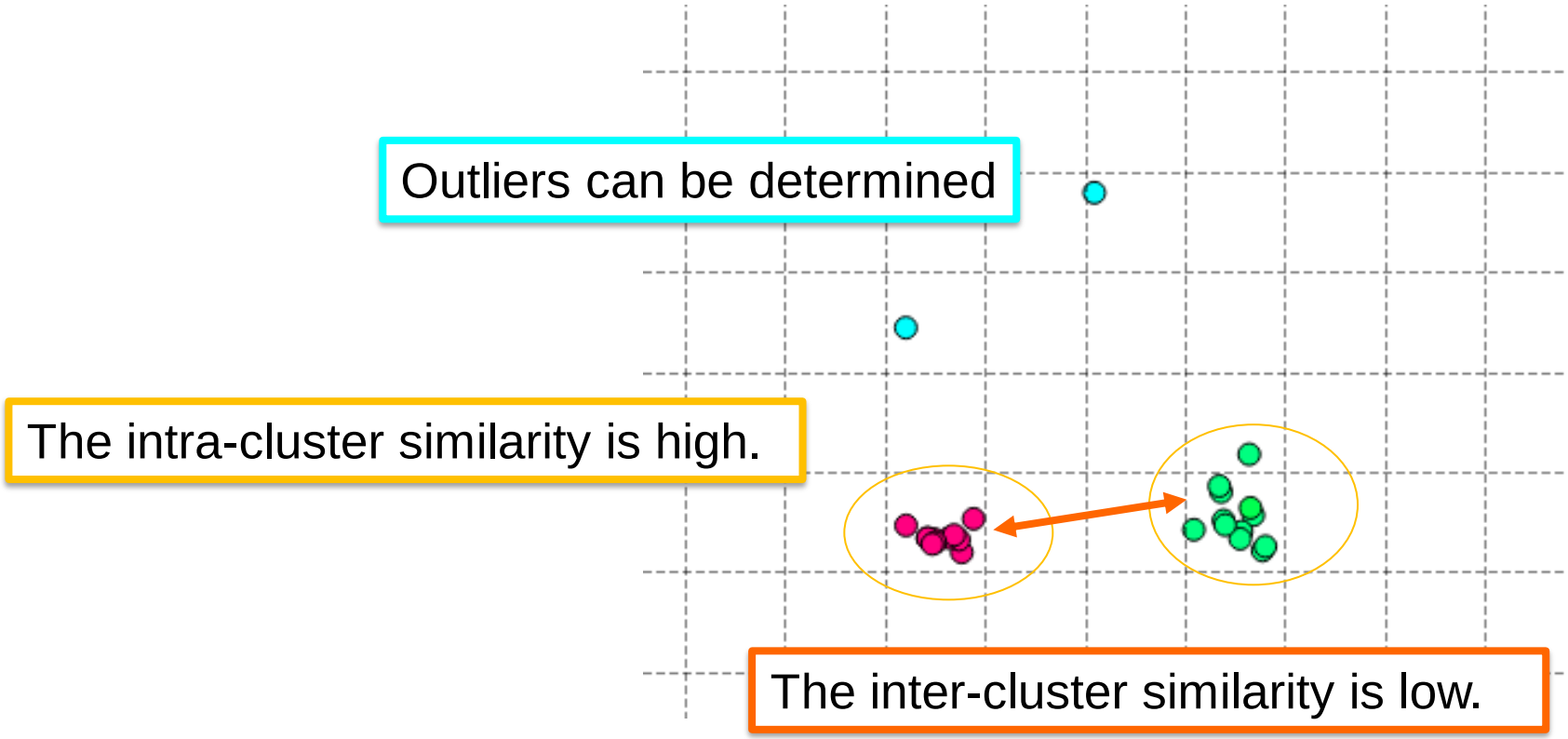
Which two subgroups of pictures are similar and why?

Dataset with outliers



What is good clustering?

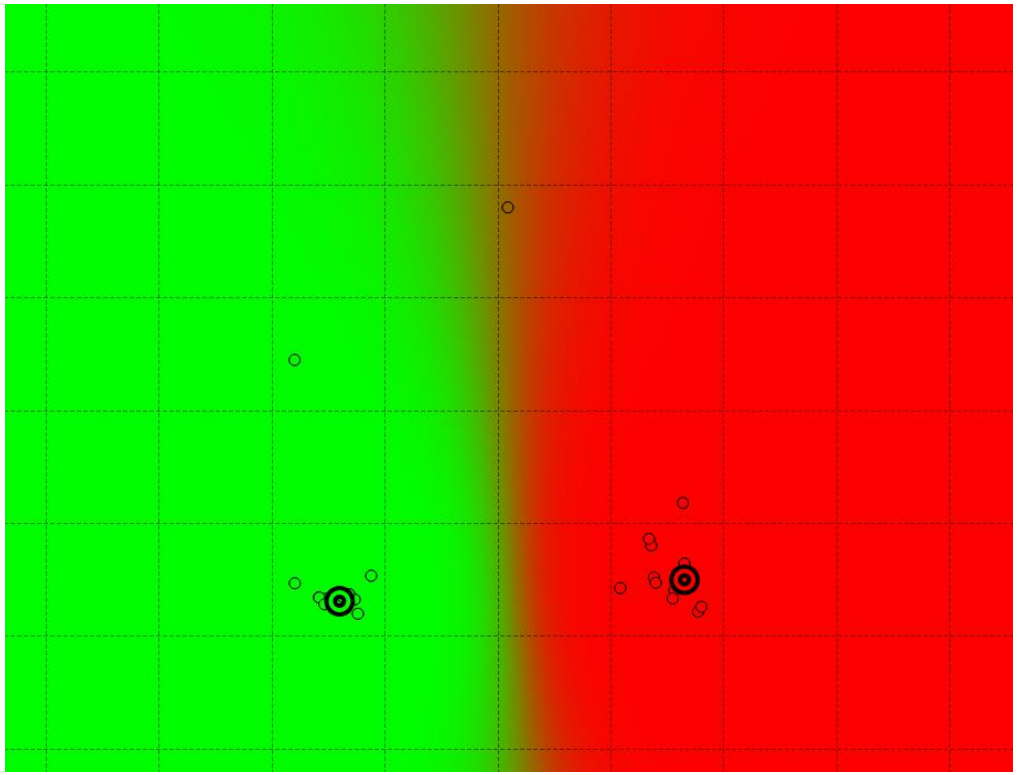
A good clustering method produces high quality clusters when:



The quality measure of a cluster depends on the similarity measure used!

The role of prior information

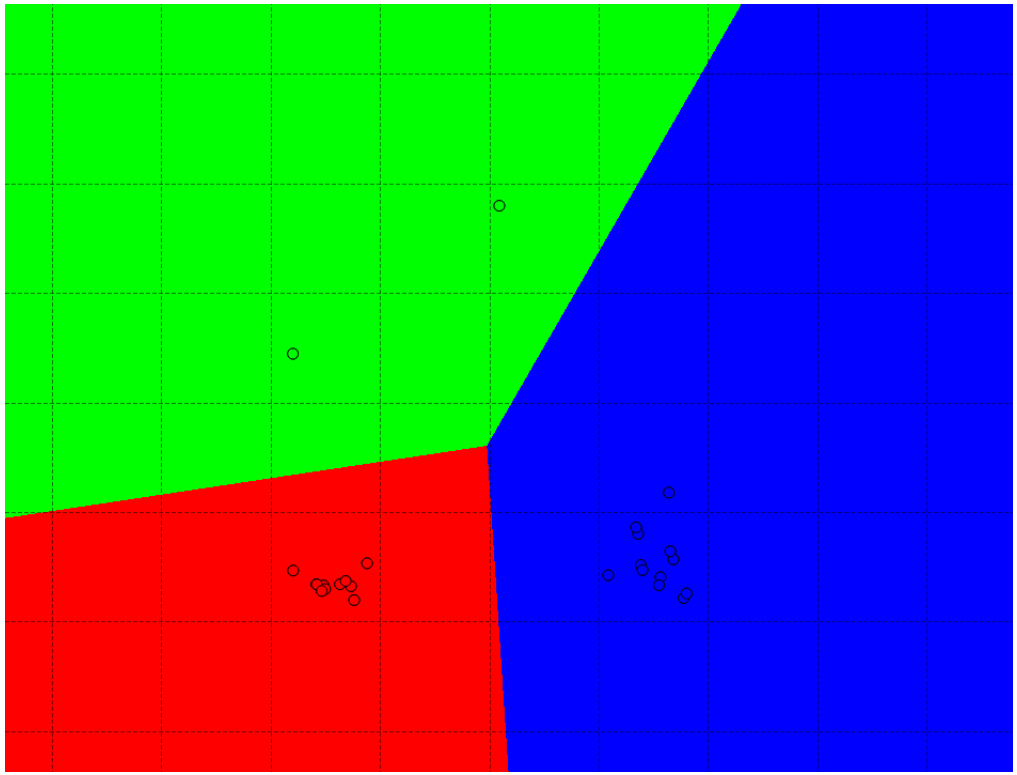
Clustering approaches often require to have a good *guess as to* how many clusters exist.



If we assume only 2 clusters, the outliers cannot be excluded.

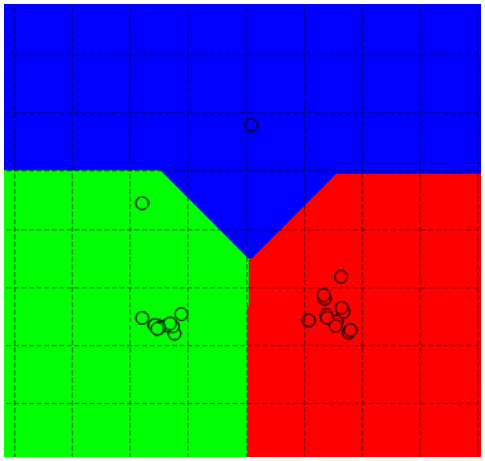
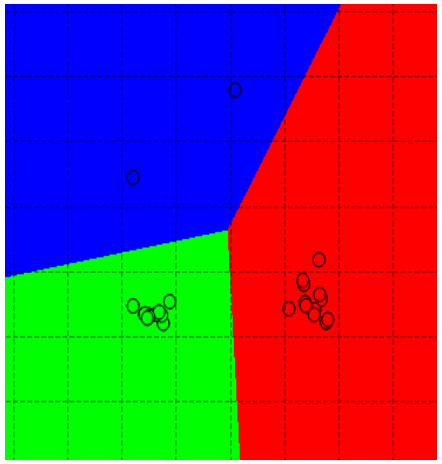
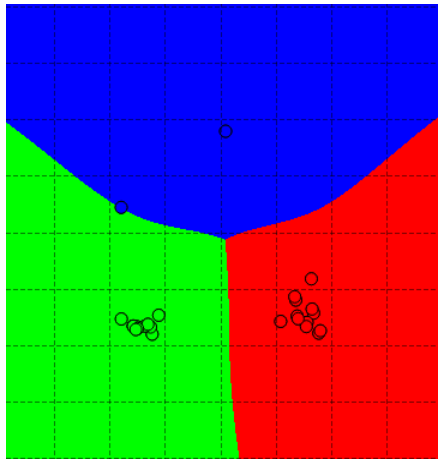
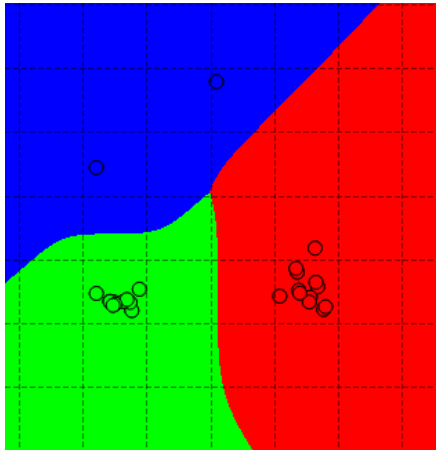
The role of prior information

Clustering approaches often require to have a good *guess as to* how many clusters exist.



If we assume 3 clusters, the outliers are grouped in one cluster.

The role of the measure of similarity

L1-norm

$$d\left(x^1, x^2\right) = \left|x^1 - x^2\right| = \sum_{i=1}^{N} \left|x_i^1 - x_i^2\right|$$

L2-norm

$$d\left(x^1, x^2\right) = \left\|x^1 - x^2\right\| = \sqrt{\sum_{i=1}^{N} \left(x_i^1 - x_i^2\right)^2}$$

$p = 3$

$p = 5$

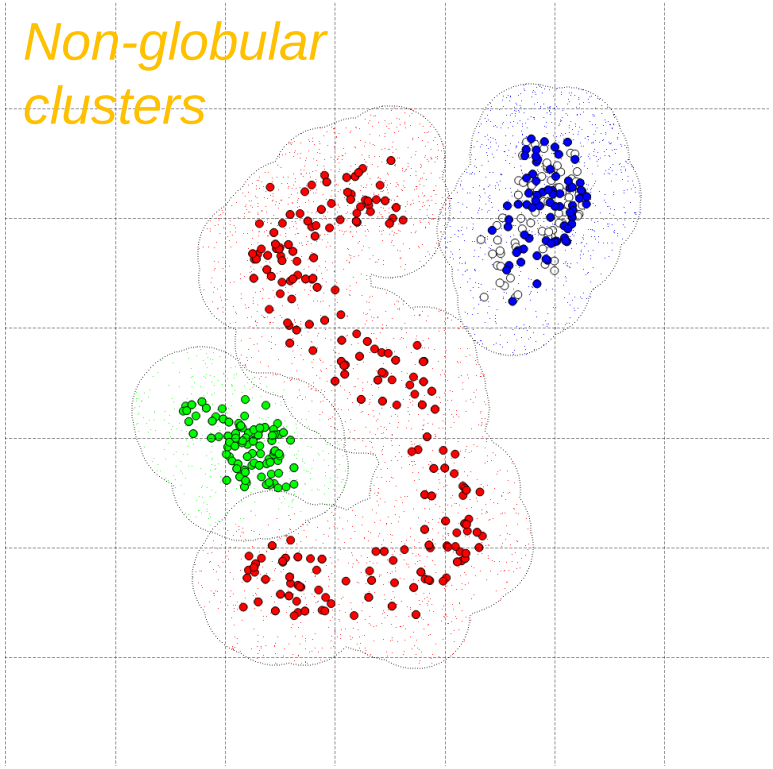
Lp-norm

$$d\left(x^1, x^2\right) = \left\|x^1 - x^2\right\|^P = \sqrt[p]{\sum_{i=1}^{N} \left|x_i^1 - x_i^2\right|^p}$$

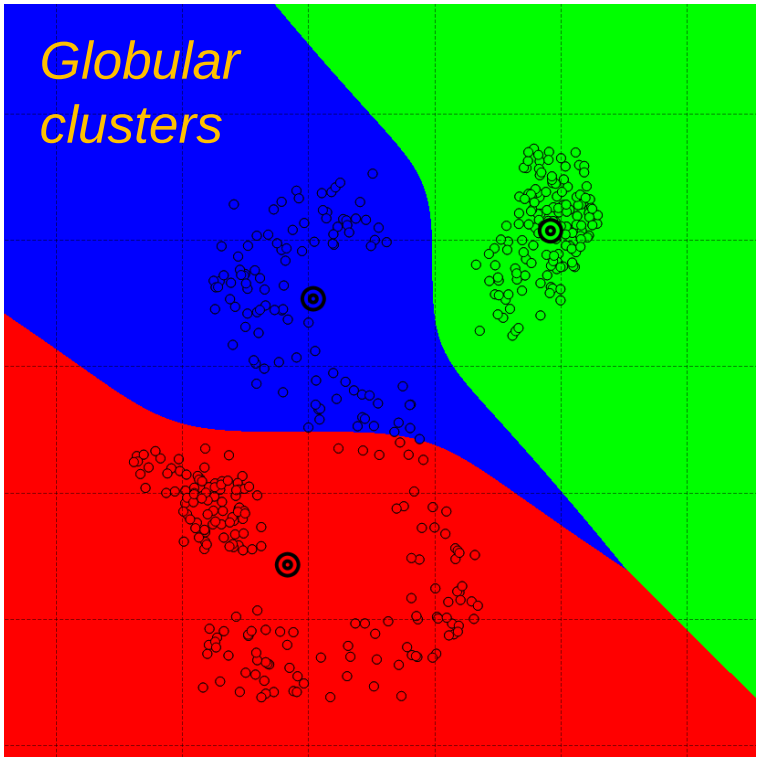
$$x^1, x^2 \in \mathbb{R}^N$$

The shape of the clusters

Clustering techniques differ in the complexity of the clusters generated.

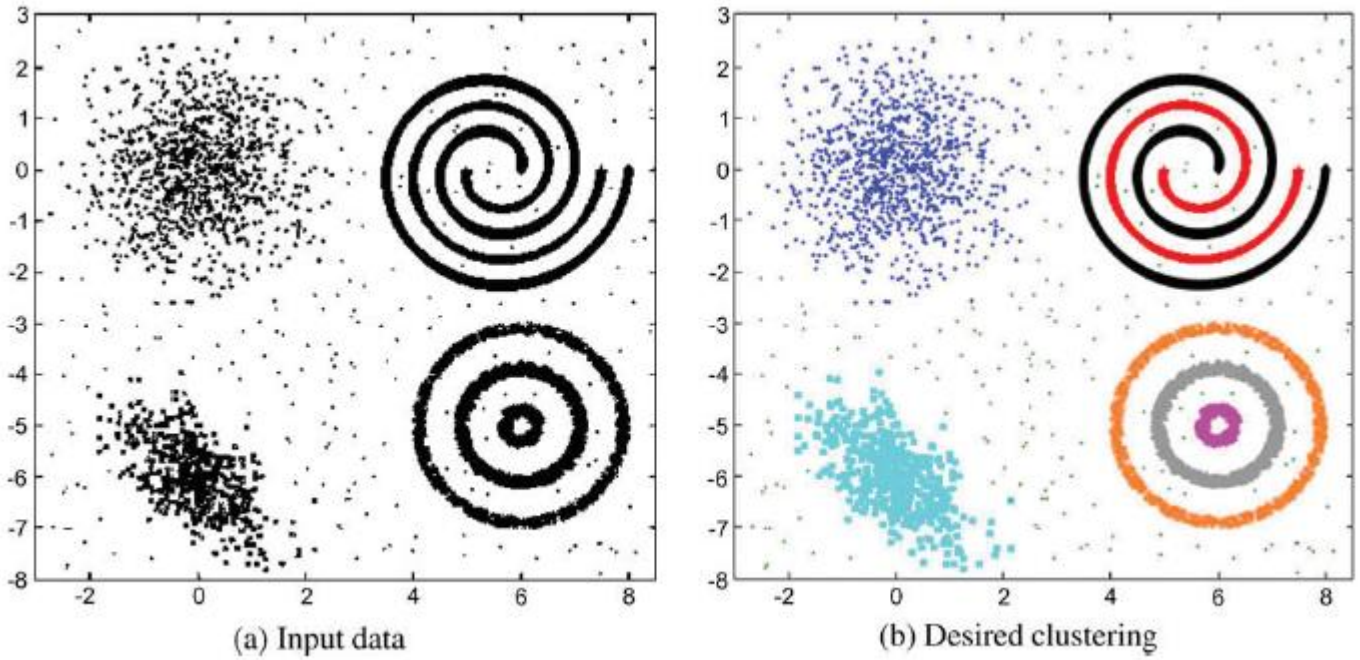


DBSCAN



K-means

Clustering: Challenges



Clustering may fail when metric changes depending on the region of space

Jain, 2010, Data clustering: 50 years beyond K-means, Pattern Recognition Letters