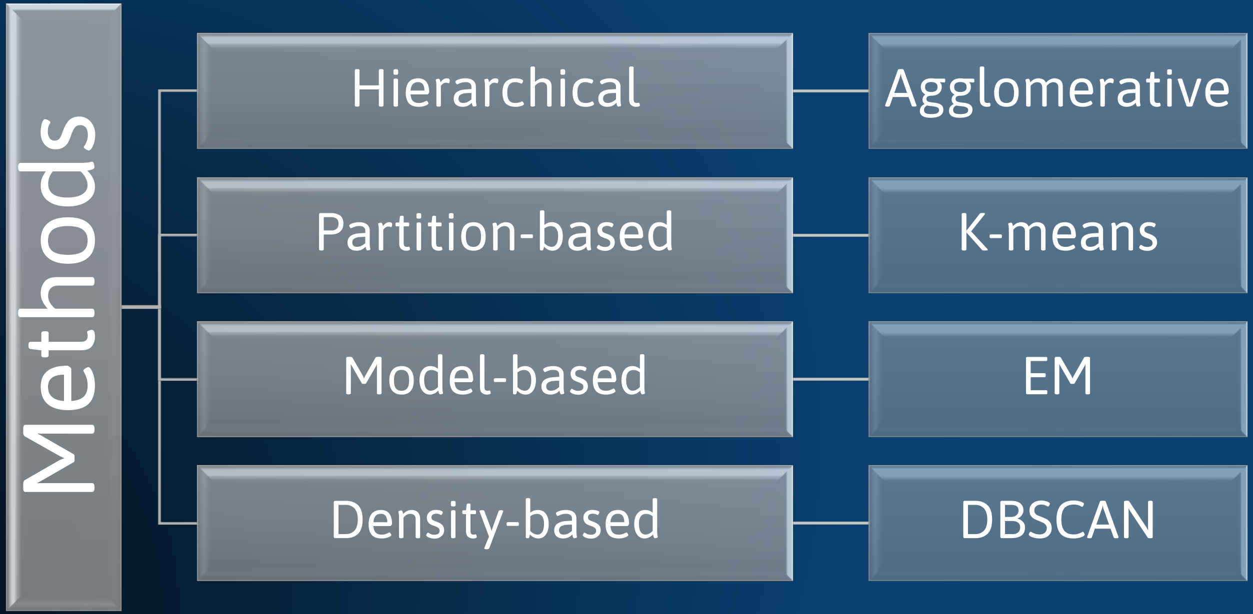
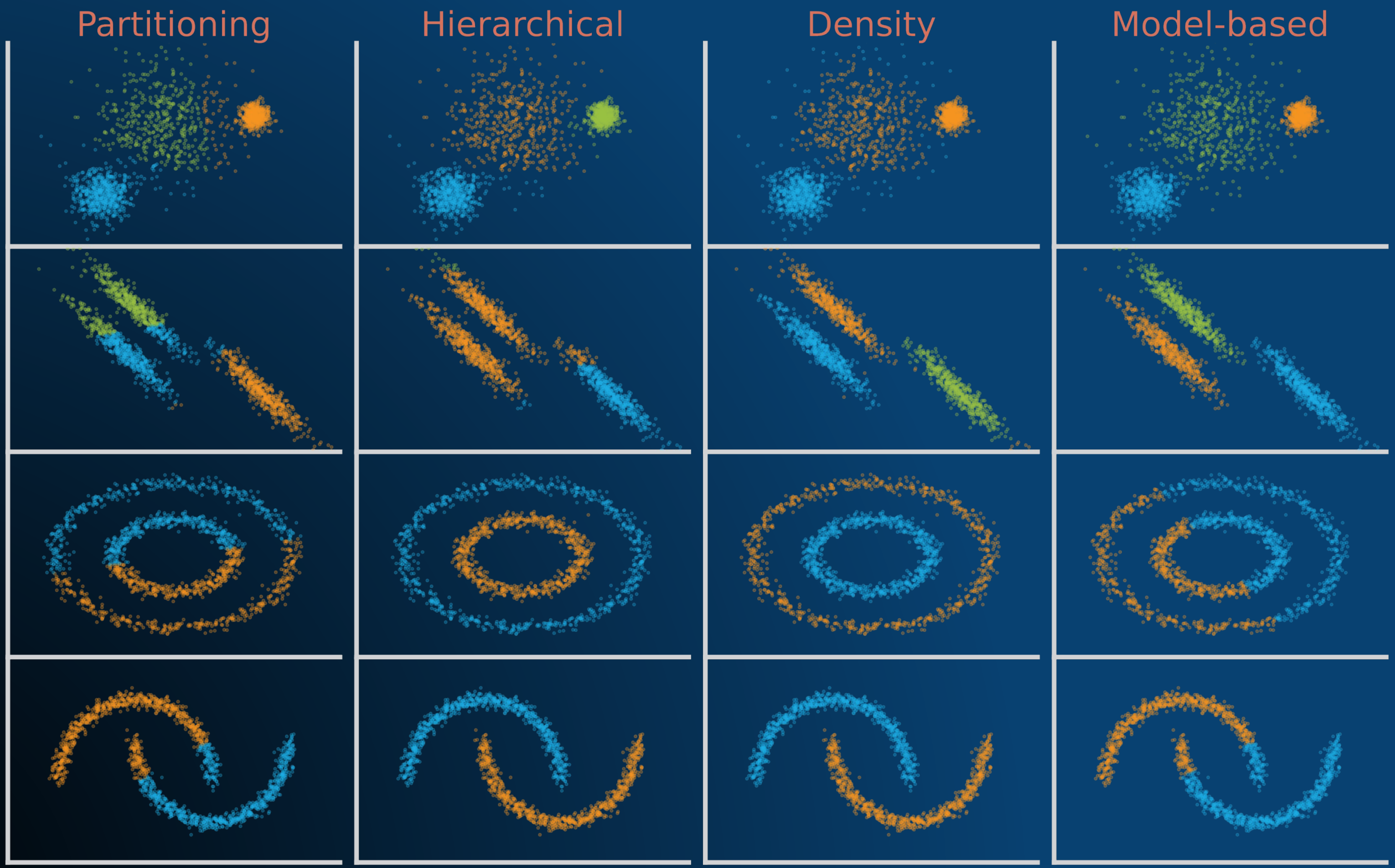
Clustering Algorithms

Cláudia Antunes

Instituto Superior Técnico – Universidade de Lisboa

APPROACHES


Methods
Hierarchical
Agglomerative
Partition-based
K-means
Model-based
EM
Density-based
DBSCAN

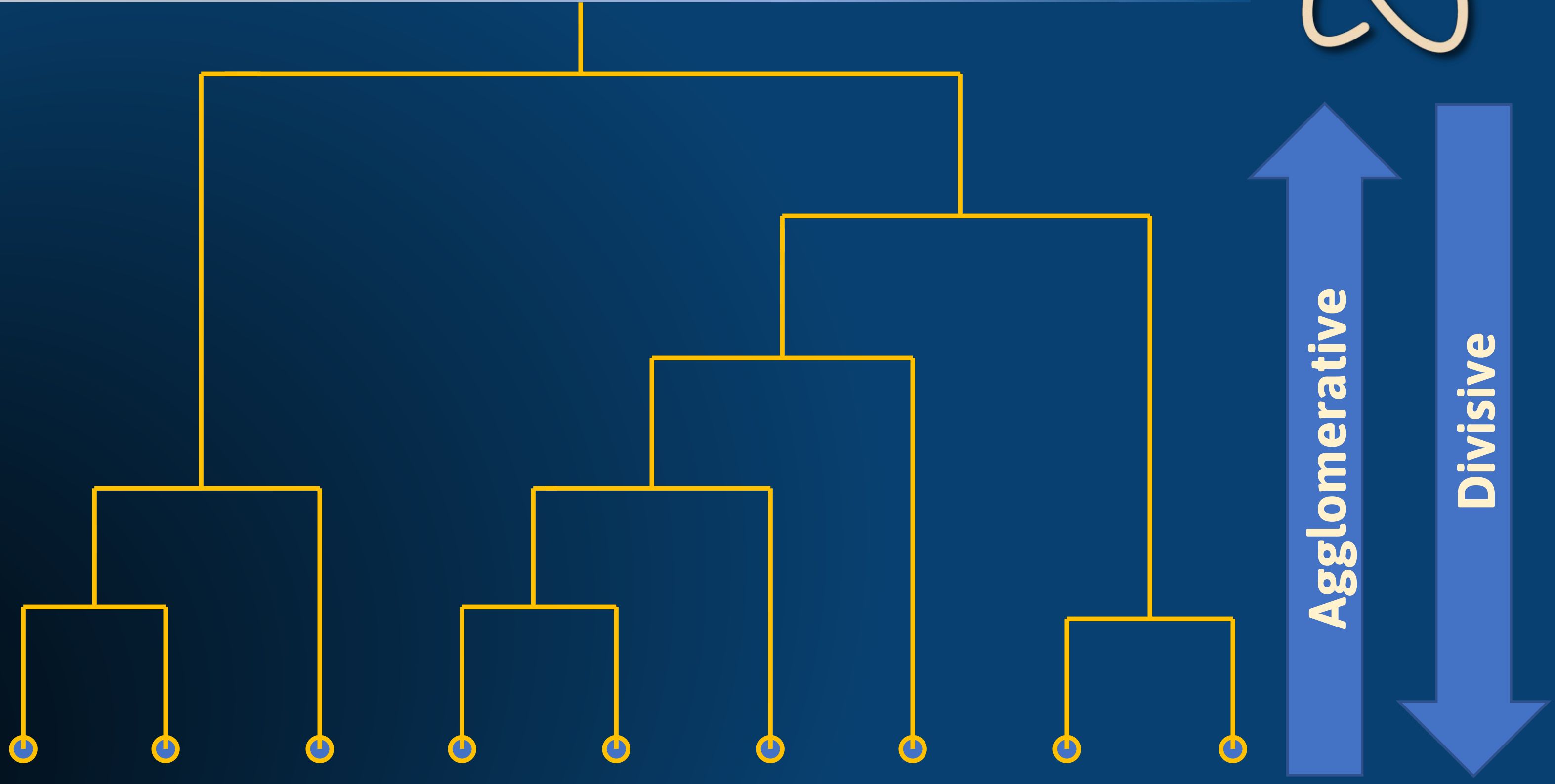


Partitioning
Hierarchical
Density
Model-based

Hierarchical Algorithms

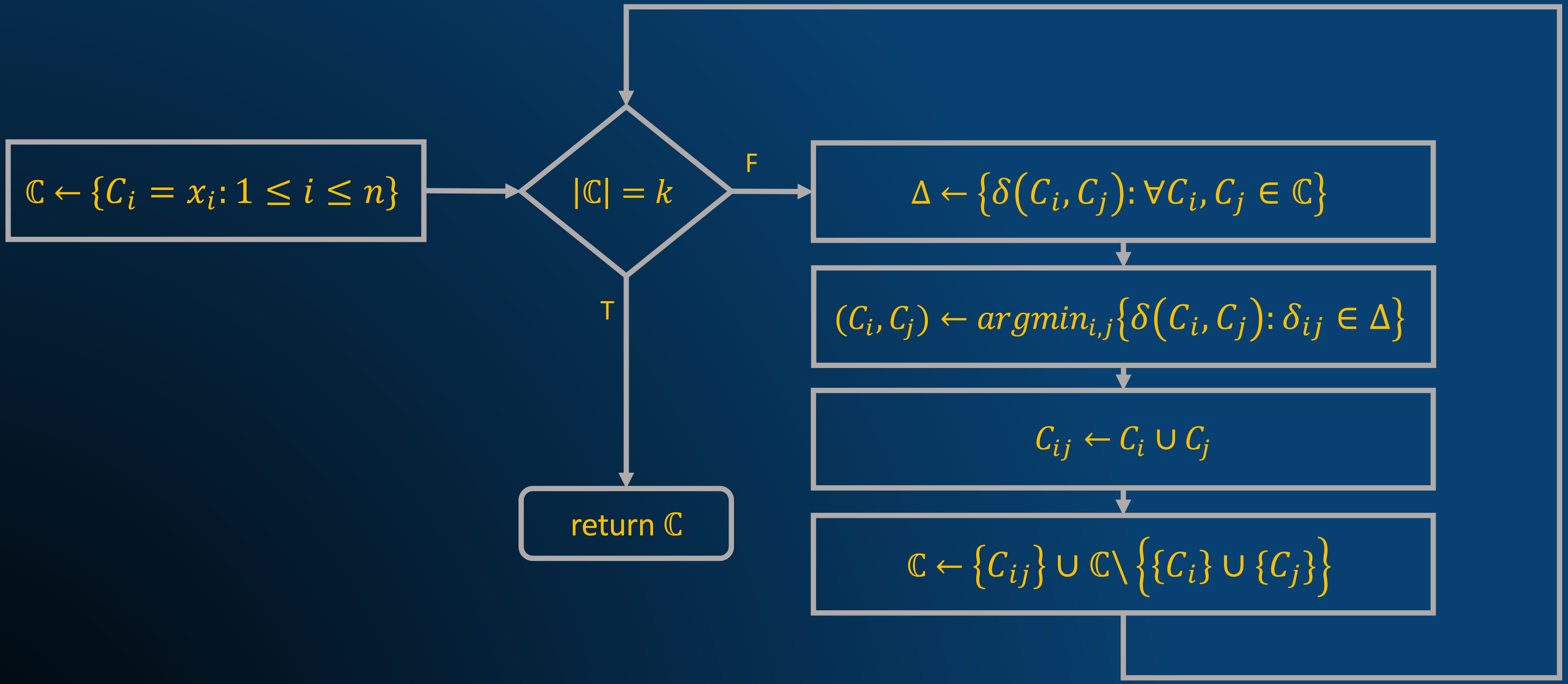
Agglomerative

Hierarchical Methods



Input $D = \{x_1, \dots, x_n\}, k$

Output $\mathbb{C}$



AgglomerativeClustering (Dataset D, int k)

 # Put each record in a separate cluster

$\mathbb{C} \leftarrow \{C_i = x_i : x_i \in D\}$

while $|\mathbb{C}| \neq k$ **do**

 # Compute distance matrix

$\Delta \leftarrow \{\delta(C_i, C_j) : \forall C_i, C_j \in \mathbb{C}\}$

 # Find the closest pair of clusters

$(C_i, C_j) \leftarrow argmin_{i,j}\{\delta(C_i, C_j) : \delta_{ij} \in \Delta\}$

 # Merge the clusters

$C_{ij} \leftarrow C_i \cup C_j$

 # Update the clustering partition

$\mathbb{C} \leftarrow \mathbb{C} \backslash \left\{\{C_i\} \cup \{C_j\}\right\} \cup \{C_{ij}\}$

return $\mathbb{C}$

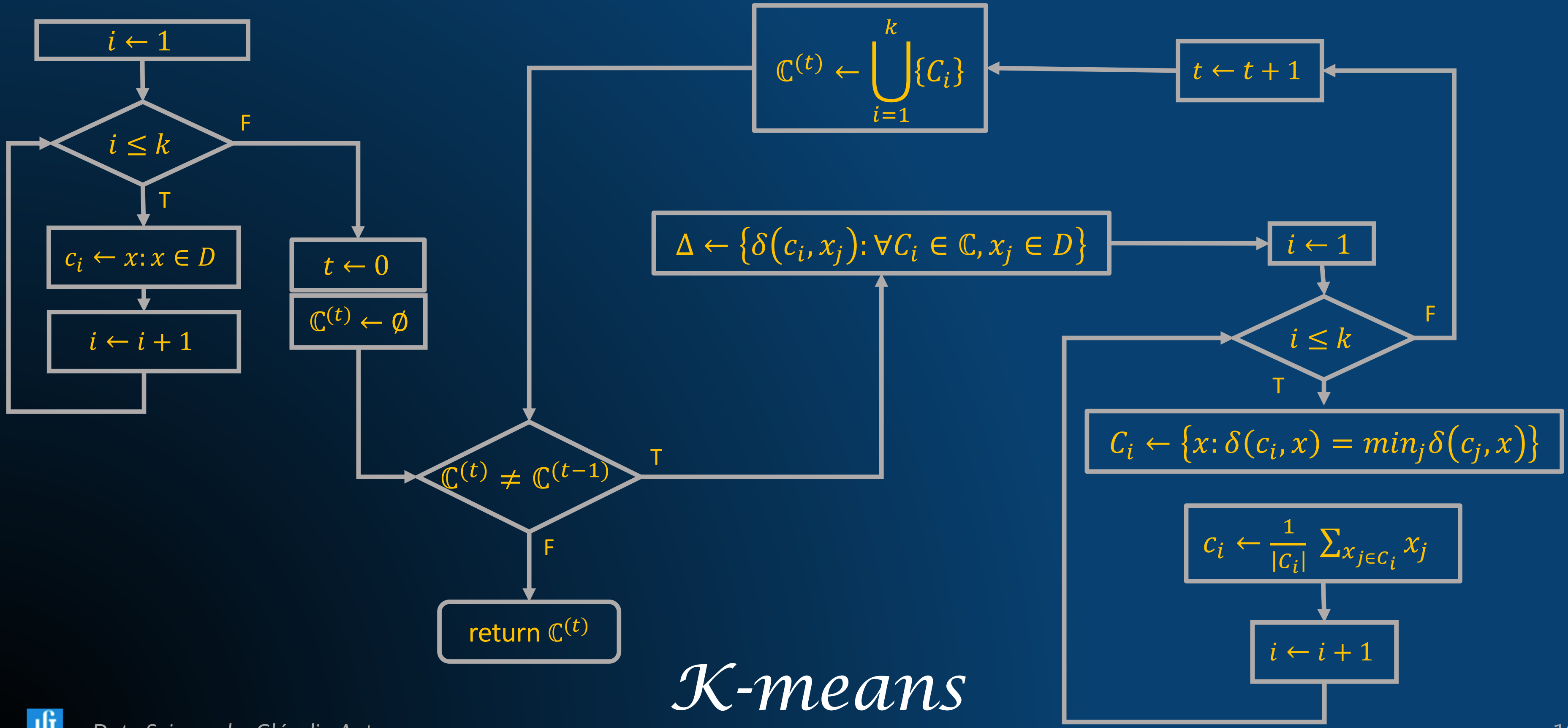
Partition-Based Algorithms

K-Means

Partition-Based Algorithms

Input $D = \{x_1, \dots, x_n\}, k, \xi$

Output $\mathbb{C}$



K-means

K-means Algorithm (Dataset D, int k , float ξ)

for each $i: 1 \leq i \leq k$ **do** *# Choose a centroid for each cluster*

$c_i \leftarrow random(x): x \in D$

$t \leftarrow 0$

$\mathbb{C}^{(0)} \leftarrow \emptyset$

do

$\Delta \leftarrow \{\delta(c_i, x_j): \forall C_i \in \mathbb{C}, x_j \in D\}$ *# Compute distance for each (record, centroid)*

for each $i: 1 \leq i \leq k$ **do**

$C_i \leftarrow \{x: \delta(c_i, x) = min_j \delta(c_j, x)\}$ # Assign each record to the closest cluster

$c_i \leftarrow \frac{1}{|C_i|} \sum_{x_{j \in C_i}} x_j$ *# Update the centroids*

$\mathbb{C}^{(t)} \leftarrow \cup_{i=1}^{k} \{C_i\}$

until $\left\|\mathbb{C}^{(t)} - \mathbb{C}^{(t-1)}\right\| \leq \xi$

return $\mathbb{C}$

Model-based Algorithms

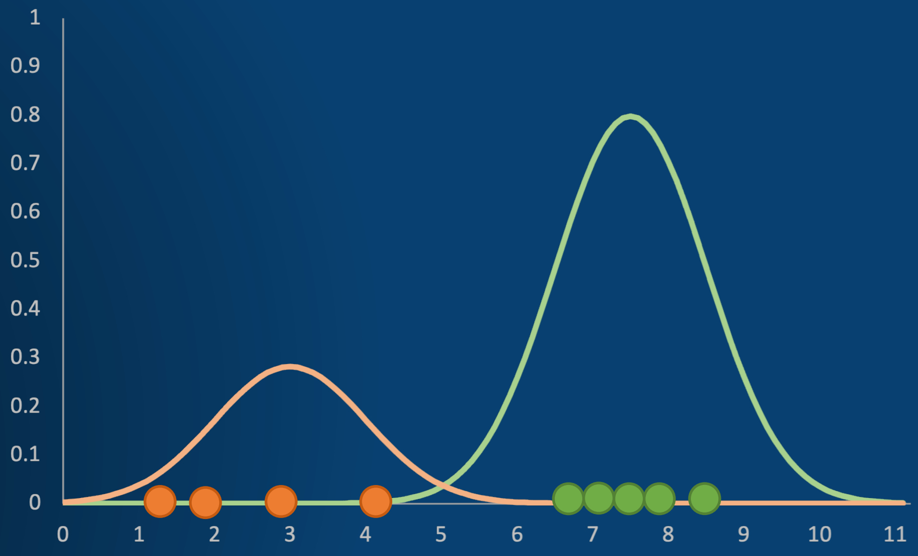
Expectation Maximization

Mixture Models

Mixture of models given by

$$P(x) = \sum_{i=1}^{k} P(C = i)P(x|C = i)$$

$$1 = \sum_{i=1}^{k} P(C = i)$$



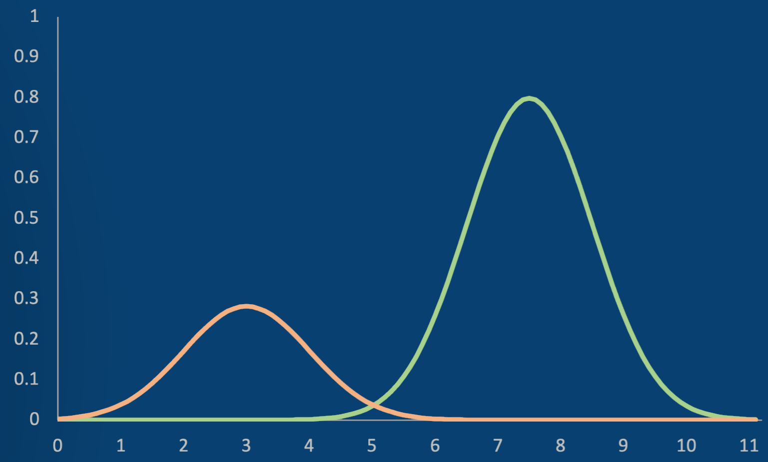
Find a set *C* of *k* probabilistic clusters where *P*(*D*|*C*) is **maximized**

Gaussian Mixture Model

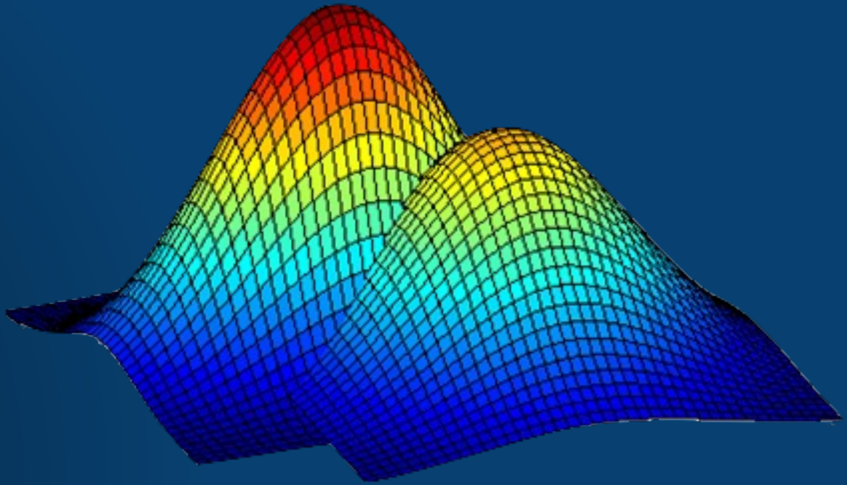
In $\mathbb{R}$:

$$C_i \sim \mathcal{N}(\mu_i, \sigma_i)$$

$$f_i(x) = f(x|\mu_i, \sigma_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$$



Gaussian Mixture Model



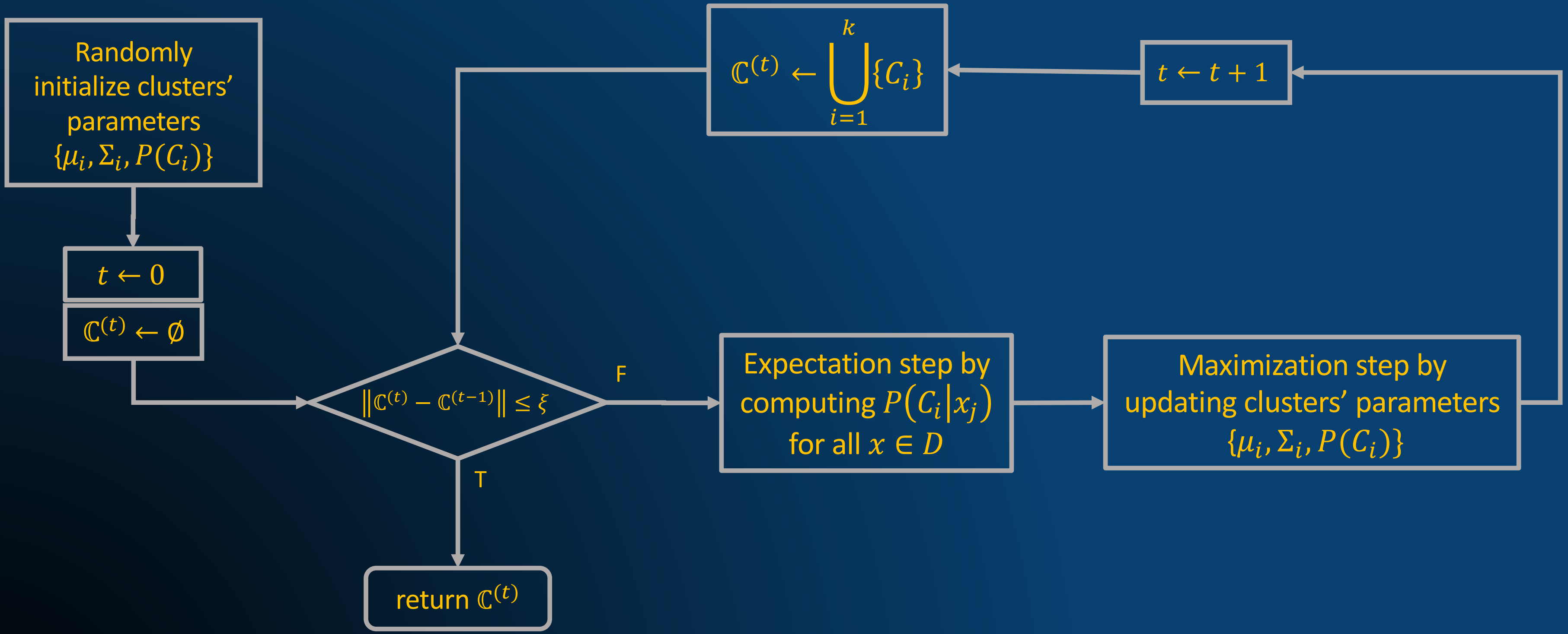
In $\mathbb{R}^d$:

$$C_i \sim \mathcal{N}(\mu_i, \Sigma_i)$$

$$f_i(x) = f(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{2\pi^d |\Sigma|}} e^{-\frac{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}{2}}$$

Input $D = \{x_1, \dots, x_n\}, k, \xi$

Output $\mathbb{C}$



Expectation-Maximization

EM Algorithm (Dataset D, int k, float ξ)

$t \leftarrow 0$

for each $i: 1 \leq i \leq k$ **do** *# Clusters initialization*

$\mu_i^t \leftarrow random()$ $\Sigma_i^t \leftarrow \mathbb{I}$ $P^t(C_i) \leftarrow \frac{1}{k}$

do

$t \leftarrow t + 1$

for $i = 1 \ldots k$ **and** $j = 1 \ldots n$ **do** *# Expectation step*

$$w_{ij} \leftarrow \frac{f(x_j|\mu_i, \Sigma_i)P(C_i)}{\sum_{a=1}^{k} f(x_j|\mu_a, \Sigma_a)P(C_a)}$$

$w_{ij} = P(C_i|x_j)$

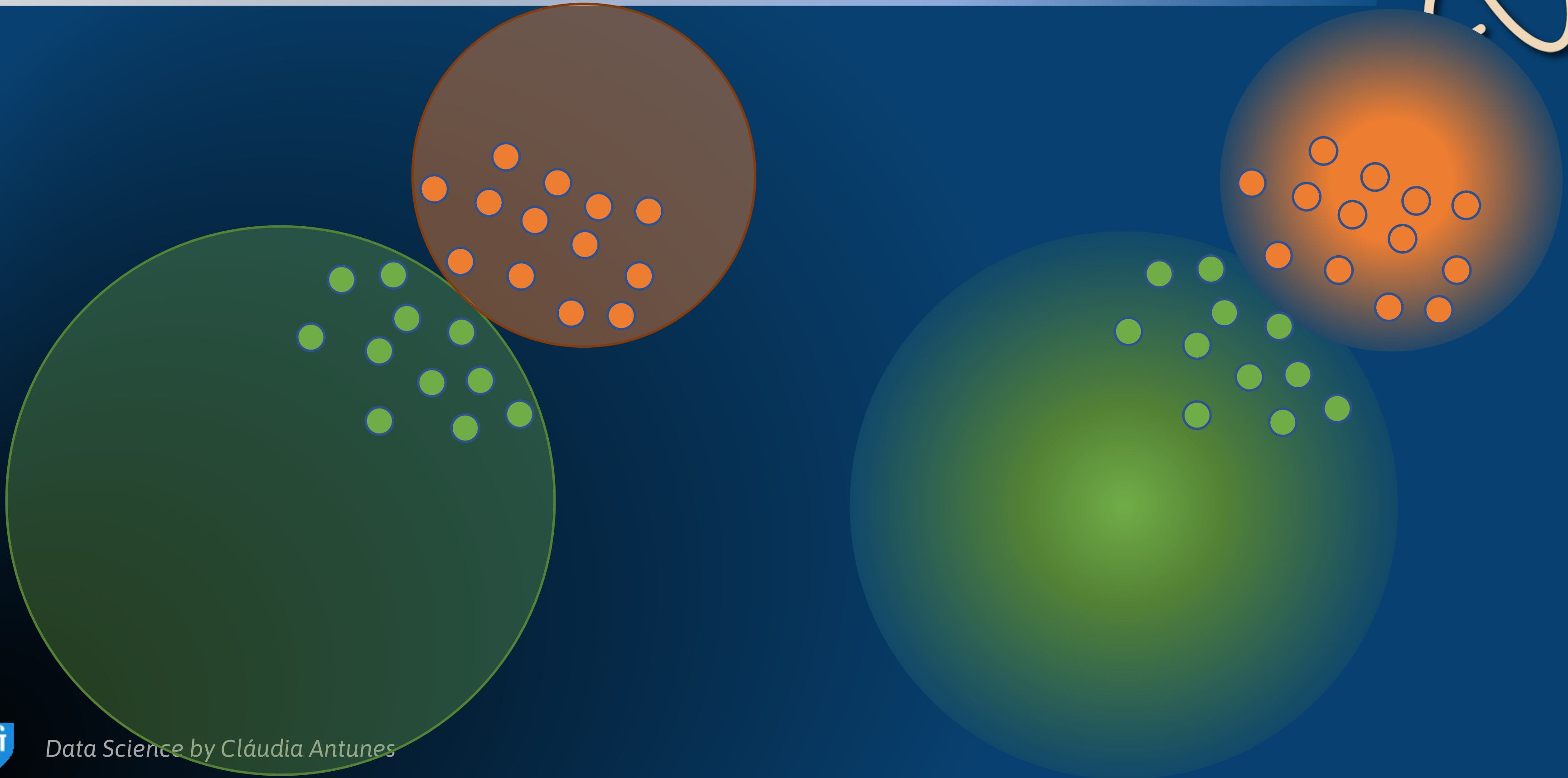
for $i = 1 \ldots k$ **do** *# Maximization step*

$$\mu_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij}x_j}{\sum_{j=1}^{n} w_{ij}} \qquad \Sigma_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij}(x_j-\mu_i))(x_j-\mu_i)^T}{\sum_{j=1}^{n} w_{ij}} \qquad P^t(C_i) \leftarrow \frac{\sum_{j=1}^{n} w_{ij}}{N}$$

until $\sum_{i=1}^{k} \left\|\mu_i^t - \mu_i^{t-1}\right\|^2 \leq \xi$

return $\cup_{i=1}^{k}\{C_i\}$

Soft x Hard Clusters

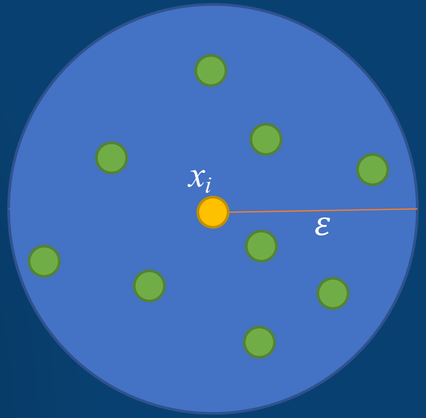


Density-based

DBSCAN

$\varepsilon - neighborhood$

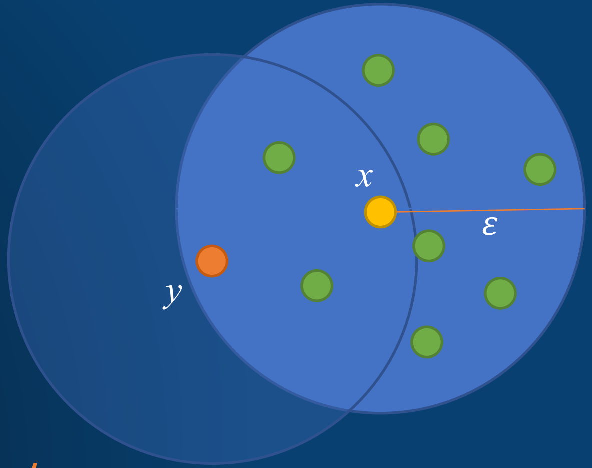
$$x \in Core \leftrightarrow |N_\varepsilon(x)| \geq minpts$$



$$N_\varepsilon(x) \leftarrow \{x' : x' \in D \wedge \delta(x, x') \leq \varepsilon\}$$

$\varepsilon - neighborhood$

$$x \in Core \leftrightarrow |N_\varepsilon(x)| \geq minpts$$

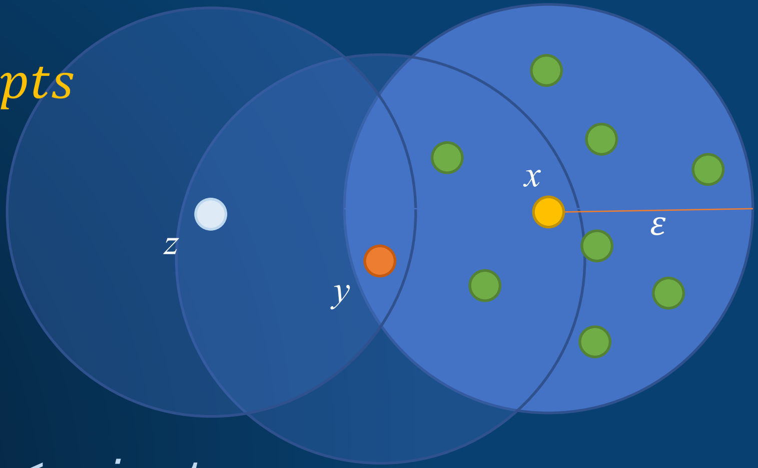


$$y \in Border$$
$$\leftrightarrow$$
$$\exists x \in Core: y \in N_\varepsilon(x) \ \wedge \ |N_\varepsilon(y)| < minpts$$

$$N_\varepsilon(x) \leftarrow \{x' : x' \in D \wedge \delta(x, x') \leq \varepsilon\}$$

$\varepsilon - neighborhood$

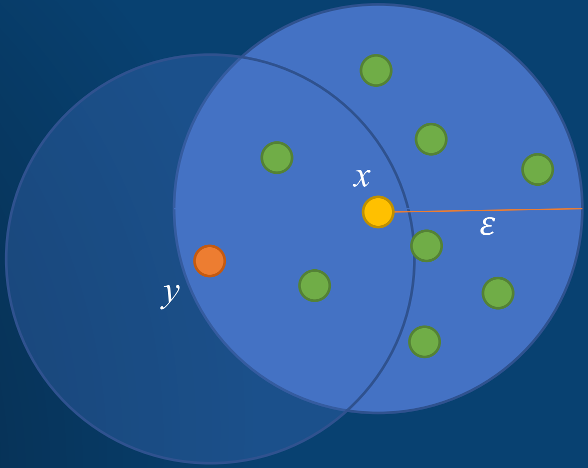
$$x \in Core \leftrightarrow |N_\varepsilon(x)| \geq minpts$$



$$z \in Noise$$
$$\leftrightarrow$$
$$\nexists x \in Core{:}\, z \in N_\varepsilon(x) \ \wedge \ |N_\varepsilon(z)| < minpts$$

$$N_\varepsilon(x) \leftarrow \{x' : x' \in D \wedge \delta(x, x') \leq \varepsilon\}$$

Directly Density Reachable

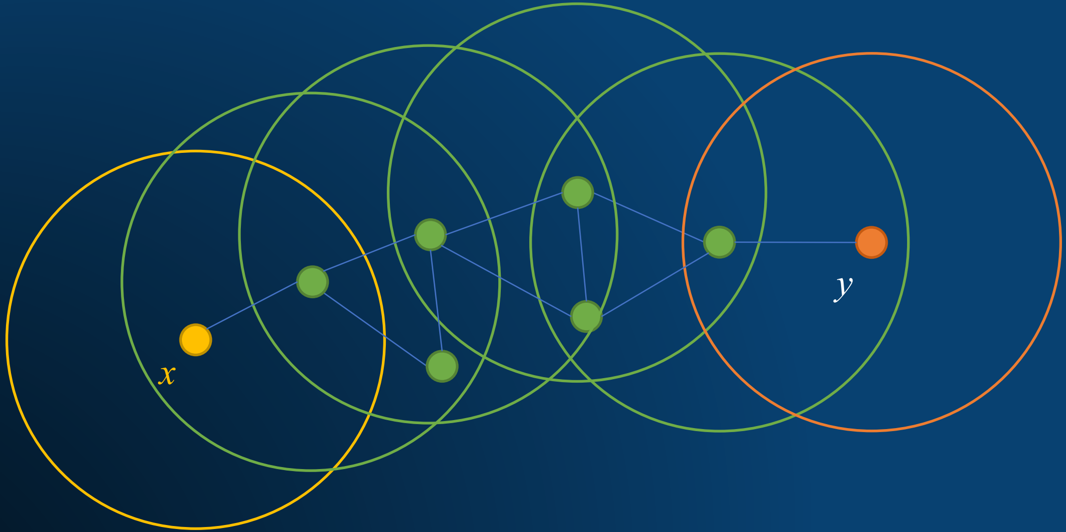


y is directly density reachable by x

$$\longleftrightarrow$$

$$y \in N_{\varepsilon}(x) \ \wedge \ x \in Core$$

Density Reachable

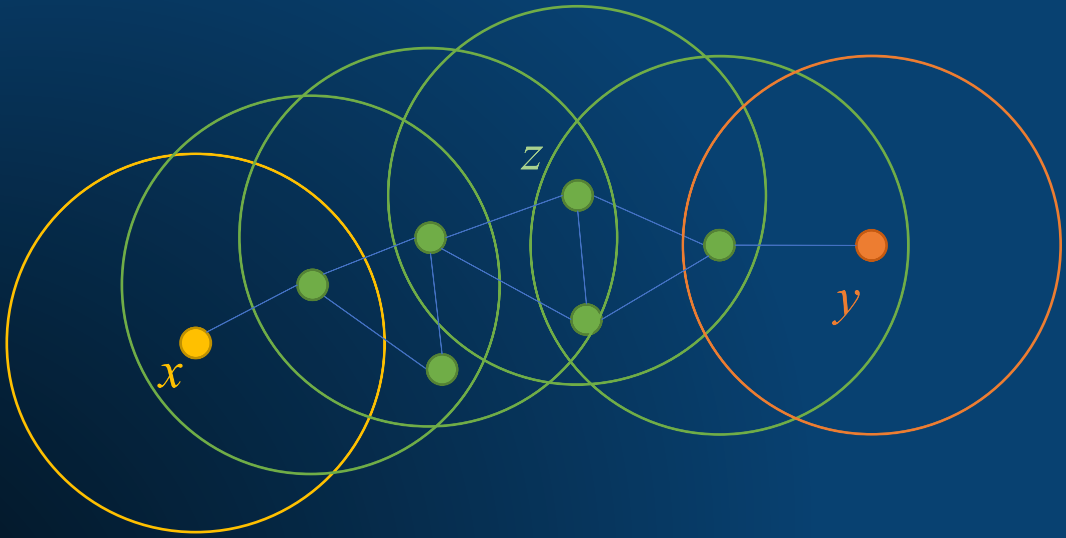


$$x \text{ is density reachable by } y$$

$$\longleftrightarrow$$

$$\exists x_0 \ldots x_m : x = x_0 \ \wedge \ y = x_m \wedge x_i \text{ is density reachable by } x_{i-1} \ \forall 0 \leq i \leq m$$

Density Connected

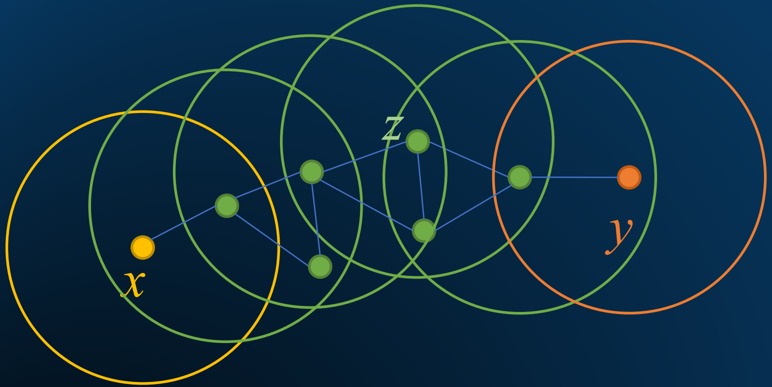


x *is density connected to* y

$$\longleftrightarrow$$

$$\exists z \in Core \wedge x \text{ is density reachable by } z \wedge y \text{ is density reachable by } z$$

DENSITY-BASED CLUSTER



A *density-based cluster* is a maximal set of density connected points.

DBSCAN Algorithm (Dataset *D*, float ε, int *minpts*)

$Core \leftarrow \emptyset$

for each $x_i \in D$ **do**

$\quad N_\varepsilon(x_i) \leftarrow \{x' : x' \in D \wedge \delta(x_i, x') \leq \varepsilon\}$

$\quad id(\boldsymbol{x_i}) \leftarrow \emptyset$

$\quad$ **if** $|N_\varepsilon(x)| \geq minpts$ **then** $Core \leftarrow Core \cup \{x_i\}$

$k \leftarrow 0$

for each $x_i \in Core \wedge id(\boldsymbol{x_i}) = \emptyset$ **do**

$\quad k \leftarrow k + 1$

$\quad id(\boldsymbol{x_i}) \leftarrow k$

$\quad DensityConnected(\boldsymbol{x_i}, k)$

$$\mathbb{C} \leftarrow \bigcup_{i=1}^{k} \{x : x \in D \wedge id(x) = i\}$$

$Noise \leftarrow \{x : x \in D \wedge id(x) = \emptyset\}$

$Border \leftarrow \boldsymbol{D} \backslash \{Core \cup Noise\}$

return $\mathbb{C}, Core, Border, Noise$

DensityConnected(*x*, *k*)

for each $y \in N_\varepsilon(x)$ **do**

$\quad id(\boldsymbol{y}) \leftarrow \boldsymbol{k}$

$\quad$ **if** $|N_\varepsilon(x)| y \in Core$ **then**

$\quad\quad DensityConnected(y, k)$

Thank you!

claudia.antunes@tecnico.ulisboa.pt