

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

**DBSCAN: Density-Based Clustering with Noise**

HUMAN CAPITAL  
HUMAN - BEST INVESTMENT

EUROPEAN UNION  
SOCIAL FUND

Project is co-financed by European Union within European Social Fund

1

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

**Basic Notions**

- *Distance metrics*
- *Eps-neighborhood of a point p*
- *Core point*
- *Cluster*
- *Noise*

2

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

**Distance Metric**

A *distance metric* is defined as a measure that satisfies the following conditions:

- $\forall p, \text{distance}(p, p) = 0$ ;
- $\forall p, q, \text{distance}(p, q) = \text{distance}(q, p)$ ;
- $\forall p, q, r, \text{distance}(p, r) \leq \text{distance}(p, q) + \text{distance}(q, r)$  /\* triangle inequality property \*/.

3

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

**Example Distance Metrics**

- Euclidean( $p, q$ ) =  $\sqrt{\sum_{i=1..n} (p_i - q_i)^2}$
- Manhattan( $p, q$ ) =  $\sum_{i=1..n} |p_i - q_i|$
- Minkowski( $p, q$ ) =  $\sqrt[m]{\sum_{i=1..n} |p_i - q_i|^m}$

4

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

**Eps-Neighborhood**

- *Eps-neighborhood of a point p* (denoted by  $N_{\text{Eps}}(p)$ ) is defined as the set of all points q in dataset D that are distant from p by no more than Eps; that is,

$$N_{\text{Eps}}(p) = \{q \in D \mid \text{distance}(p, q) \leq \text{Eps}\}.$$

5

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

**Example: Eps-Neighborhood**

$|N_{\text{Eps}}(p)| = 4.$

6

## Core Points

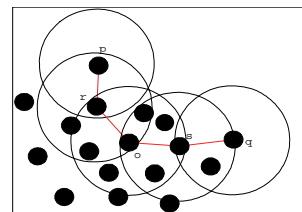
- A point  $p$  is defined as a *core point* if its *Eps-neighborhood* contains at least  $\text{MinPts}$  points; that is, if  $|N_{\text{Eps}}(p)| \geq \text{MinPts}$ .

7

## Example: Core Points

For  $\text{MinPts} = 6$ :

- $r$  is a core point;
- $p$  is not a core.



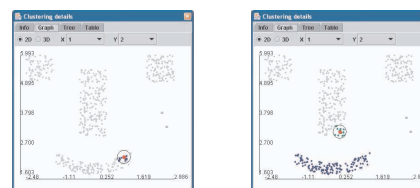
8

## Clusters in DBSCAN...

- A core point is treated as a seed from which a cluster is built.
- Whenever any core point is included in the cluster, all points in its Eps-neighborhood are also included in the cluster.

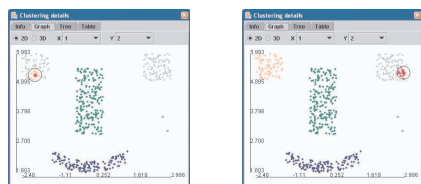
9

## Clusters in DBSCAN...



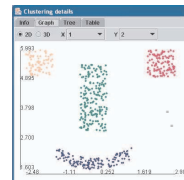
10

## Clusters in DBSCAN...



11

## Clusters in DBSCAN

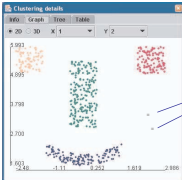


12

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

## Clusters and Noise in DBSCAN

- The points that are not included in any cluster constitute *noise*.



13

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

## Major Challenge in DBSCAN

- Efficient calculation of Eps-neighborhood for each point.
- To this end, DBSCAN uses the R\*-tree index.
- The use of such indices helps in the case of low dimensional data only.

14

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

## Efficient Calculation of Eps-Neighborhoods

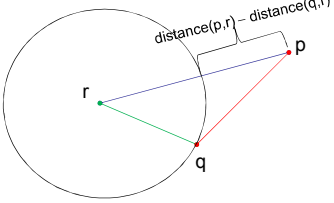
- Use the **triangle inequality property (TI)** to reduce the number of candidates for being a member of Eps-neighborhood of a given point.

15

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

## Equivalent Form of TI

For any three points p, q, r:

$$\text{distance}(p,q) \geq \text{distance}(p,r) - \text{distance}(q,r).$$


16

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

## TI & Eps-Neighborhood...

**Lemma.** Let D be a set of points. For any two points p, q in D and any point r:

$$\text{distance}(p,r) - \text{distance}(q,r) > \text{Eps} \Rightarrow$$

$$\text{distance}(p,q) \geq \text{distance}(p,r) - \text{distance}(q,r) > \text{Eps} \Rightarrow$$

by TI

$$q \notin N_{\text{Eps}}(p) \wedge p \notin N_{\text{Eps}}(q).$$

17

WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

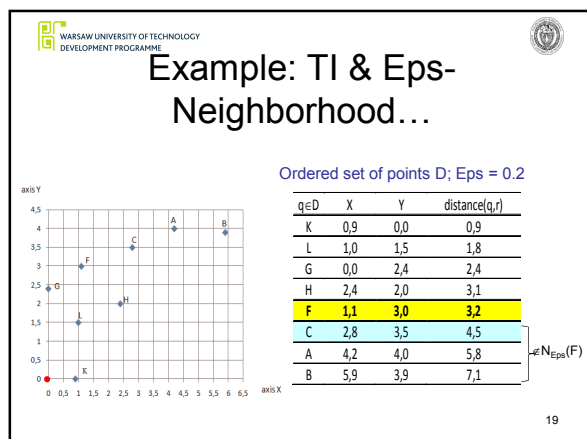
## TI & Eps-Neighborhood...

**Theorem.** Let:

- r be any point,
- D be a set of points ordered in a non-decreasing way wrt. their distances to r;
- p be any point in D;
- q be a point following point p in D such that  $\text{distance}(q,r) - \text{distance}(p,r) > \text{Eps}$ .

Then q and all points following q in D do not belong to  $N_{\text{Eps}}(p)$ .

18



WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

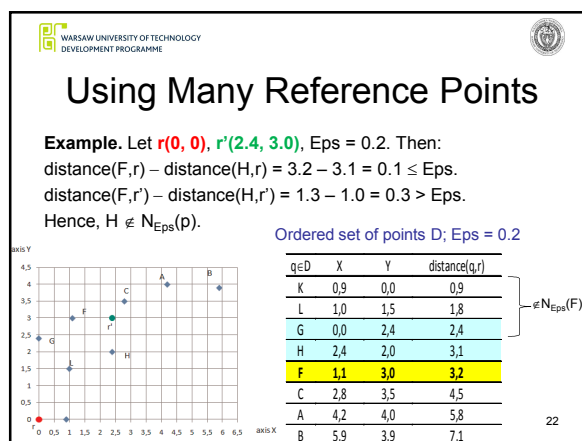
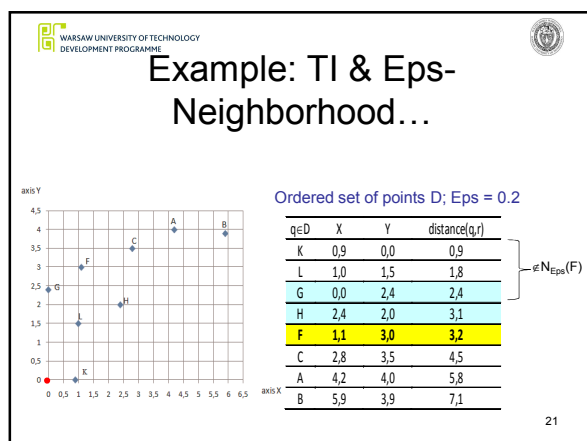
### TI & Eps-Neighborhood...

**Theorem.** Let:

- $r$  be any point,
- $D$  be a set of points ordered in a non-decreasing way wrt. their distances to  $r$ ;
- $p$  be any point in  $D$ ;
- $q$  be a point preceding point  $p$  in  $D$  such that  $\text{distance}(p,r) - \text{distance}(q,r) > \text{Eps}$ .

Then  $q$  and all points preceding  $q$  in  $D$  do not belong to  $N_{Eps}(p)$ .

20



WARSAW UNIVERSITY OF TECHNOLOGY  
DEVELOPMENT PROGRAMME

### References

- Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. [KDD 1996](#): 226-231
- Marzena Kryszkiewicz, Piotr Lasek: TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality. [RSCTC 2010](#): 60-69

23