DBSCAN CLUSTERING

What is DBSCAN Clustering Algorithm?

DBSCAN--- Density **B**ased **S**patial **C**lustering of **A**pplications with **N**oise

DBSCAN is a density-based clustering method used in machine learning to separate clusters of high density from clusters of low density.

DBSCAN Clustering

ML Labs Pvt Ltd

- A dense cluster is a region in which the density of points is greater than a minimum.

- Since DBSCAN algorithm expand clusters based on density, they form clusters of arbitrary shapes.

DBSCAN Clustering:
Core, Border, Noise points

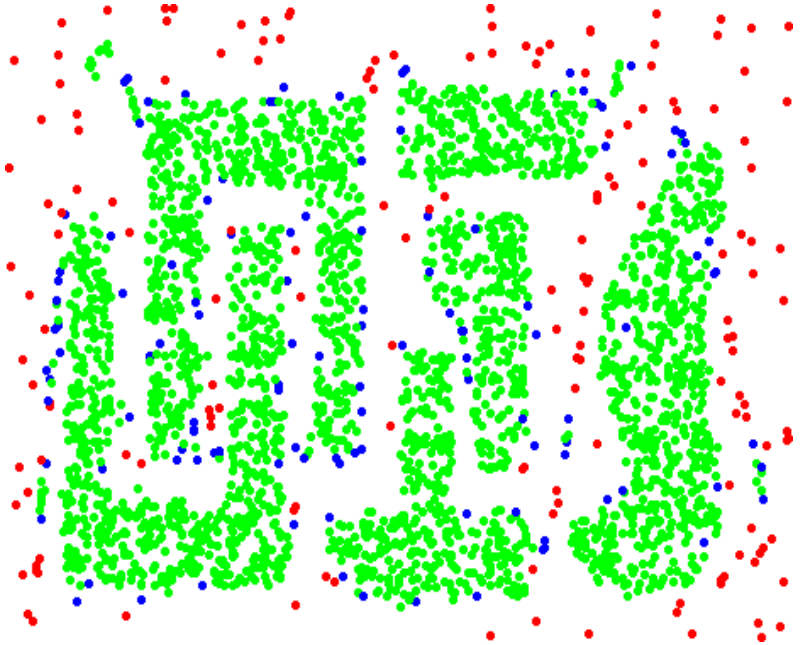
ML Labs Pvt Ltd

- **A Core point** has more than a certain number of points (MinPts) within Eps.Core points are at the interior of a cluster
- A **border point** has fewer than MinPts within Eps, but is in the neighborhood of a core point
- A **noise point** is any point that is neither a core point nor a border point

DBSCAN Clustering: Core, Border, Noise points

ML Labs Pvt Ltd

Original Points



Point types: core, border and noise

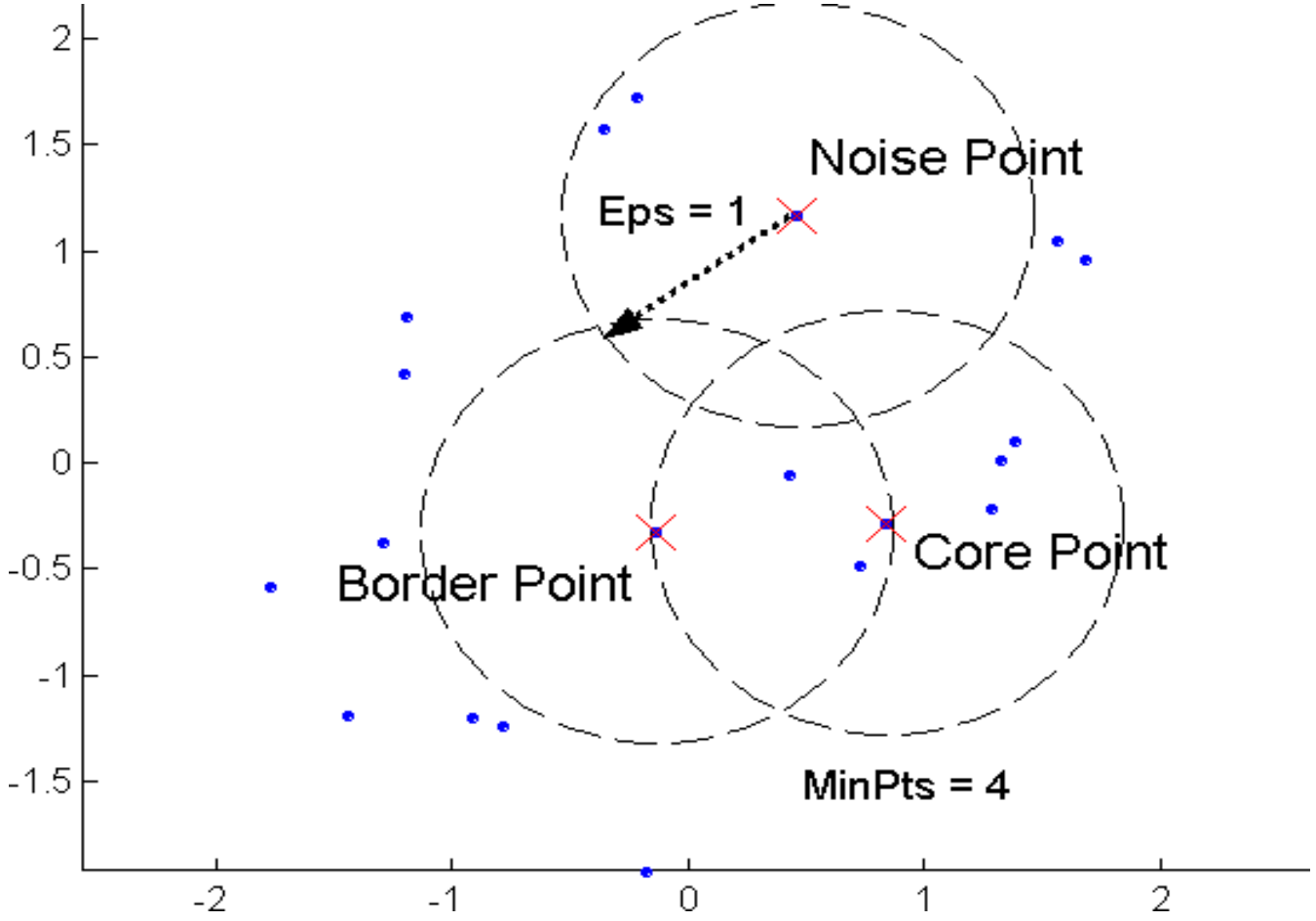
DBSCAN Clustering: Process

ML Labs Pvt Ltd

- Any two core points close enough– within a distance *Eps* of one another – are put in the same cluster
- Any border point that is close enough to a core point is put in the same cluster as the core point
- Noise points are discarded

DBSCAN Clustering: Core, Border, Noise points

ML Labs Pvt Ltd



Parameter Estimation

ML Labs Pvt Ltd

parameters must be specified by the user.

ε = physical distance(radius),

minPts = desired minimum cluster size

minPts

- derived from the number of dimensions *D* in the data set, as $minPts \geq D + 1$
- $minPts = 1$ does not make sense, as then every point on its own will already be a cluster
- *minPts* must be chosen at least 3. larger is better.
- larger the data set, the larger the value of *minPts* should be chosen.

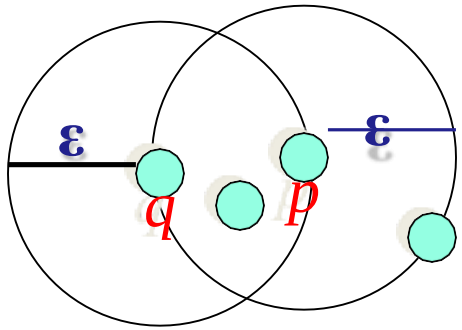
ε

- value can be chosen by using a k-distance graph.
- if ε is chosen much too small, a large part of the data will not be clustered.
- if too high value ,majority of objects will be in the same cluster
- In general, small values of ε are preferable.

Concepts: ε-Neighborhood

ML Labs Pvt Ltd

- ε-Neighborhood - Objects within a radius of ε from an object. (epsilon-neighborhood)
- Core objects - ε-Neighborhood of an object contains at least MinPts of objects



ε-Neighborhood of *p*

ε-Neighborhood of *q*

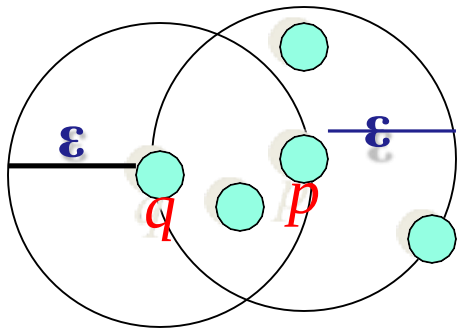
p is a core object (MinPts = 4)

q is not a core object

DBSCAN : Reachability

ML Labs Pvt Ltd

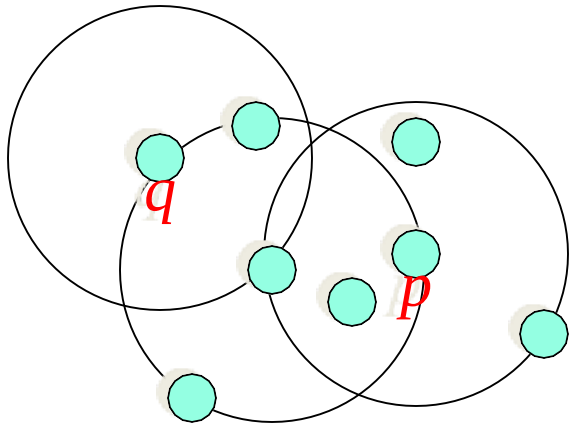
- **Directly density-reachable**
  - An object q is directly density-reachable from object p if q is within the ε-Neighborhood of p and p is a core object.



- q is directly density- reachable from p
- p is not directly density-reachable from q.

ML Labs Pvt Ltd

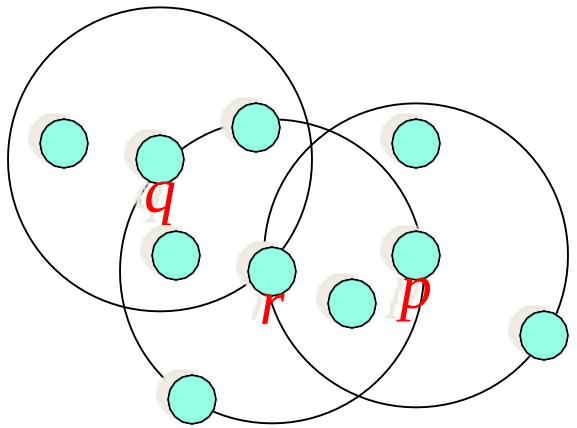
DBSCAN : Reachability



DBSCAN :Connectivity

ML Labs Pvt Ltd

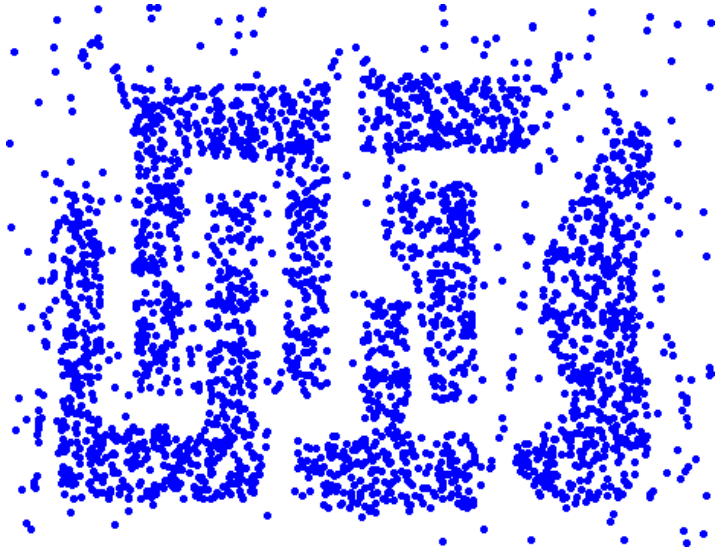
- **Density-connectivity**
  - Object *p* is density-connected to object *q* w.r.t ε and *MinPts* if there is an object *r* such that both *p* and *q* are density-reachable from *r* w.r.t ε and *MinPts*



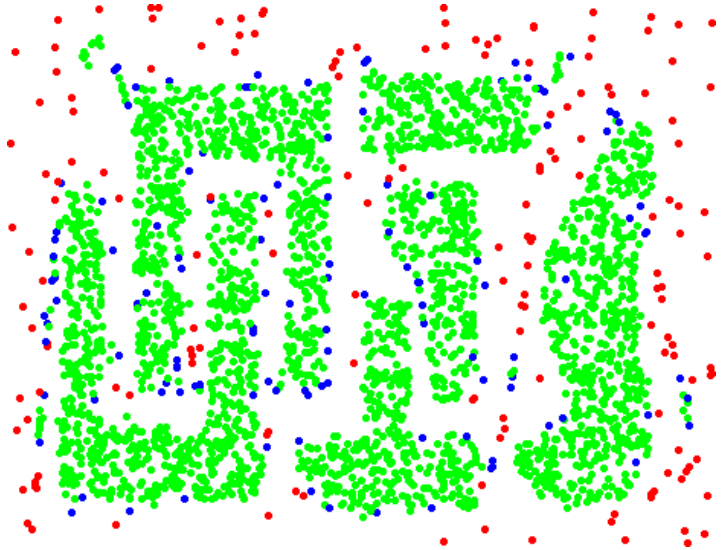
- *P* and *q* are density-connected to each other by *r*
- Density-connectivity is symmetric

Core, Border, Noise points representation

ML Labs Pvt Ltd



Original Points



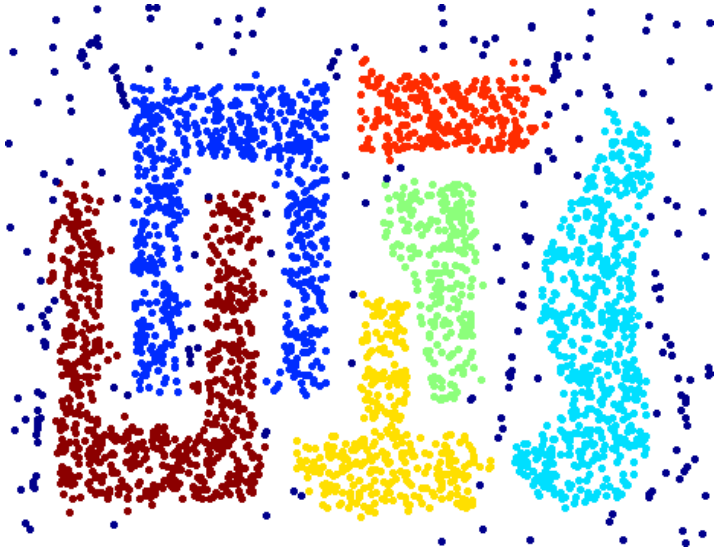
Point types: core, border and noise

Eps = 10, MinPts = 4

DBSCAN Clustering

ML Labs Pvt Ltd

Original Points



Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

DBScan Algorithm

Input: N objects to be clustered and global parameters *Eps*, *MinPts*.

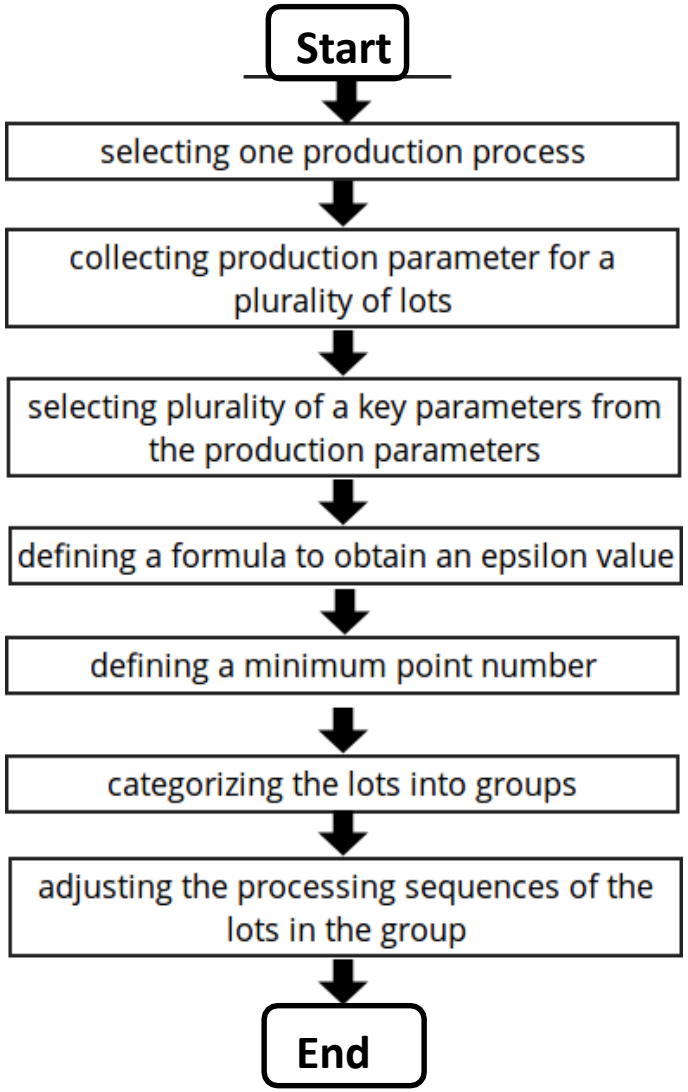
Output: Clusters of objects.

Algorithm:

1) Arbitrary select a point *P*.
2) Retrieve all points density-reachable from *P* wrt ***Eps*** and ***MinPts***.
3) If *P* is a core point, a cluster is formed.
4) If *P* is a border point, no points are density-reachable from *P* and **DBSCAN** visits the next point of the database.
5) Continue the process until all of the points have been processed.

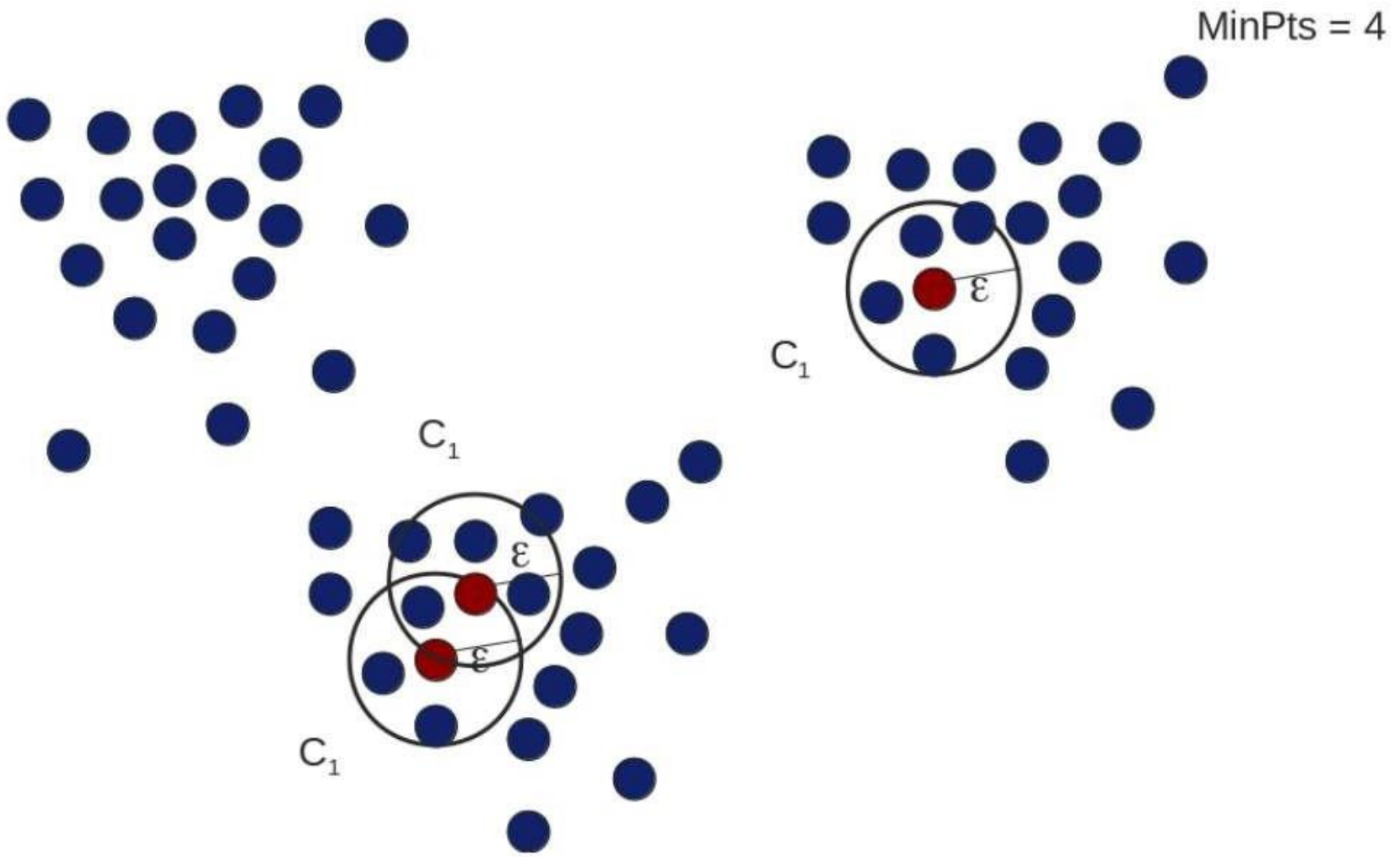
DBScan :Flowchart

ML Labs Pvt Ltd



DBScan : Example

ML Labs Pvt Ltd



DBSCAN : Advantages

- Does not require one to specify the number of clusters in the data
- Can find arbitrarily shaped clusters. even find a cluster completely surrounded by a different cluster.
- Has a notion of noise, and is robust to outliers.
- Requires just two parameters and is mostly insensitive to the ordering of the points in the database.
- Designed for accelerate region queries.
- minPts and ε can be set by a domain expert

DBSCAN : Disadvantages

ML Labs Pvt Ltd

- DBSCAN is not entirely deterministic: Border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data is processed.

- The quality of DBSCAN depends on the distance measure used in the function regionQuery. (such as Euclidean distance)

- If the data and scale are not well understood, choosing a meaningful distance threshold ε can be difficult.

DBSCAN : Complexity

ML Labs Pvt Ltd

- Time Complexity: $O(n^2)$
  - for each point it has to be determined if it is a core point.
  - can be reduced to O(n*log(n)) in lower dimensional spaces by using efficient data structures (n is the number of objects to be clustered);
- Space Complexity: O(n).

Summary of DBSCAN

ML Labs Pvt Ltd

Advantages:

- can detect arbitrary shapes
- not very sensitive to noise
- supports outlier detection
- complexity is kind of okay
- the second most used clustering algorithm after K-means

Summary of DBSCAN

ML Labs Pvt Ltd

Disadvantages:

- does not work well in high-dimensional datasets
- parameter selection is tricky
- has problems of identifying clusters of varying densities (SSN algorithm)
- density estimation is kind of simplistic (does not create a real density function, but rather a graph of density-connected points)