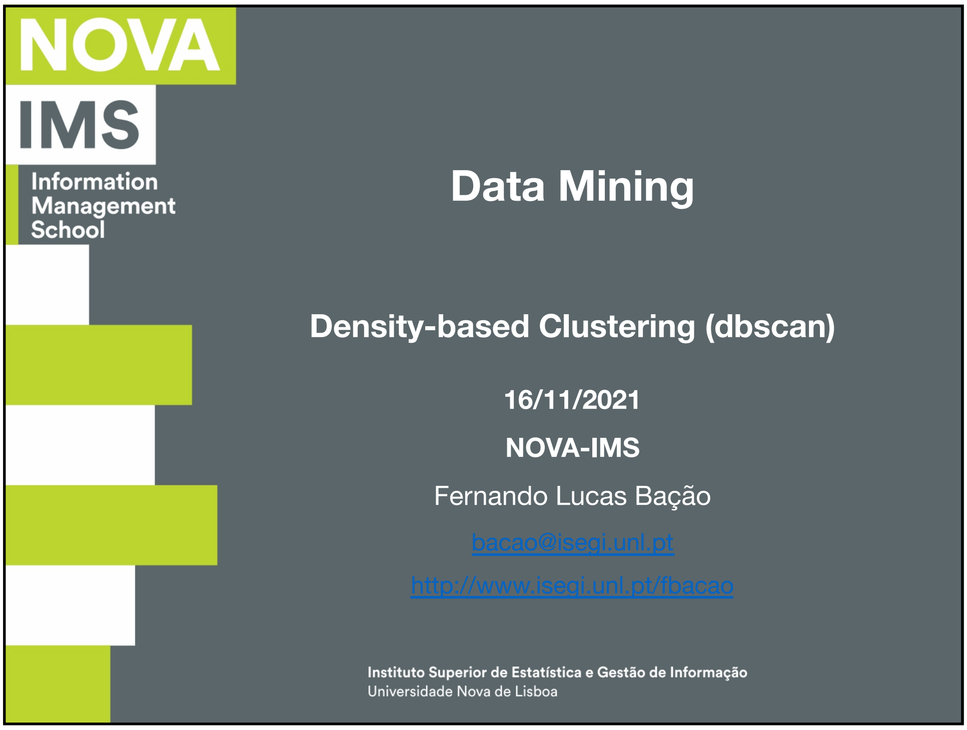


Data Mining

Density-based Clustering (dbscan)

16/11/2021

NOVA-IMS

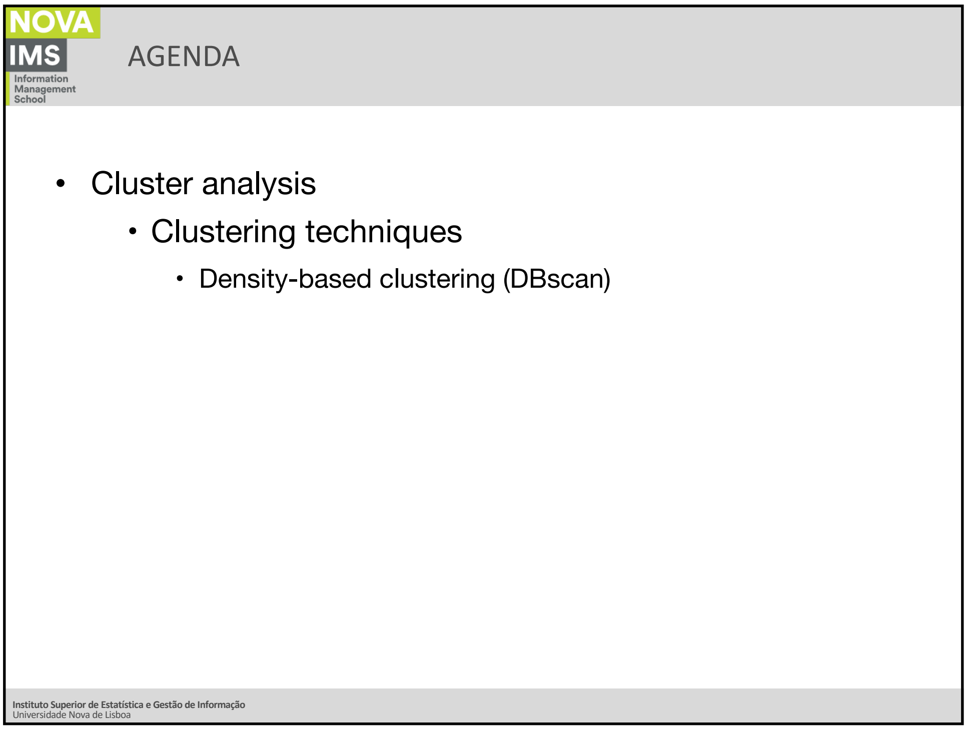
Fernando Lucas Bação

bacao@isegi.unl.pt

http://www.isegi.unl.pt/fbacao

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1

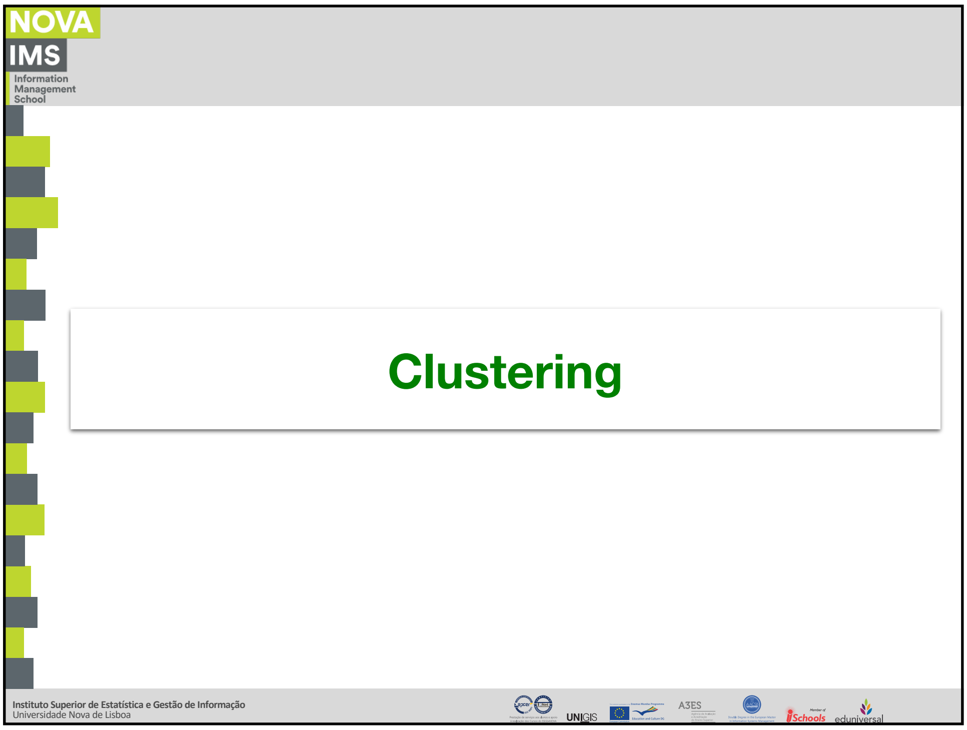


AGENDA

- Cluster analysis
  - Clustering techniques
    - Density-based clustering (DBscan)

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2

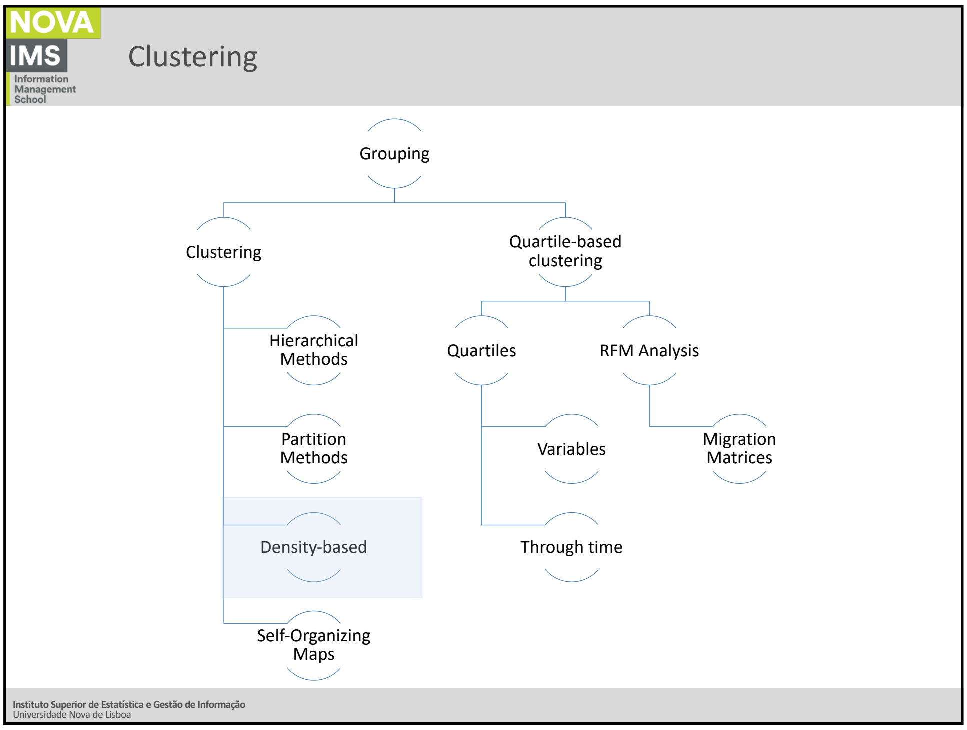


NOVA IMS Information Management School

Clustering

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

3



NOVA IMS Information Management School

Clustering

Grouping

Clustering

Quartile-based clustering

Hierarchical Methods

Partition Methods

Density-based

Self-Organizing Maps

Quartiles

RFM Analysis

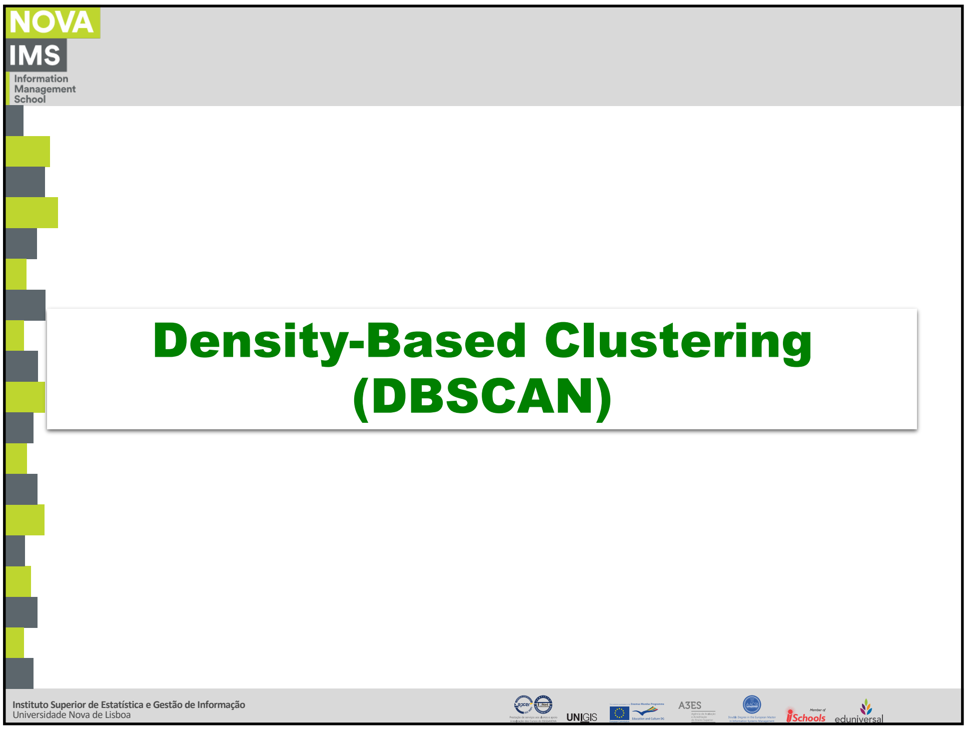
Variables

Through time

Migration Matrices

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

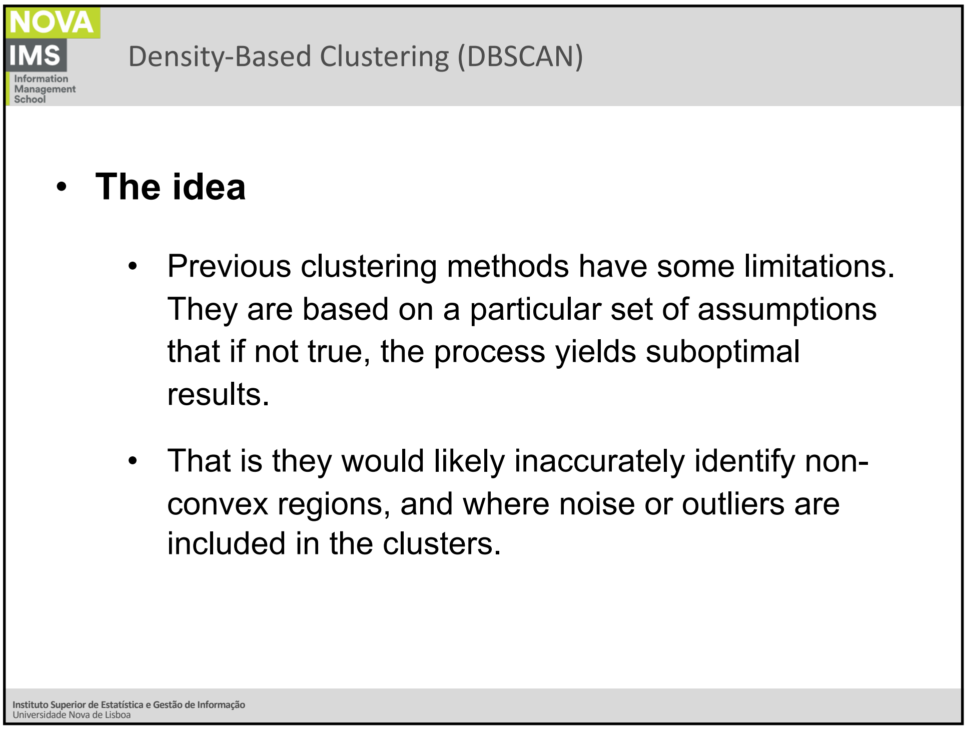
4



Density-Based Clustering (DBSCAN)

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

5



Density-Based Clustering (DBSCAN)

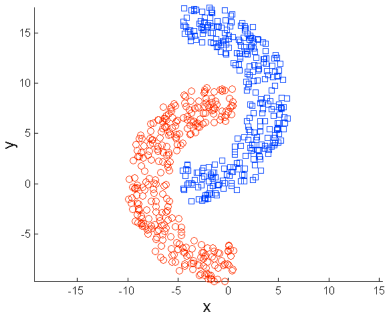
- **The idea**
  - Previous clustering methods have some limitations. They are based on a particular set of assumptions that if not true, the process yields suboptimal results.
  - That is they would likely inaccurately identify non-convex regions, and where noise or outliers are included in the clusters.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

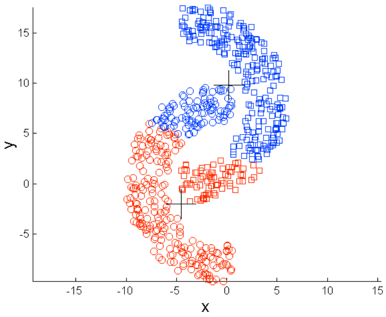
6

NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)



Original Points

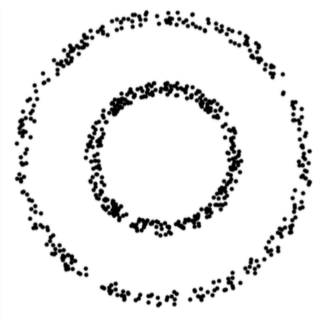


K-means (2 Clusters)

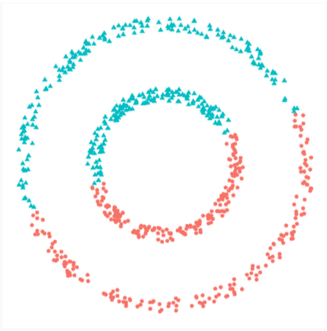
7

NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)



Original Points



K-means (2 Clusters)

8

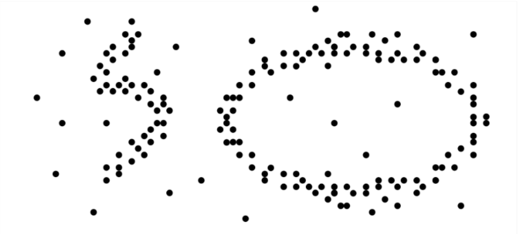
Density-Based Clustering (DBSCAN)

- **The idea**
  - Density-based clustering algorithms try to **find clusters in data without assuming a particular shape**. Thus solving correctly cases like the ring example.
  - To find clusters of **arbitrary shape**, these methods model clusters as **dense regions in data space**, separated by sparse regions.

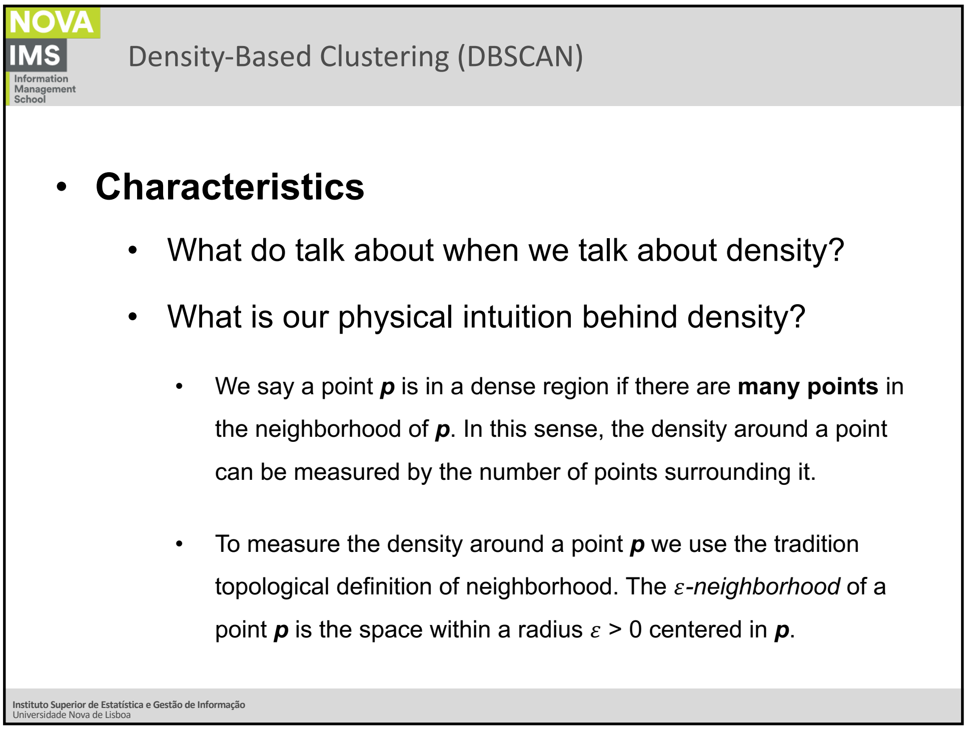
9

Density-Based Clustering (DBSCAN)

- **The idea**
  - We want to be able to cluster data like this:



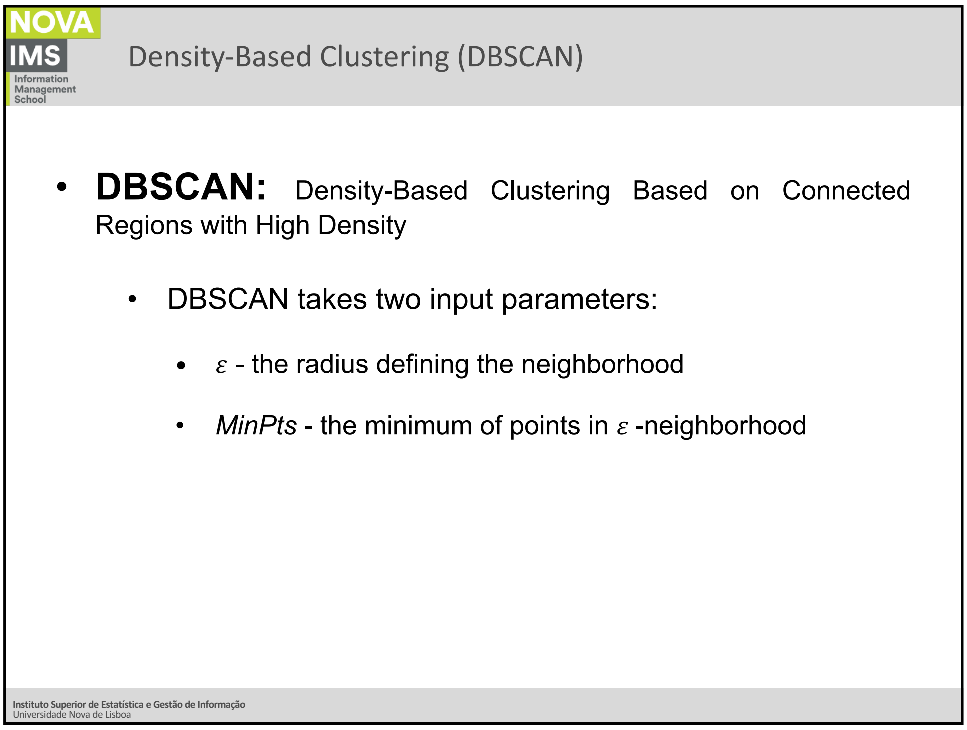
10



Density-Based Clustering (DBSCAN)

- **Characteristics**
  - What do talk about when we talk about density?
  - What is our physical intuition behind density?
    - We say a point ***p*** is in a dense region if there are **many points** in the neighborhood of ***p***. In this sense, the density around a point can be measured by the number of points surrounding it.
    - To measure the density around a point ***p*** we use the tradition topological definition of neighborhood. The *ε-neighborhood* of a point ***p*** is the space within a radius $\varepsilon > 0$ centered in ***p***.

11



Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - DBSCAN takes two input parameters:
    - ε - the radius defining the neighborhood
    - *MinPts* - the minimum of points in ε -neighborhood

12

NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - DBSCAN importante concepts:
    - If ***z*** is a point that have at least *MinPts* in its ε -neighborhood is called **core point**.
    - ***x*** is **border point**, if the number of its neighbors is less than *MinPts*, but it belongs to the ε -neighborhood of some core point ***z***
    - If a point is neither a **core** nor a **border** point, then it is called a **noise point** or an **outlier**.

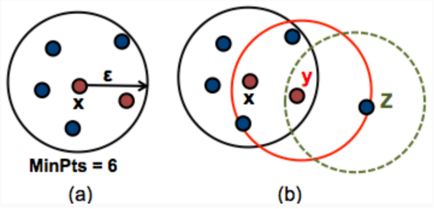
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

13

NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

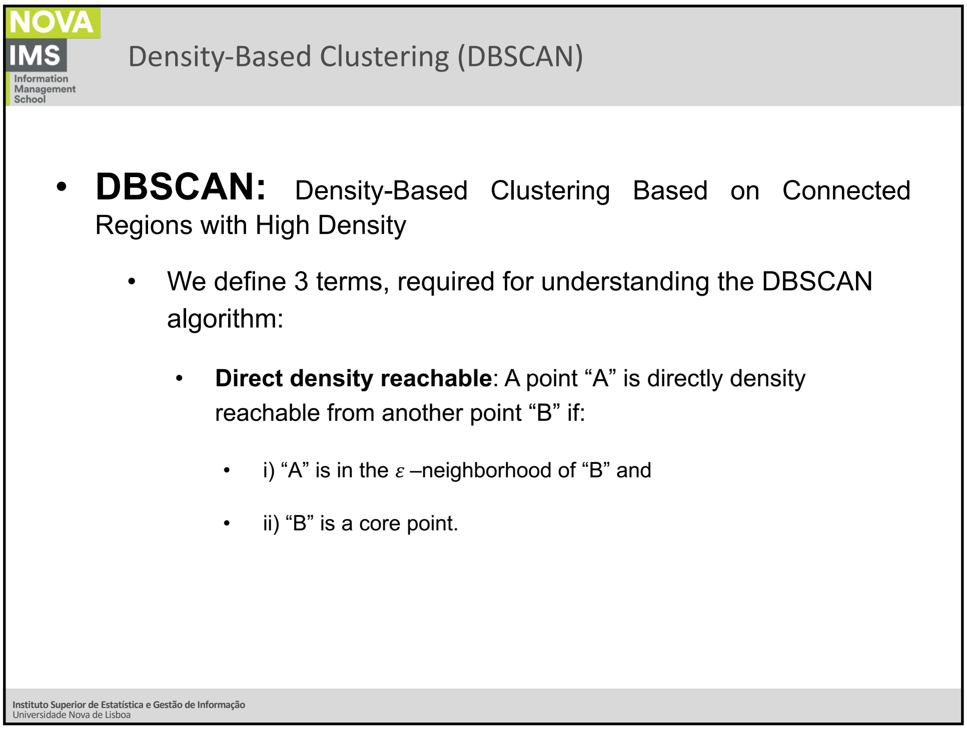
- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density



  - Assuming MinPts = 6.
    - **x** is a **core point** because ε -neighborhood (x) = 6,
    - **y** is a **border point** because ε –neighborhood (y) < *MinPts*, but it belongs to the ε -neighborhood of the core point x.
    - **z** is a **noise point**.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

14



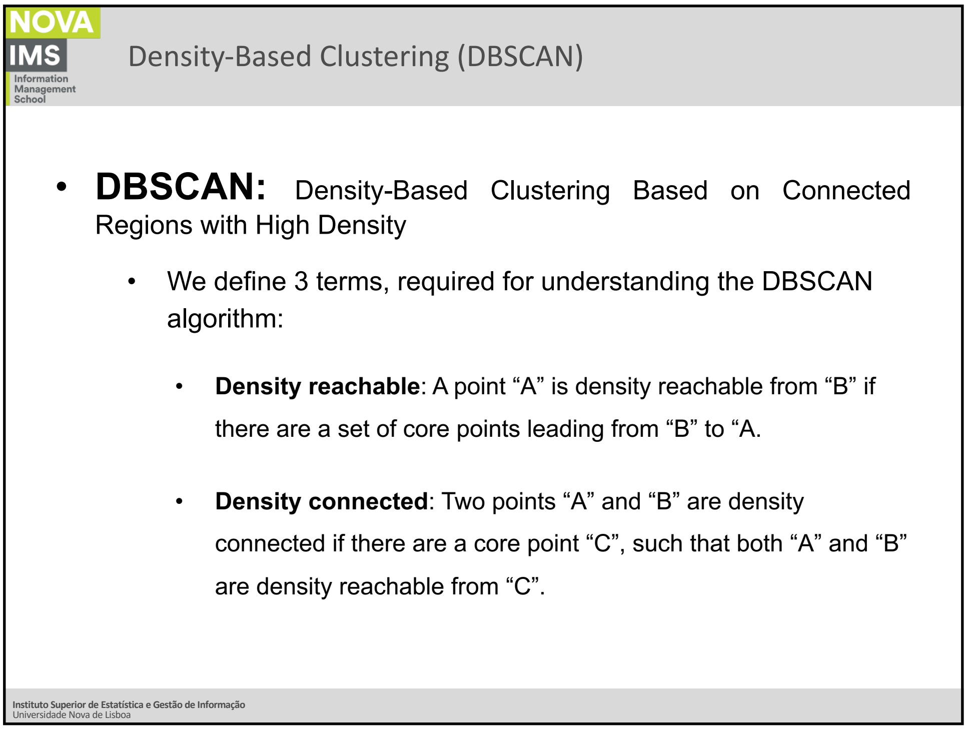
NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - We define 3 terms, required for understanding the DBSCAN algorithm:
    - **Direct density reachable**: A point “A” is directly density reachable from another point “B” if:
      - i) “A” is in the ε –neighborhood of “B” and
      - ii) “B” is a core point.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

15



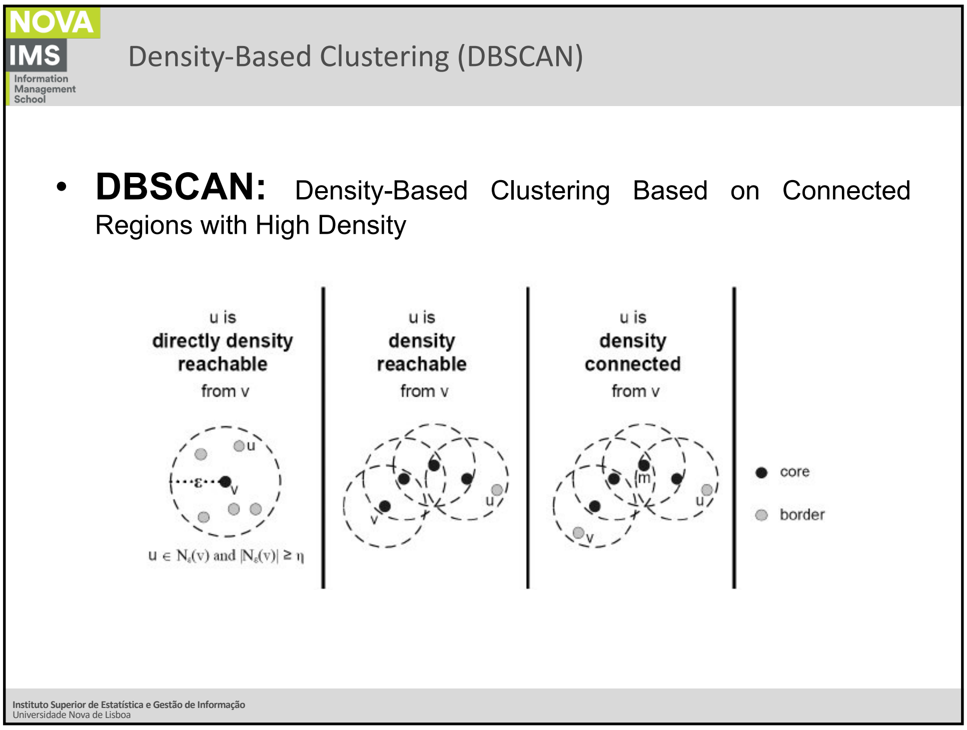
NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - We define 3 terms, required for understanding the DBSCAN algorithm:
    - **Density reachable**: A point “A” is density reachable from “B” if there are a set of core points leading from “B” to “A.
    - **Density connected**: Two points “A” and “B” are density connected if there are a core point “C”, such that both “A” and “B” are density reachable from “C”.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

16



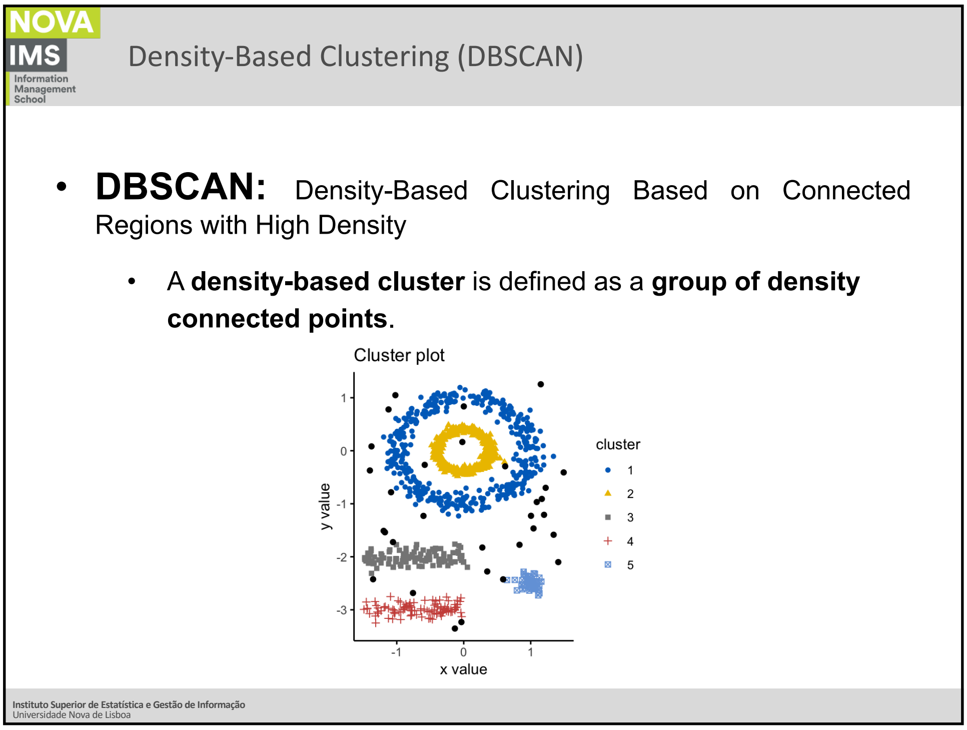
NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

17



NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - A **density-based cluster** is defined as a **group of density connected points**.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

18

NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

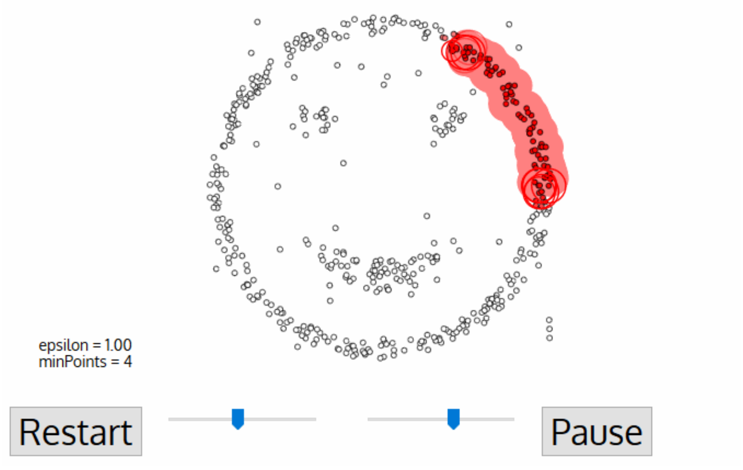
- **Algorithm**

1. DBSCAN begins with an arbitrary data point that has not been visited. The ε –neighborhood of this point is extracted (all points which are within the ε distance are neighborhood points).
2. If there are a sufficient number of points (according to *MinPts*) within this neighborhood then the clustering process starts and the current data point becomes the first point in the new cluster. Otherwise, the point will be labeled as noise. In both cases that point is marked as "visited".
3. For this first point in the new cluster, the points within its ε distance neighborhood also become part of the same cluster. This procedure of making all points in the ε –neighborhood belong to the same cluster is then repeated for all of the new points that have been just added to the cluster group.
4. This process of steps 2 and 3 is repeated until all points in the cluster are determined i.e all points within the ε –neighborhood of the cluster have been visited and labeled.

19

NOVA IMS Information Management School

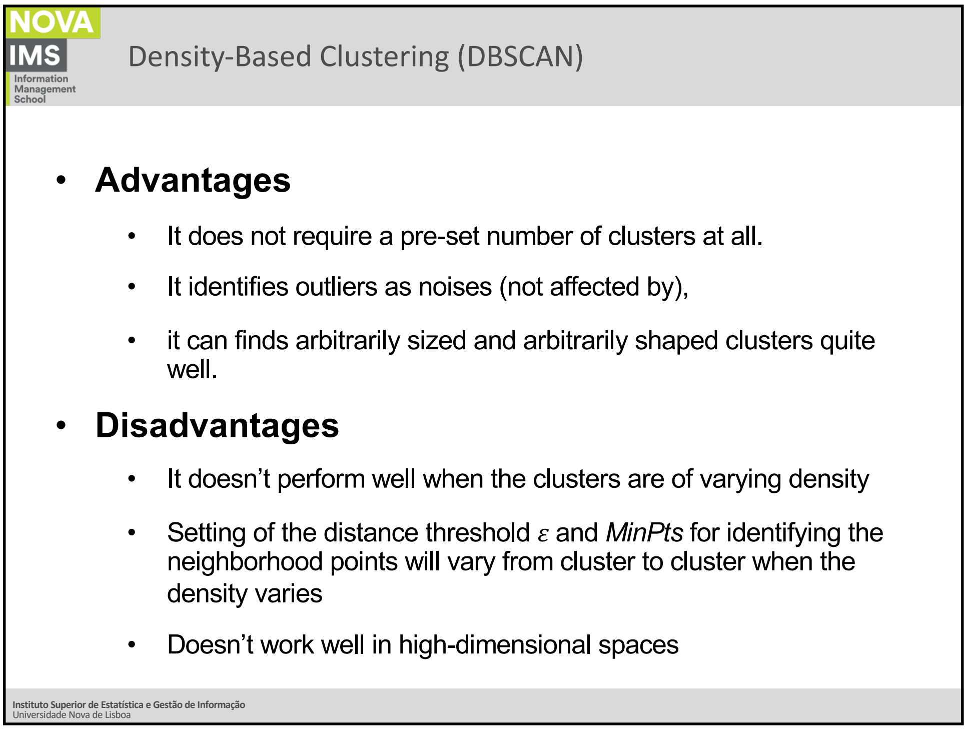
Density-Based Clustering (DBSCAN)



epsilon = 1.00
minPoints = 4

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

20



NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

- **Advantages**
  - It does not require a pre-set number of clusters at all.
  - It identifies outliers as noises (not affected by),
  - it can finds arbitrarily sized and arbitrarily shaped clusters quite well.
- **Disadvantages**
  - It doesn't perform well when the clusters are of varying density
  - Setting of the distance threshold ε and *MinPts* for identifying the neighborhood points will vary from cluster to cluster when the density varies
  - Doesn't work well in high-dimensional spaces

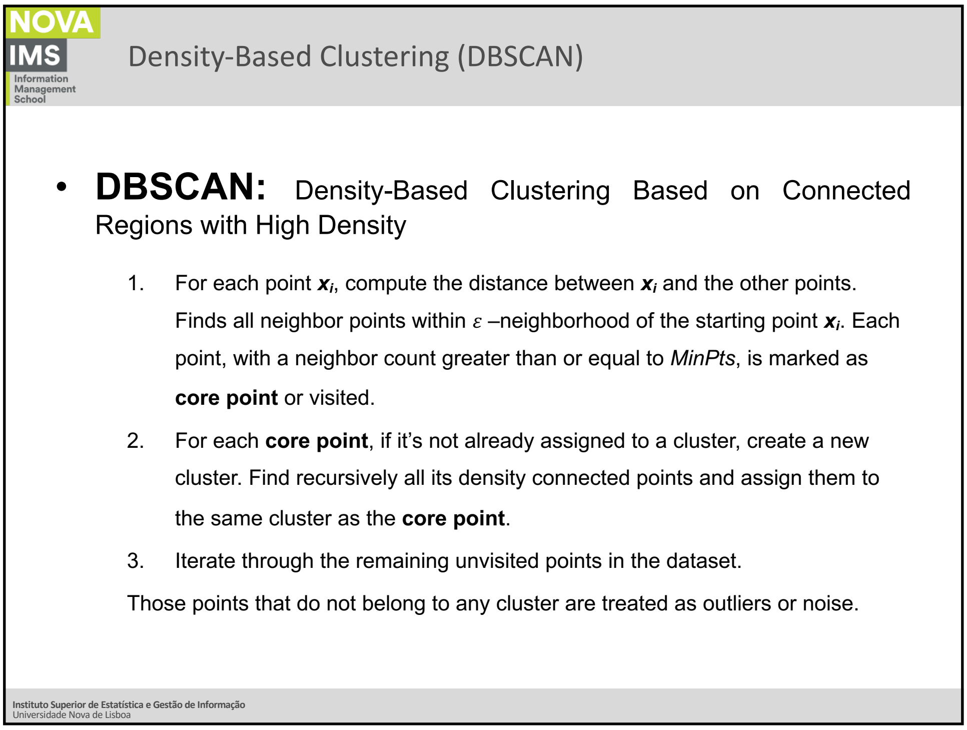
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

21

Questions?

22

22



Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density

  1. For each point $\boldsymbol{x_i}$, compute the distance between $\boldsymbol{x_i}$ and the other points. Finds all neighbor points within ε –neighborhood of the starting point $\boldsymbol{x_i}$. Each point, with a neighbor count greater than or equal to *MinPts*, is marked as **core point** or visited.
  2. For each **core point**, if it's not already assigned to a cluster, create a new cluster. Find recursively all its density connected points and assign them to the same cluster as the **core point**.
  3. Iterate through the remaining unvisited points in the dataset.

  Those points that do not belong to any cluster are treated as outliers or noise.