DBSCAN
Density-Based Spatial Clustering of Applications with Noise

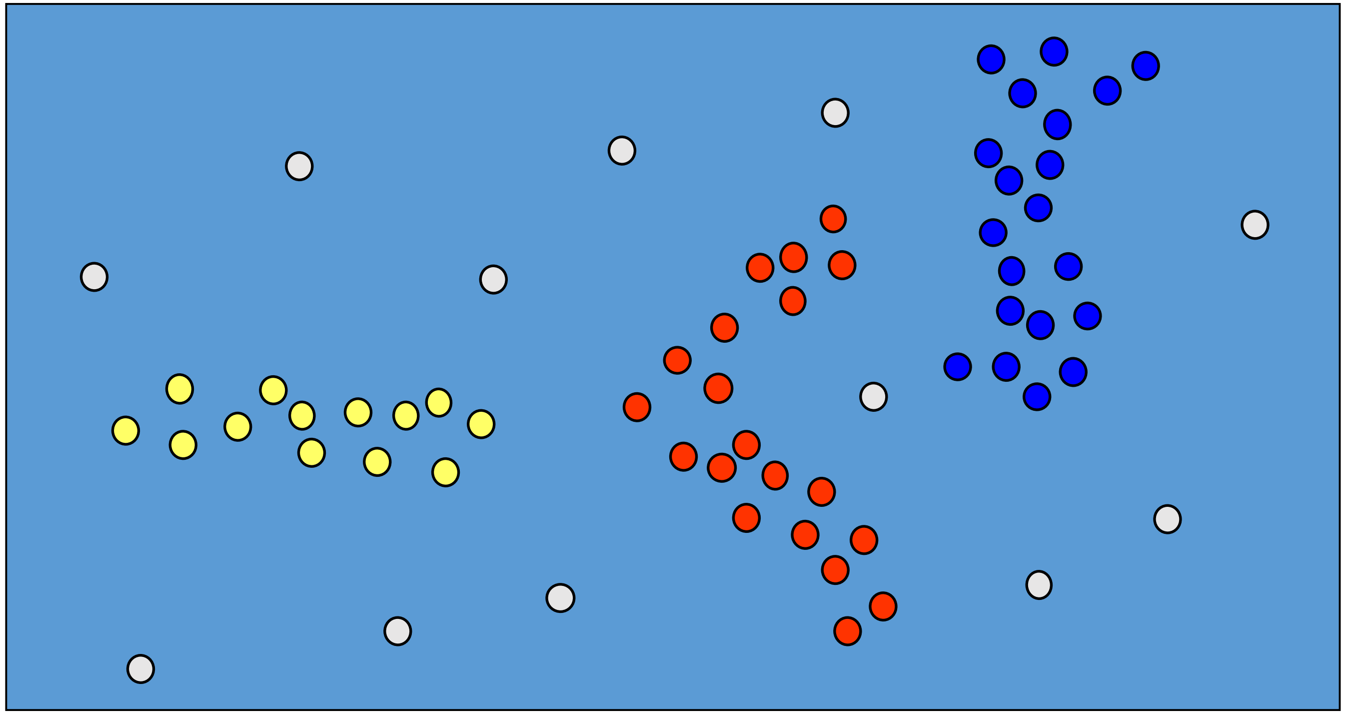
Reference:

M.Ester, H.P.Kriegel, J.Sander and Xu.

A density-based algorithm for discovering clusters in large spatial databases, Aug 1996

Density-Based Clustering – Basic Idea

Clusters are dense regions in the data space, separated by regions of lower object density



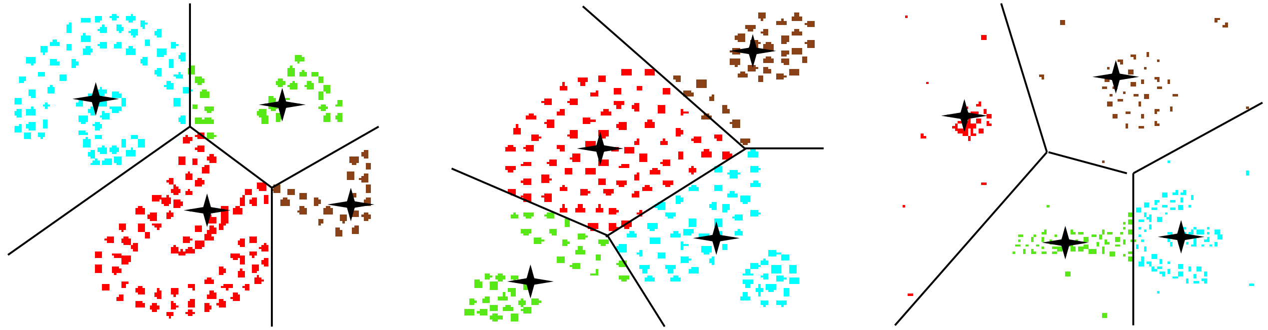
Density-based Approaches

Why Density-Based Clustering methods?

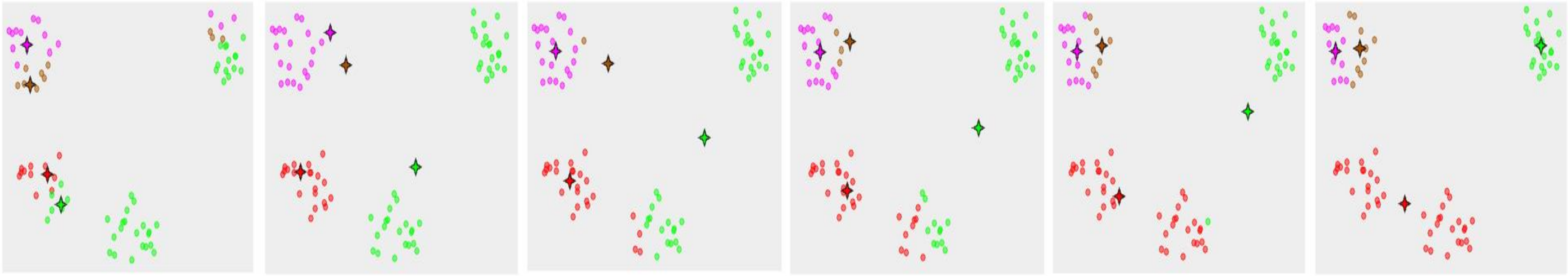
- Clusters can have arbitrary shapes and sizes, can even identify clusters contained within other clusters
- The number of clusters is determined automatically
- Can separate clusters from surrounding noise
- Can be supported by a spatial index structures

- DBSCAN – the first density based clustering
- OPTICS – density based cluster-ordering
- DENCLUE – a general density-based description of cluster and clustering

Why Use Density-Based Clustering?



Results of a *k*-medoid algorithm for *k*=4



Results of a *k*-means algorithm for *k*=4, wrongly converging in 6 steps

DBSCAN in a nutshell

Intuition for the formalization of the basic idea:

- For any point in a cluster, the local point density around that point has to exceed some threshold
- The set of points from one cluster is spatially connected

Local point density at a point *p* is defined by two parameters:

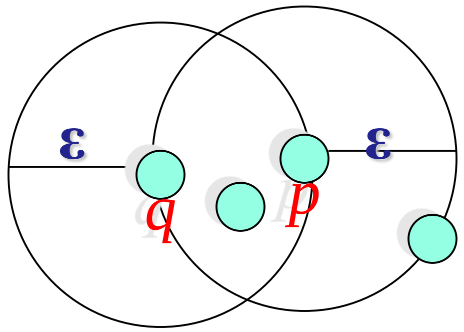
- ε – radius for the neighborhood of point p:
  $N_\varepsilon(p) := \{q \text{ in data set } D \mid distance(p, q) \leq \varepsilon\}$
- *MinPts* – minimum number of points in the given neighborhood $N(p)$

Definitions

- ε-Neighborhood – Objects within a radius of ε from an object.

$$N_\varepsilon(p):\{q \mid d(p,q) \le \varepsilon\}$$

- "High density" - ε-Neighborhood of an object contains at least *MinPts* of objects.



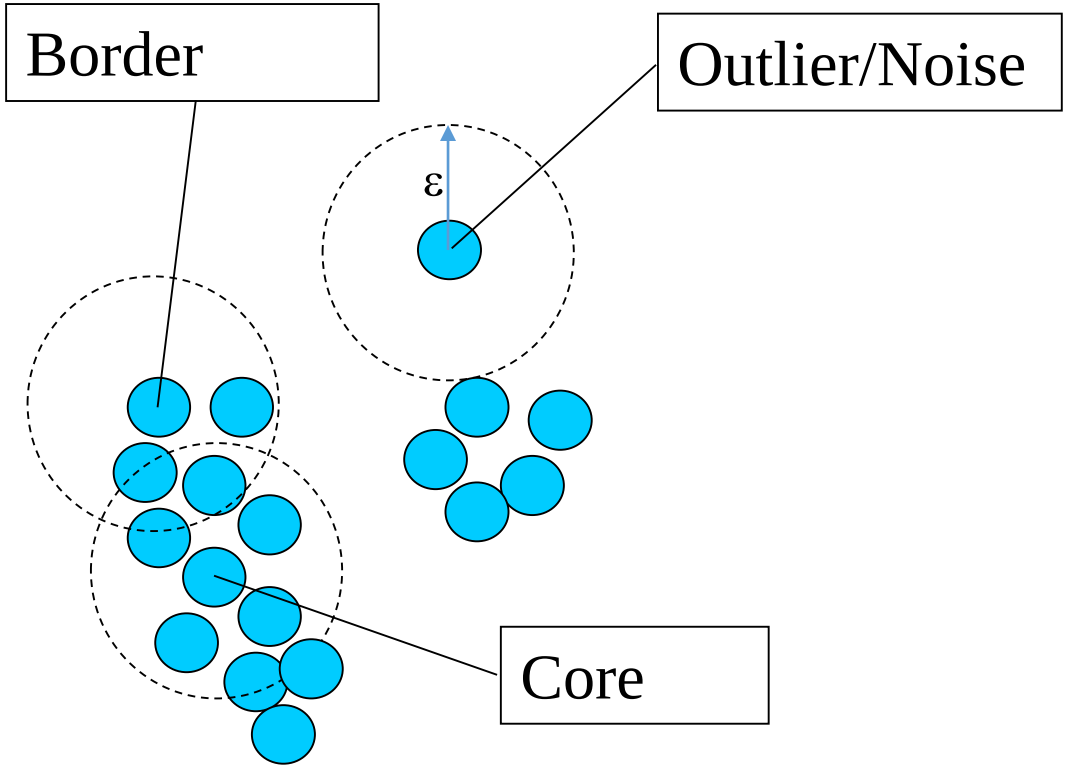
For MinPts = 4

Density of p is "high": N ε(p) = 4

Density of q is "low": N ε(q) = 3

Definitions

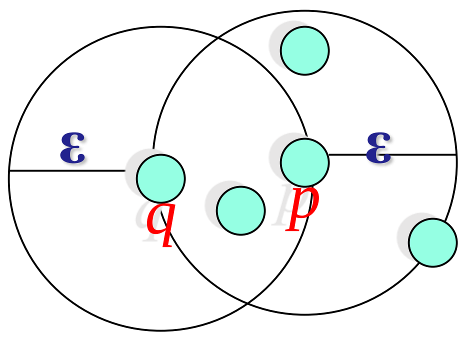
- A point is a core point if it has more than MinPts within ε. These are points at the interior of a cluster
- A border point has fewer than MinPts within ε, but is within ε of a core point
- A noise point is any point that is not a core point nor a border point



ε = 1 unit, MinPts = 5

Definitions

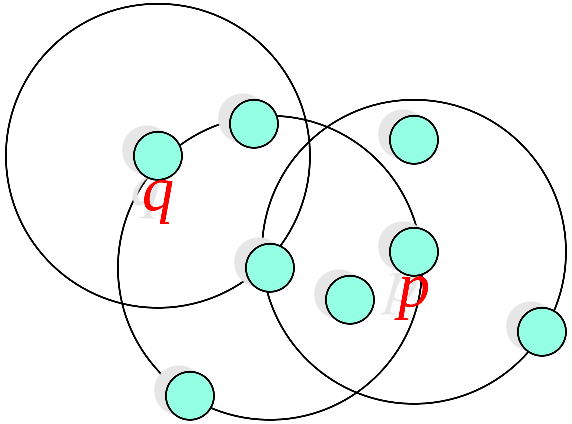
- **Directly density-reachable**
  - An object q is directly density-reachable from object p if q is within the ε-Neighborhood of p and p is a core object.



- q is directly density-reachable from p
- p is not directly density-reachable from q (for MinPts > 3)
- Direct density reachability is asymmetric.

Definitions

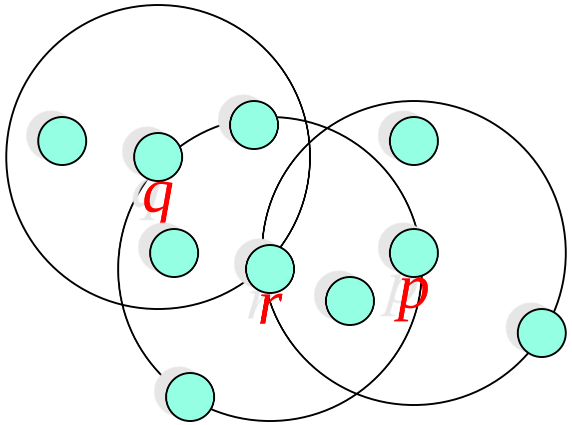
- **Density-reachable:**
  - An object p is density-reachable from q w.r.t ε and *MinPts* if there is a chain of objects $p_1,\ldots,p_n$, with $p_1=q$, $p_n=p$ such that p_{i+1} is directly density-reachable from p_i w.r.t ε and *MinPts* for all 1 <= i <= n



- q is density-reachable from p
- p is not density- reachable from q (for MinPts > 3)
- Transitive closure of direct density-reachability, still asymmetric

Definitions

- **Density-connectivity**
  - Object *p* is density-connected to object *q* w.r.t ε and *MinPts* if there is an object *o* such that both *p* and *q* are density-reachable from *o* w.r.t ε and *MinPts*



- *P* and *q* are density-connected to each other by *r*
- Density-connectivity is symmetric

Definitions

- **Cluster**: A cluster ***C*** is defined as a maximal set of density-connected points. The set ***C*** satisfies:
  - Maximality: For all p, q if $p \in$ ***C*** and if q is density-reachable from p w.r.t ε and *MinPts*, then also $q \in$ ***C.***
  - Connectivity: for all p, $q \in$ ***C***, p is density-connected to q w.r.t ε and *MinPts* in ***D.***
  - ***Note:*** cluster contains *core objects* as well as *border objects*
- **Noise:** objects which are not directly density-reachable from any core object

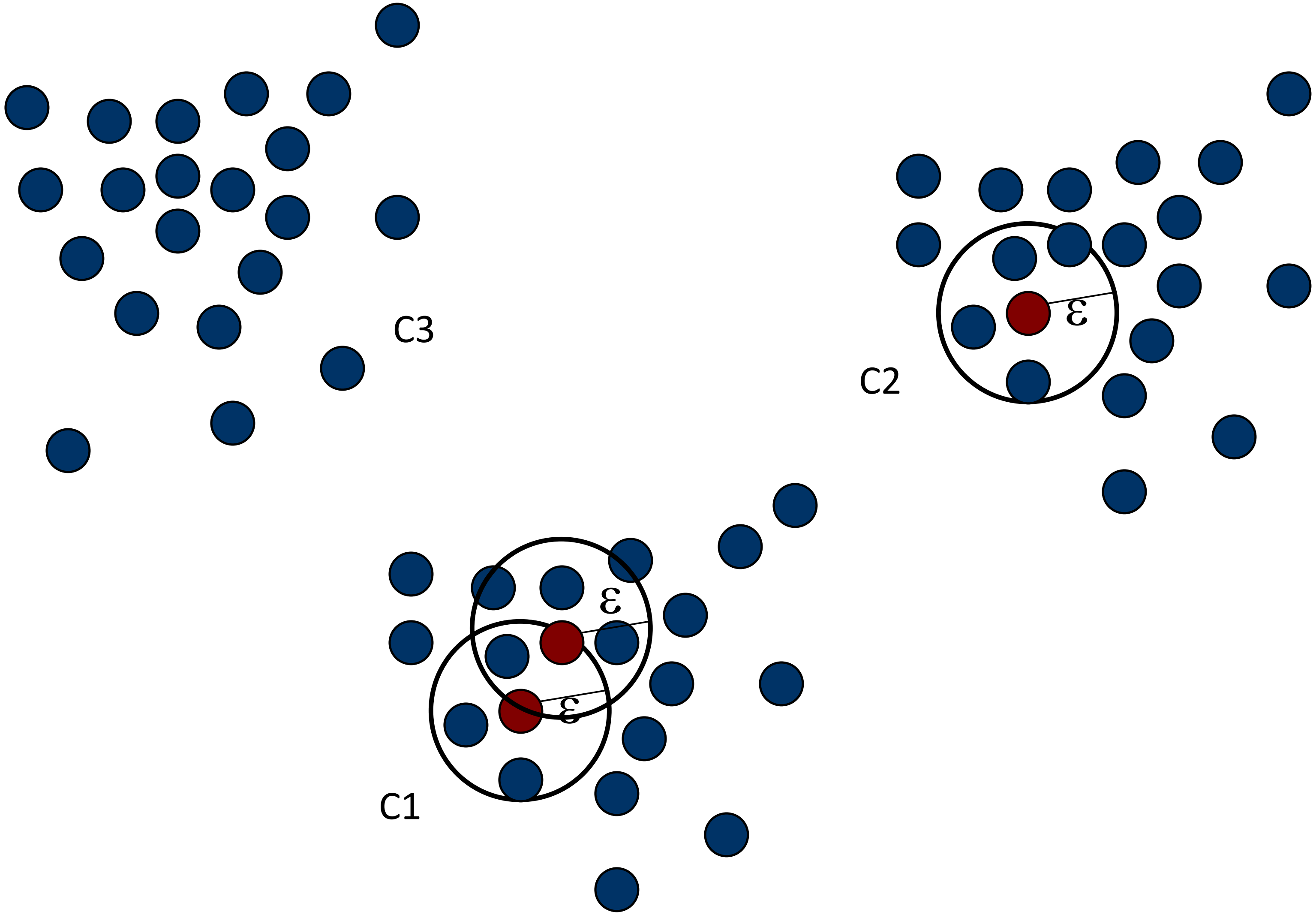
DBSCAN pseudo code

- select a point ***p***
- Retrieve all points density-reachable from ***p*** w.r.t ε and ***MinPts***
- If ***p*** is a core point, a cluster is formed
- If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point in the dataset
- Continue the process until all of the points have been processed

Result is independent of the order of processing the points.

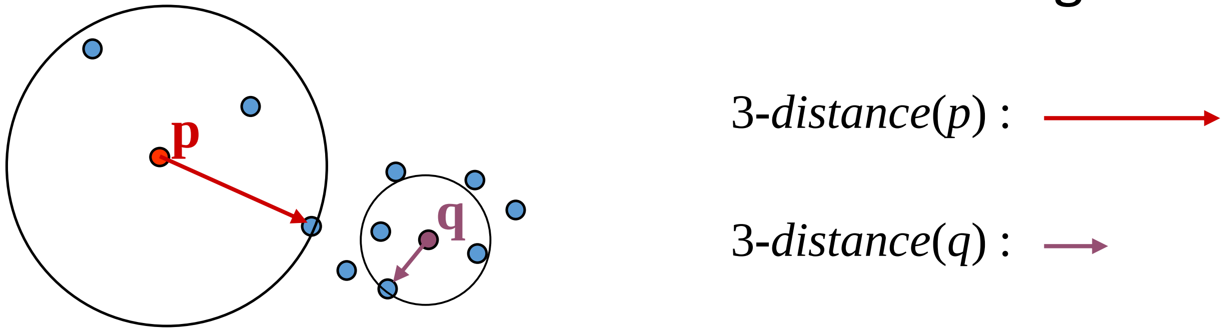
Example

MinPts = 4



Determining the Parameters ε and *MinPts*

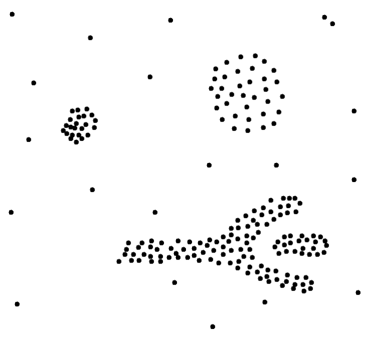
- Cluster: Point density higher than specified by ε and *MinPts*
- Idea: use the point density of the least dense cluster in the dataset as parameters – but how to determine this?
- Heuristic: look at the distances to the *k*-nearest neighbor

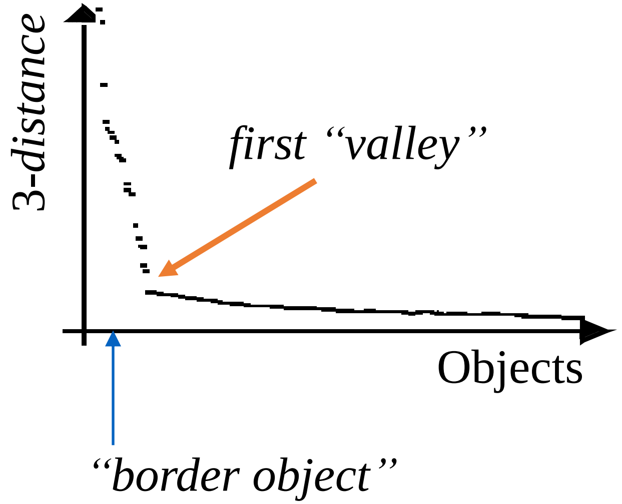


- Function *k-distance*(*p*): distance from *p* to the its *k*-nearest neighbor
- *k-distance plot*: *k*-distances of all objects, sorted in decreasing order

Determining the Parameters ε and *MinPts*

- Example of a *k*-distance plot





- Heuristic method:
  - Fix a value for *MinPts*
  - User selects "border object" *o* from the *MinPts-distance* plot; ε is set to *MinPts-distance*(o)