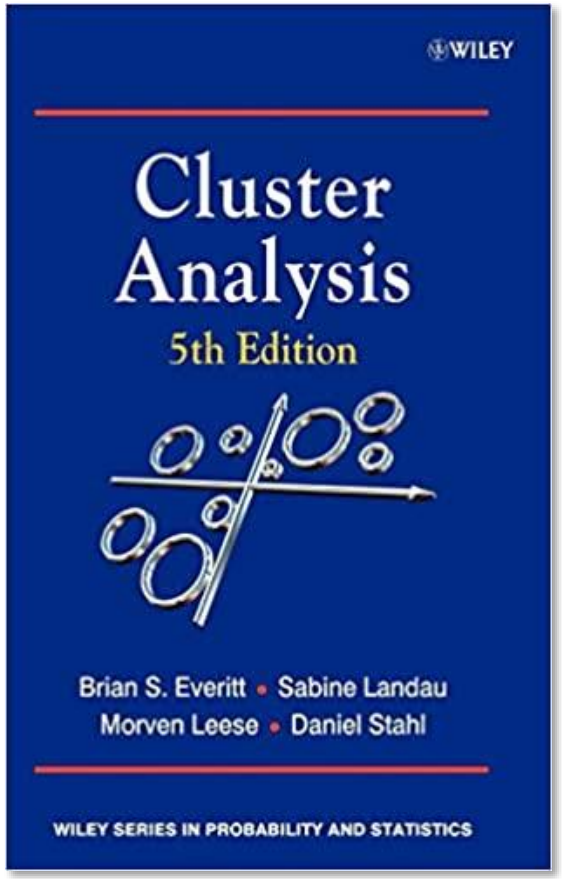
Fundamentals of AI

Clustering

General principles and classification of clustering methods

Some books



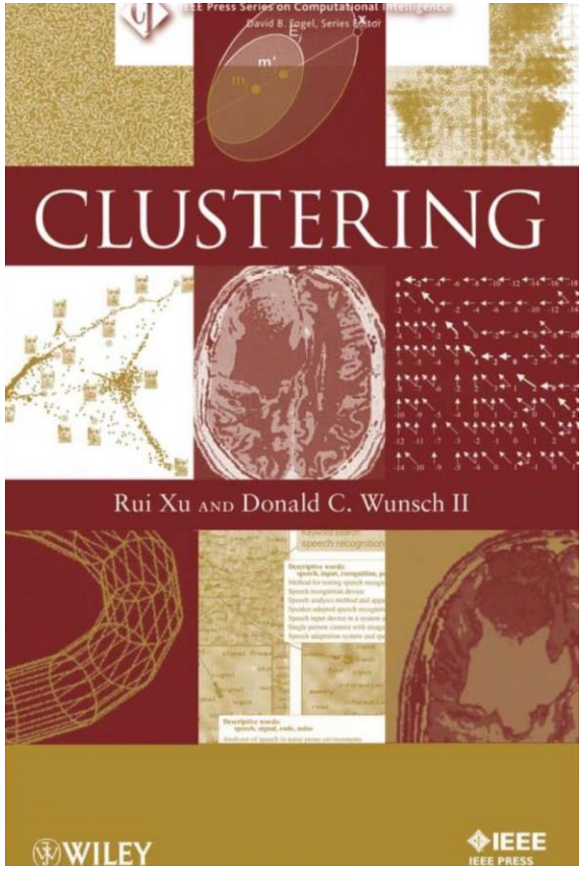
WILEY

Cluster Analysis

5th Edition

Brian S. Everitt • Sabine Landau
Morven Leese • Daniel Stahl

WILEY SERIES IN PROBABILITY AND STATISTICS



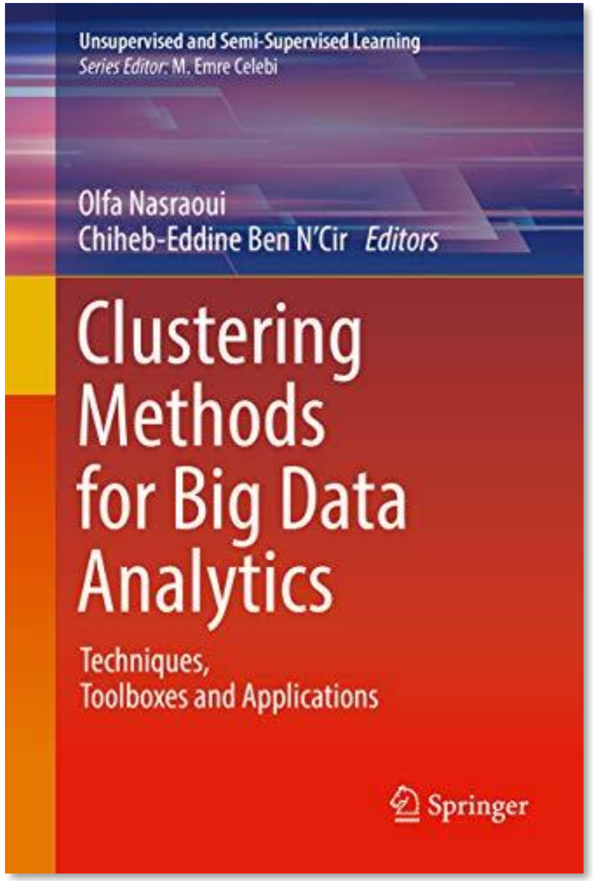
IEEE Press Series on Computational Intelligence
David B. Fogel, Series Editor

CLUSTERING

Rui Xu and Donald C. Wunsch II

WILEY

IEEE
IEEE PRESS



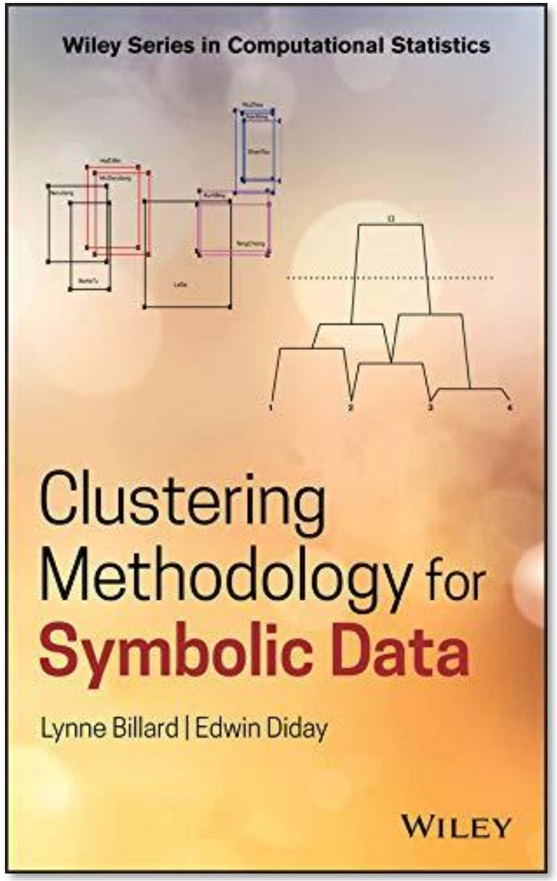
Unsupervised and Semi-Supervised Learning
Series Editor: M. Emre Celebi

Olfa Nasraoui
Chiheb-Eddine Ben N'Cir *Editors*

Clustering Methods for Big Data Analytics

Techniques, Toolboxes and Applications

Springer



Wiley Series in Computational Statistics

Clustering Methodology for Symbolic Data

Lynne Billard | Edwin Diday

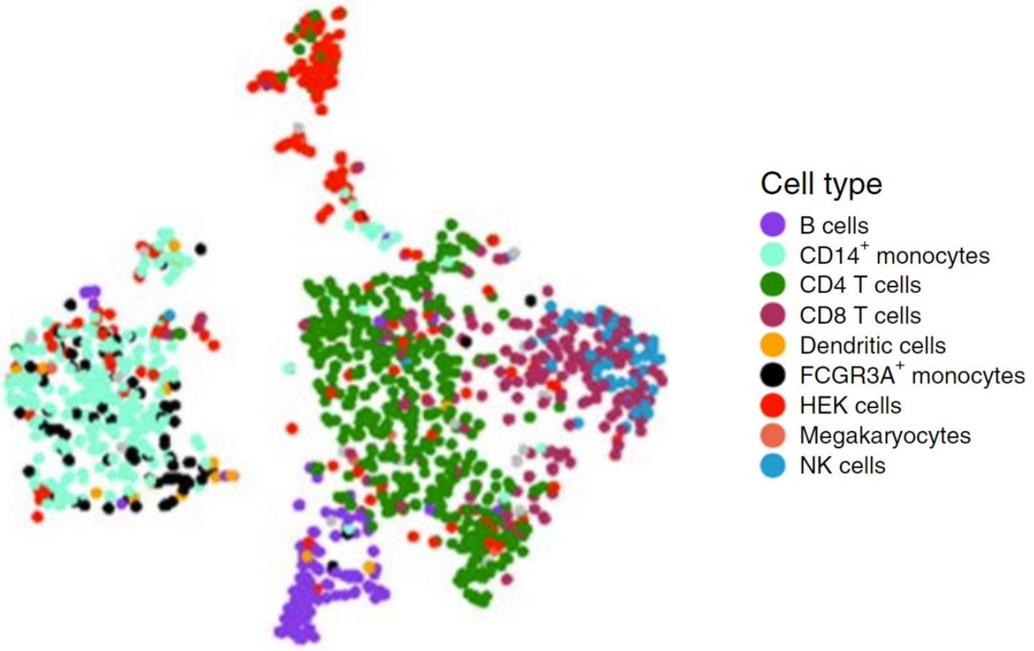
WILEY

Clustering problem in machine learning

- Part of unsupervised classification and exploratory data analysis
- Pure case : no labeled data are available (Everitt et al., 2001 ; Jain and Dubes, 1988)
- The goal of clustering is to separate a finite, unlabeled data set into a finite and discrete set of “natural”, “hidden” data structures

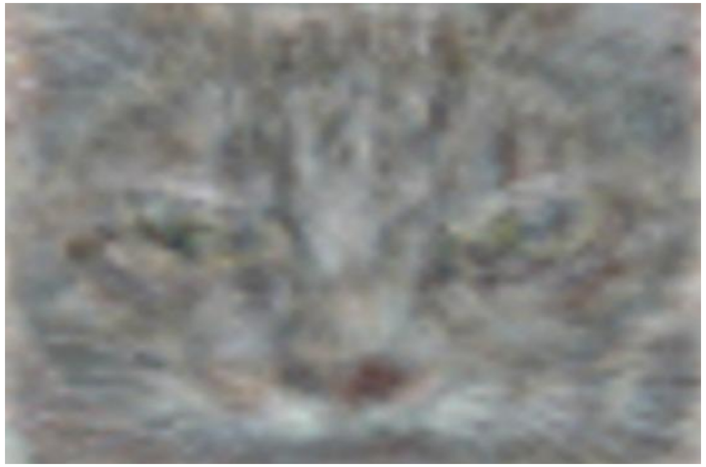
Distinguish *classes* and *clusters*!!!

- **Class** = set of data points with the same pre-defined label
- **Cluster** = result of solving a clustering problem

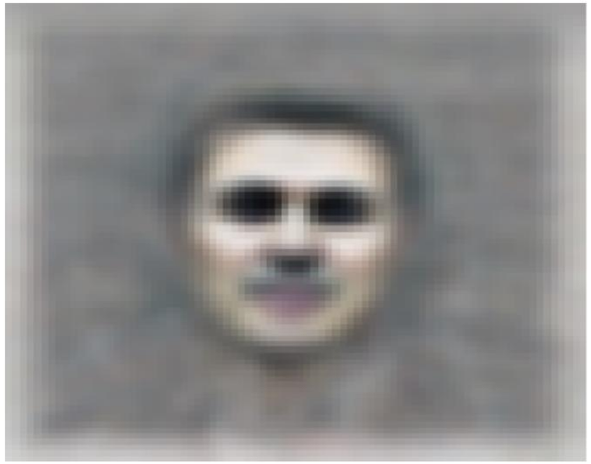


From Mereu et al, Nature Biotech , 2020

Example of complex clustering problem

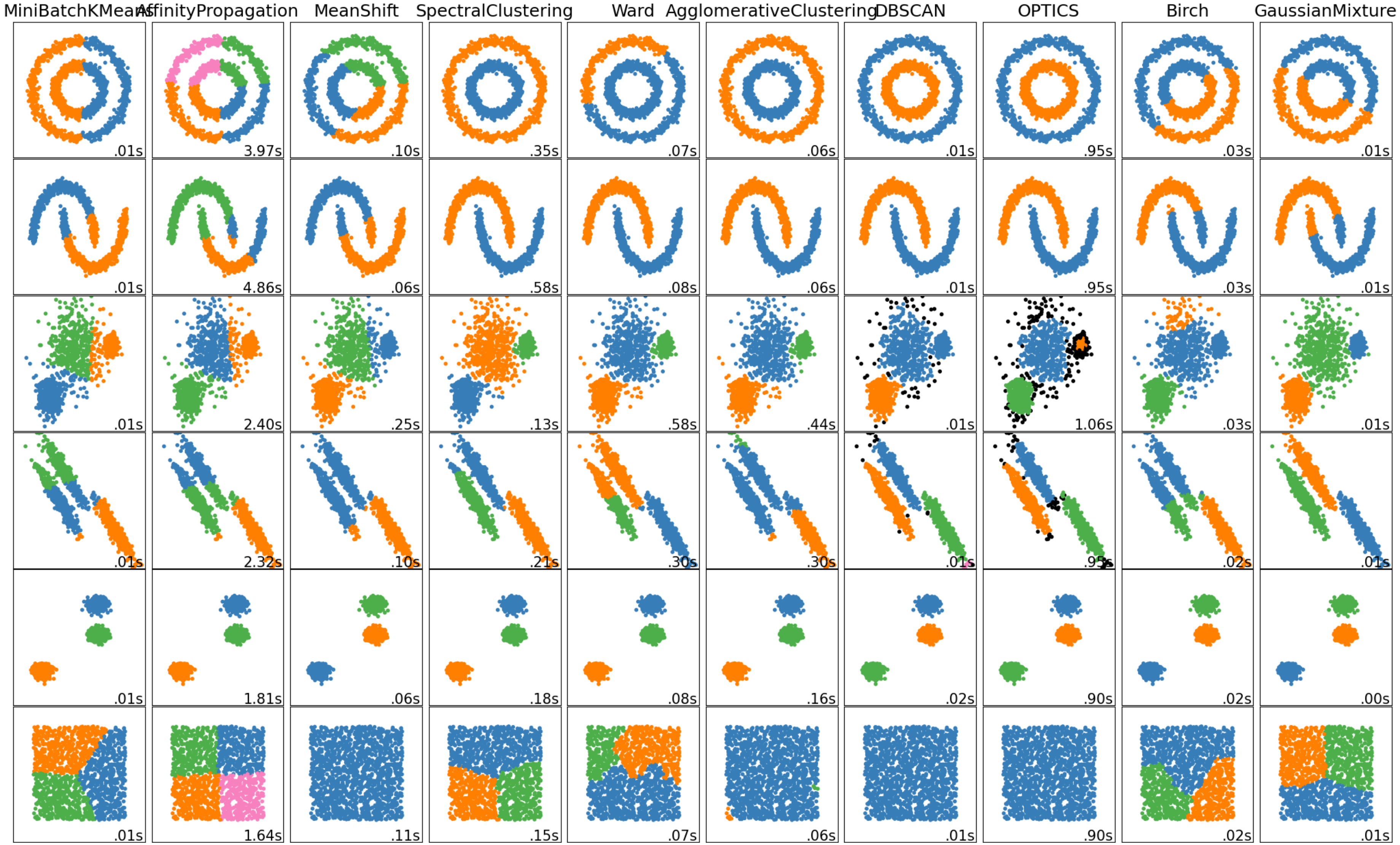


“Google cat”

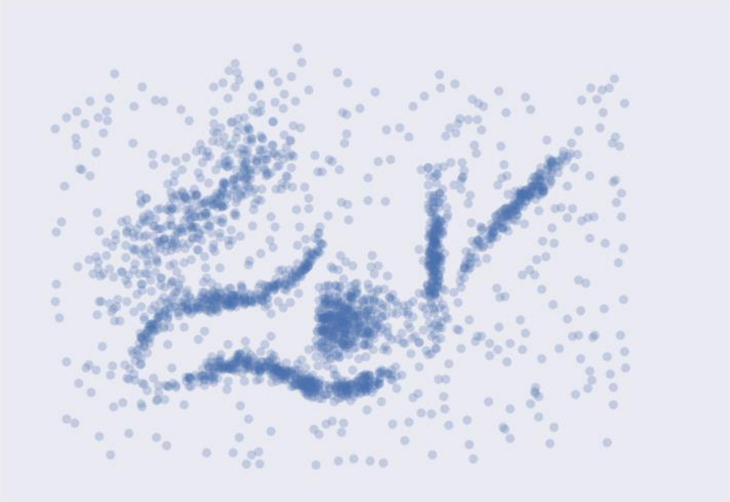


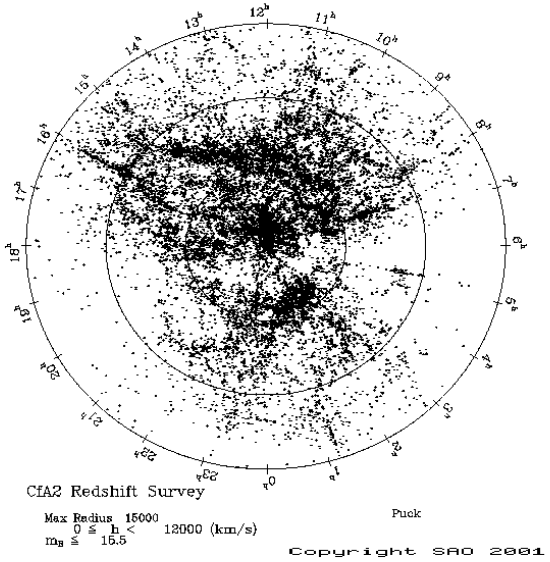
Le et al. Building High-level Features Using Large Scale Unsupervised Learning. ICML-2012

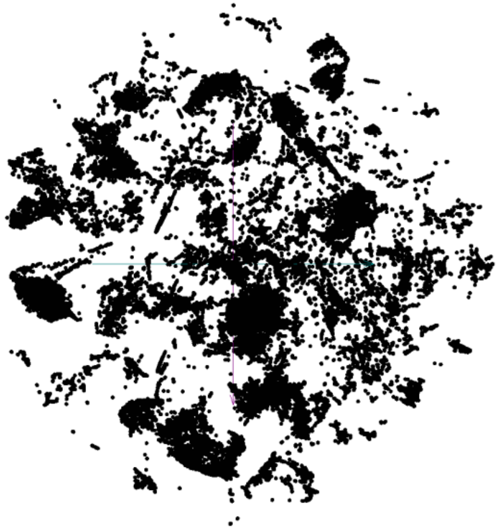
Clustering in scikit-learn in 2D



What about this?







https://amp.pharm.mssm.edu/archs4/data.html

Immediate questions

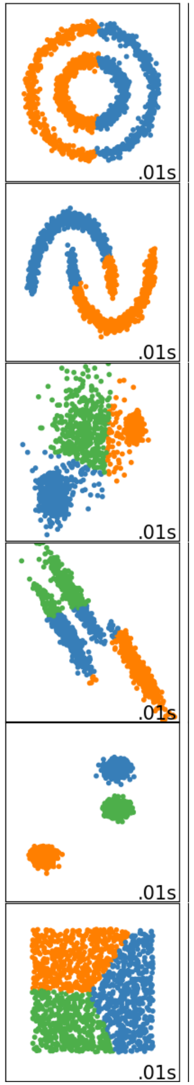
- What is cluster?
- How many clusters?
- Typology of cluster methods
- How to choose the best method?

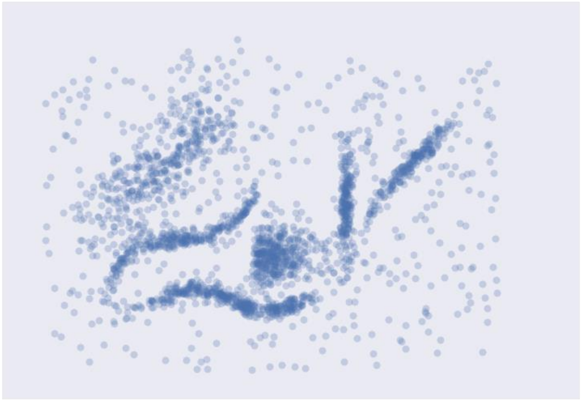
What is cluster?

“ formal definition of cluster is not only difficult but may even be misplaced. ” (Everit, 2001)

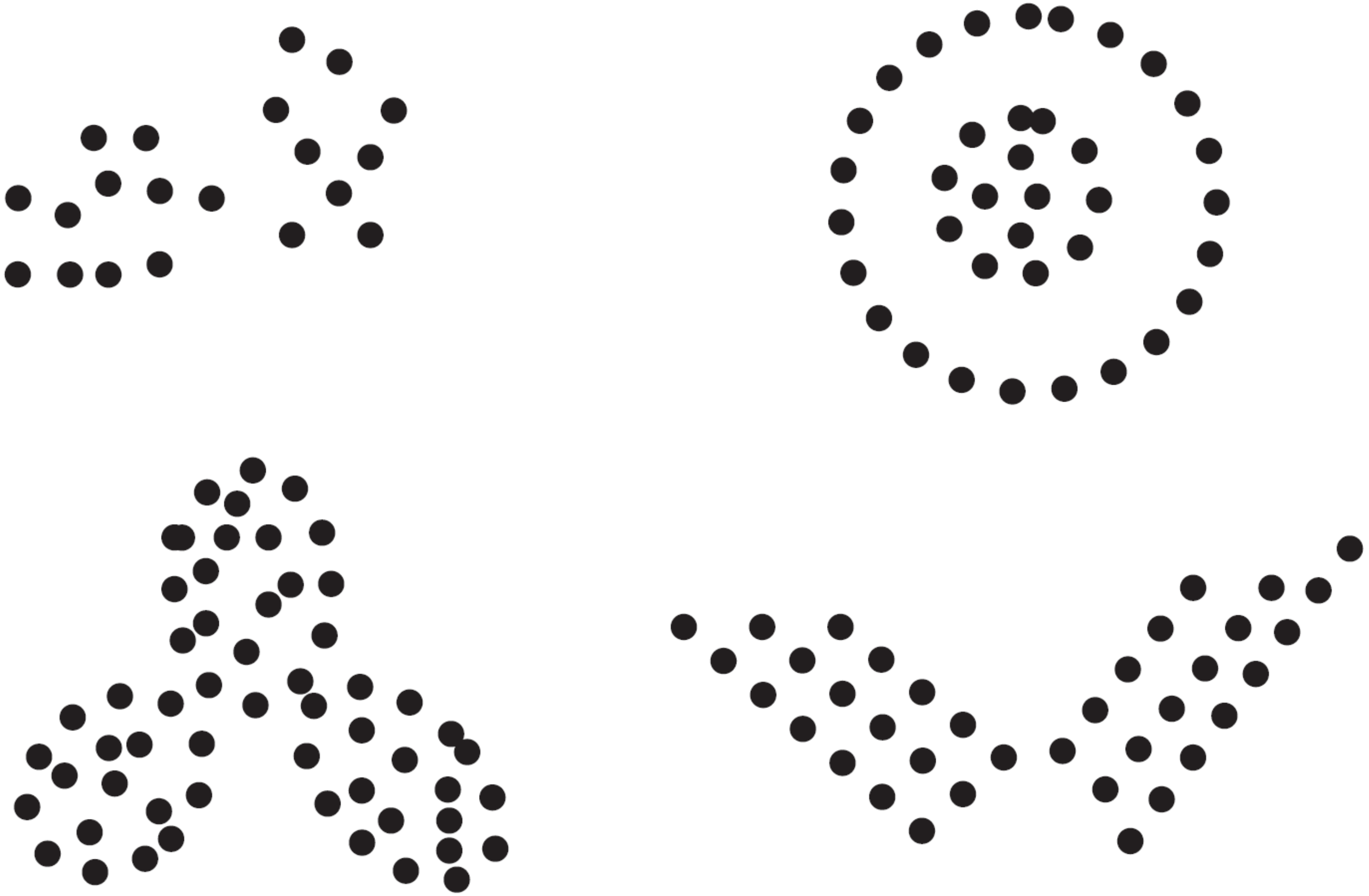
Collection of definitions (Everitt (1980)) :

- “ A cluster is a set of entities which are alike, and entities from different clusters are not alike. ” (**VAGUE!**)
- A cluster is “ an aggregate of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it. ” (**TOO RESTRICTIVE!**)
- “ Clusters may be described as continuous regions of this space (*d* -dimensional feature space) containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points. ” (**COOL, BUT HOW TO COMPUTE DENSITY IN HIGH-DIMENSIONS?**)





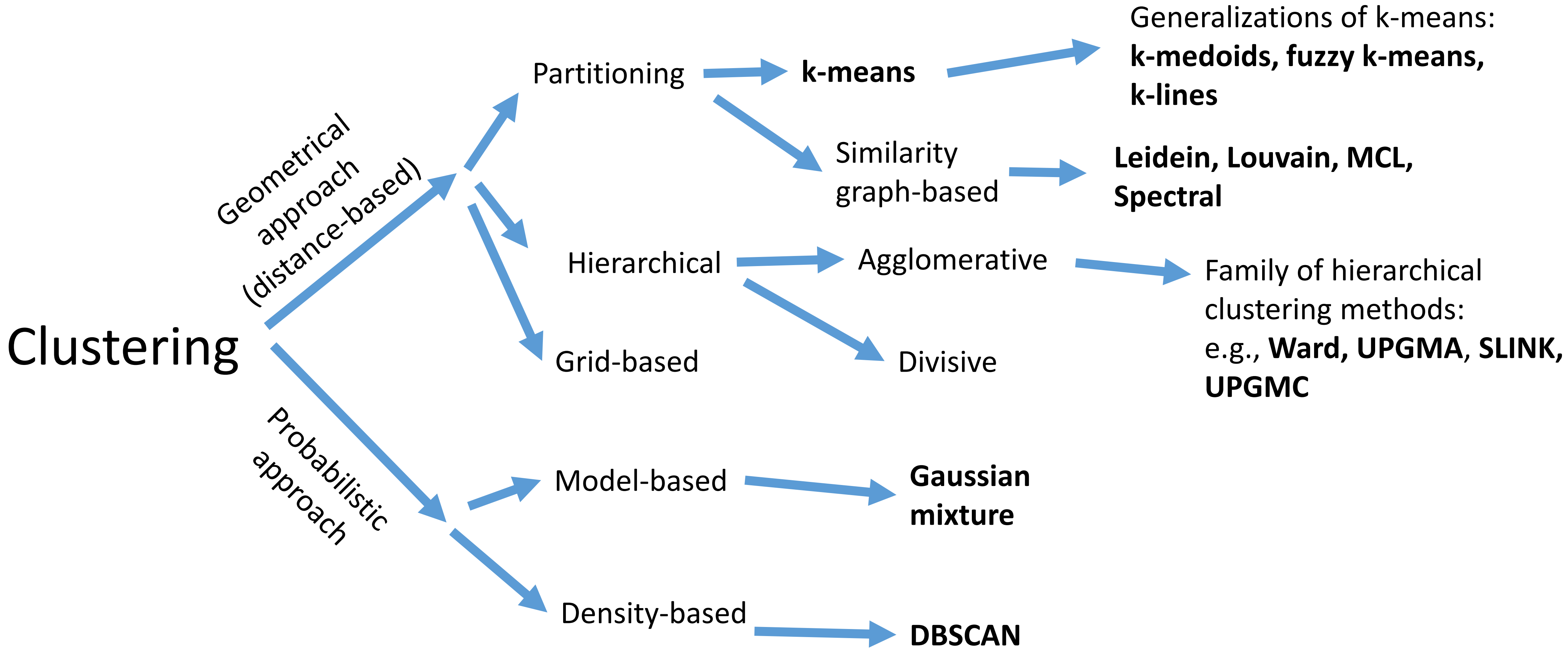
How many clusters?



ANY clustering method requires specifying the number of clusters as a parameter

- Sometimes it is done explicitly
- Sometimes it is done through some kind of ‘scale’ or ‘resolution’ parameter

Typology of cluster methods



Clustering

Geometrical approach (distance-based)

Partitioning

k-means

Generalizations of k-means: **k-medoids, fuzzy k-means, k-lines**

Similarity graph-based

Leidein, Louvain, MCL, Spectral

Hierarchical

Agglomerative

Family of hierarchical clustering methods: e.g., **Ward**, **UPGMA**, **SLINK, UPGMC**

Divisive

Grid-based

Probabilistic approach

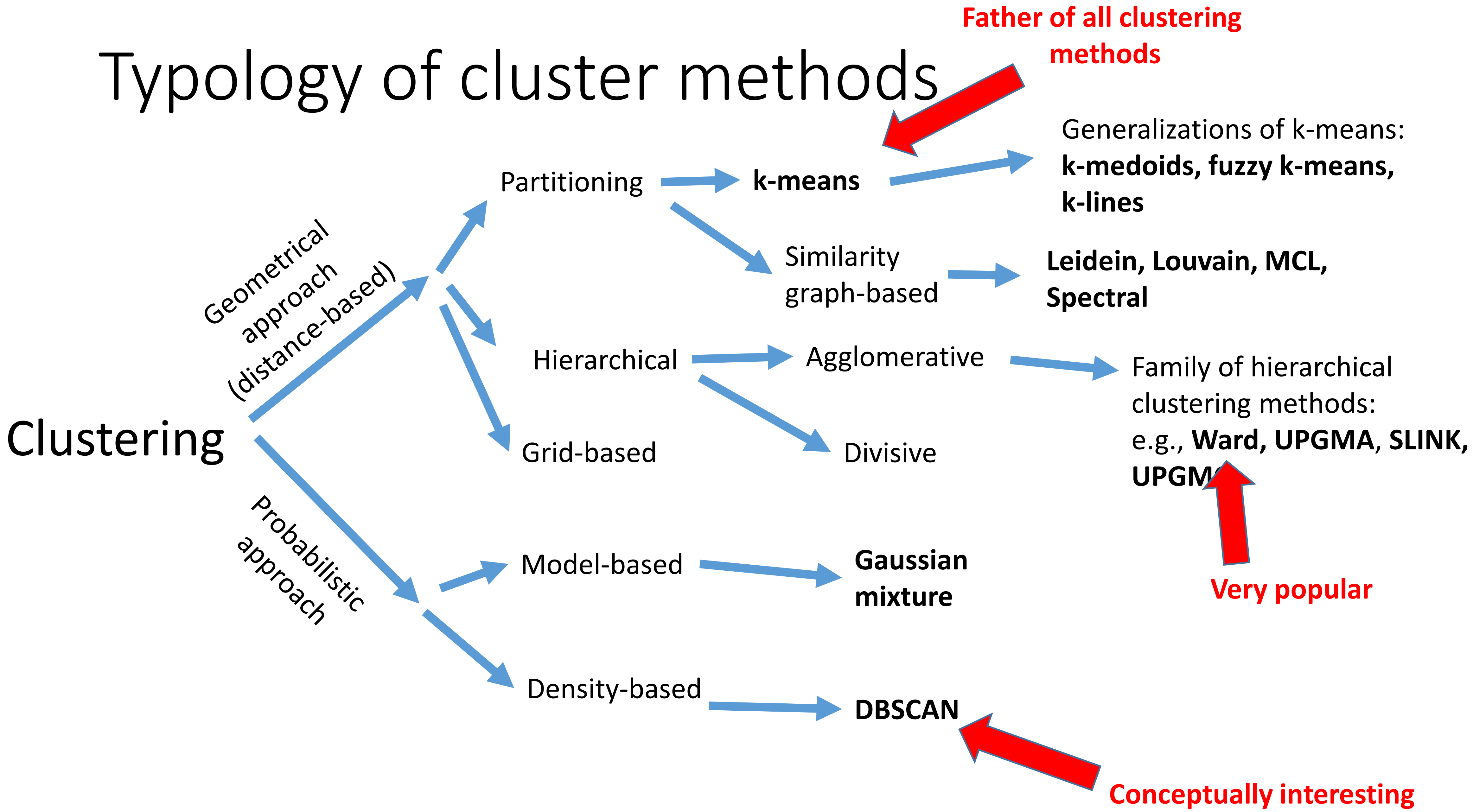
Model-based

Gaussian mixture

Density-based

DBSCAN

Typology of cluster methods



Father of all clustering methods

Clustering

Geometrical approach (distance-based)

Partitioning

k-means

Generalizations of k-means: **k-medoids, fuzzy k-means, k-lines**

Similarity graph-based

Leidein, Louvain, MCL, Spectral

Hierarchical

Agglomerative

Family of hierarchical clustering methods: e.g., **Ward, UPGMA**, **SLINK, UPGMA**

Divisive

Grid-based

Probabilistic approach

Model-based

Gaussian mixture

Very popular

Density-based

DBSCAN

Conceptually interesting

What you should ask about a clustering method

Base level

- Input information (data table, distance table, KNN-graph, …)
- Computational complexity (time and memory requirements), scalability for big data ($O(K^r m^s N^p)$, K – number of clusters, m – number of dimensions, N – number of data points)
- Constraints on cluster shapes (spherical, convex, …)
- Key parameters and requirements for domain knowledge to determine

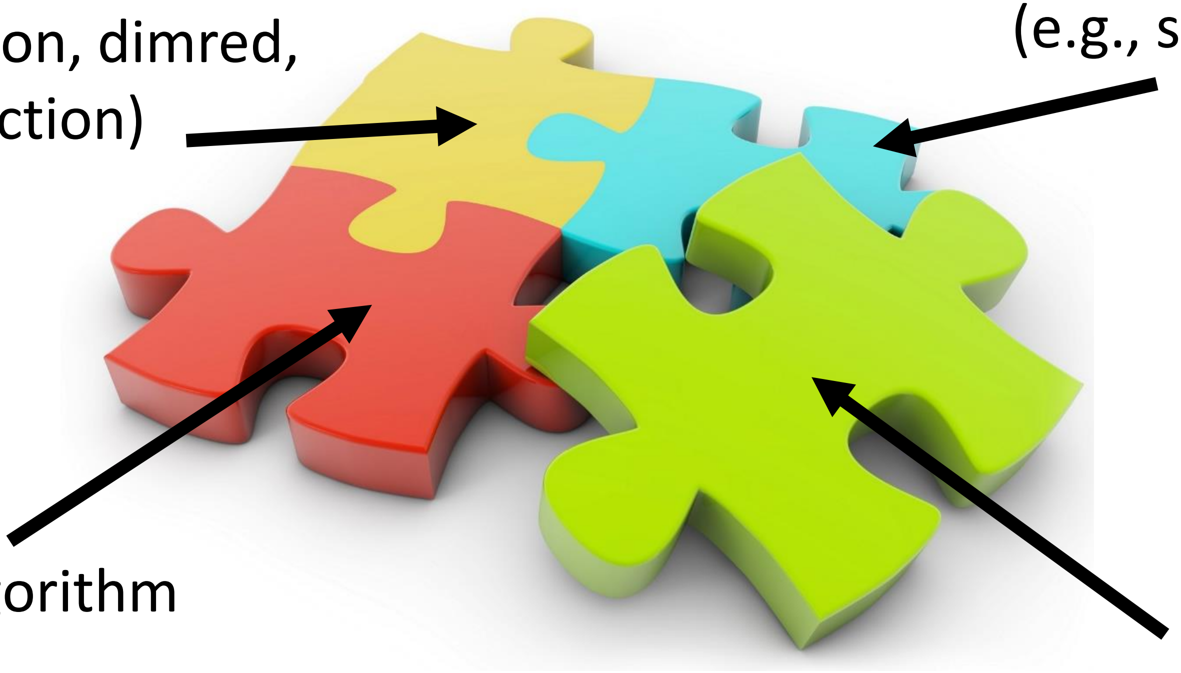
Technicality

- Possibility to work with various distance metrics
- Sensitivity to noise and outliers
- Ability to work in high-dimensional spaces

Flexibility

- Ability for online learning
- Incorporation of user-specified constraints
- Interpretability and usability

Solving clustering problem



Data preprocessing (normalization, dimred, feature selection)

Cluster validation (e.g., stability tests)

Clustering algorithm

Cluster interpretation