Introduction
Cluster Algorithms
Density-based Algorithms
Analysis
Summary

Density-based Cluster Algorithms in Low-dimensional and High-dimensional Applications

Benno Stein
Bauhaus University Weimar
Web-based Information Systems

Michael Busch
IBM Silicon Valley Laboratory
(WebSphere II Project)

Introduction

Given a set of objects (documents) $D = \{\mathbf{d_1}, \ldots, \mathbf{d_n}\}$.

Clustering is the *unsupervised* classification of $\mathbf{d_i}$ into groups.

Result is a partitioning $\mathcal{C}$ of D.

Introduction

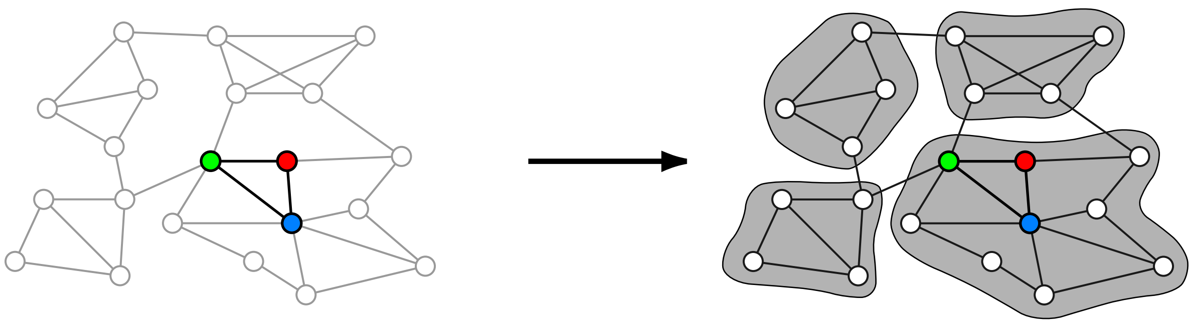
Given a set of objects (documents) $D = \{\mathbf{d_1}, \ldots, \mathbf{d_n}\}$.

Clustering is the *unsupervised* classification of $\mathbf{d_i}$ into groups.

Result is a partitioning $\mathcal{C}$ of D.

Objective: Maximize intra-group similarity.

Minimize inter-group similarity.



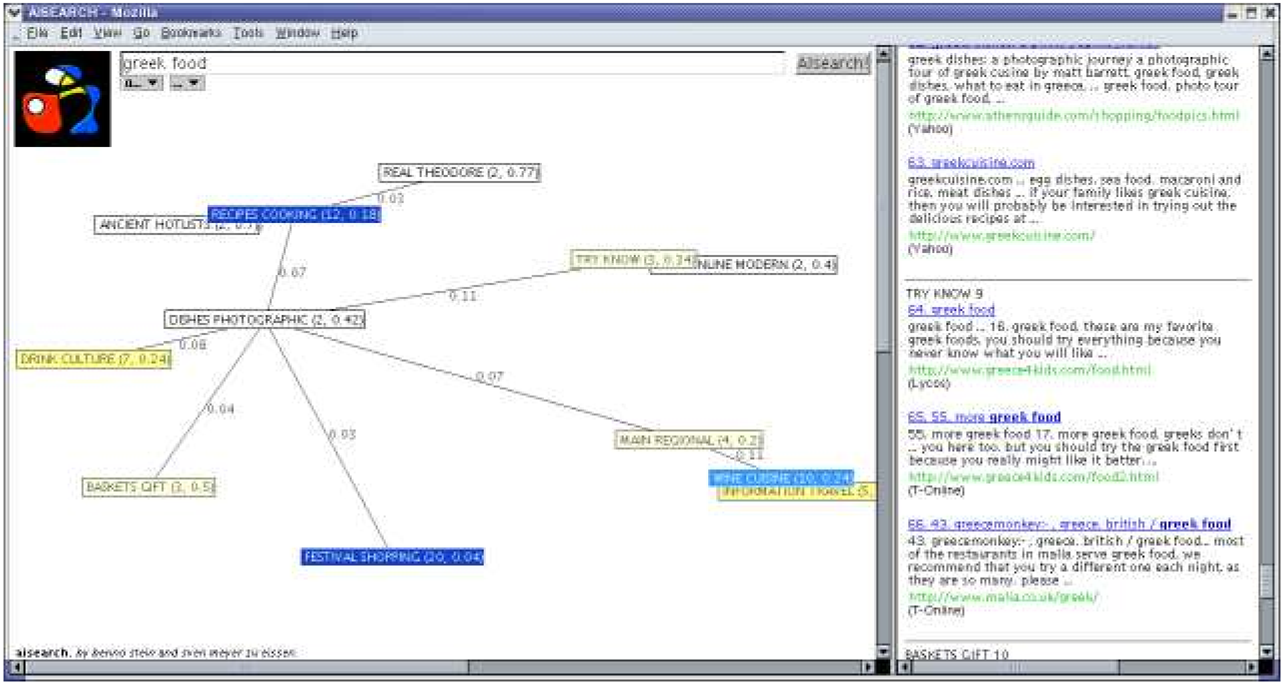
Similarity graph

Clustered graph

Introduction

Cluster algorithms form the backbone of document categorization.

Example AIsearch [www.aisearch.de] :



Introduction

Documents → Feature extraction → Similarity computation → Cluster analysis → Categories

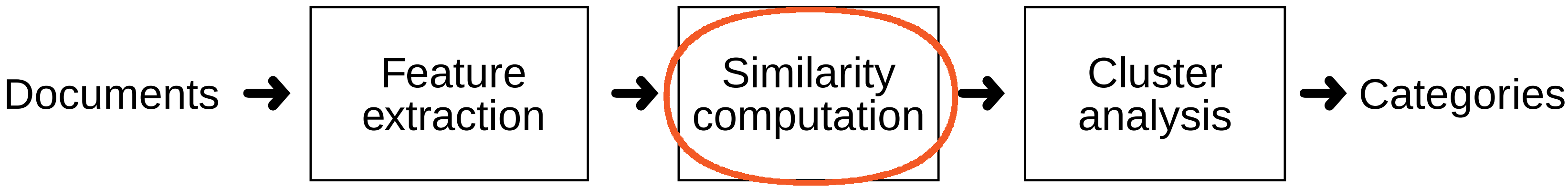
Indexing (includes parsing, stopword elimination, stemming):

Vector representation
with weighting scheme:

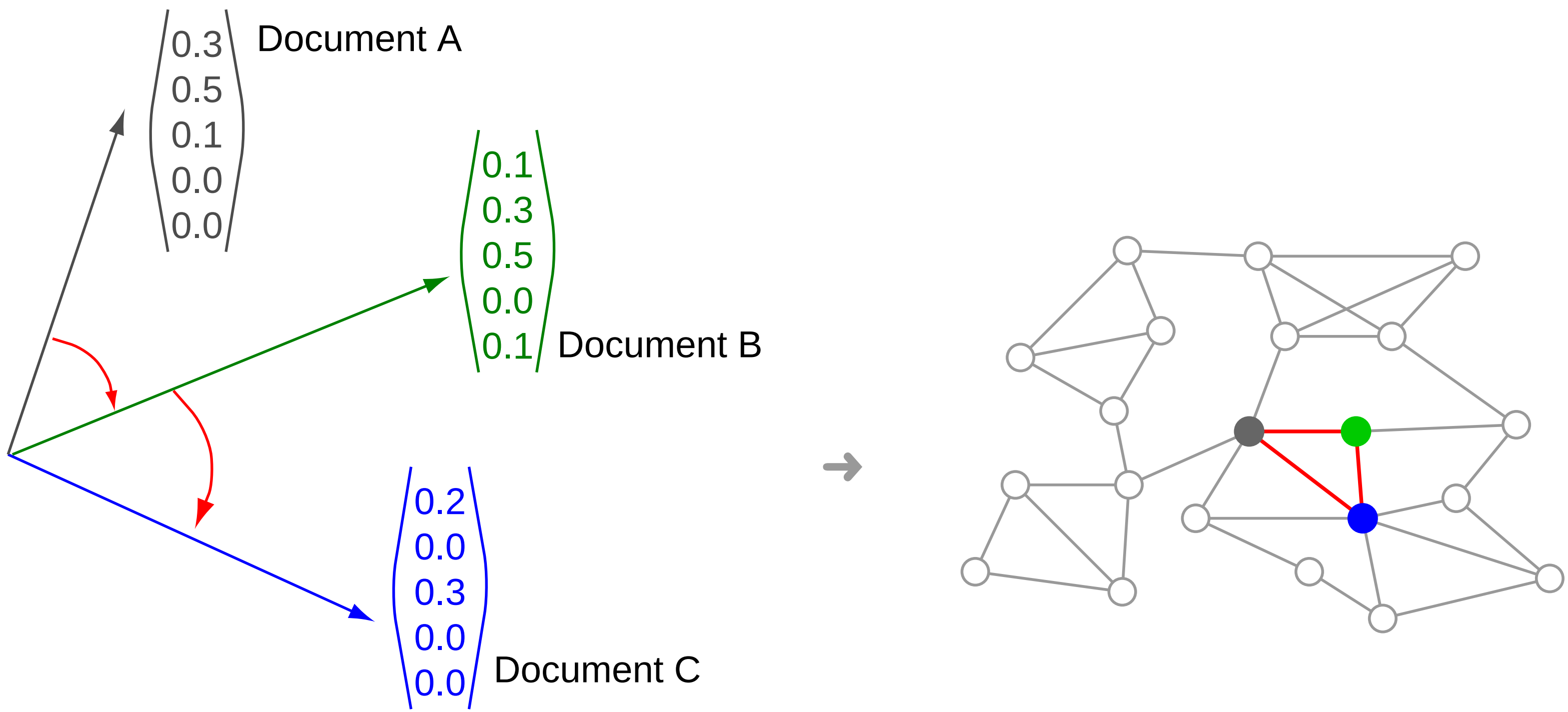
$$\underbrace{\#abs}_{tf} \cdot \underbrace{\log \frac{|X|}{\#docs}}_{idf}$$

$$\begin{pmatrix} \texttt{chrysler} & 0.12 \\ \texttt{deal} & 0.2 \\ \texttt{leav} & 0.1 \\ \texttt{amc} & 0.01 \\ \texttt{cat} & 0.0 \\ \texttt{sal} & 0.01 \\ \texttt{dog} & 0.0 \\ \vdots & \vdots \end{pmatrix}$$

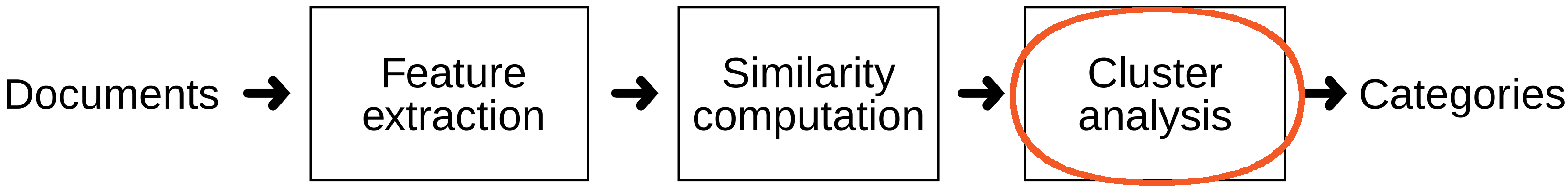
Introduction

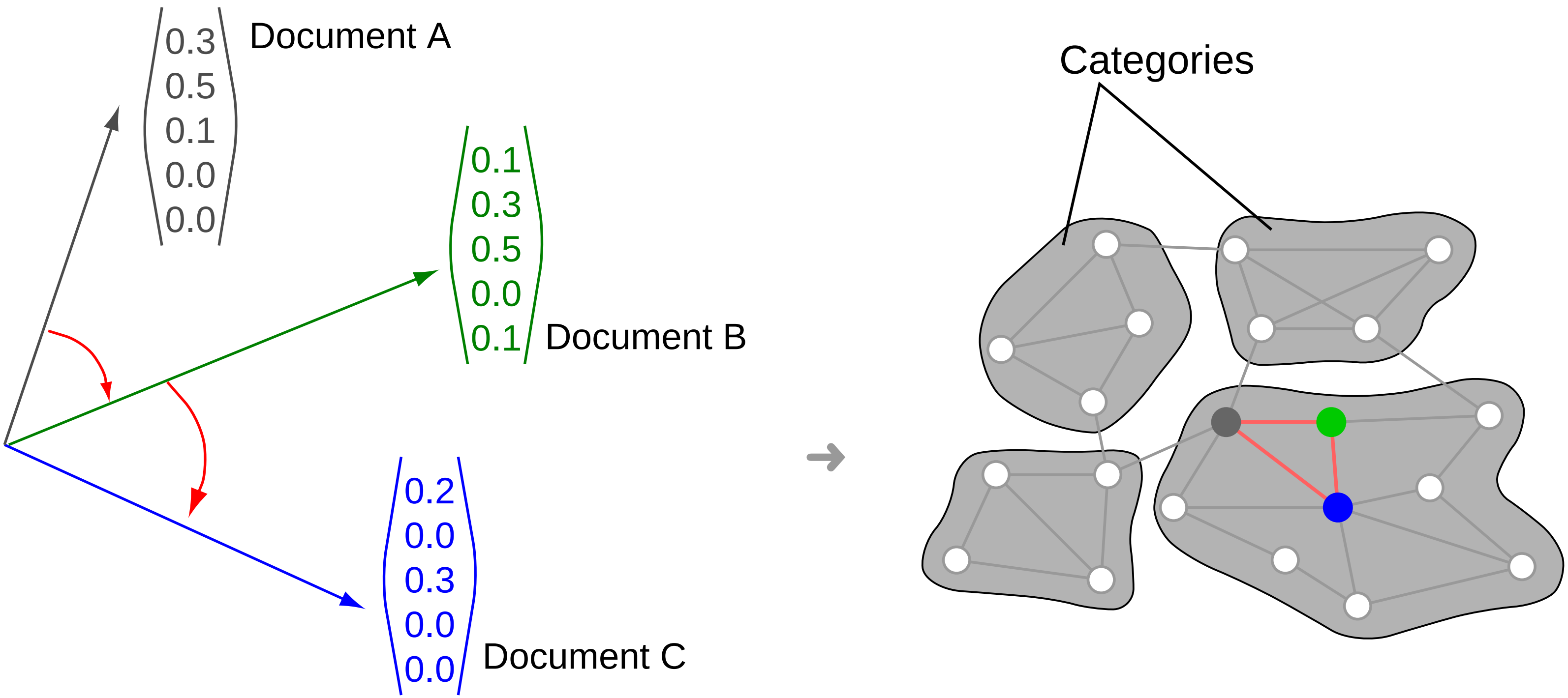


e. g. under the vector space model:

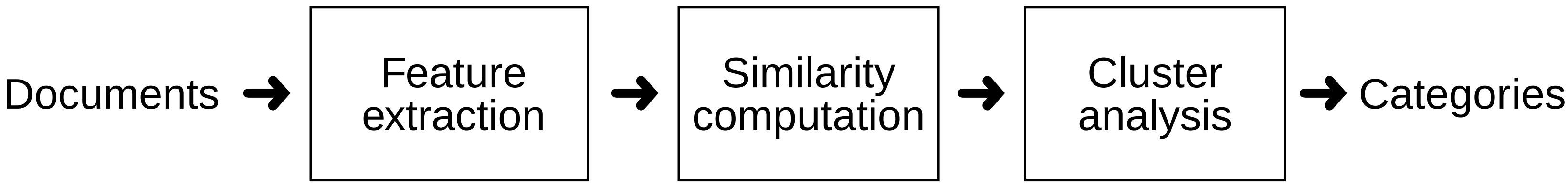


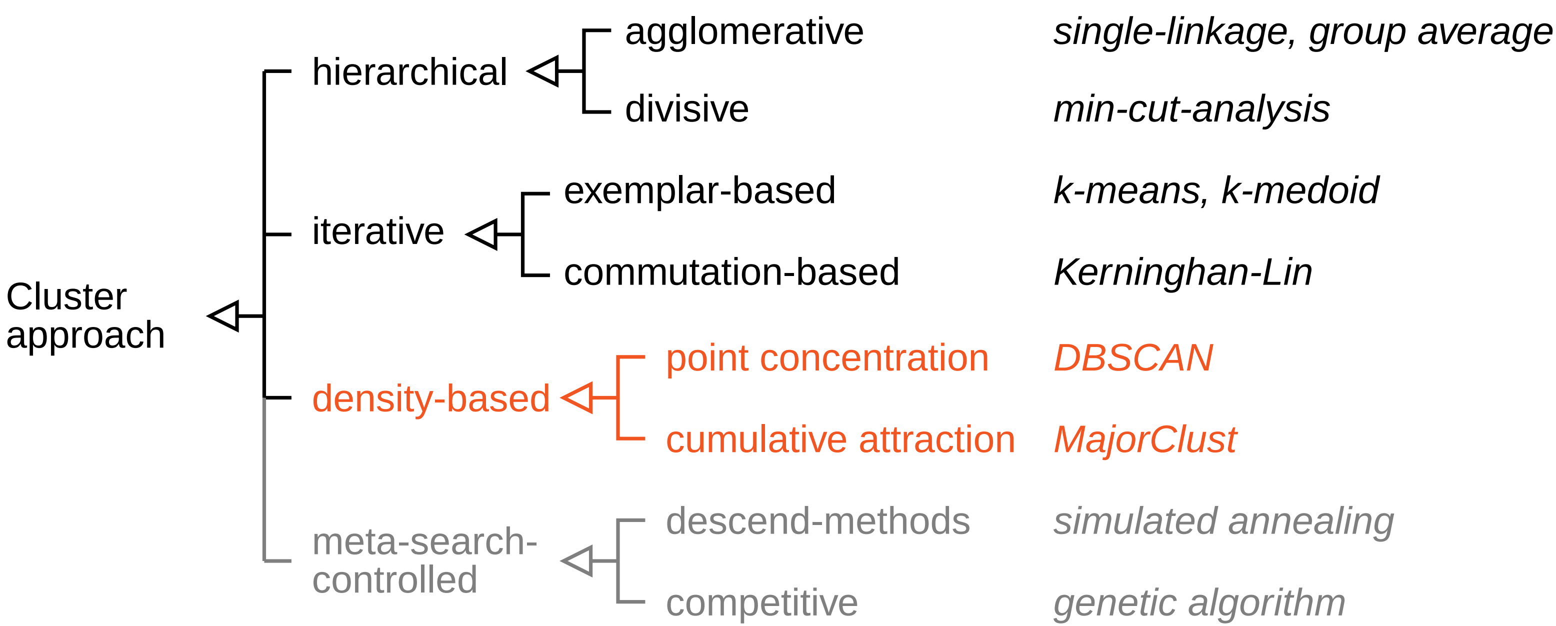
Introduction



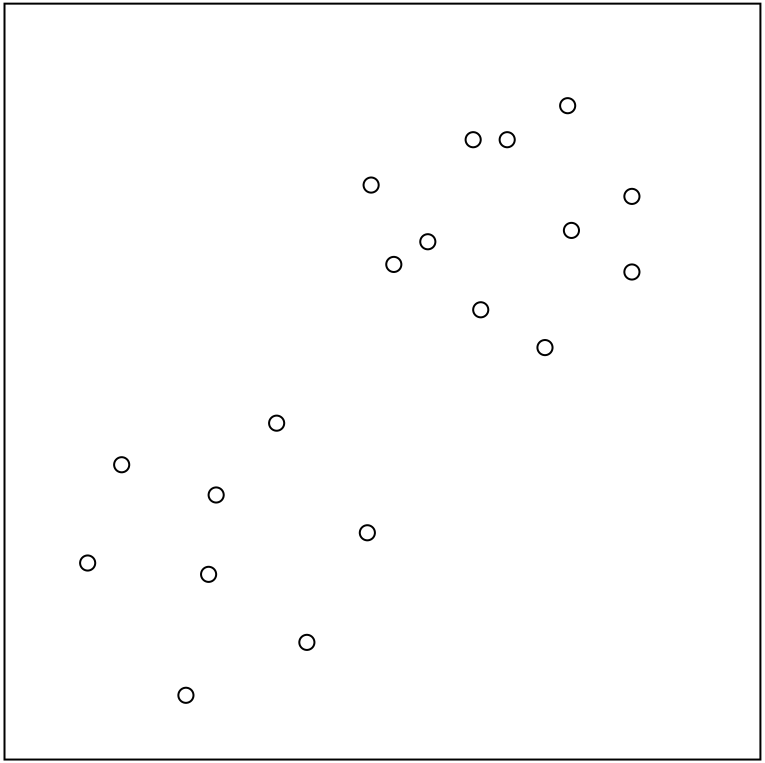


Introduction

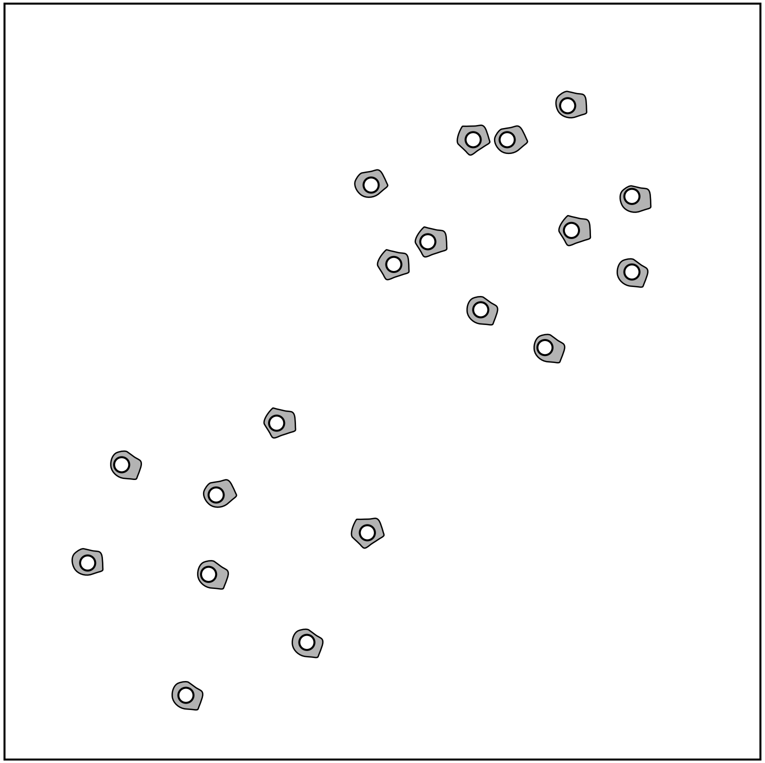




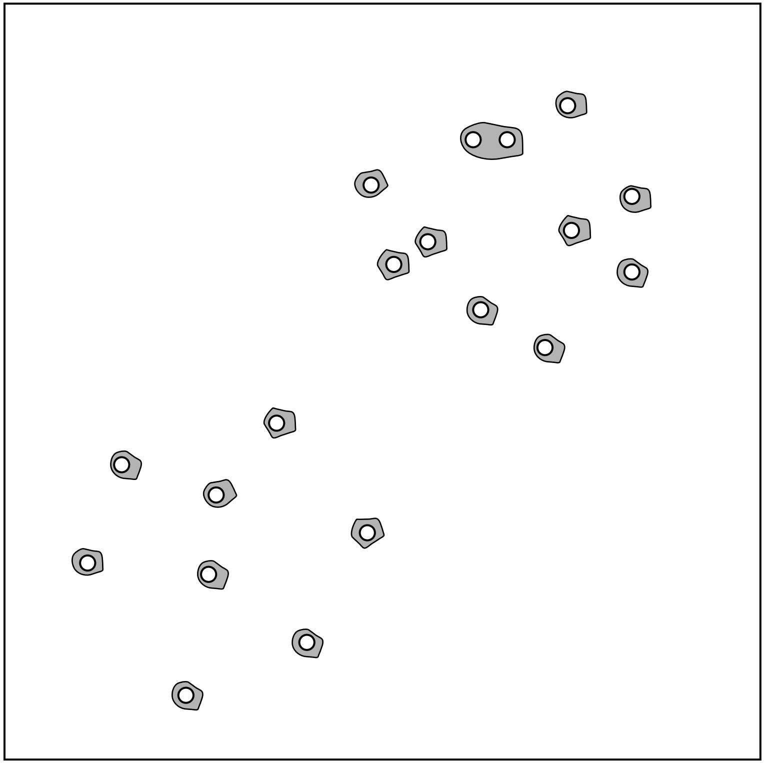
Hierarchical agglomerative: Single-linkage



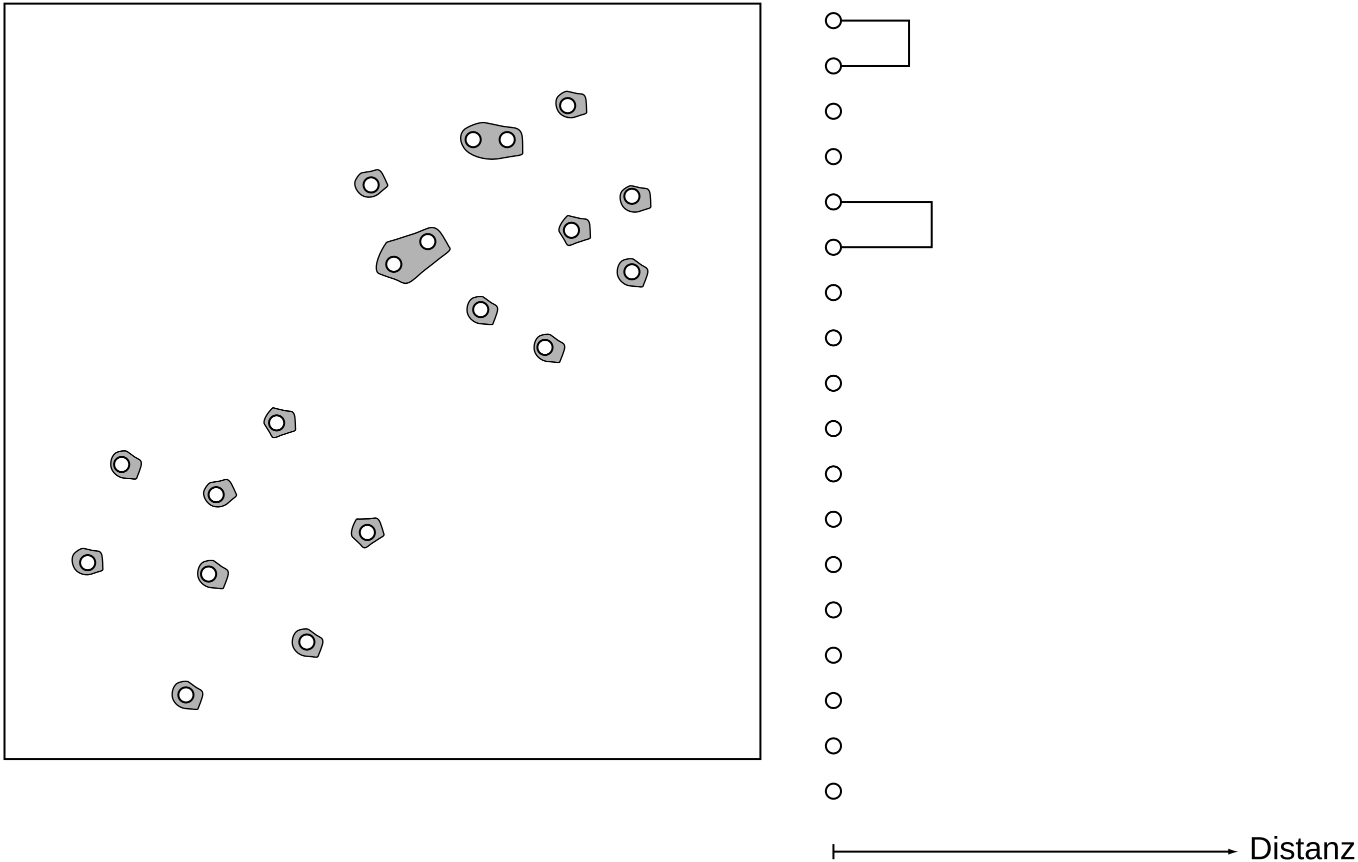
Hierarchical agglomerative: Single-linkage



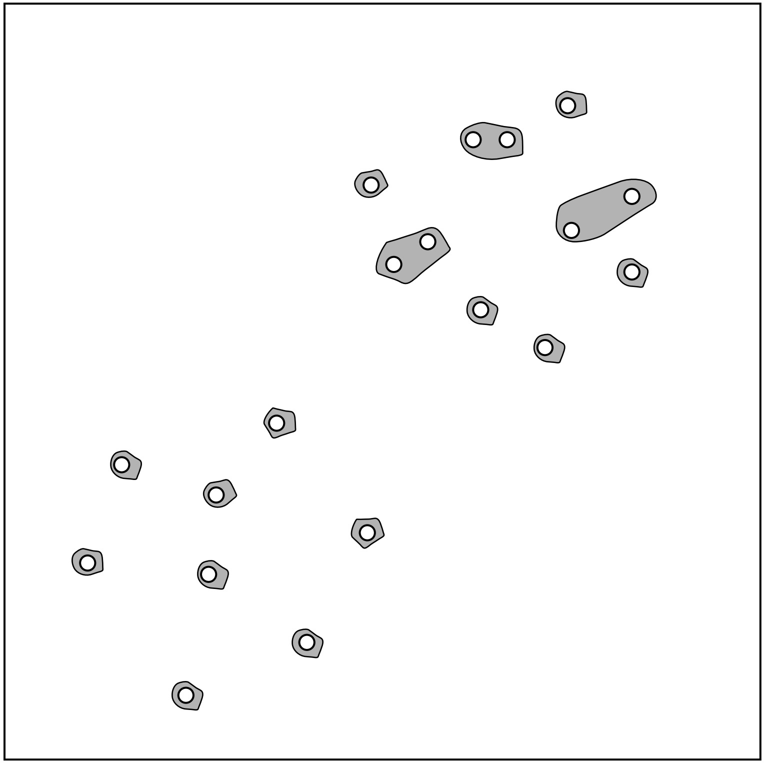
Hierarchical agglomerative: Single-linkage

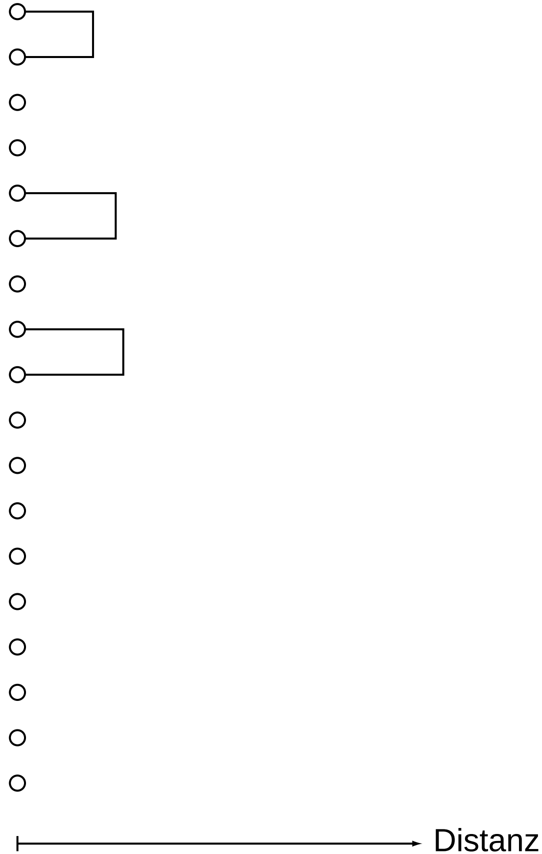


Hierarchical agglomerative: Single-linkage

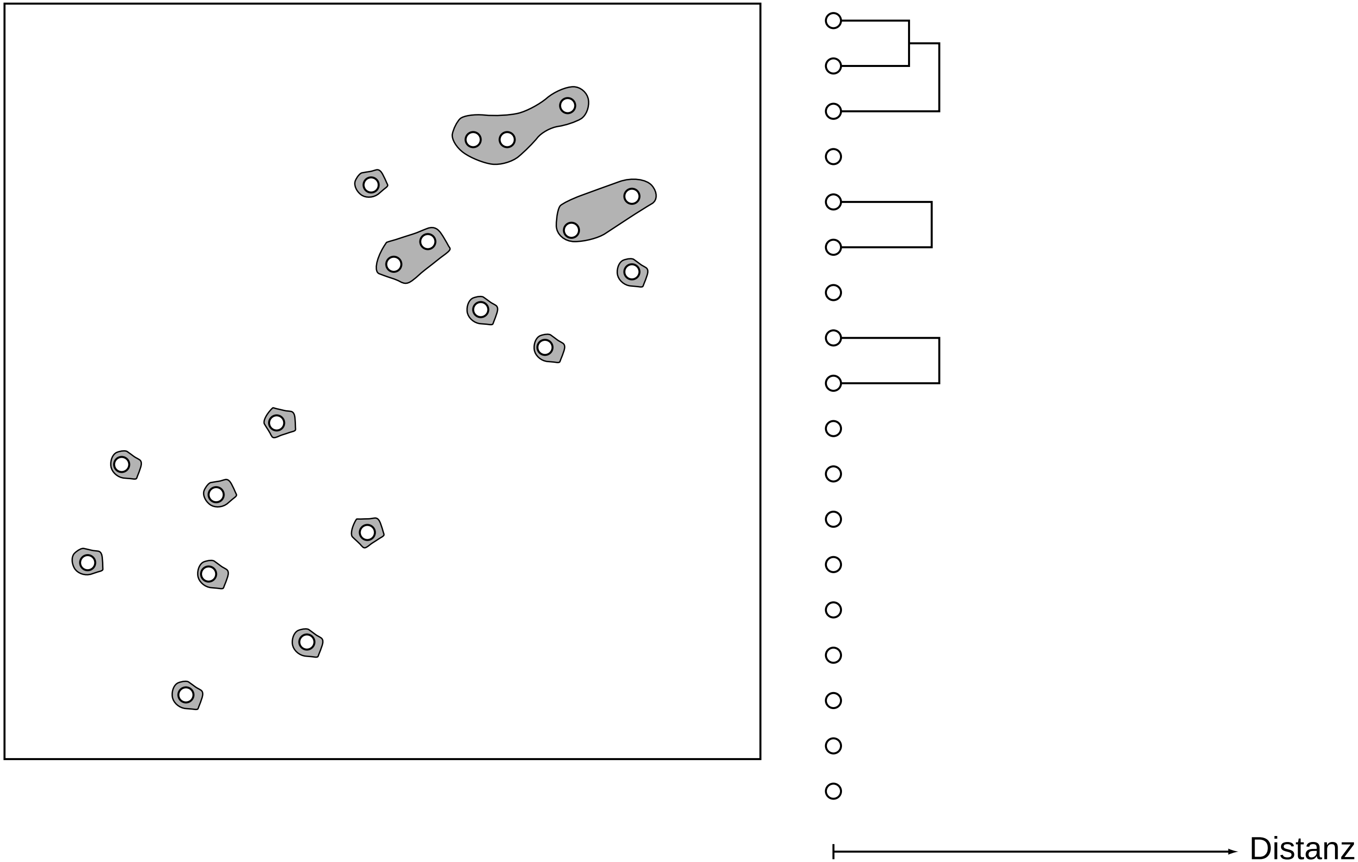


Hierarchical agglomerative: Single-linkage

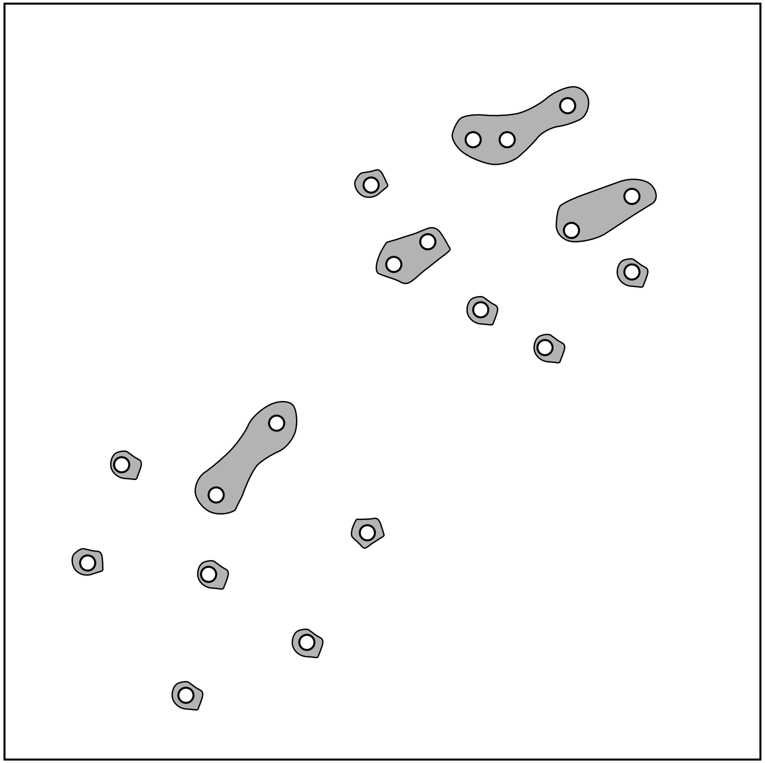


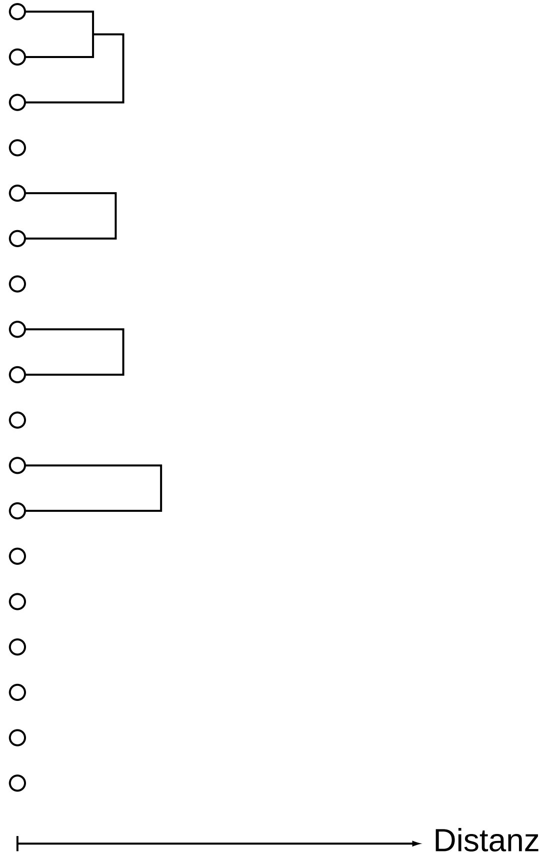


Hierarchical agglomerative: Single-linkage

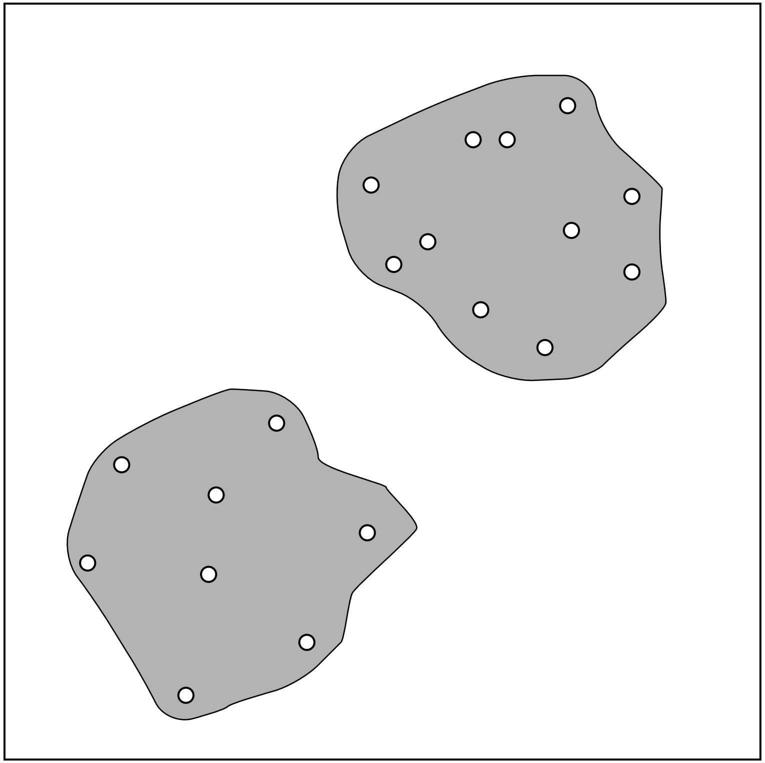


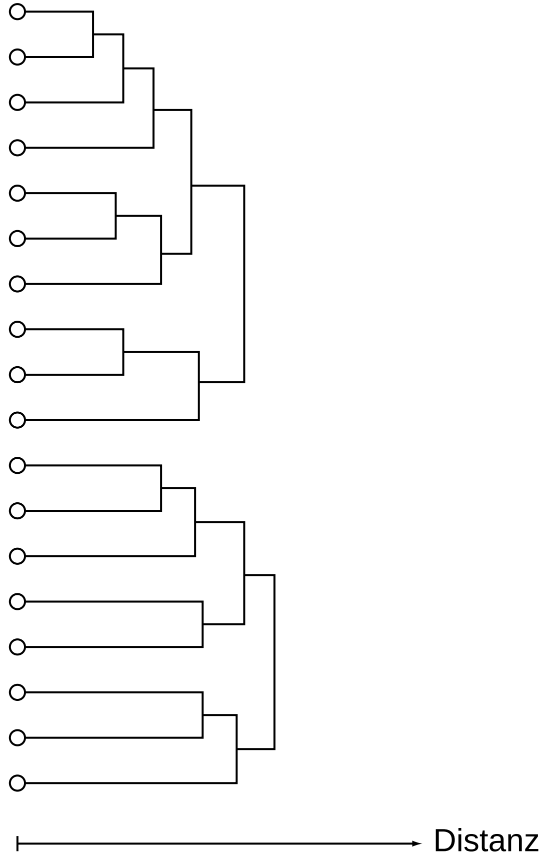
Hierarchical agglomerative: Single-linkage



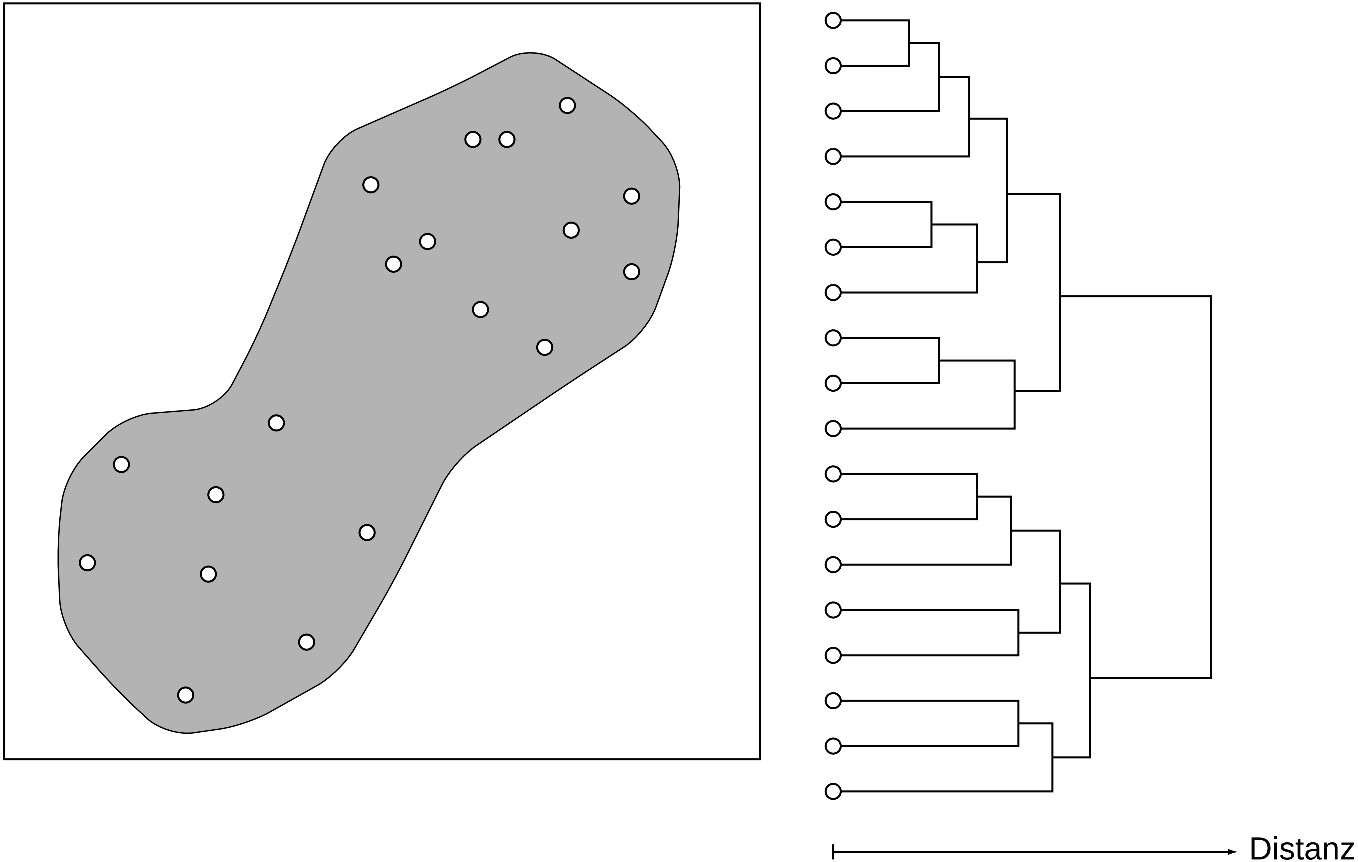


Hierarchical agglomerative: Single-linkage

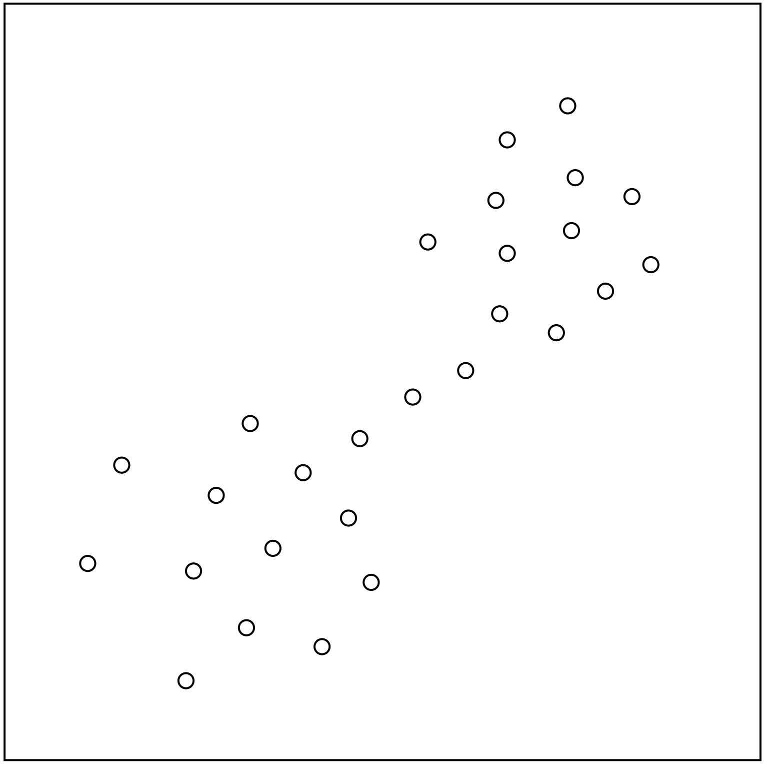




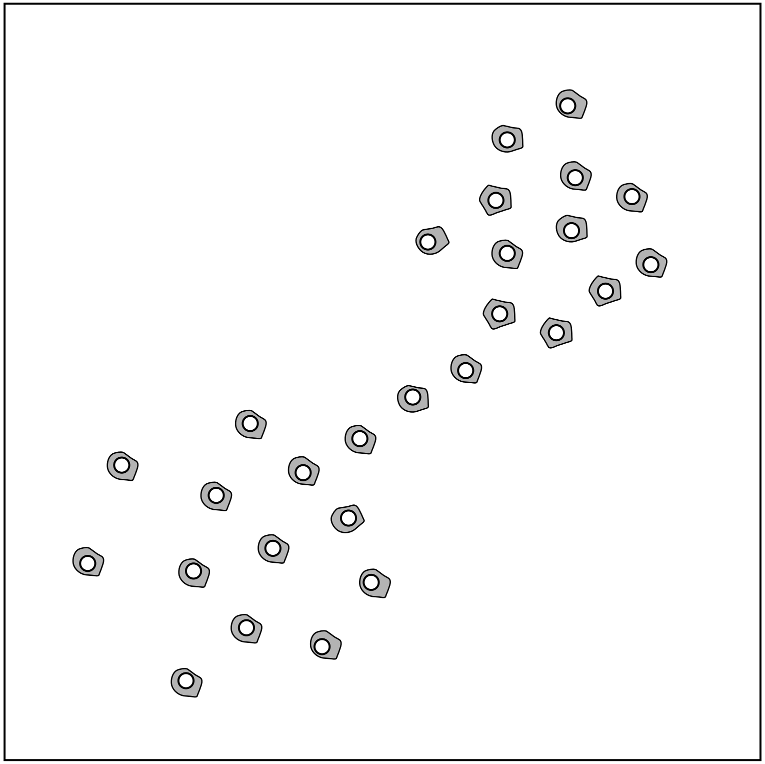
Hierarchical agglomerative: Single-linkage



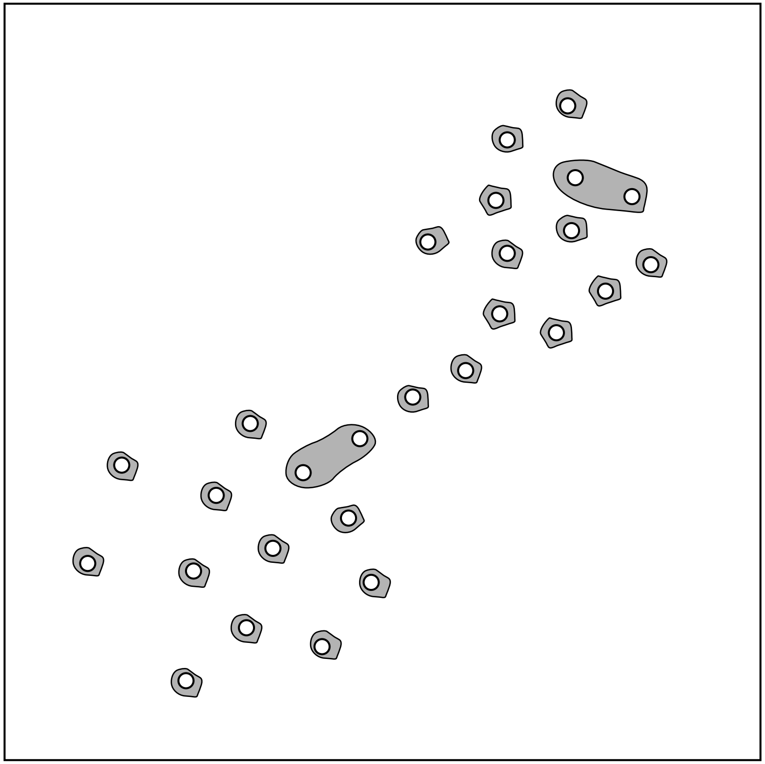
Single-linkage: Chaining Problem



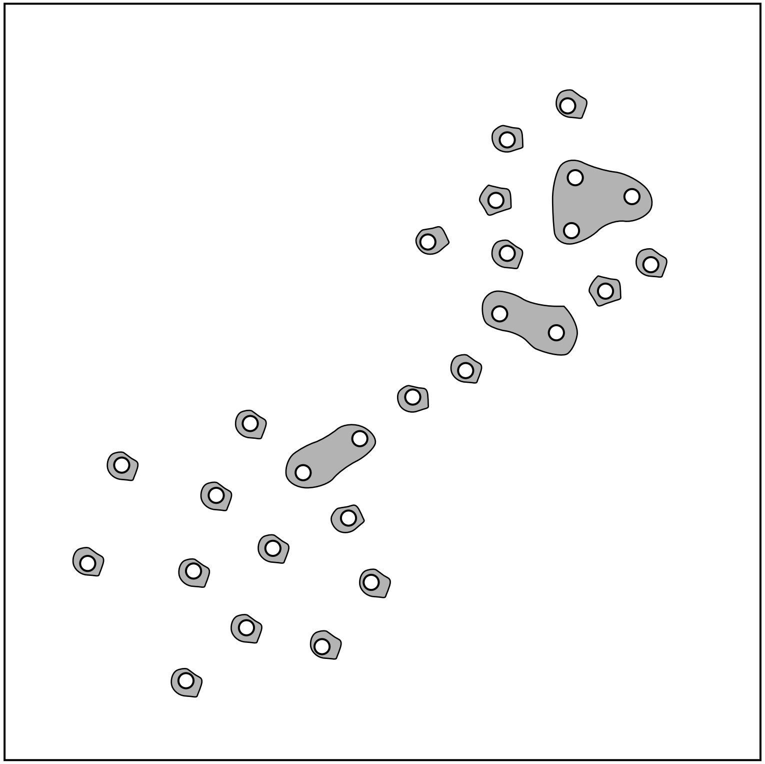
Single-linkage: Chaining Problem



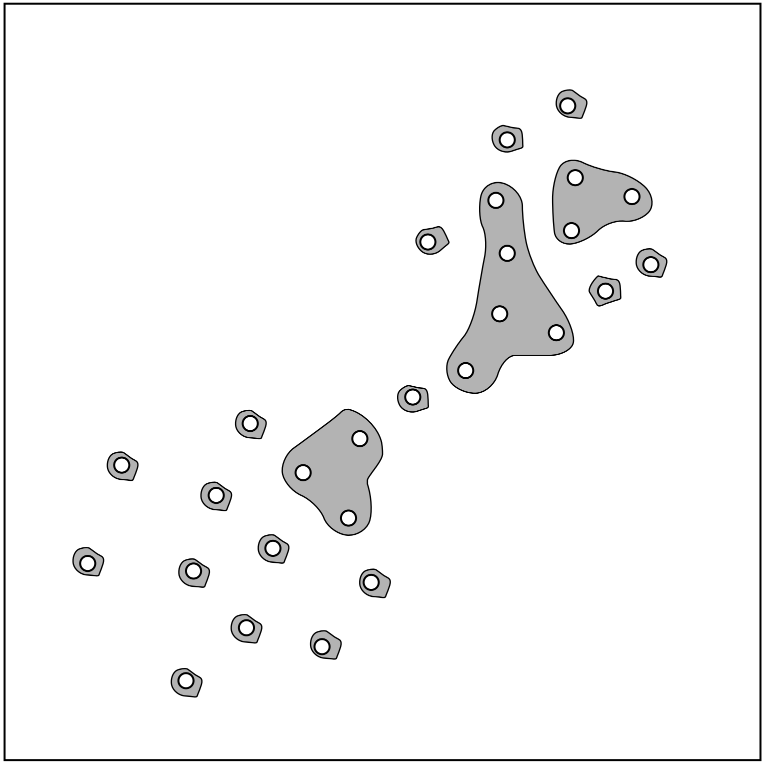
Single-linkage: Chaining Problem



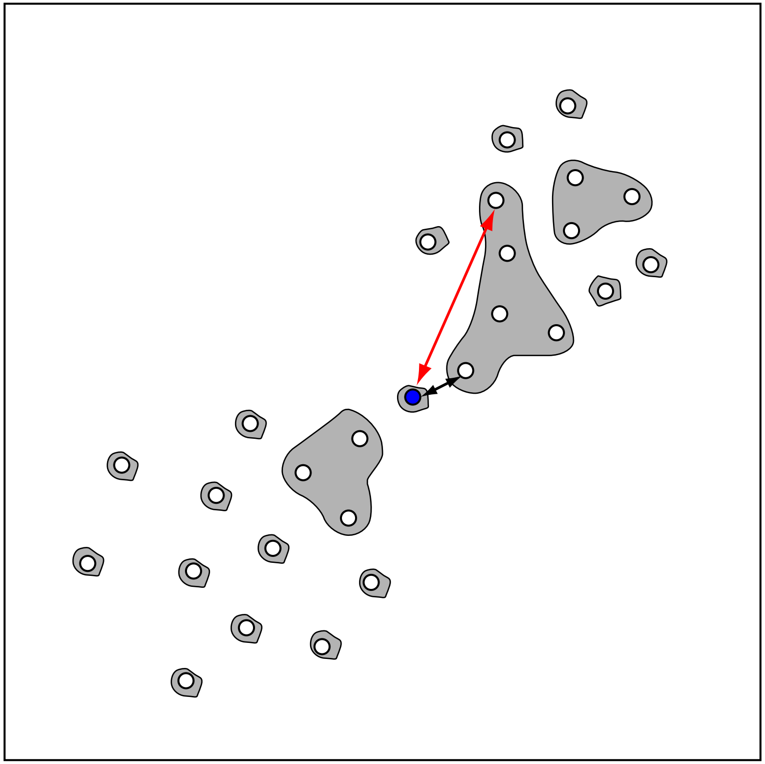
Single-linkage: Chaining Problem



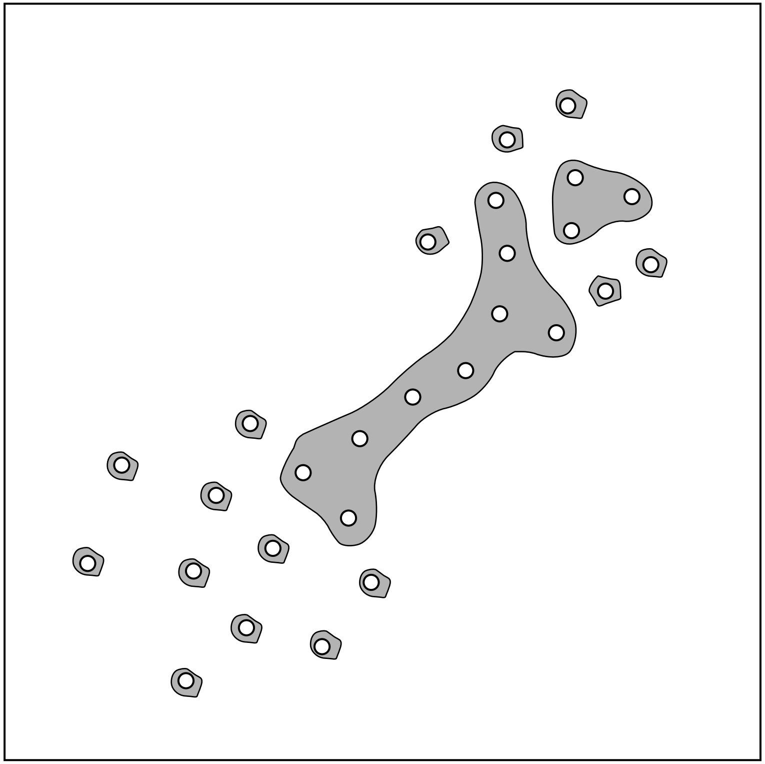
Single-linkage: Chaining Problem



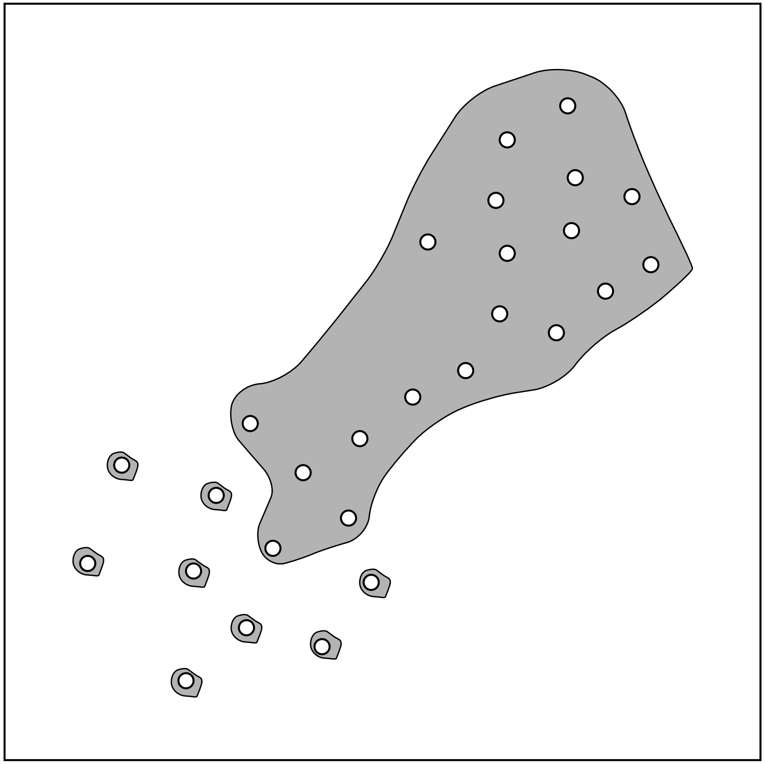
Single-linkage: Chaining Problem



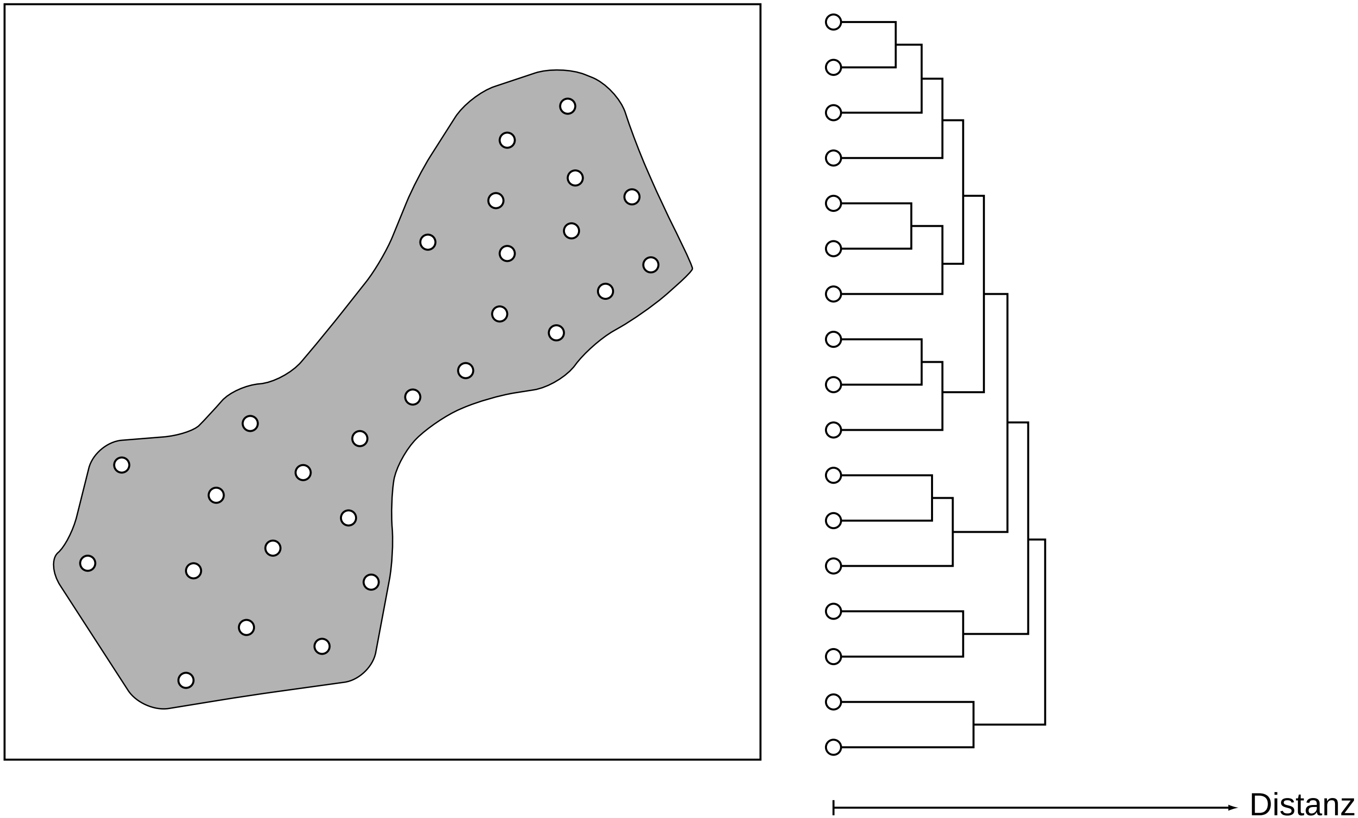
Single-linkage: Chaining Problem



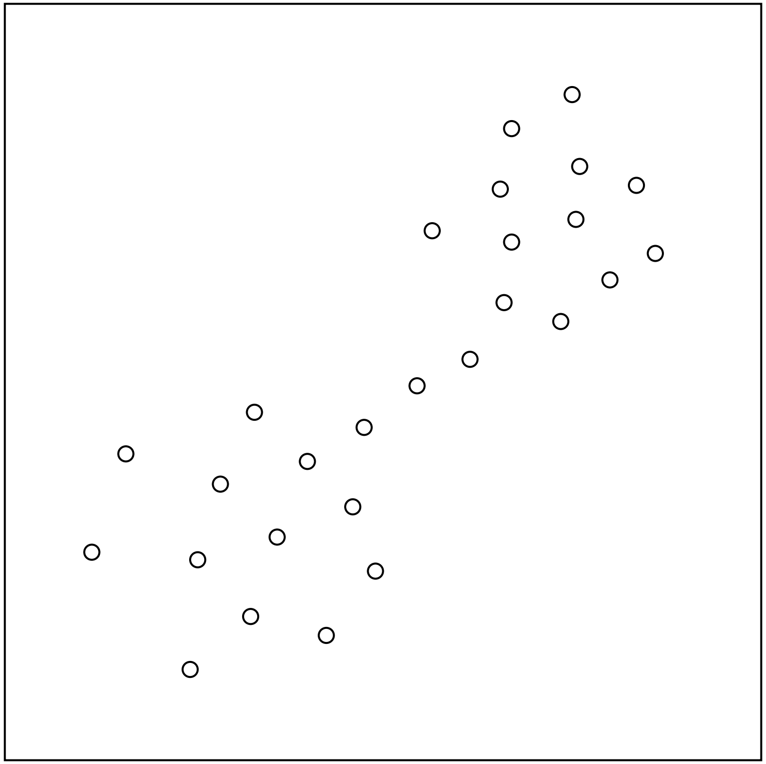
Single-linkage: Chaining Problem



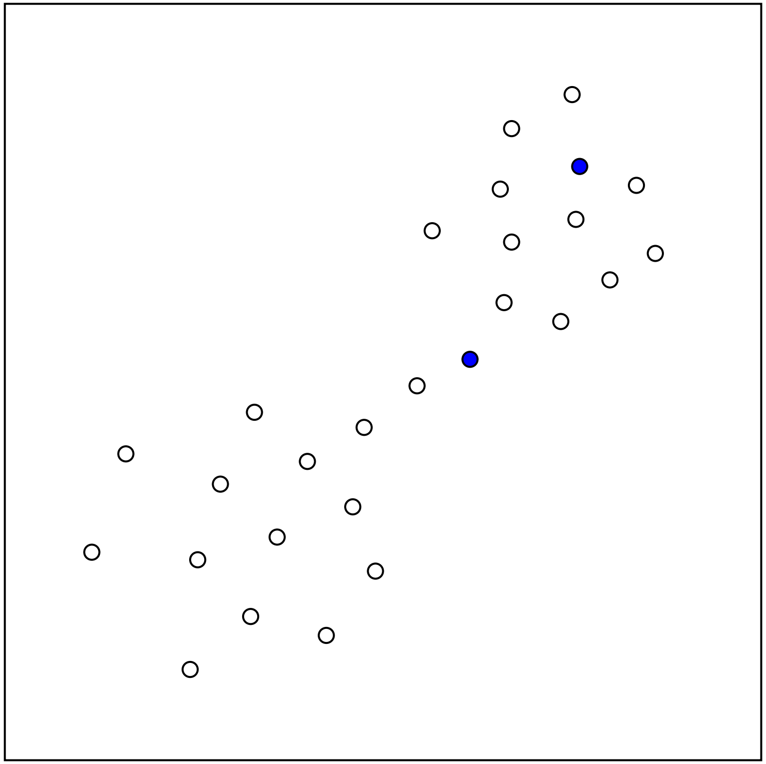
Single-linkage: Chaining Problem



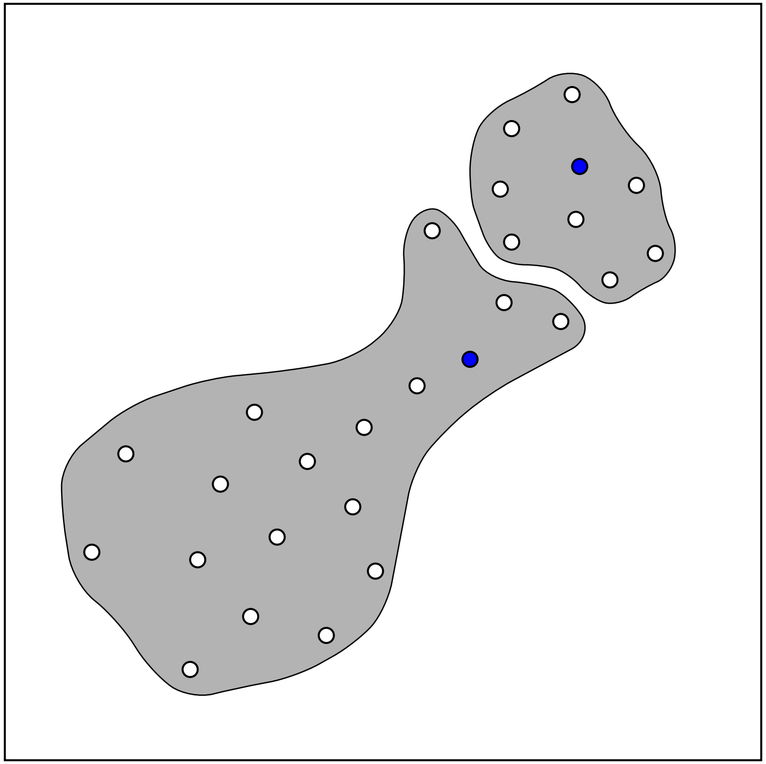
Exemplar-based algorithm: k-Means



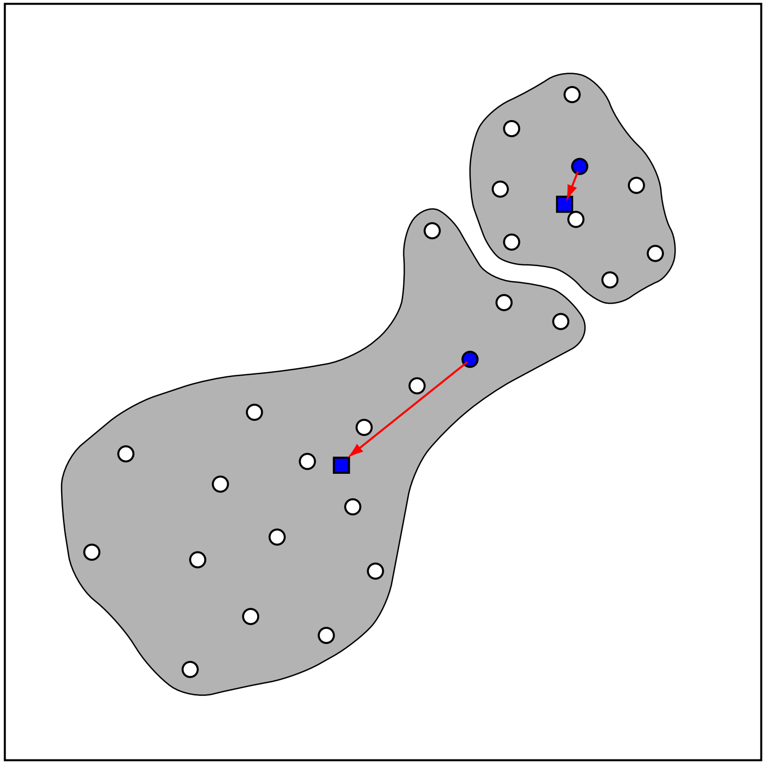
Exemplar-based algorithm: k-Means



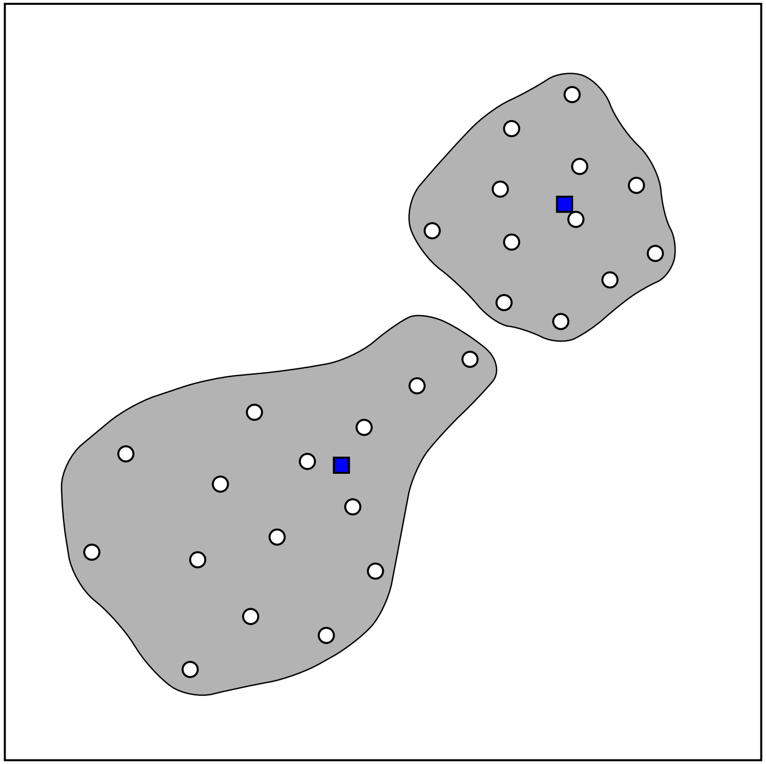
Exemplar-based algorithm: k-Means



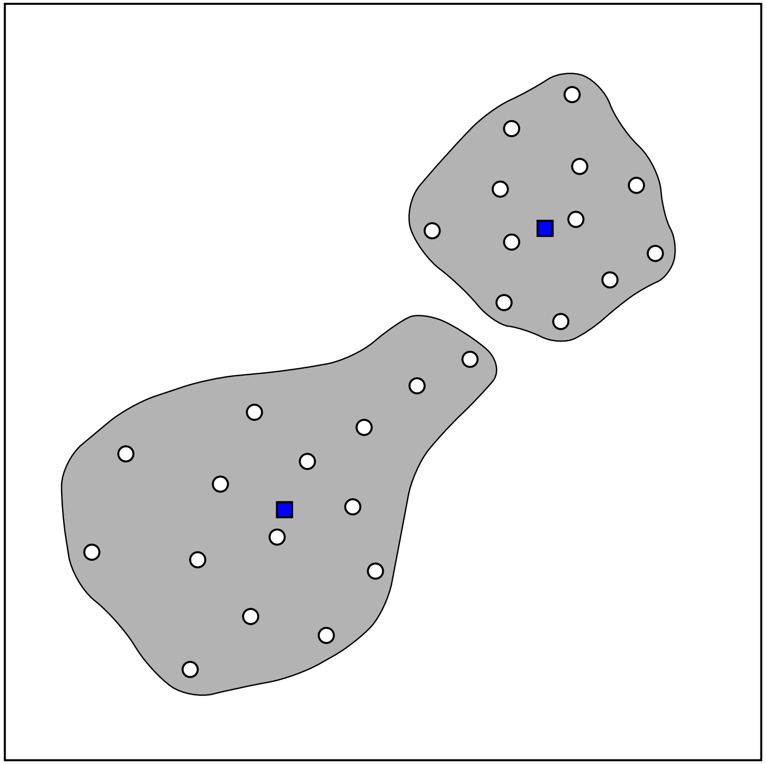
Exemplar-based algorithm: k-Means



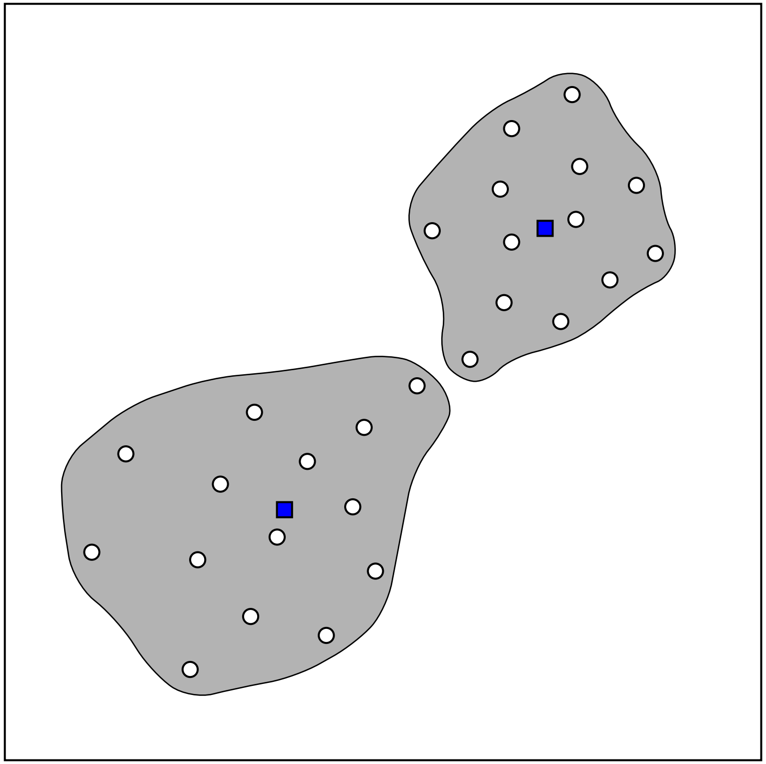
Exemplar-based algorithm: k-Means



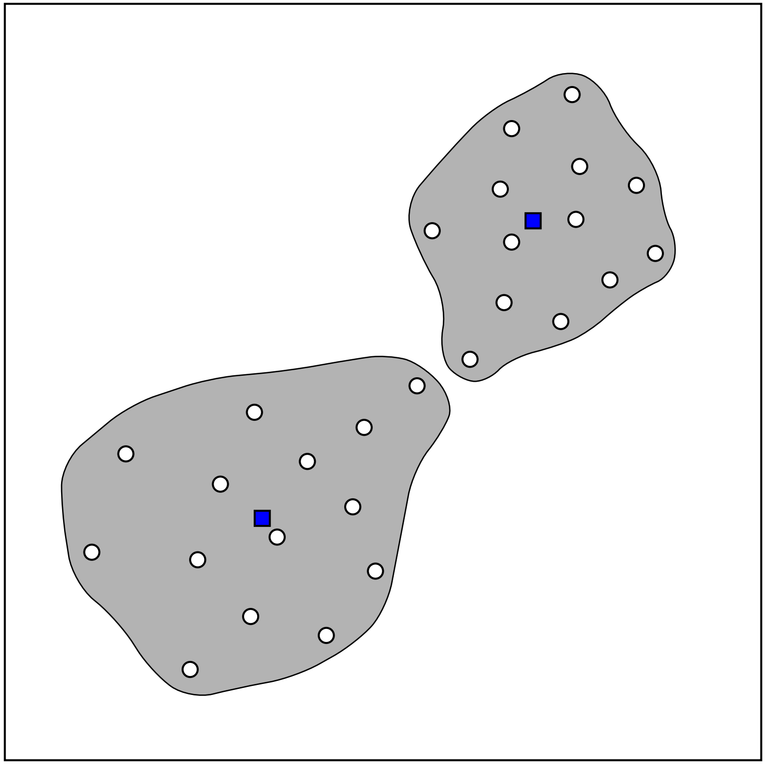
Exemplar-based algorithm: k-Means



Exemplar-based algorithm: k-Means

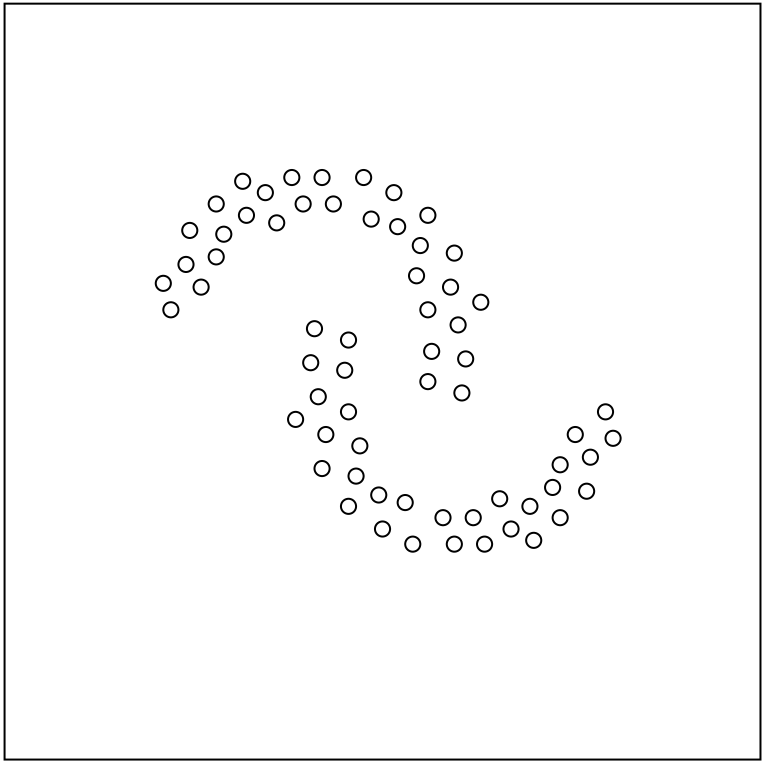


Exemplar-based algorithm: k-Means



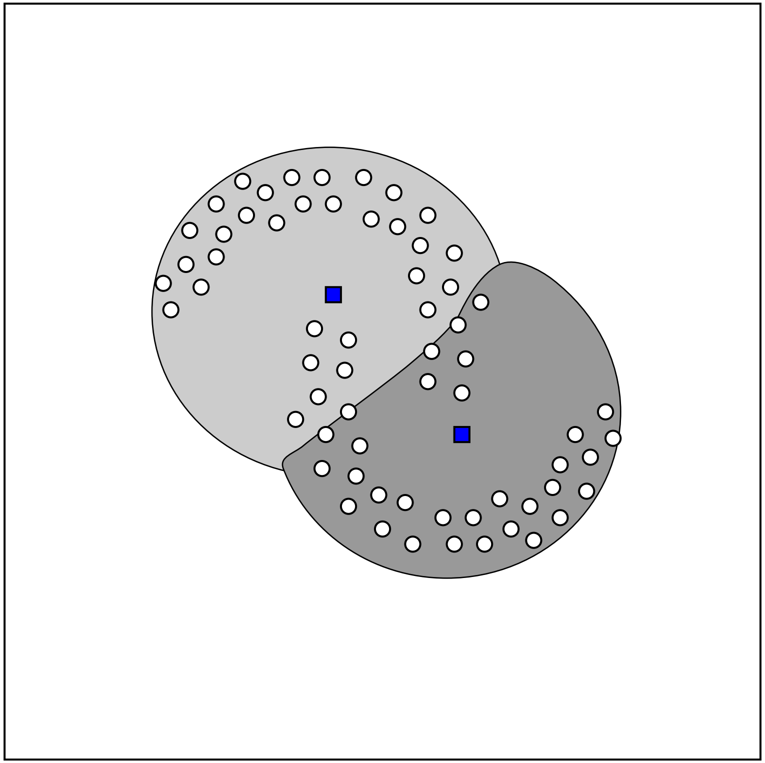
Exemplar-based versus Linkage

Exemplar-based algorithms fail with large differences in size.



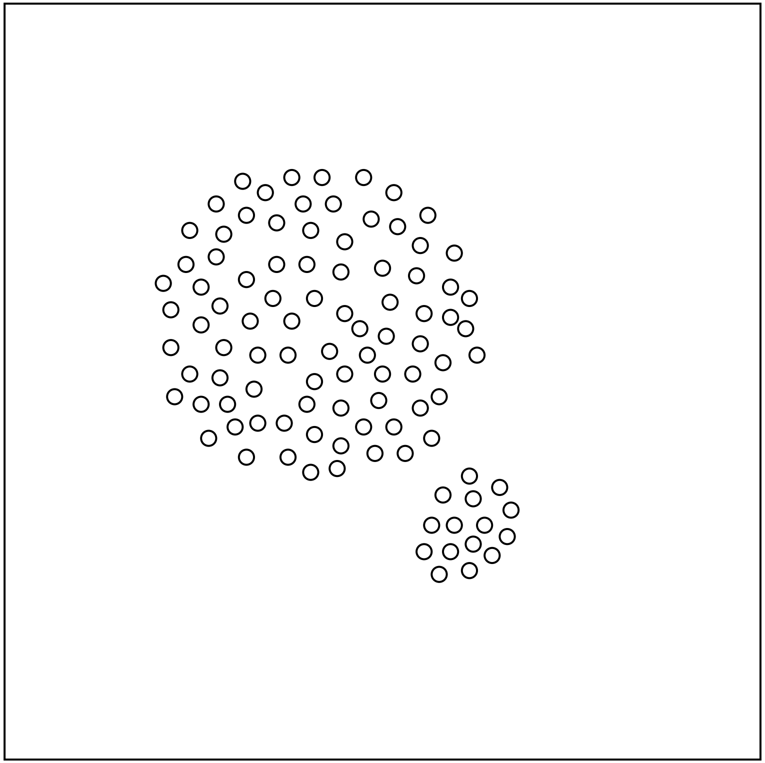
Exemplar-based versus Linkage

Exemplar-based algorithms fail with entwined clusters.



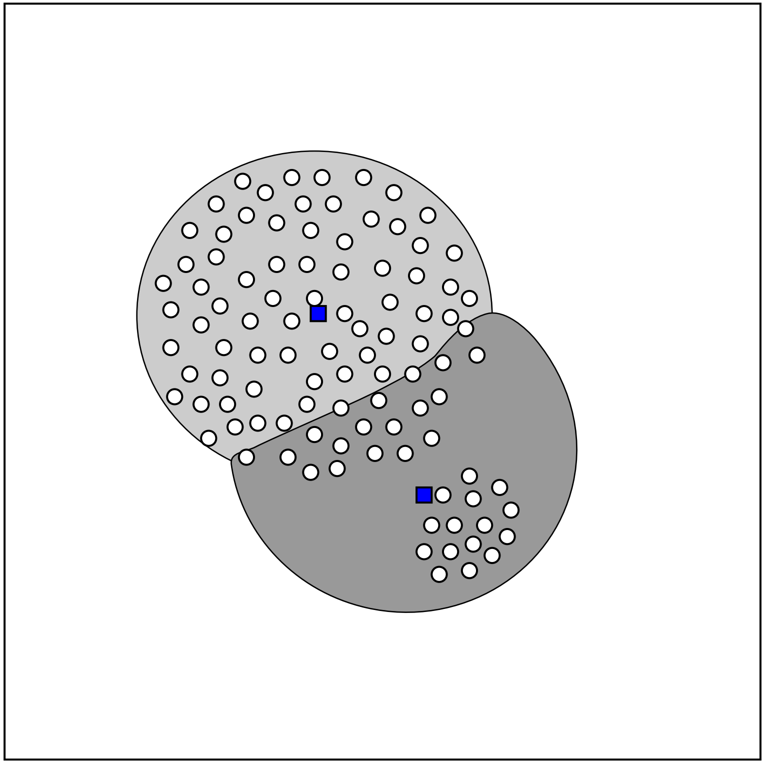
Exemplar-based versus Linkage

Exemplar-based algorithms fail with entwined clusters.



Exemplar-based versus Linkage

Exemplar-based algorithms fail with entwined clusters.



Density-based Cluster Analysis

Density-based algorithms try to separate the set D into subsets of similar densities.

Density estimation can happen

- parameter-based: the underlying distribution is known
- parameter-less: histogramm, kernel function
  (construct barcharts, superimpose continuous functions)

Density-based Cluster Analysis

Density-based algorithms try to separate the set D into subsets of similar densities.

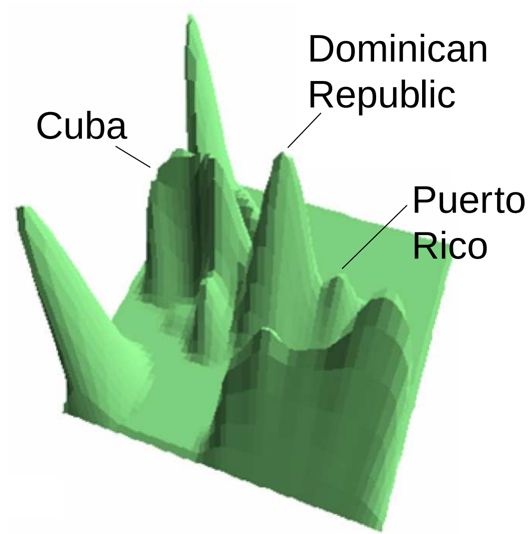
Density estimation can happen

- parameter-based: the underlying distribution is known
- parameter-less: histogramm, kernel function
  (construct barcharts, superimpose continuous functions)

Example (Carribean Isands):

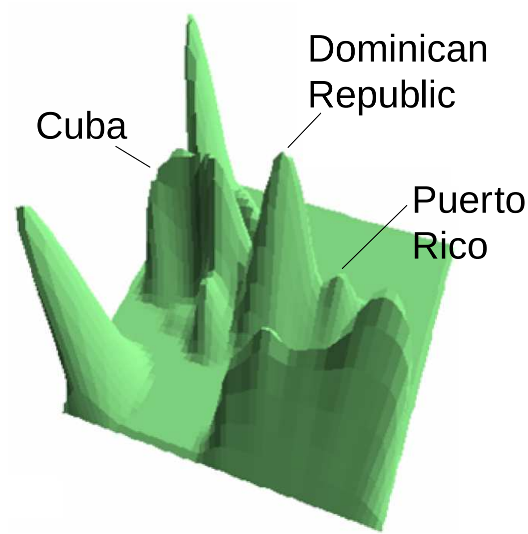
Density-based Cluster Analysis

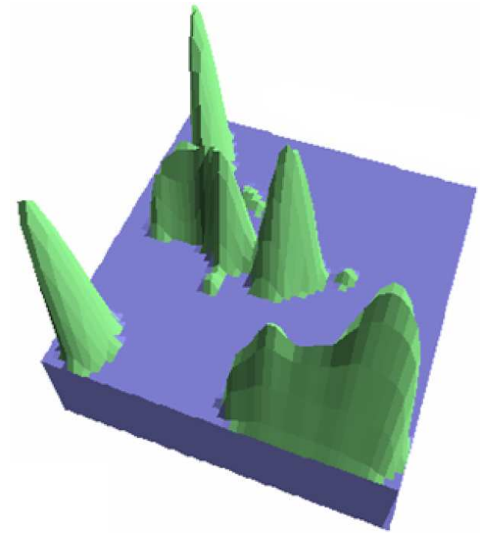
Density estimation with Gaussian Kernel for the example.



Density-based Cluster Analysis

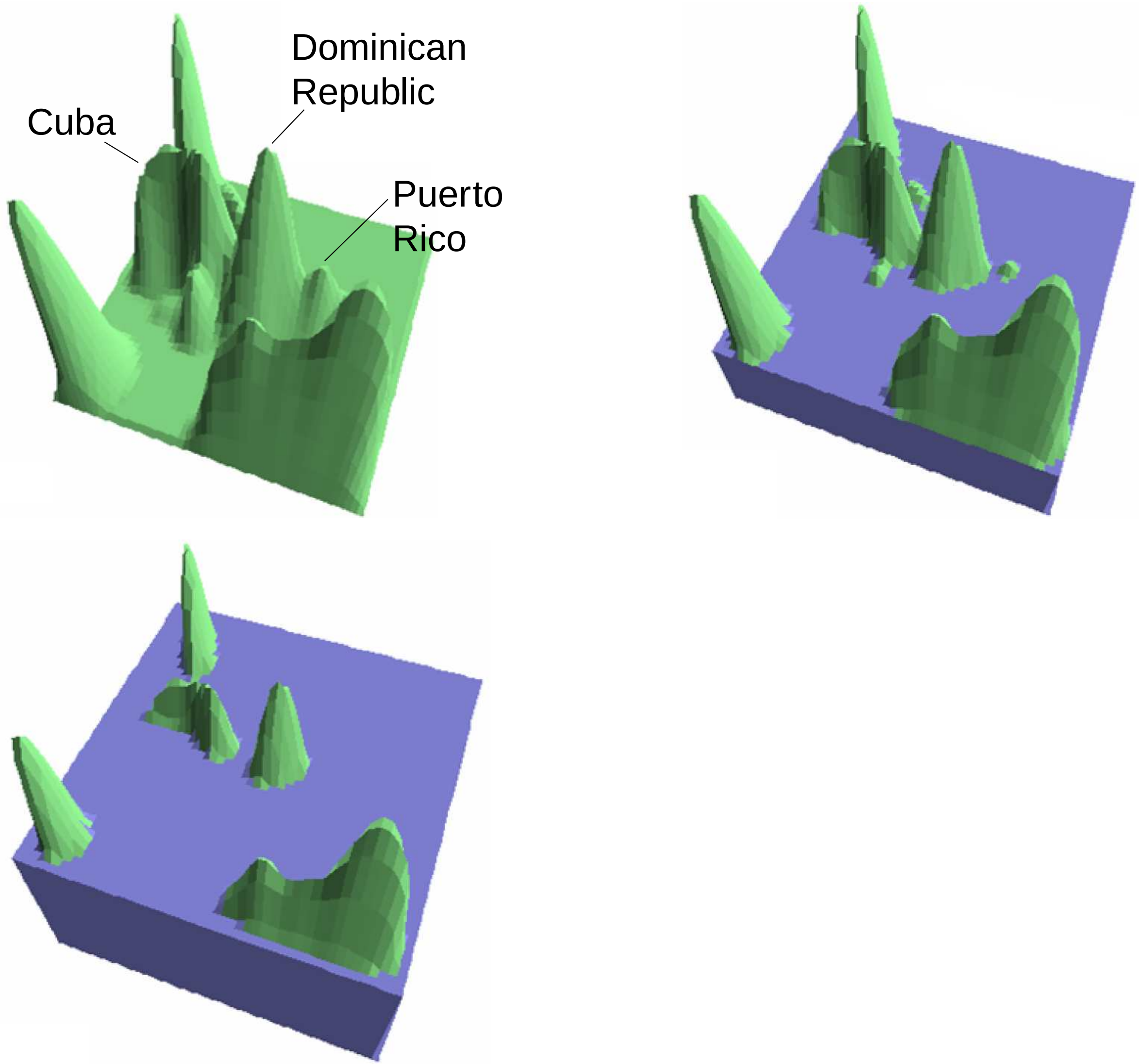
Density estimation with Gaussian Kernel for the example.





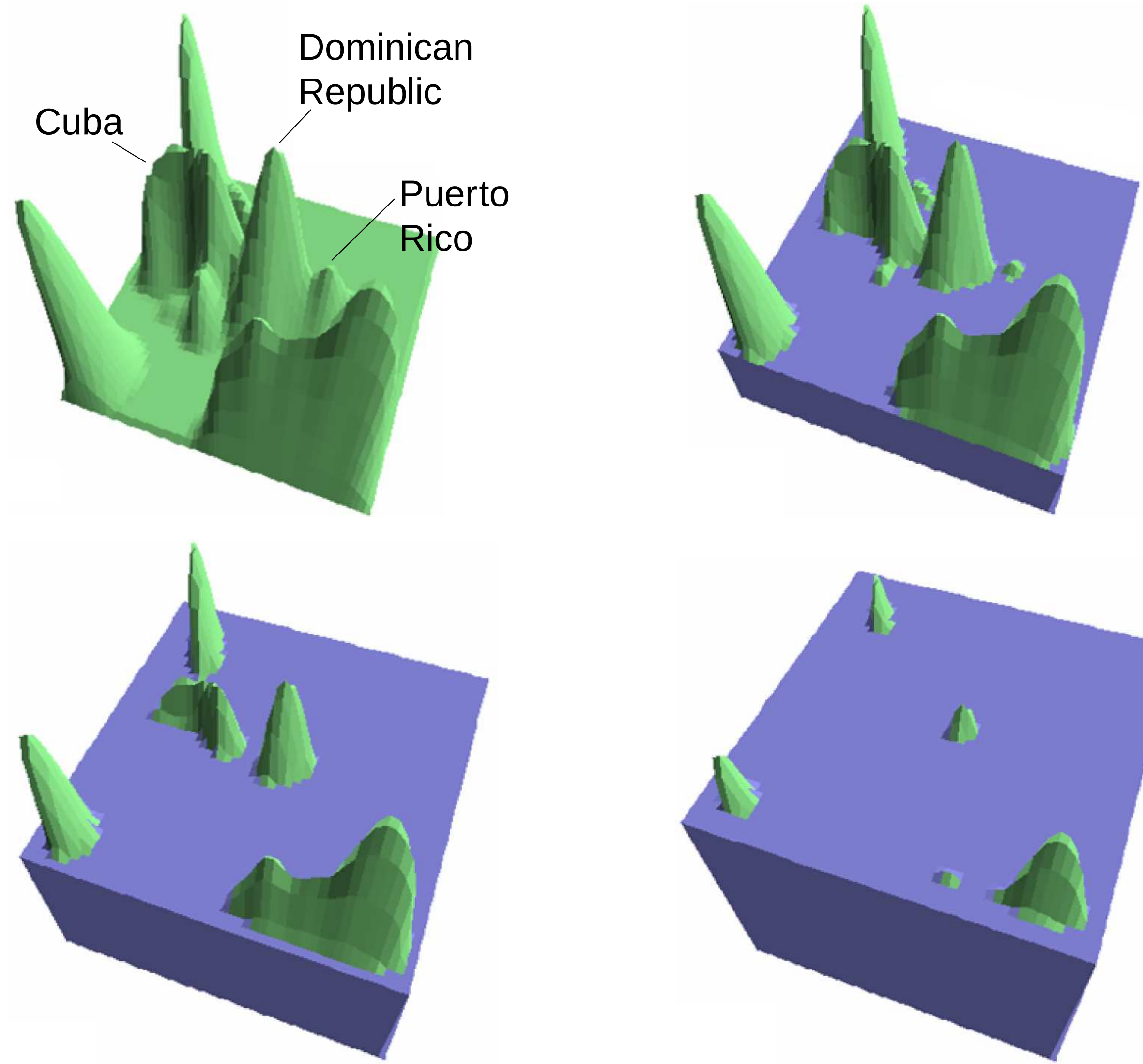
Density-based Cluster Analysis

Density estimation with Gaussian Kernel for the example.



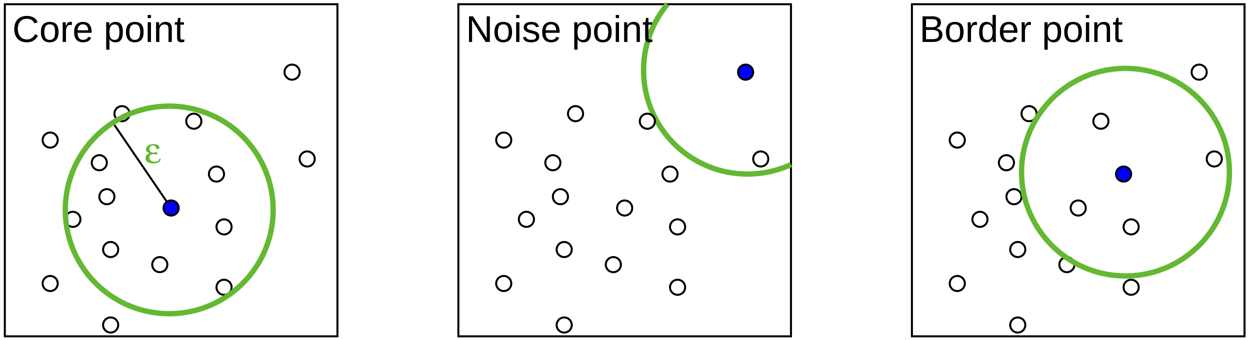
Density-based Cluster Analysis

Density estimation with Gaussian Kernel for the example.



Density-based Algorithm: DBSCAN [Ester et al. 1996]

Density-based Algorithm: DBSCAN [Ester et al. 1996]

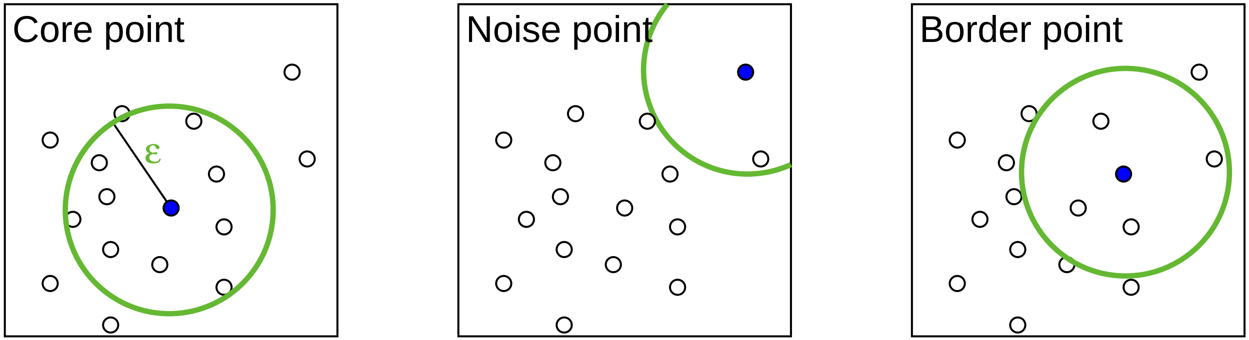


p is core point: $|N_\varepsilon(p)| \geq$ *MinPts*.

p is noise point: p is not density-reachable from a core point.

p is border point: otherwise.

Density-based Algorithm: DBSCAN



p is core point: $|N_\varepsilon(p)| \geq$ *MinPts*.

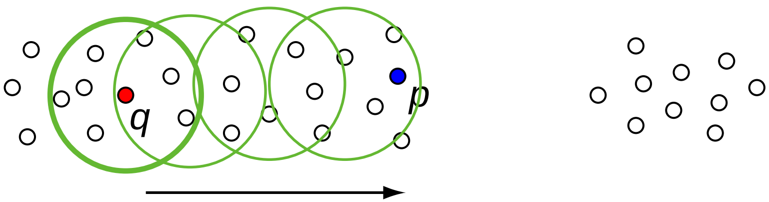
p is noise point: p is not density-reachable from a core point.

p is border point: otherwise.

p is density-reachable from q:

(a) $p \in |N_\varepsilon(q)|$, where q is a corepoint

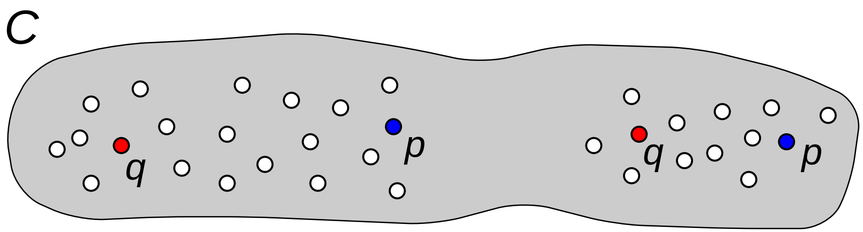
(b) transitive application of condition (a):



Density-based Algorithm: DBSCAN

A cluster $C \subseteq D$ satisfies the following conditions:

1. $\forall p, q$: If $p \in C$ and q is density-reachable from p then $q \in C$.

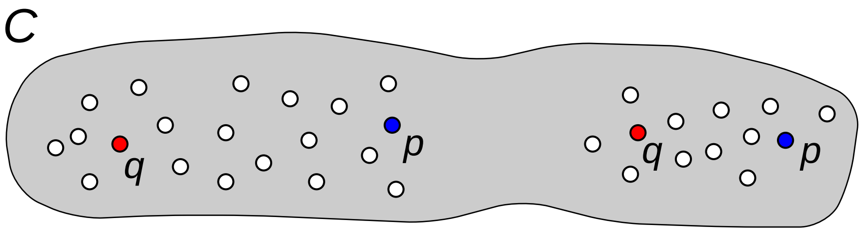


Maximality condition

Density-based Algorithm: DBSCAN

A cluster $C \subseteq D$ satisfies the following conditions:

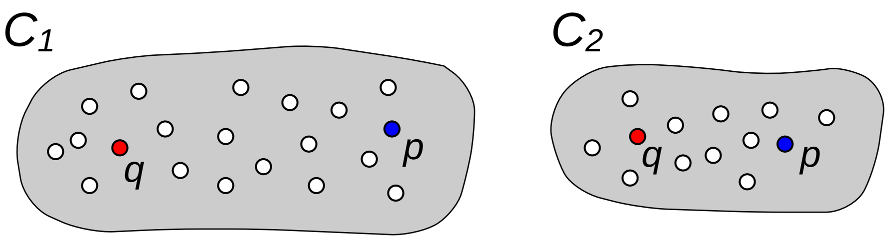
1. $\forall p, q$: If $p \in C$ and q is density-reachable from p then $q \in C$.



Maximality condition

2. $\forall p, q$: p is density-connected to q.

   There is a point o such that both, p and q are density-reachable from o.



Connectivity condition

Density-based Algorithm: DBSCAN

Overall cluster procedure:

1. Select unclassified point $p \in D$.

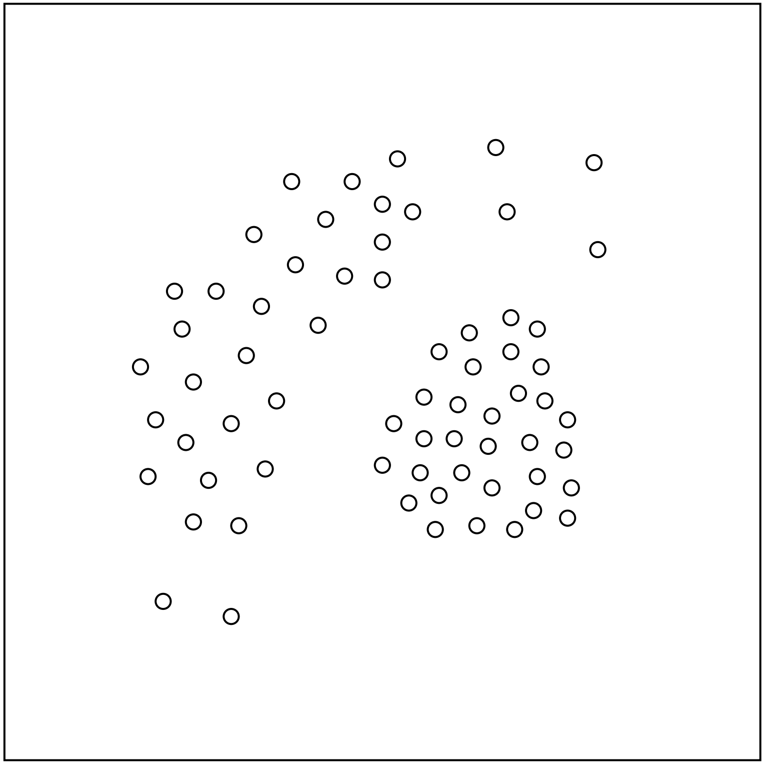
2. Construct ε-neighborhood $N_\varepsilon(p)$.

3. **If** p is a core point

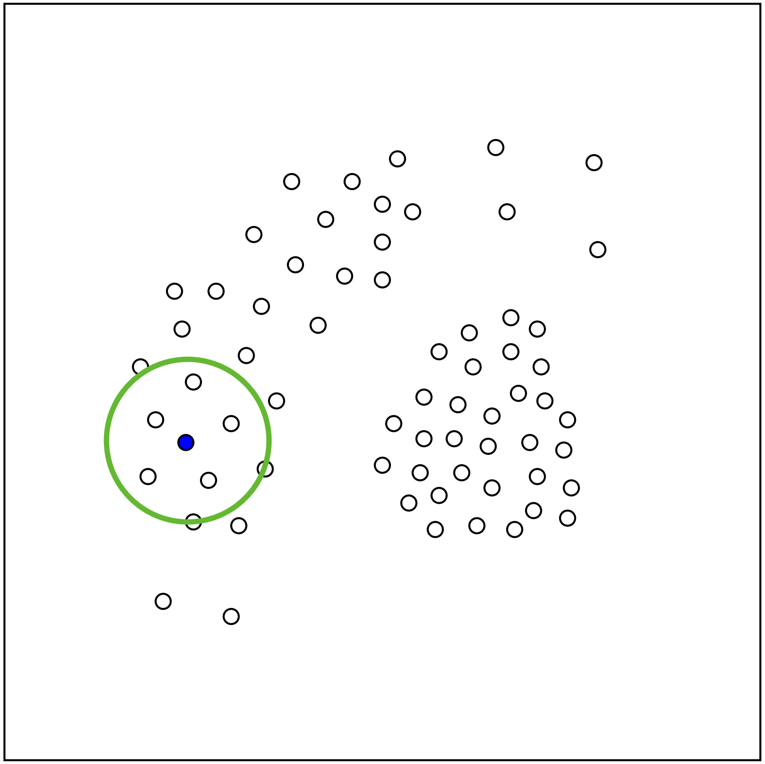
   Then Insert $N_\varepsilon(p)$ into new cluster C.
   Recursively analyze the ε-neighborhoods of $q \in N_\varepsilon(p)$ and insert all density-reachable points into C.

   Else Classify p as noise.

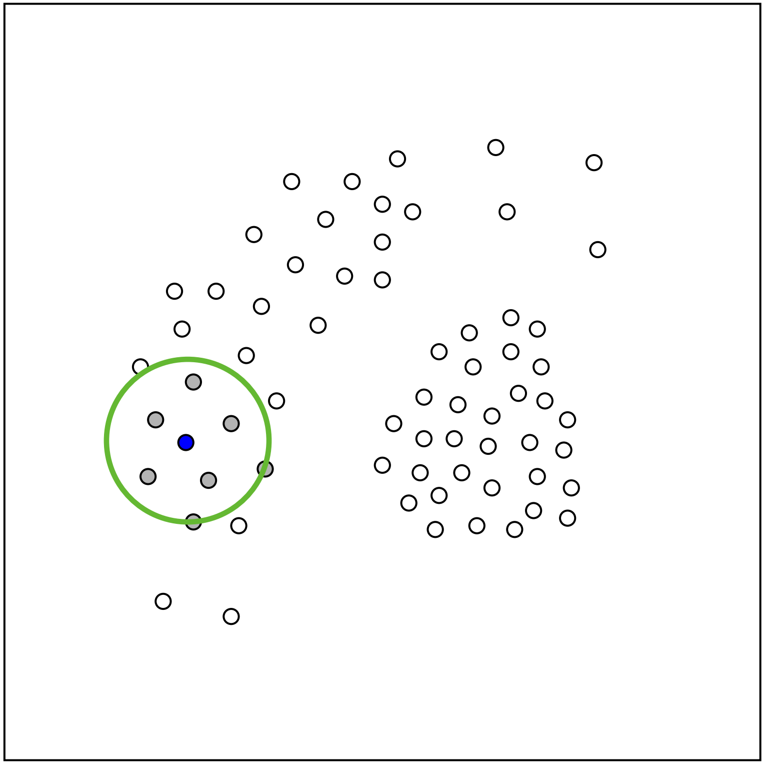
Density-based Algorithm: DBSCAN



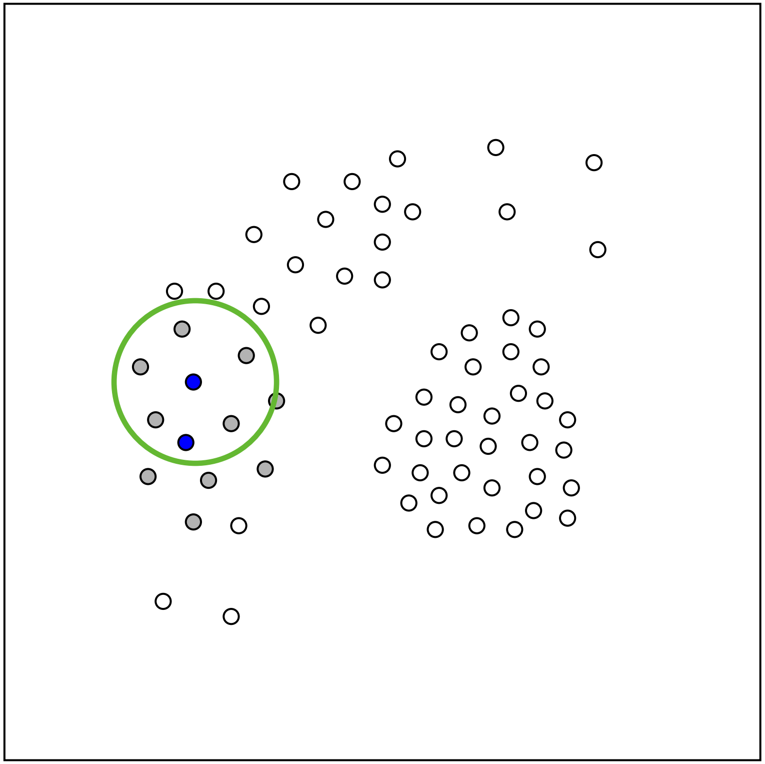
Density-based Algorithm: DBSCAN



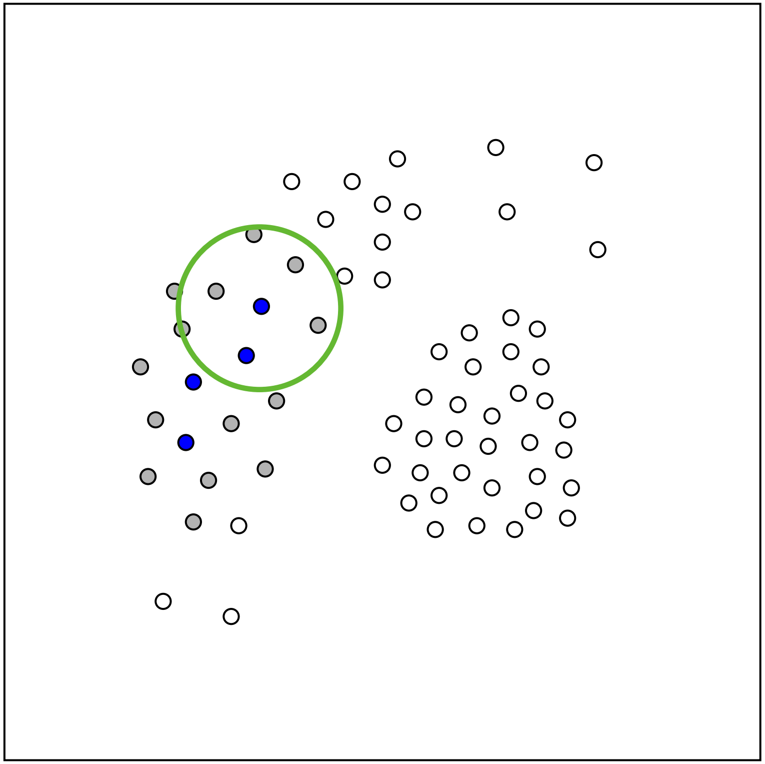
Density-based Algorithm: DBSCAN



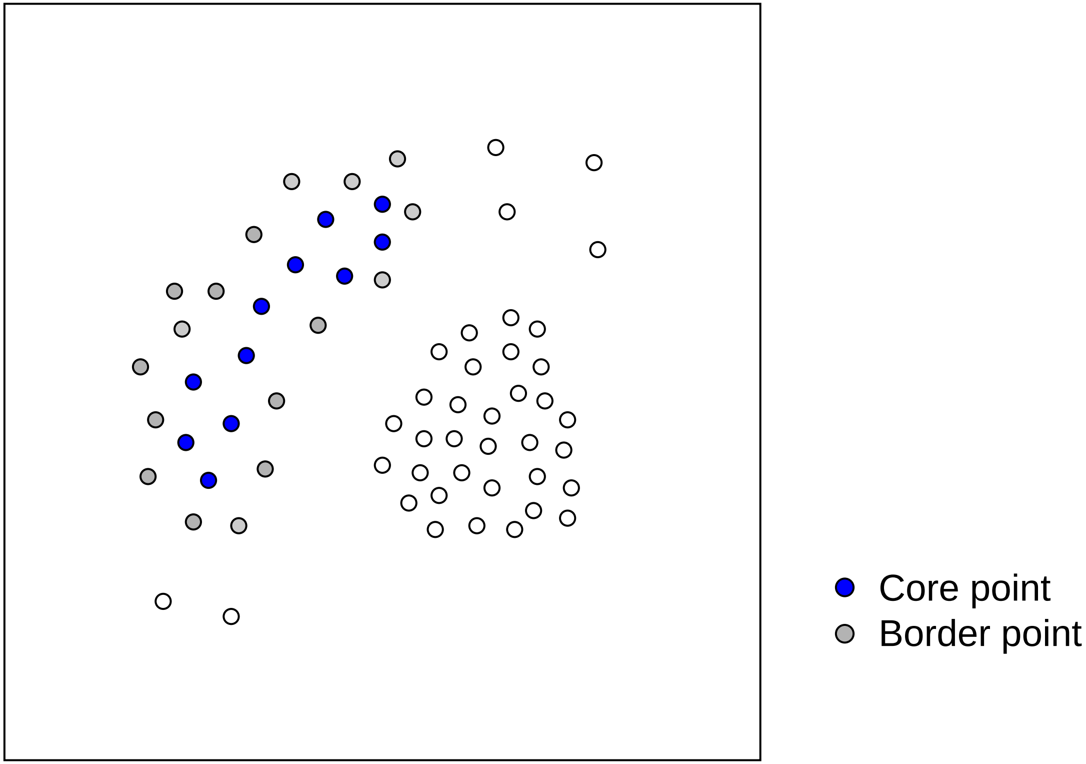
Density-based Algorithm: DBSCAN



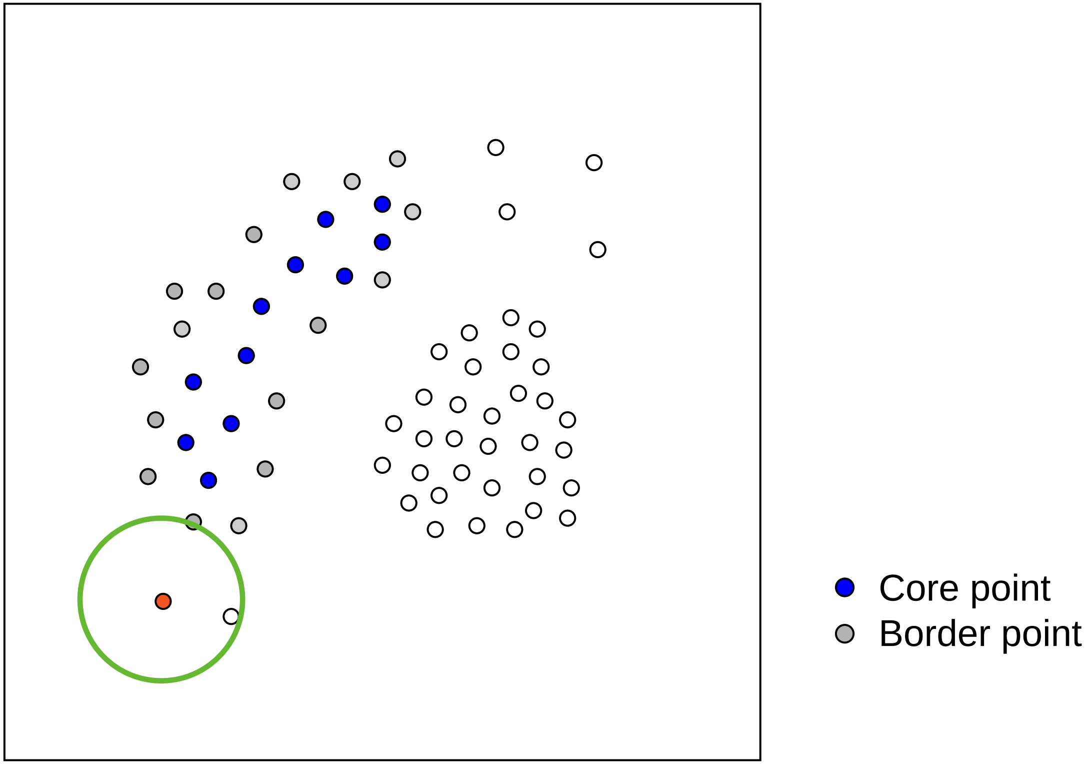
Density-based Algorithm: DBSCAN



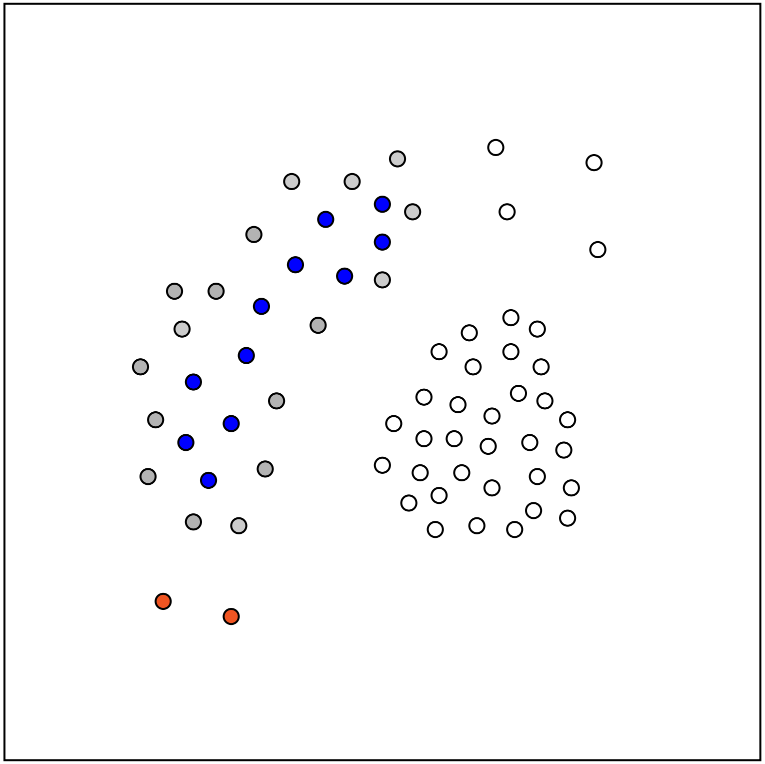
Density-based Algorithm: DBSCAN



Density-based Algorithm: DBSCAN

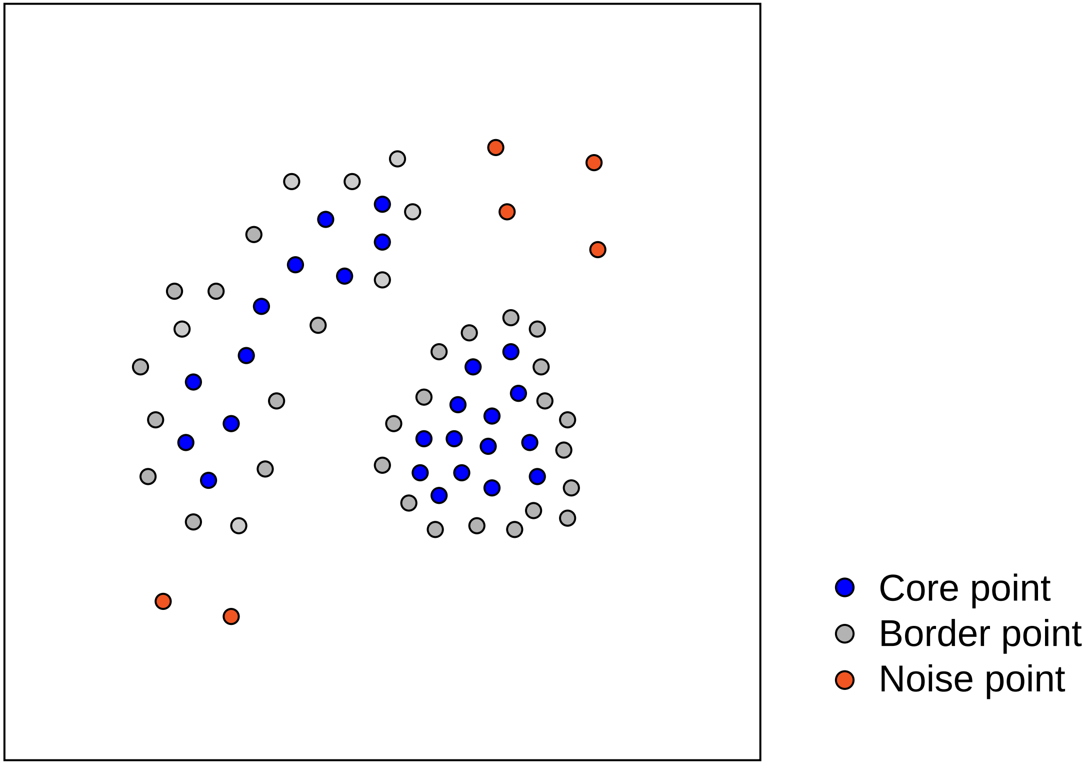


Density-based Algorithm: DBSCAN

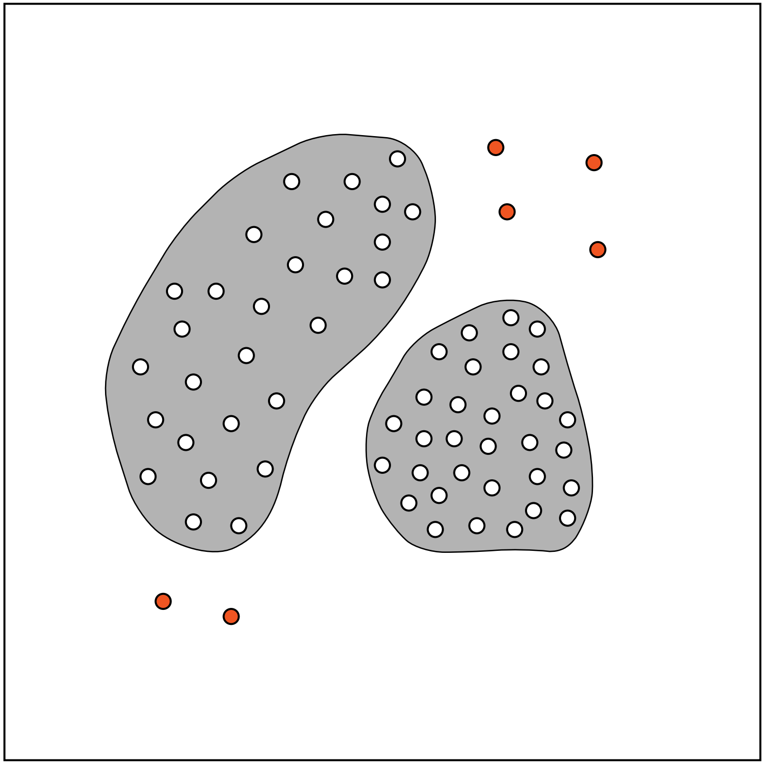


- Core point
- Border point
- Noise point

Density-based Algorithm: DBSCAN



Density-based Algorithm: DBSCAN

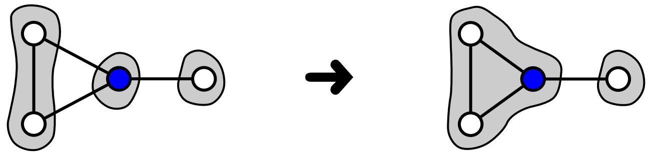


Noise point

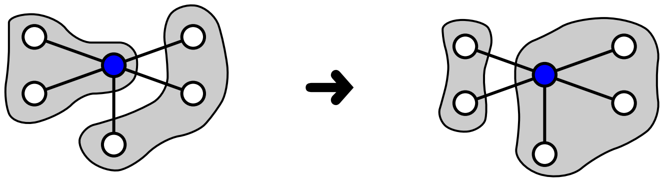
Density-based Algorithm: MajorClust

Density-based Algorithm: MajorClust

- Definite majority decision (agglomeration):

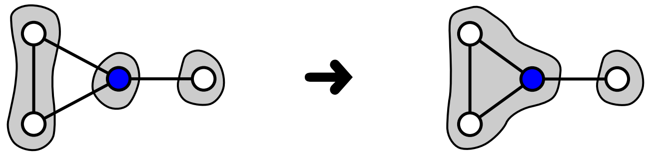


- Definite majority decision (node changes cluster):

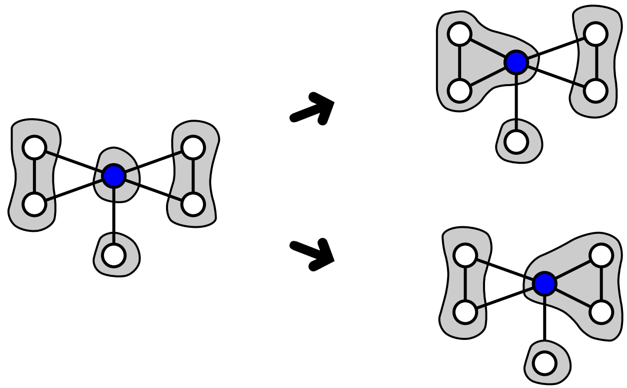


Density-based Algorithm: MajorClust

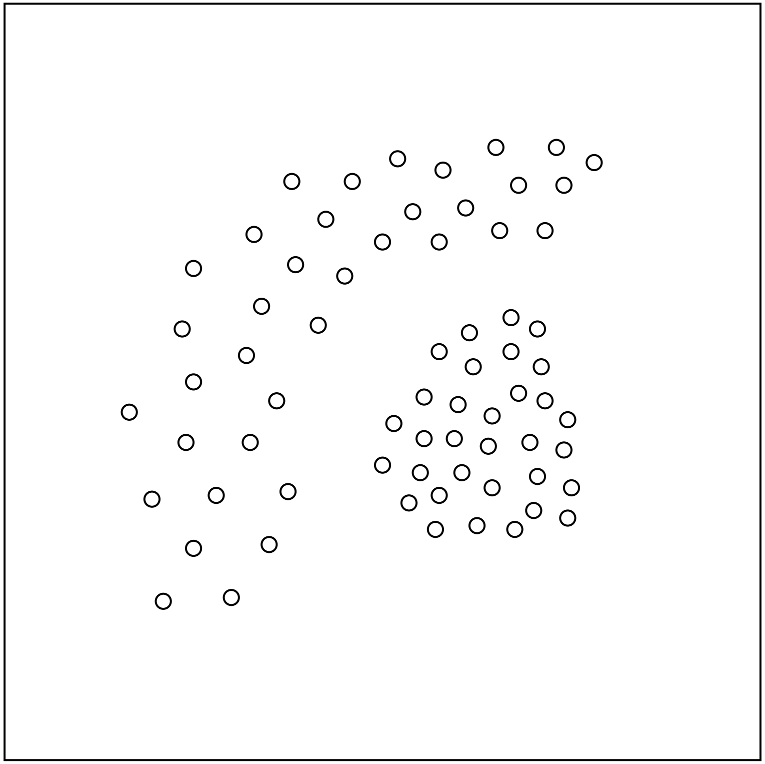
- Definite majority decision (agglomeration):



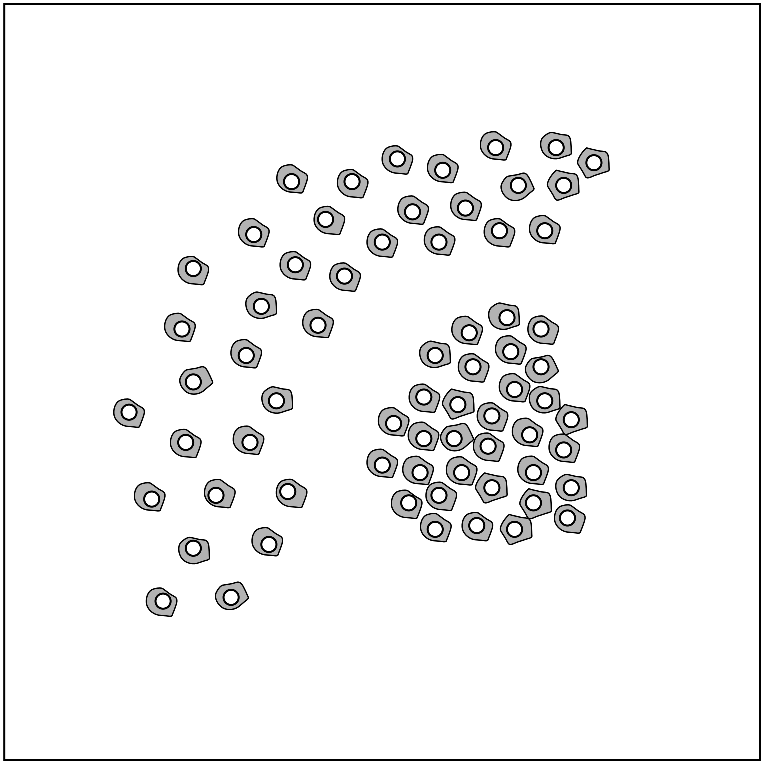
- Indefinite majority decision:



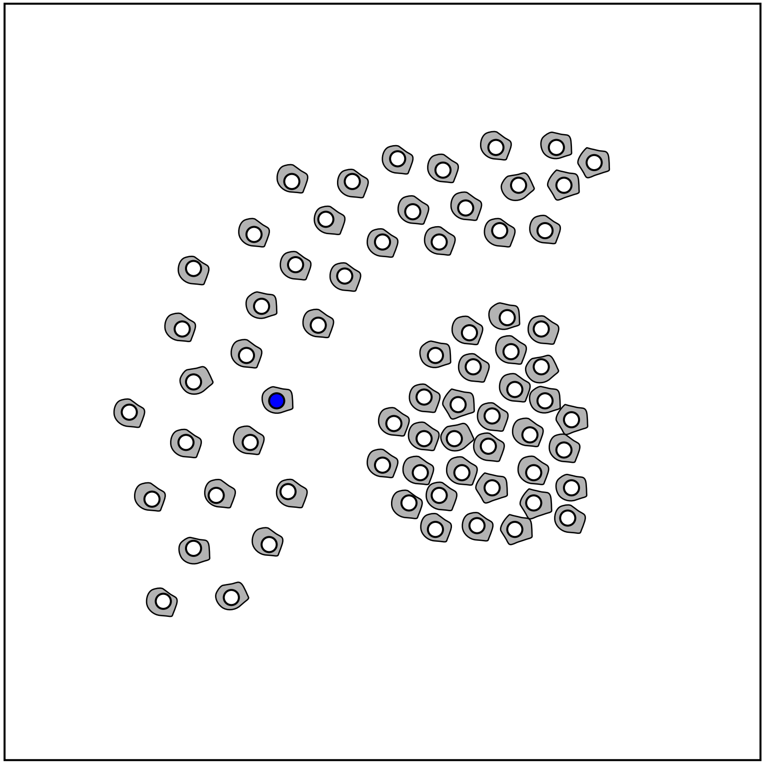
Density-based Algorithm: MajorClust



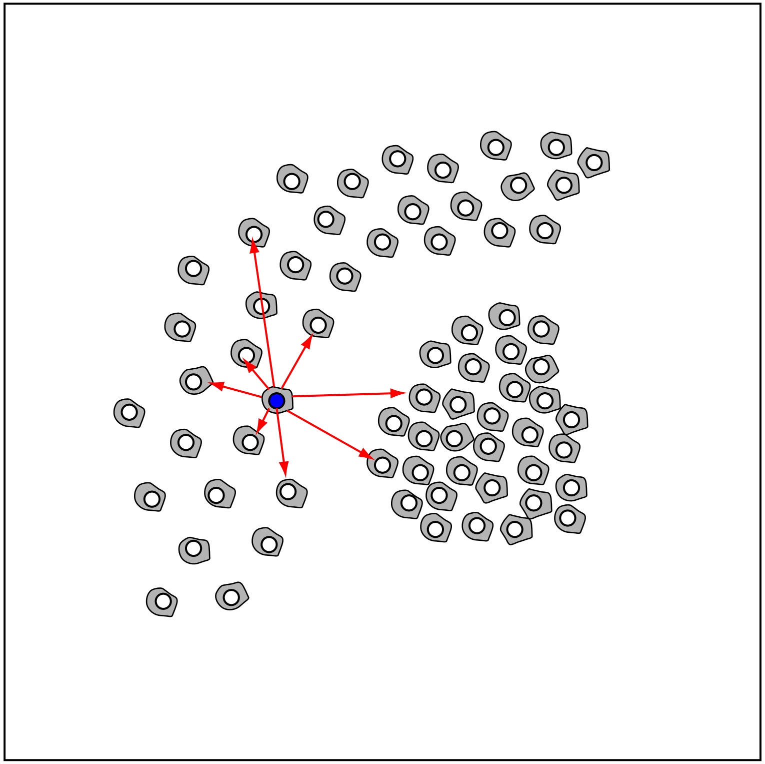
Density-based Algorithm: MajorClust



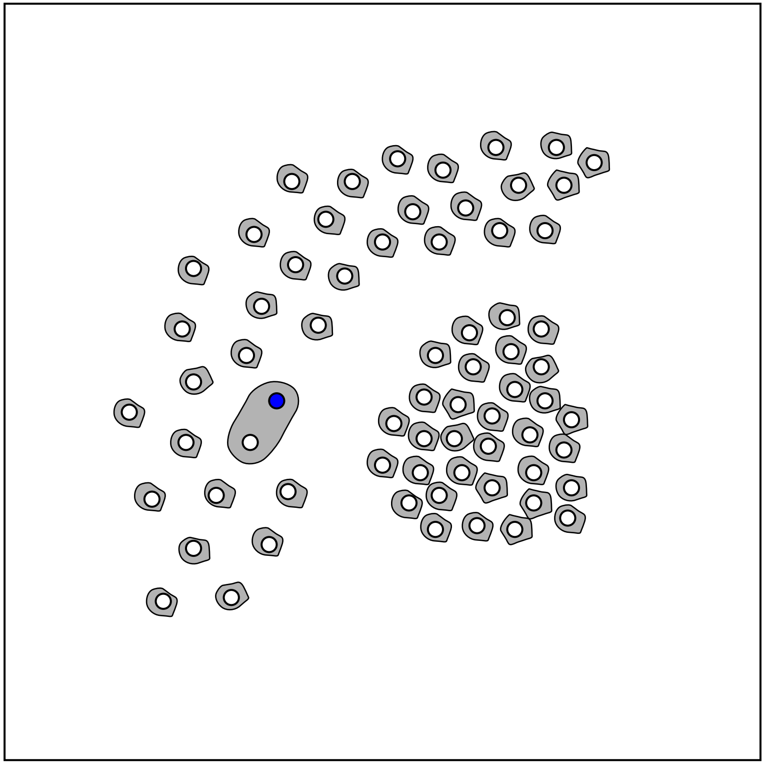
Density-based Algorithm: MajorClust



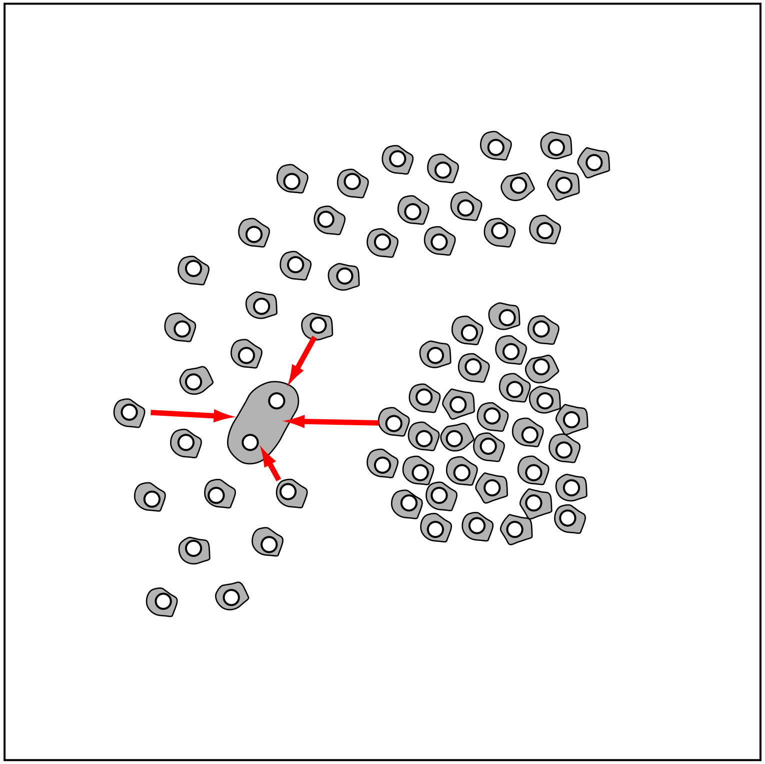
Density-based Algorithm: MajorClust



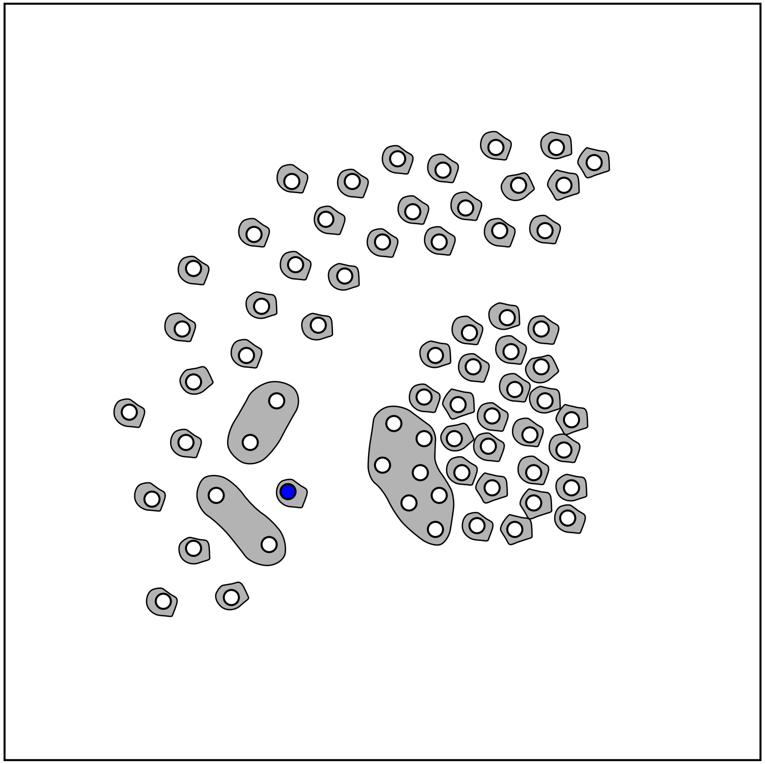
Density-based Algorithm: MajorClust



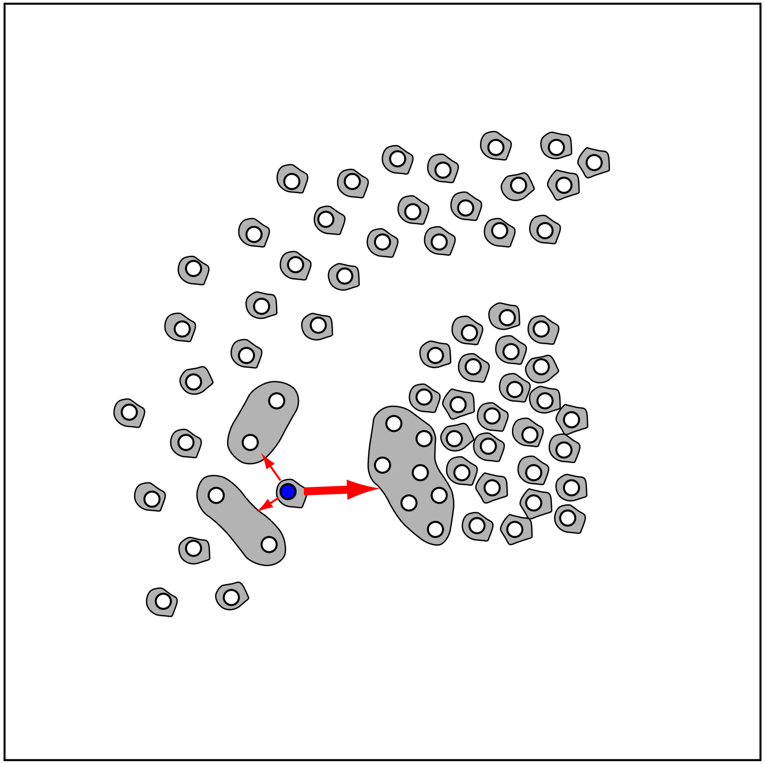
Density-based Algorithm: MajorClust



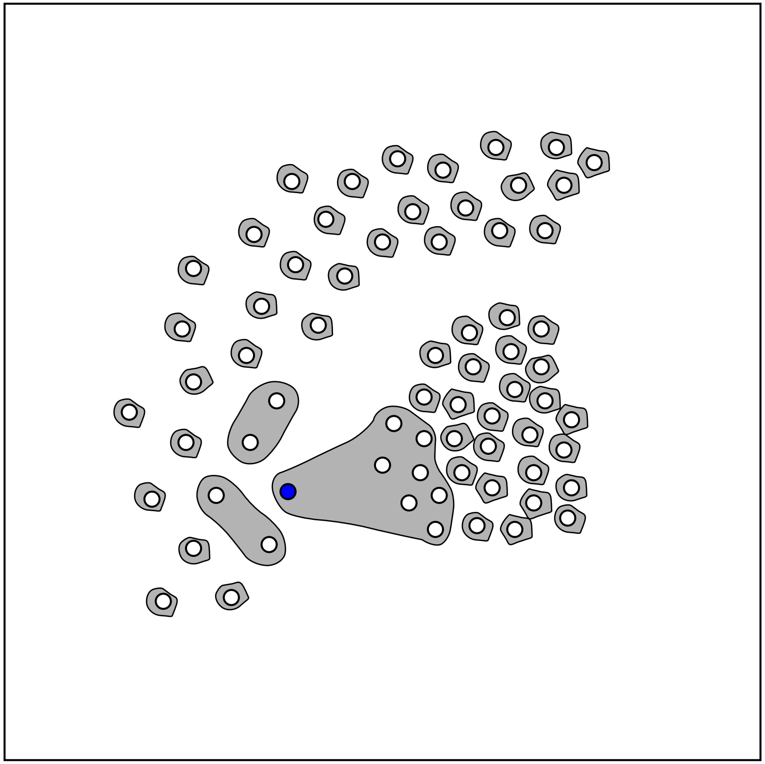
Density-based Algorithm: MajorClust



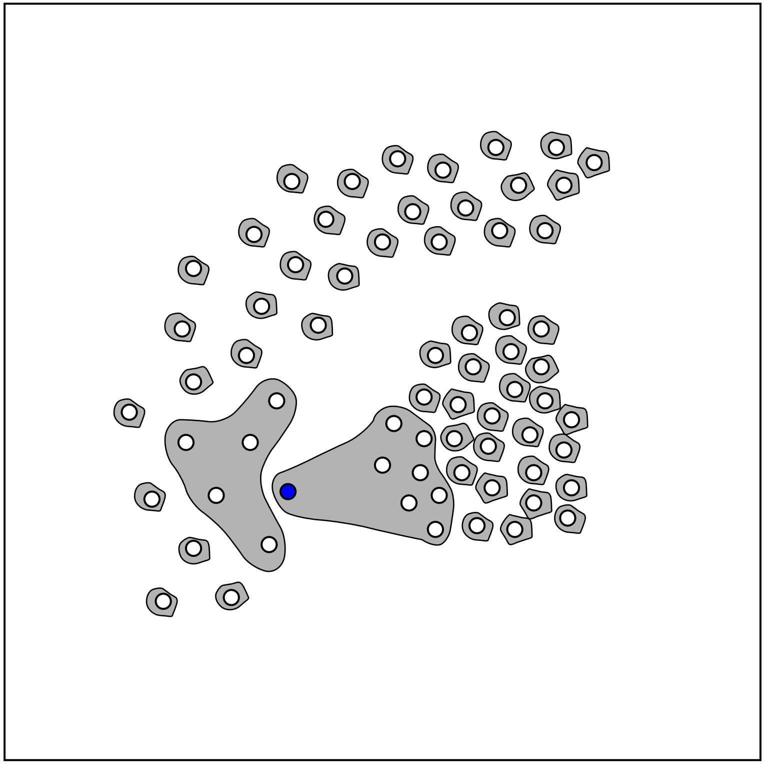
Density-based Algorithm: MajorClust



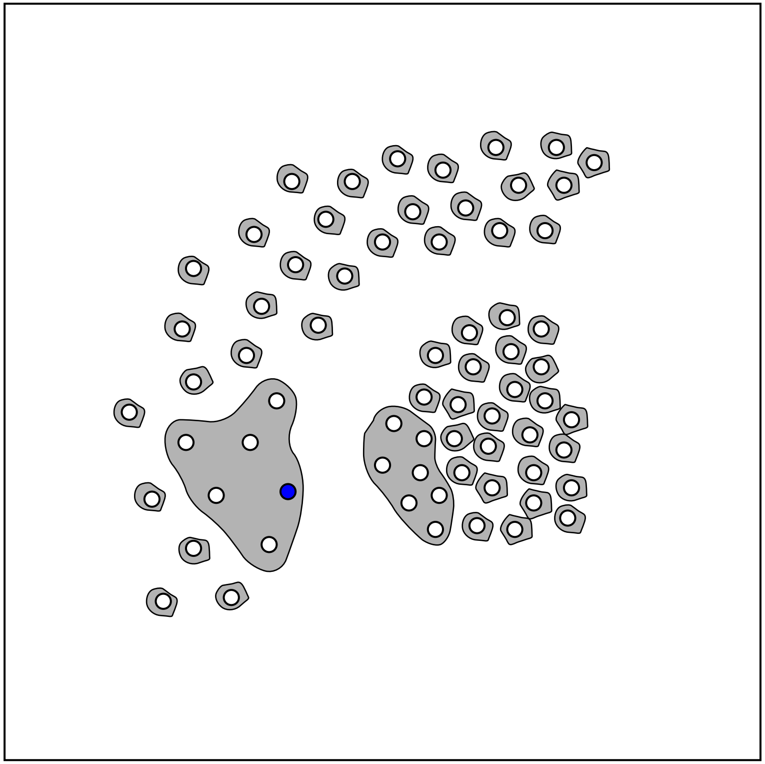
Density-based Algorithm: MajorClust



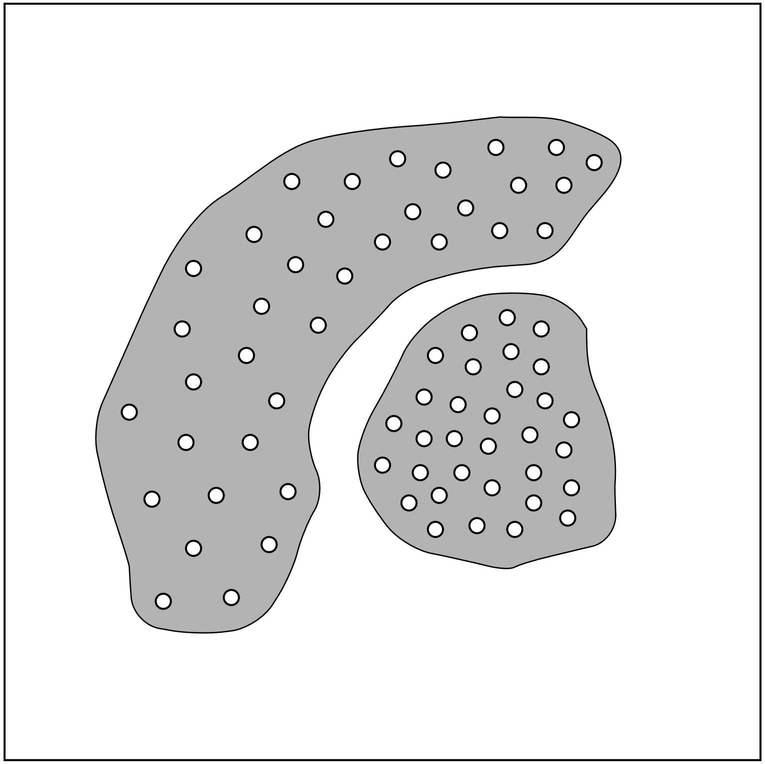
Density-based Algorithm: MajorClust



Density-based Algorithm: MajorClust



Density-based Algorithm: MajorClust



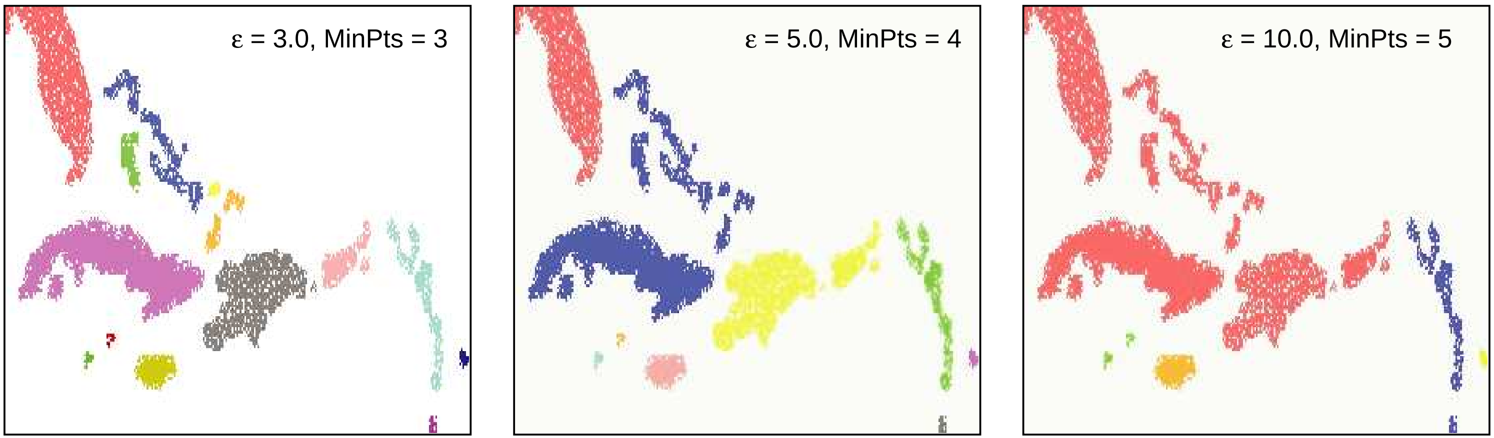
Analysis I (low-dimensional)

Geometrical Data—map of the Caribbean Islands (approx. 20,000 points) :

Analysis I (low-dimensional)

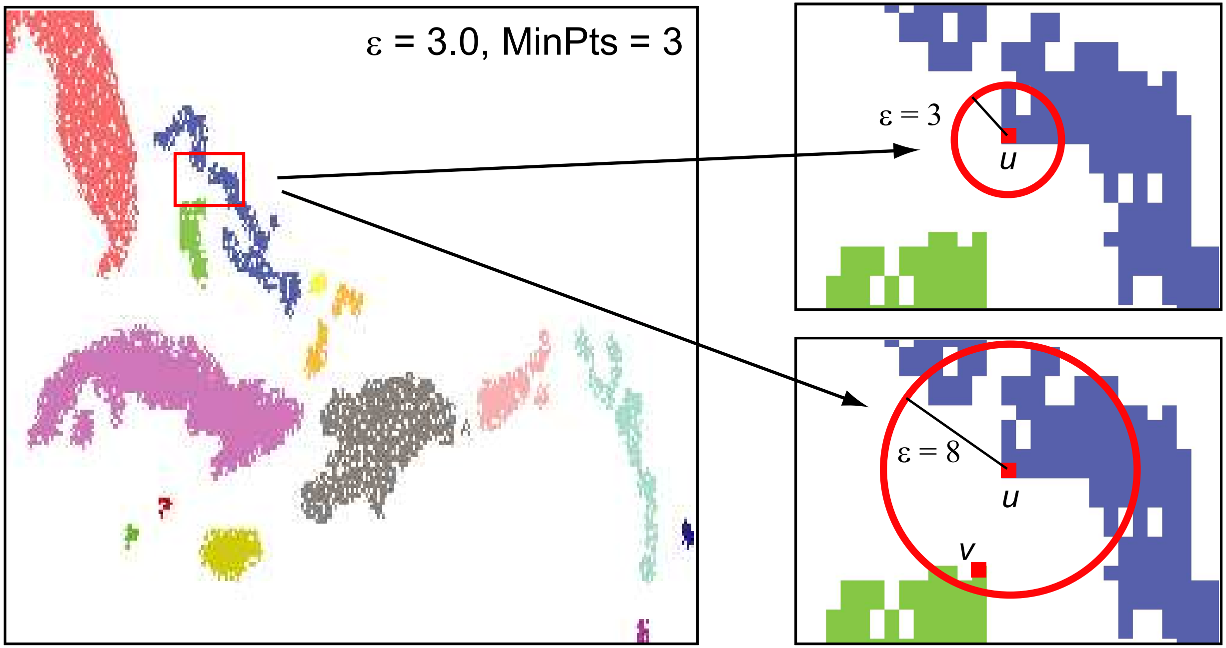
Geometrical Data—map of the Caribbean Islands (approx. 20,000 points) :

DBSCAN:



Analysis I (low-dimensional)

The problem of choosing a good ε-value in DBSCAN.



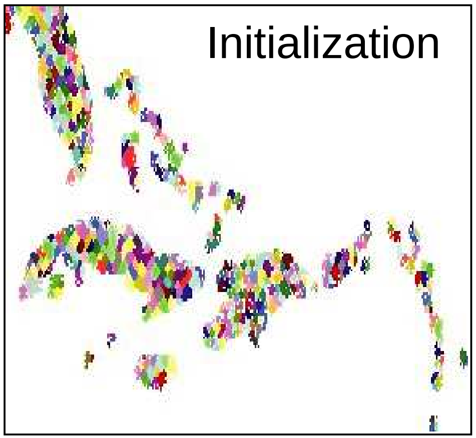
Two separate clusters are found.

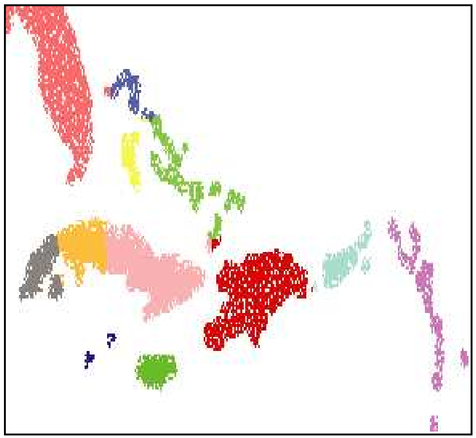
Clusters are merged.

Analysis I (low-dimensional)

Geometrical Data—map of the Caribbean Islands (approx. 20,000 points) :

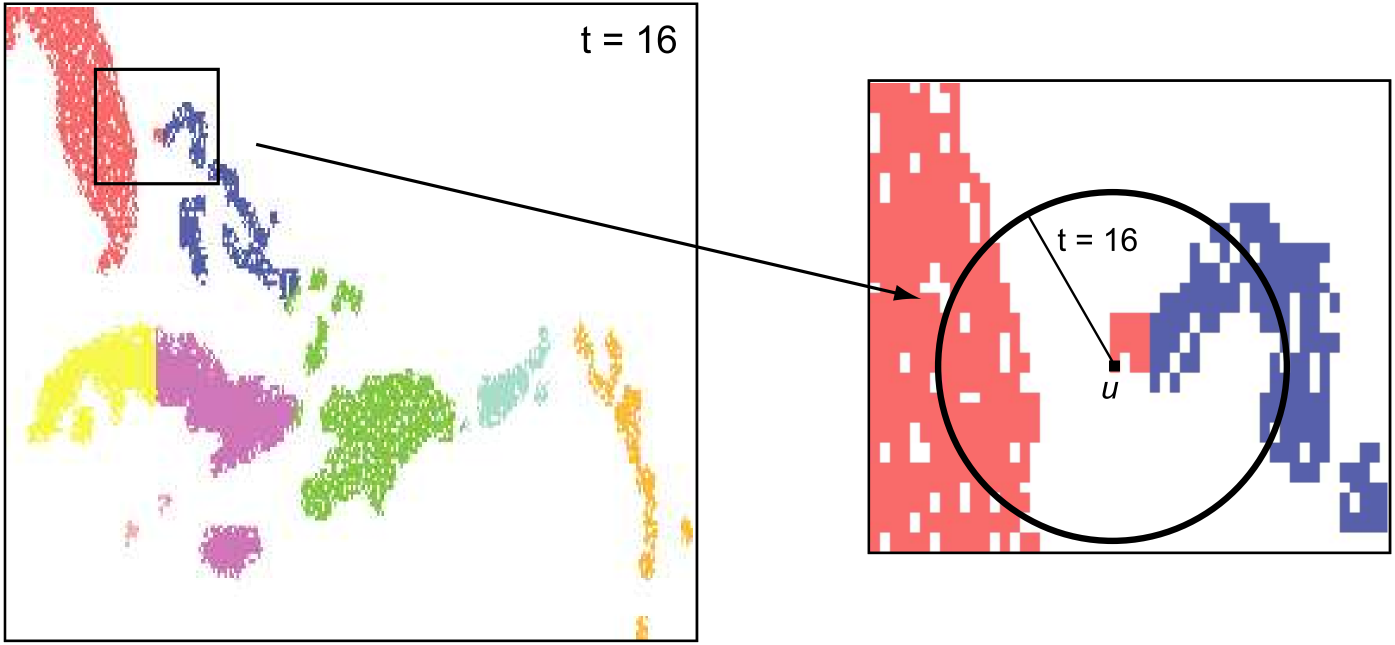
MajorClust:





Analysis I (low-dimensional)

The problem of a global analysis (no ε-neighborhood restriction) in MajorClust.



Analysis II (high-dimensional)

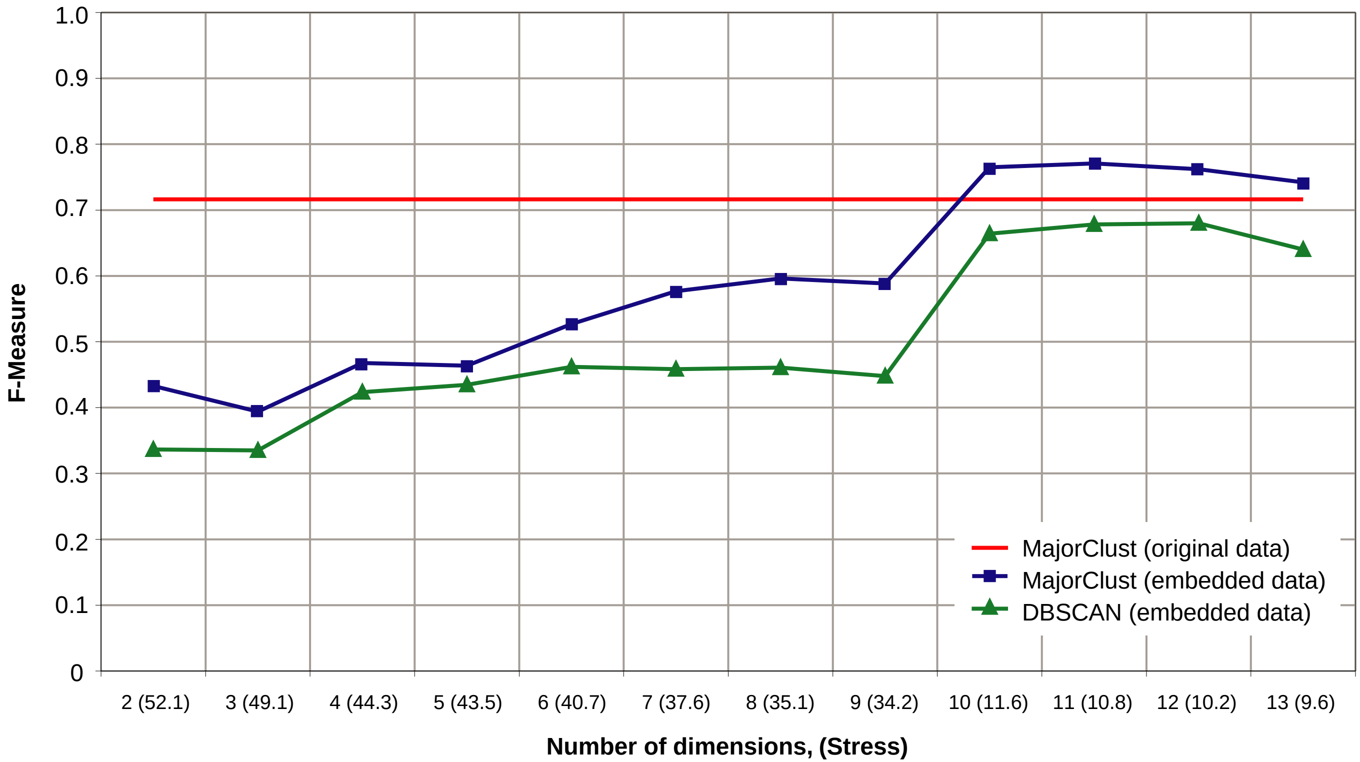
Document categorization with the Reuters corpus.

- 1000 documents
- 10 categories: politics, culture, economics, etc.
- uniformly distributed, exclusive membership
- > 10,000 dimensions

Analysis II (high-dimensional)

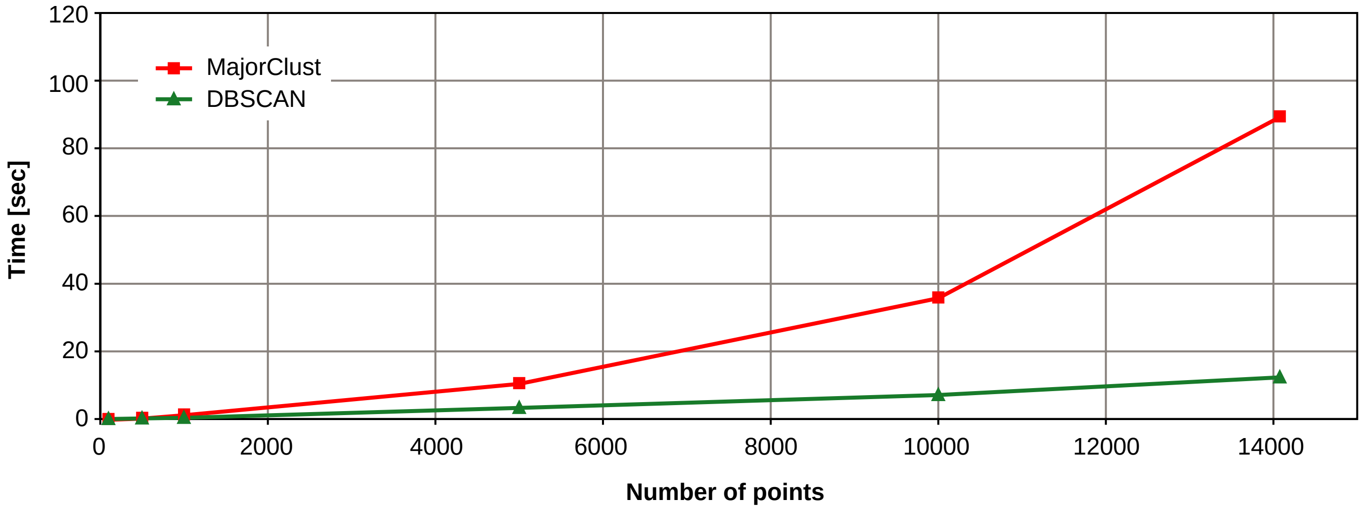
DBSCAN requires embedding of data in low-dimensional space.

Classification results (F-Measure) over dimensionality:



Analysis (runtime)

Runtime-behavior on the geometrical data:



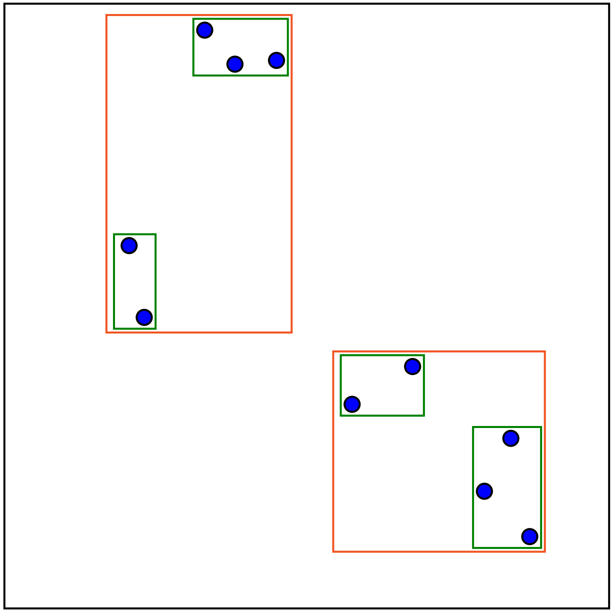
Note:

The embeddding of data in a low-dimensional space (MDS) is computationally very expensive:

I. e., most cluster algorithms will be faster than DBSCAN + MDS.

Analysis (runtime)

DBSCAN employes the R-tree data structure for region queries, which constructs minimum bounding regions for inserted points:



“Existing methods are outperformed on on average by a simple sequential scan, if the number of dimensions exceeds around 10.”

[Weber 99, Gionis/Indyk/Motwani 99-04]

Summary

An alternative categorization scheme :

	Cluster approach			
	hierarchical	iterative	density-based	meta-search controlled
Analysis strategy	relative comparison based on two items	absolute comparison based on k items	relative comparison based on k items	absolute comparison based on all items
Recovery characteristics	irrevocable	revocable	revocable	revocable

Orthogonal to this scheme is the concept for similarty computation:

- distance (neighborhood) analysis in low-dimensional space
- similarity predicate in arbitrary (high-dimensional) space

Summary

The strengths and weaknesses of density-based cluster algorithms can be explained with the dimensionality of the data.

- DBSCAN usually outperfoms other cluster algorithms on low-dimensional data.
- MajorClust usually outperfoms other cluster algorithms on high-dimensional data, in particular in the document categorization field.

Current work:

How fingerprints can be utilized for efficient region queries in high-dimensional spaces.