NOVA IMS Information Management School

# Data Mining

**NOVA-IMS**

**10/11/2021**

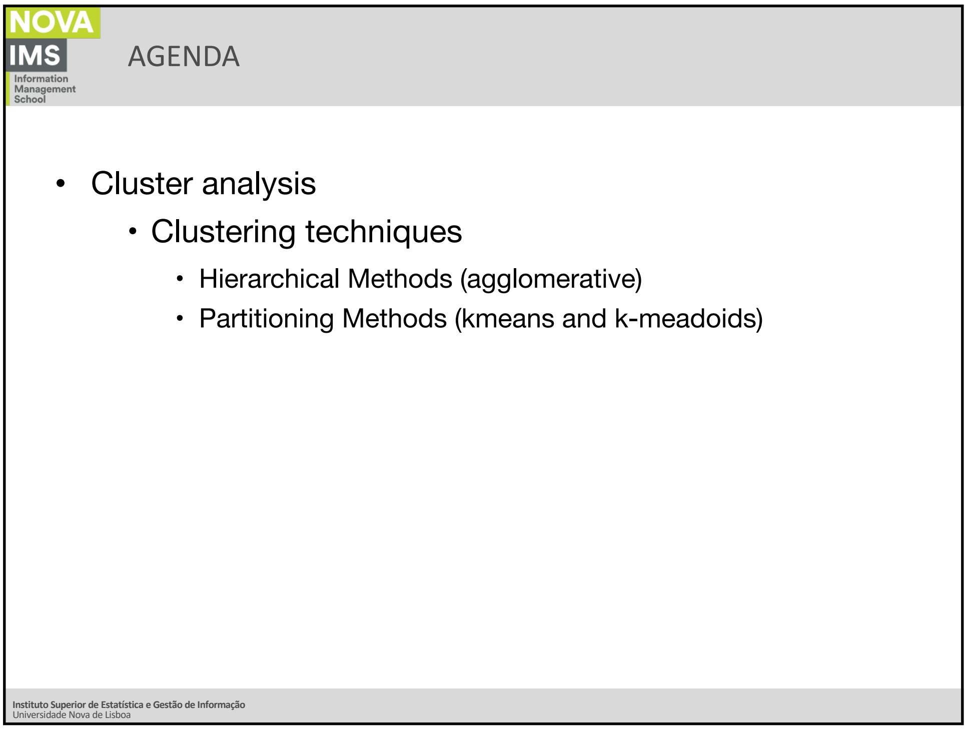
Fernando Lucas Bação

bacao@isegi.unl.pt

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

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

1



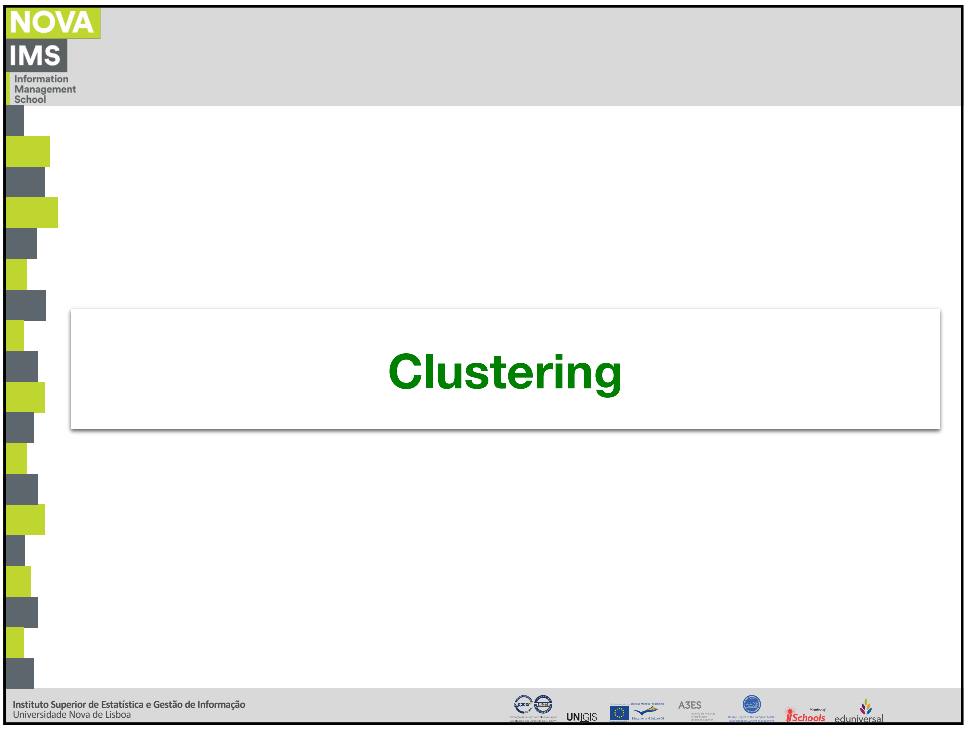
NOVA IMS Information Management School

## AGENDA

- Cluster analysis
  - Clustering techniques
    - Hierarchical Methods (agglomerative)
    - Partitioning Methods (kmeans and k-meadoids)

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

2

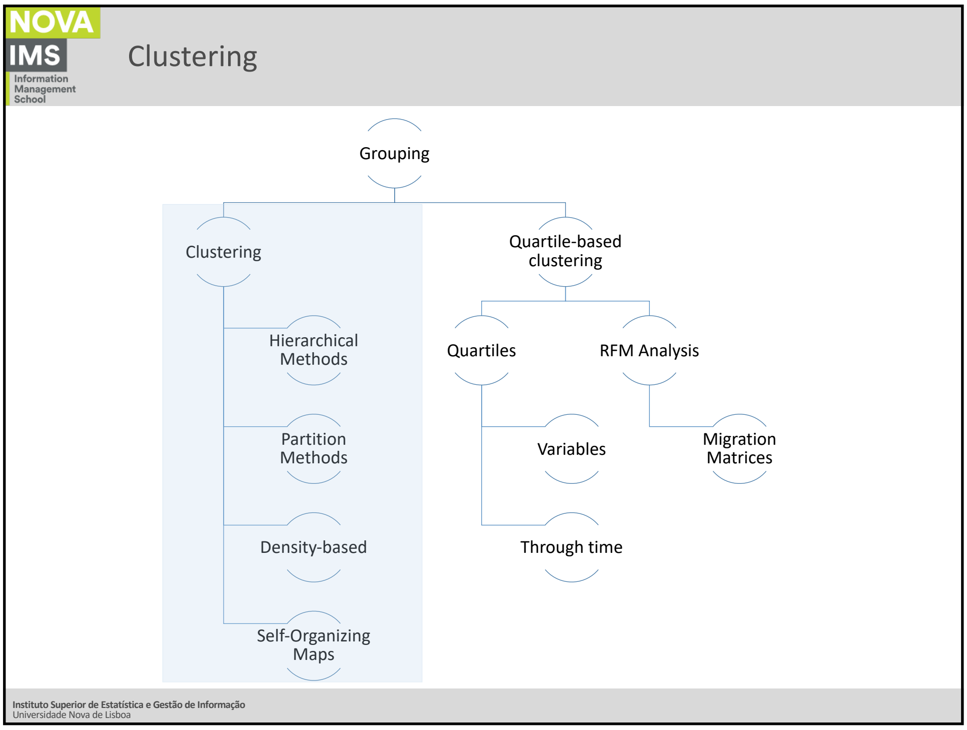


NOVA IMS Information Management School

# Clustering

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

3

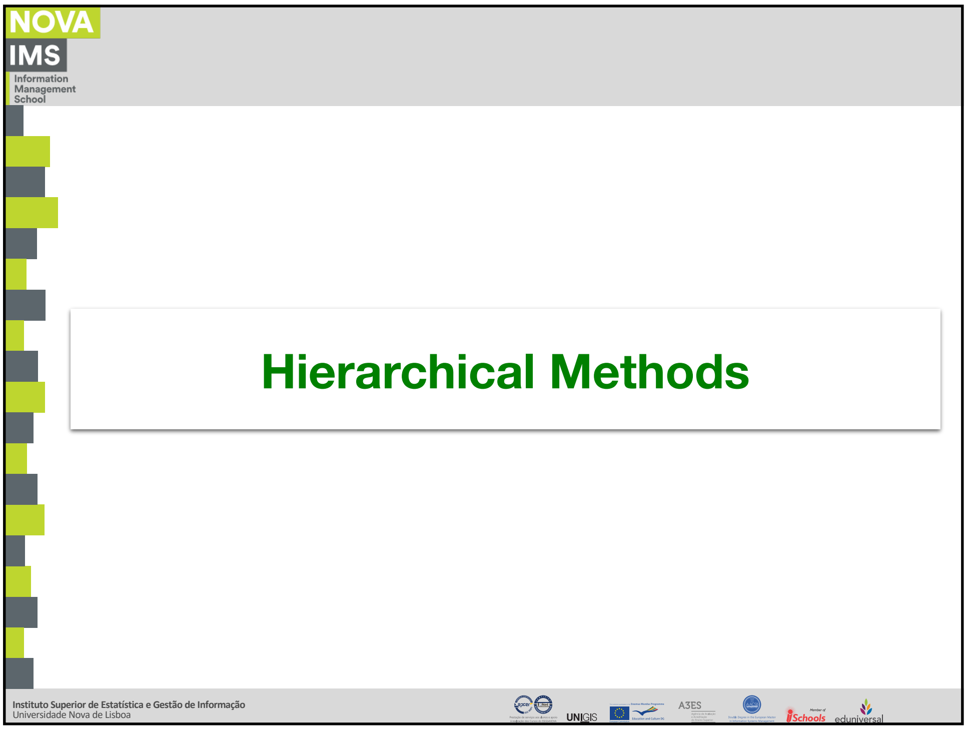


NOVA IMS Information Management School

## Clustering

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

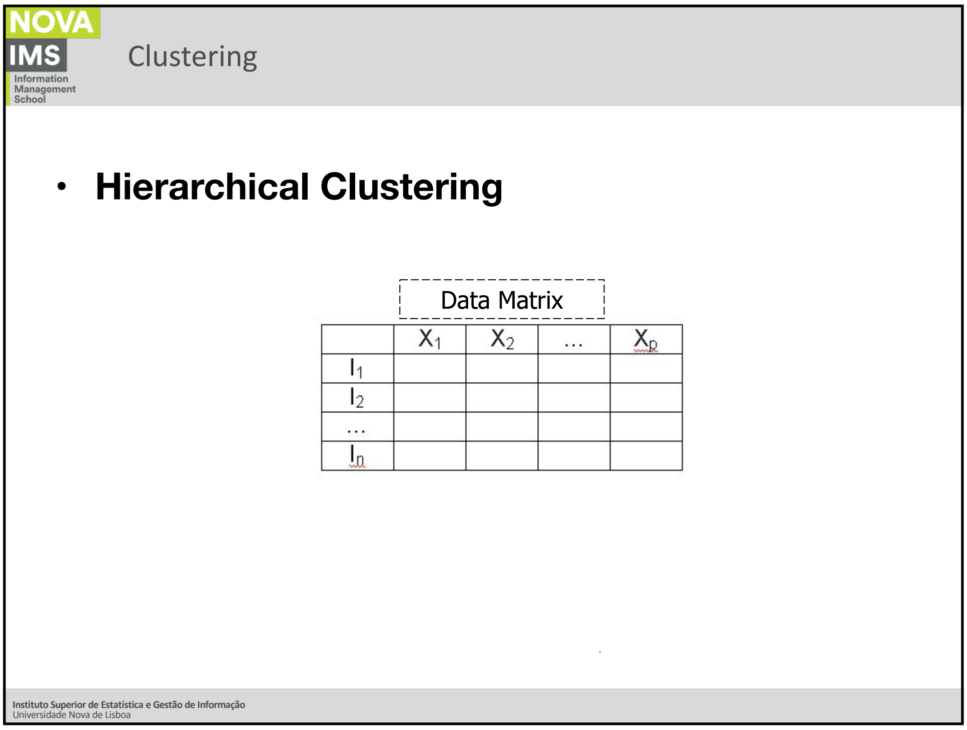
4



NOVA IMS Information Management School

# Hierarchical Methods

5



NOVA IMS Information Management School

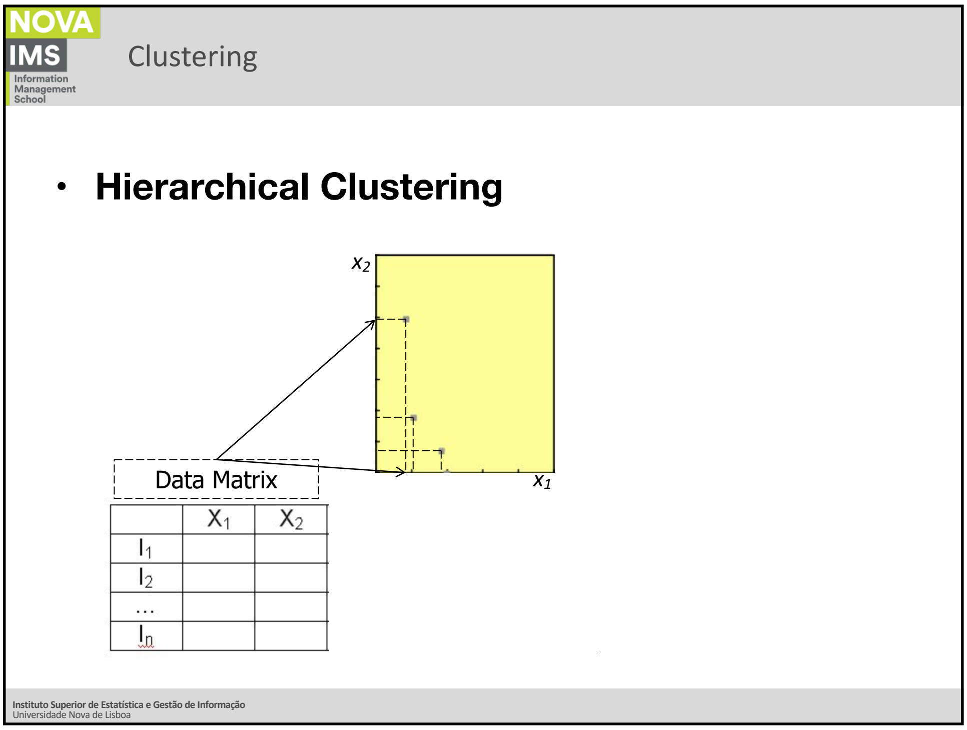
## Clustering

- **Hierarchical Clustering**

Data Matrix

| | $X_1$ | $X_2$ | … | $X_p$ |
|---|---|---|---|---|
| $I_1$ | | | | |
| $I_2$ | | | | |
| … | | | | |
| $I_n$ | | | | |

6



NOVA IMS Information Management School

## Clustering

- **Hierarchical Clustering**

$x_2$

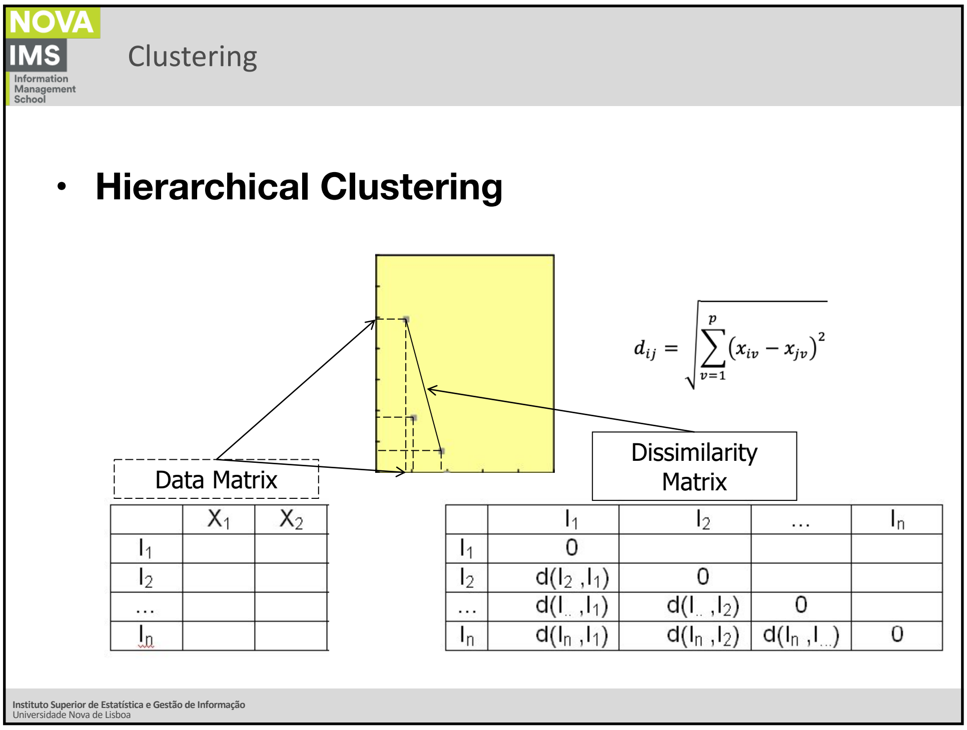
$x_1$

Data Matrix

| | $X_1$ | $X_2$ |
|---|---|---|
| $I_1$ | | |
| $I_2$ | | |
| … | | |
| $I_n$ | | |

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

7



NOVA IMS Information Management School

## Clustering

- **Hierarchical Clustering**

$$d_{ij} = \sqrt{\sum_{v=1}^{p}\left(x_{iv} - x_{jv}\right)^2}$$

Data Matrix

| | $X_1$ | $X_2$ |
|---|---|---|
| $I_1$ | | |
| $I_2$ | | |
| … | | |
| $I_n$ | | |

Dissimilarity Matrix

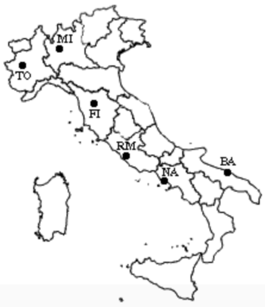
| | $I_1$ | $I_2$ | … | $I_n$ |
|---|---|---|---|---|
| $I_1$ | 0 | | | |
| $I_2$ | $d(I_2, I_1)$ | 0 | | |
| … | $d(I_{..}, I_1)$ | $d(I_{..}, I_2)$ | 0 | |
| $I_n$ | $d(I_n, I_1)$ | $d(I_n, I_2)$ | $d(I_n, I_{...})$ | 0 |

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

8

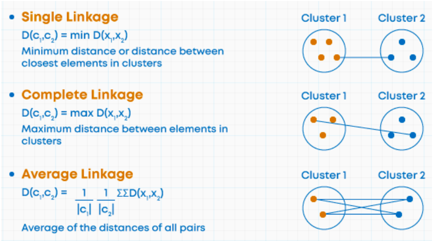
## Clustering

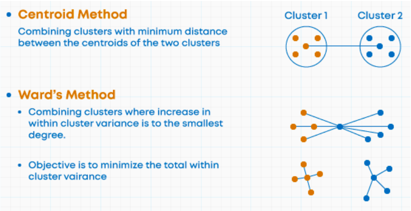
| | BA | FI | MI | NA | RM | TO |
|---|---|---|---|---|---|---|
| **BA** | 0 | 662 | 877 | 255 | 412 | 996 |
| **FI** | 662 | 0 | 295 | 468 | 268 | 400 |
| **MI** | 877 | 295 | 0 | 754 | 564 | 138 |
| **NA** | 255 | 468 | 754 | 0 | 219 | 869 |
| **RM** | 412 | 268 | 564 | 219 | 0 | 669 |
| **TO** | 996 | 400 | 138 | 869 | 669 | 0 |



## Clustering

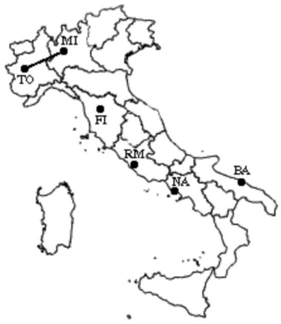
- **Hierarchical Clustering**
  - Linkage or Aggregation Rules


- **Single Linkage**
  $D(c_1,c_2) = \min D(x_1,x_2)$
  Minimum distance or distance between closest elements in clusters
- **Complete Linkage**
  $D(c_1,c_2) = \max D(x_1,x_2)$
  Maximum distance between elements in clusters
- **Average Linkage**
  $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\Sigma\Sigma D(x_1,x_2)$
  Average of the distances of all pairs


- **Centroid Method**
  Combining clusters with minimum distance between the centroids of the two clusters
- **Ward's Method**
  - Combining clusters where increase in within cluster variance is to the smallest degree.
  - Objective is to minimize the total within cluster vairance

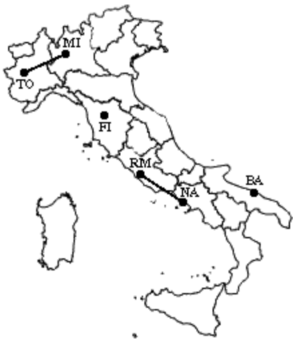
## Clustering

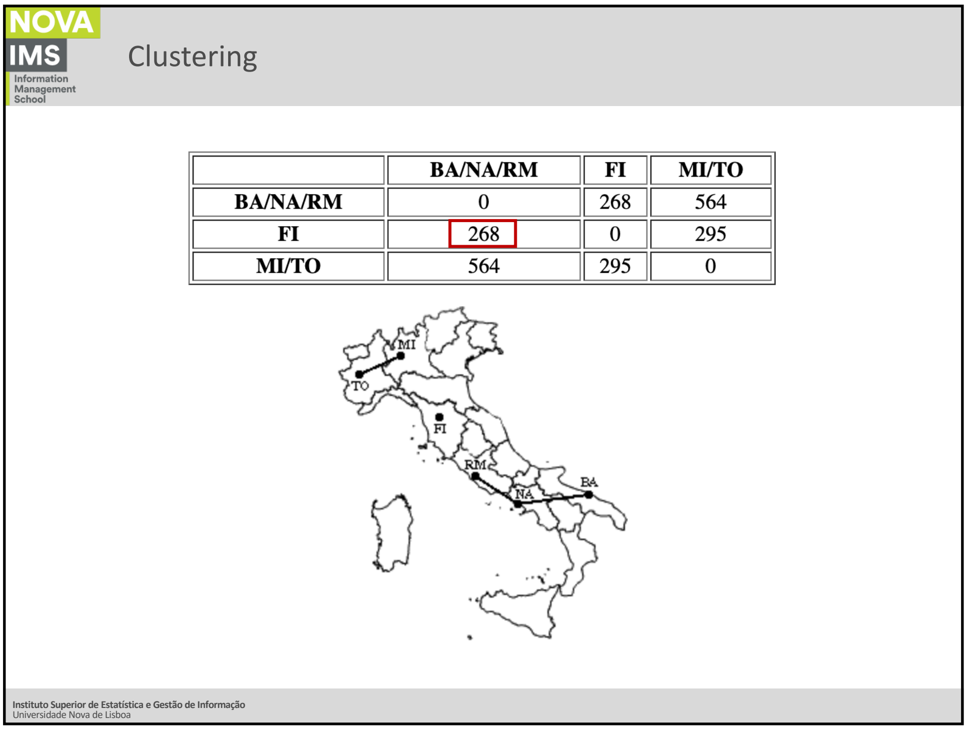
| | BA | FI | MI/TO | NA | RM |
|---|---|---|---|---|---|
| **BA** | 0 | 662 | 877 | 255 | 412 |
| **FI** | 662 | 0 | 295 | 468 | 268 |
| **MI/TO** | 877 | 295 | 0 | 754 | 564 |
| **NA** | 255 | 468 | 754 | 0 | 219 |
| **RM** | 412 | 268 | 564 | 219 | 0 |



## Clustering

| | BA | FI | MI/TO | NA/RM |
|---|---|---|---|---|
| **BA** | 0 | 662 | 877 | 255 |
| **FI** | 662 | 0 | 295 | 268 |
| **MI/TO** | 877 | 295 | 0 | 564 |
| **NA/RM** | 255 | 268 | 564 | 0 |

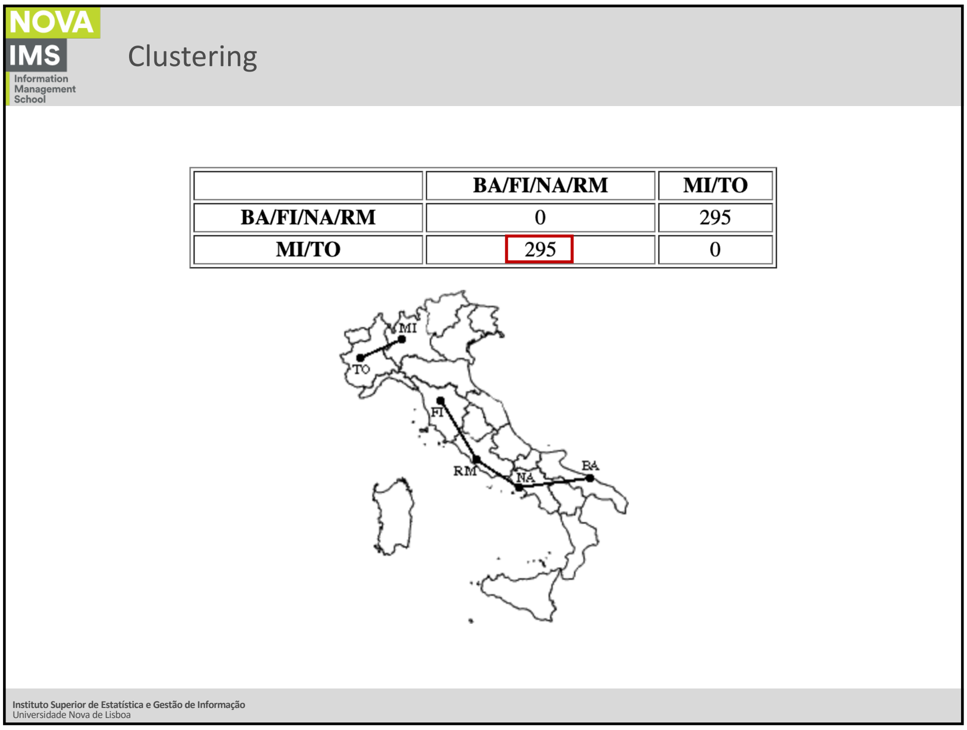




## Clustering

| | BA/NA/RM | FI | MI/TO |
|---|---|---|---|
| **BA/NA/RM** | 0 | 268 | 564 |
| **FI** | 268 | 0 | 295 |
| **MI/TO** | 564 | 295 | 0 |

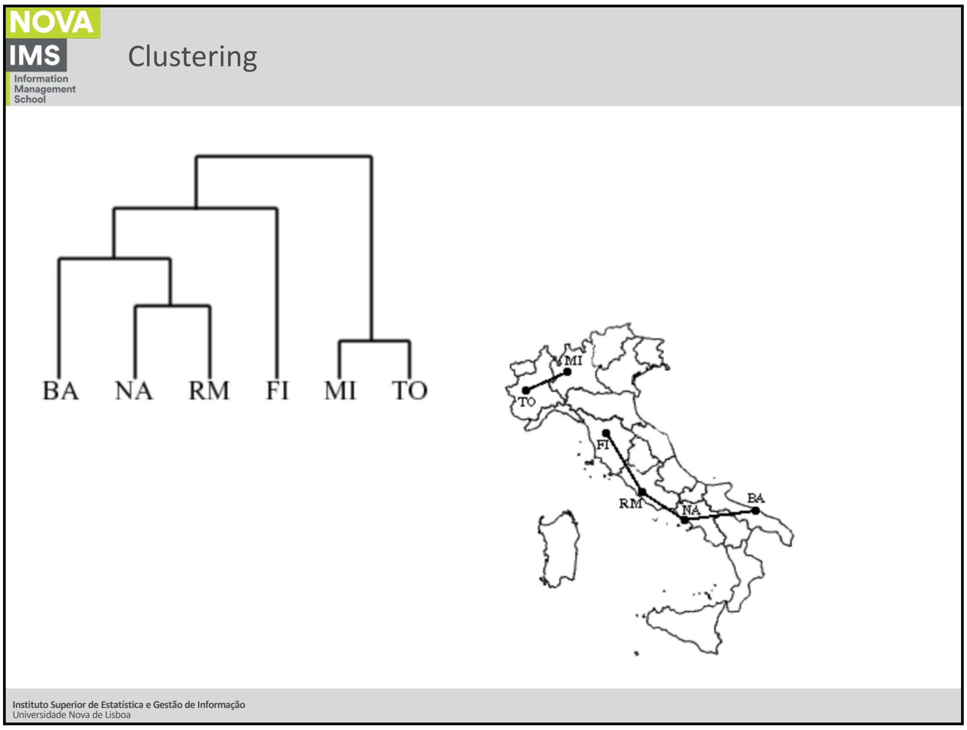
13



## Clustering

| | BA/FI/NA/RM | MI/TO |
|---|---|---|
| **BA/FI/NA/RM** | 0 | 295 |
| **MI/TO** | 295 | 0 |

14

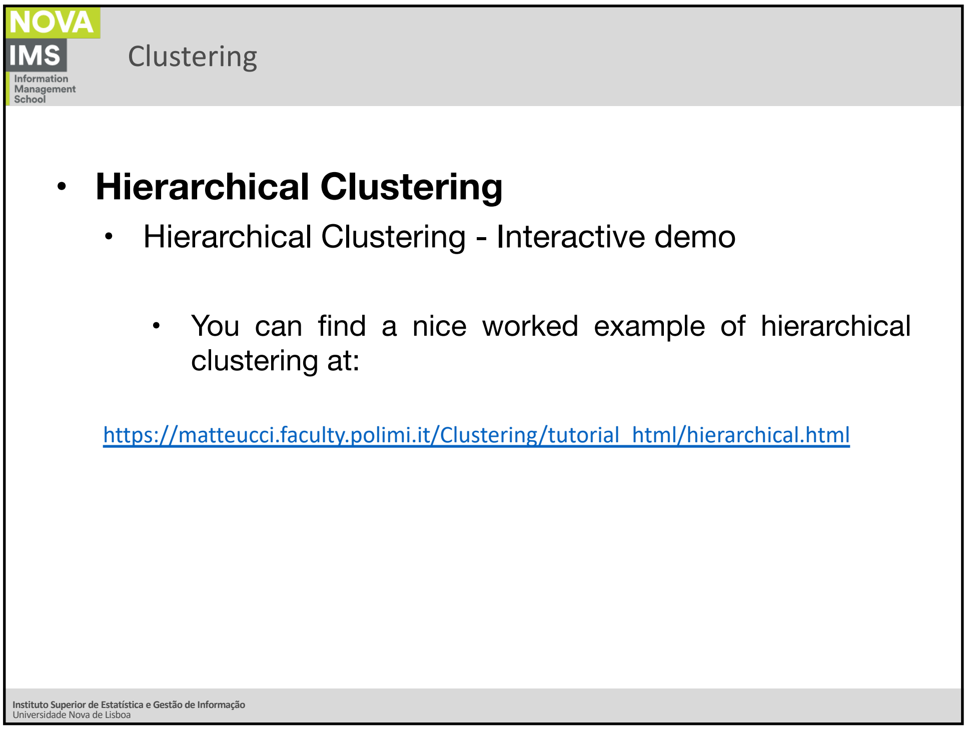


NOVA IMS Information Management School

Clustering

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

15



NOVA IMS Information Management School

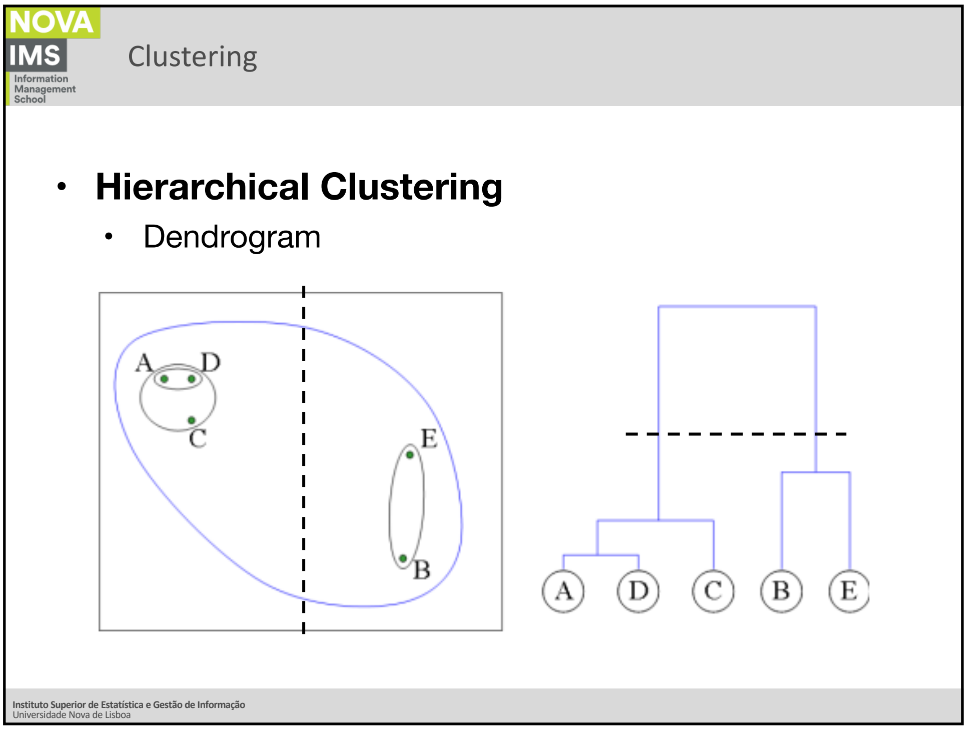
Clustering

- **Hierarchical Clustering**
  - Hierarchical Clustering - Interactive demo
    - You can find a nice worked example of hierarchical clustering at:

https://matteucci.faculty.polimi.it/Clustering/tutorial_html/hierarchical.html

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

16



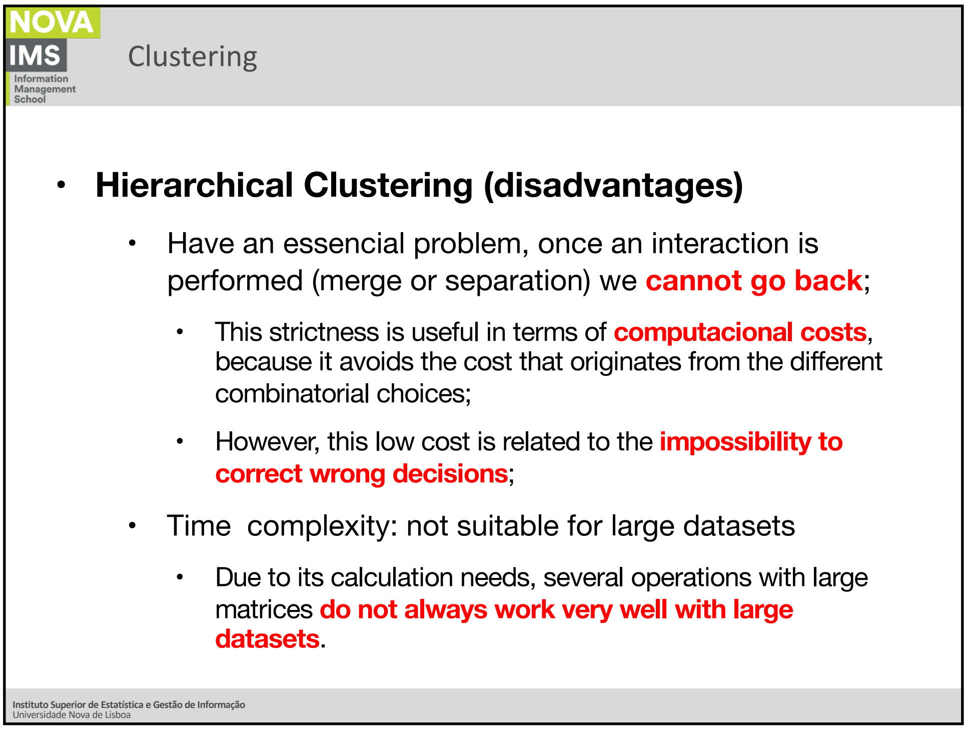
NOVA IMS Information Management School

Clustering

- **Hierarchical Clustering**
  - Dendrogram

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

17



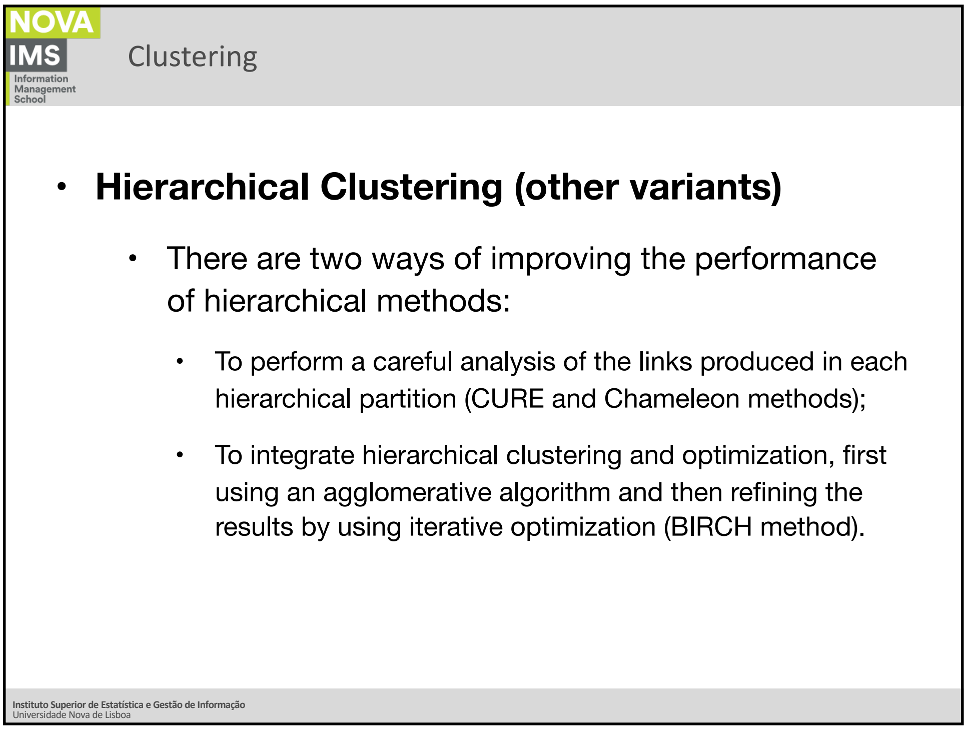
NOVA IMS Information Management School

Clustering

- **Hierarchical Clustering (disadvantages)**
  - Have an essencial problem, once an interaction is performed (merge or separation) we **cannot go back**;
    - This strictness is useful in terms of **computacional costs**, because it avoids the cost that originates from the different combinatorial choices;
    - However, this low cost is related to the **impossibility to correct wrong decisions**;
  - Time complexity: not suitable for large datasets
    - Due to its calculation needs, several operations with large matrices **do not always work very well with large datasets**.

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

18



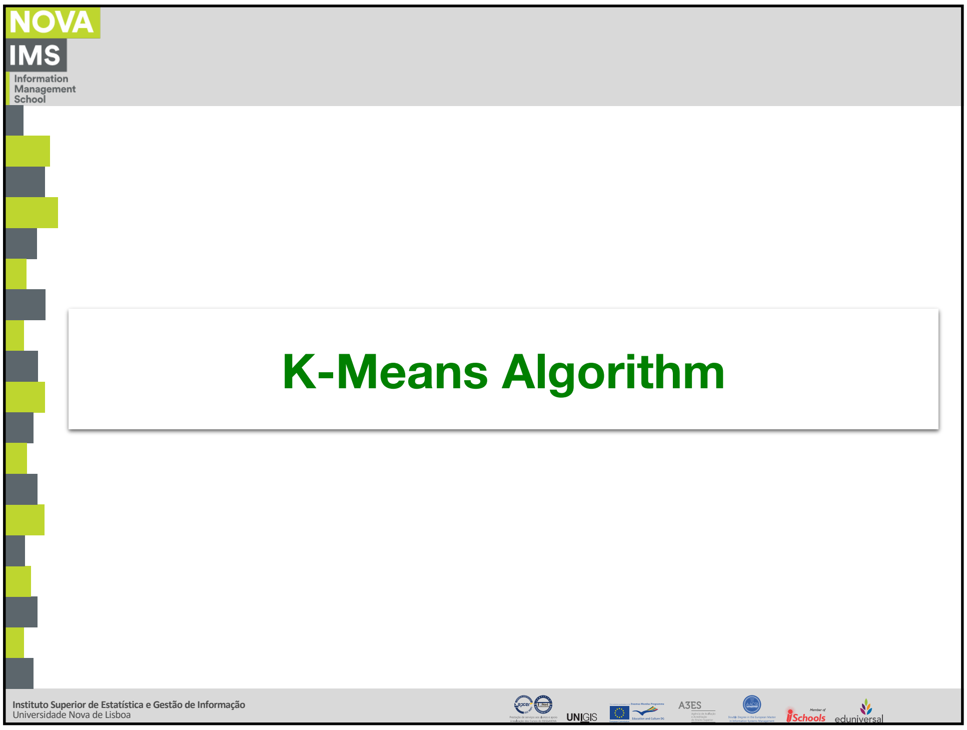
NOVA IMS Information Management School

Clustering

- **Hierarchical Clustering (other variants)**
  - There are two ways of improving the performance of hierarchical methods:
    - To perform a careful analysis of the links produced in each hierarchical partition (CURE and Chameleon methods);
    - To integrate hierarchical clustering and optimization, first using an agglomerative algorithm and then refining the results by using iterative optimization (BIRCH method).

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

19

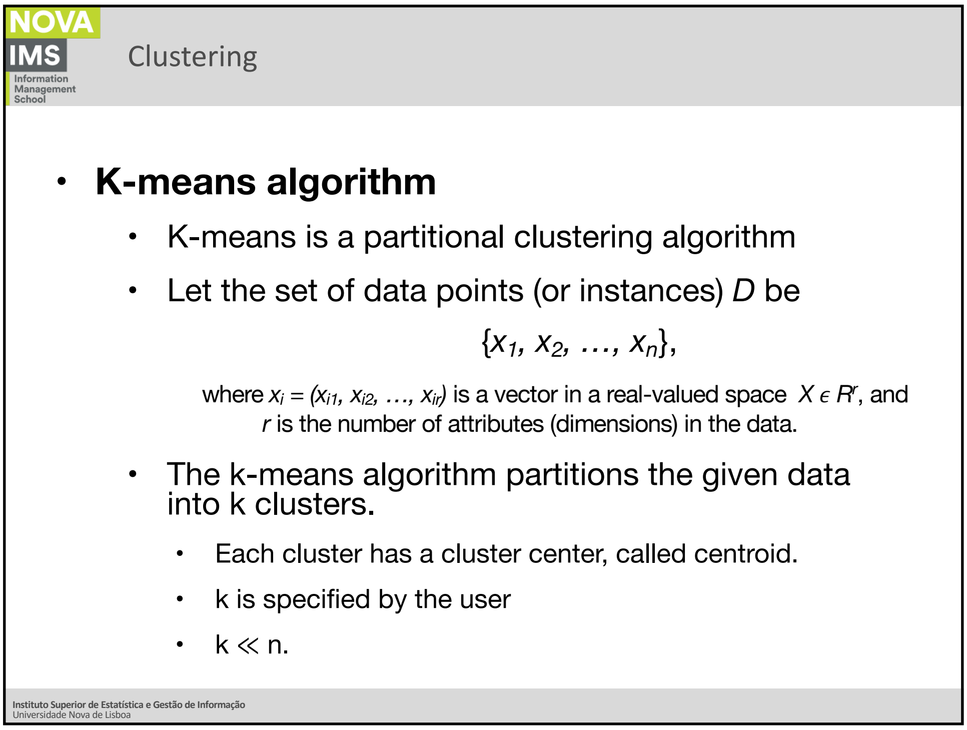


NOVA IMS Information Management School

## K-Means Algorithm

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

20



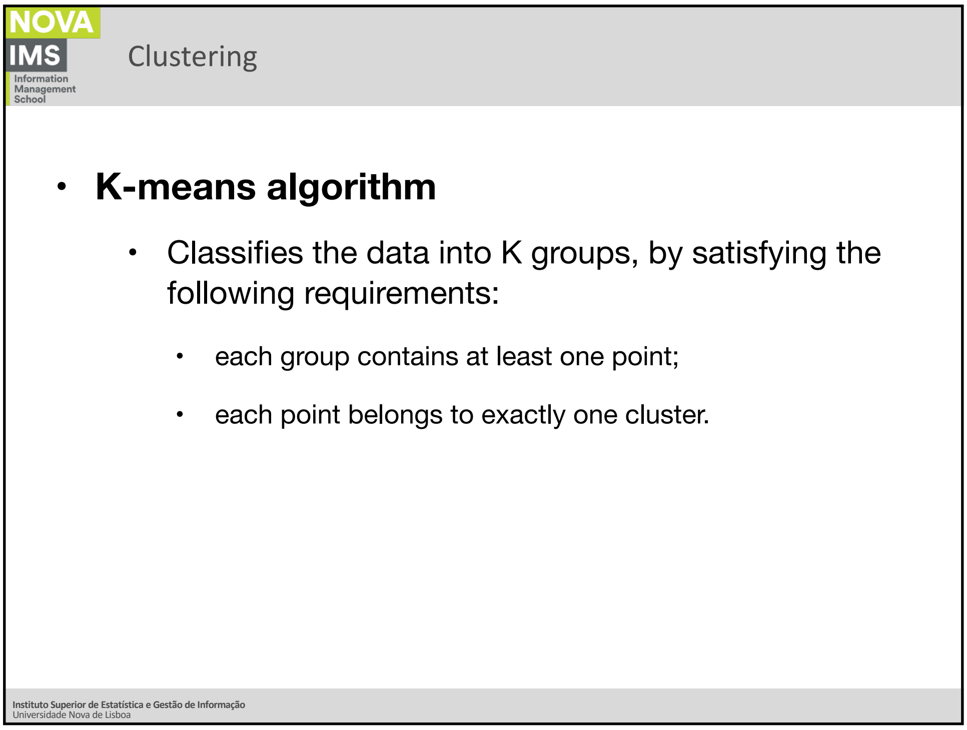
## Clustering

- **K-means algorithm**
  - K-means is a partitional clustering algorithm
  - Let the set of data points (or instances) *D* be

$$\{x_1, x_2, \ldots, x_n\},$$

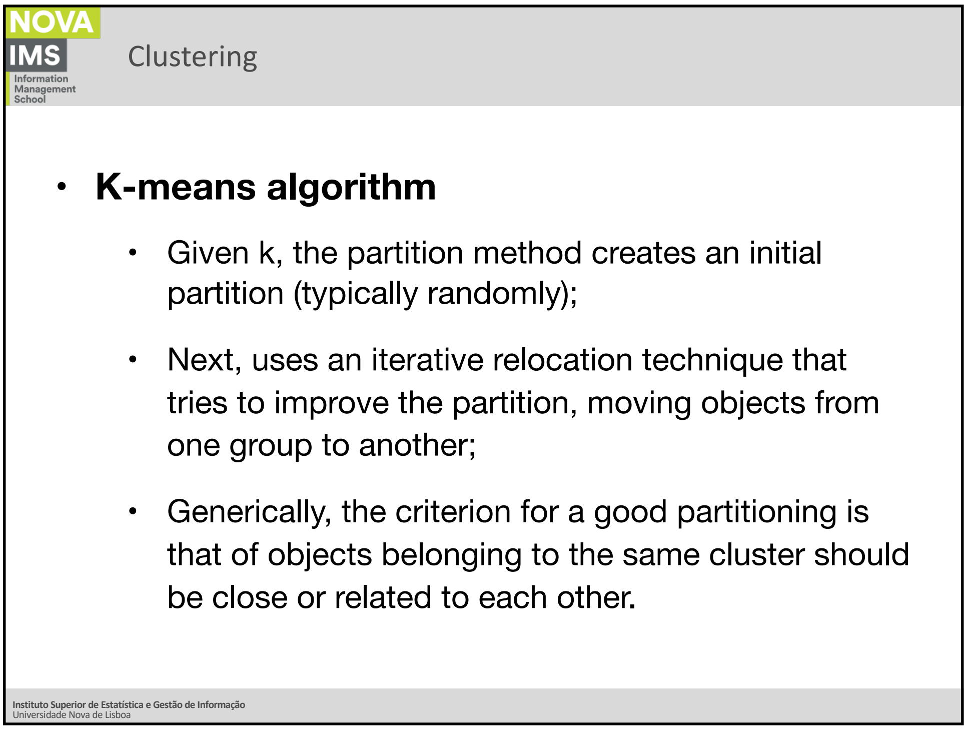
where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $X \in R^r$, and $r$ is the number of attributes (dimensions) in the data.

  - The k-means algorithm partitions the given data into k clusters.
    - Each cluster has a cluster center, called centroid.
    - k is specified by the user
    - k ≪ n.



## Clustering

- **K-means algorithm**
  - Classifies the data into K groups, by satisfying the following requirements:
    - each group contains at least one point;
    - each point belongs to exactly one cluster.

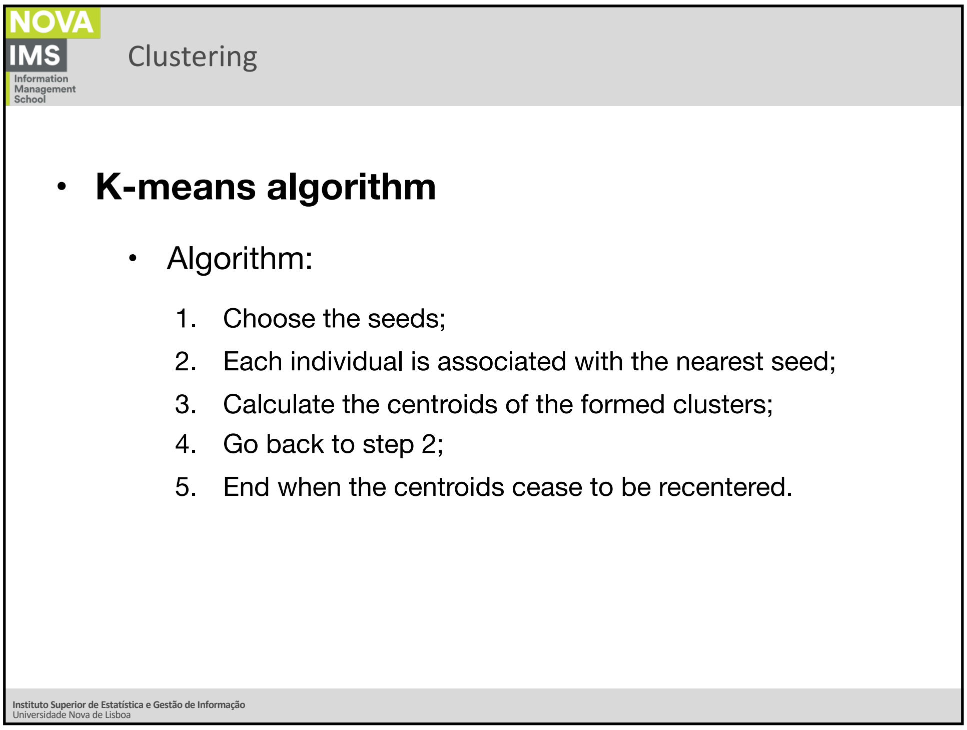


Clustering

- **K-means algorithm**
  - Given k, the partition method creates an initial partition (typically randomly);
  - Next, uses an iterative relocation technique that tries to improve the partition, moving objects from one group to another;
  - Generically, the criterion for a good partitioning is that of objects belonging to the same cluster should be close or related to each other.

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

23

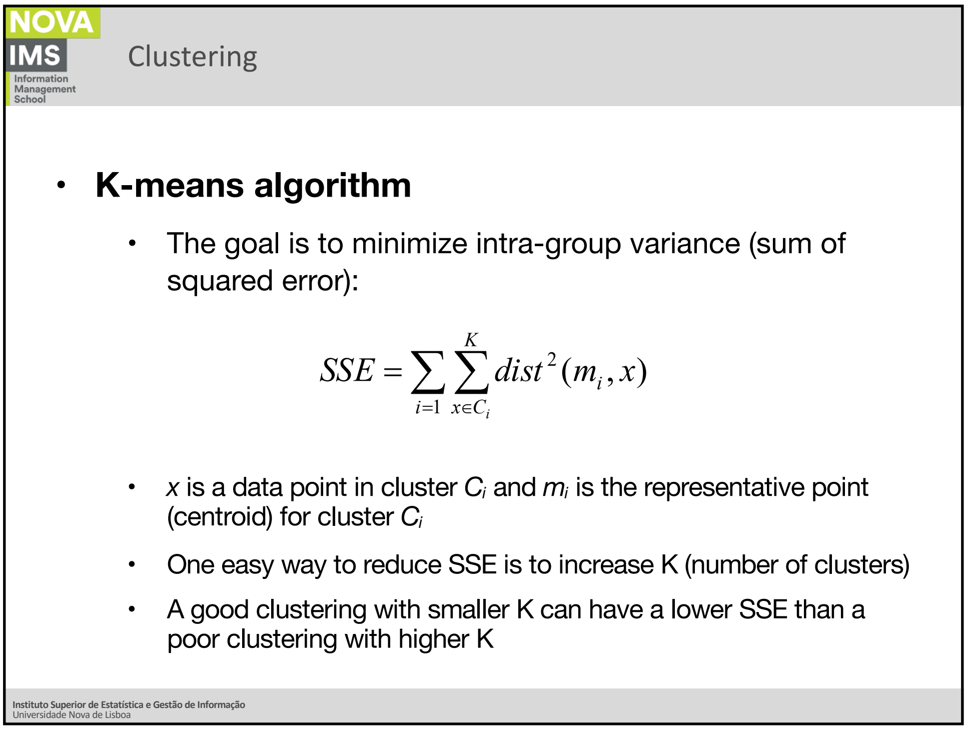


Clustering

- **K-means algorithm**
  - Algorithm:
    1. Choose the seeds;
    2. Each individual is associated with the nearest seed;
    3. Calculate the centroids of the formed clusters;
    4. Go back to step 2;
    5. End when the centroids cease to be recentered.

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

24



## Clustering

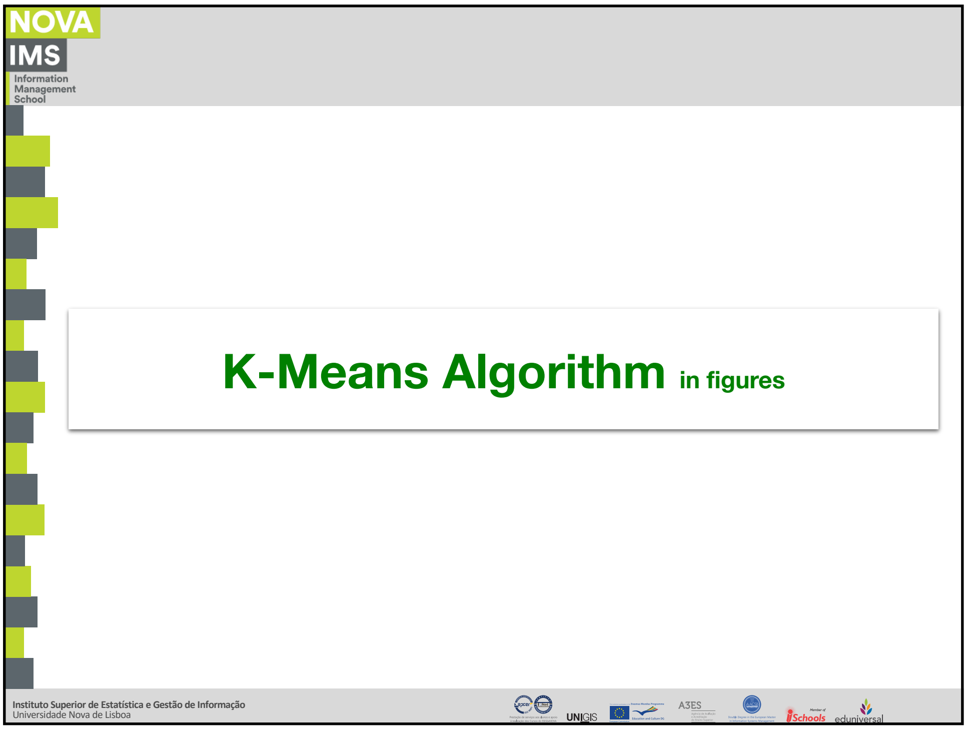
- **K-means algorithm**
  - The goal is to minimize intra-group variance (sum of squared error):

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - *x* is a data point in cluster $C_i$ and $m_i$ is the representative point (centroid) for cluster $C_i$
  - One easy way to reduce SSE is to increase K (number of clusters)
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

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

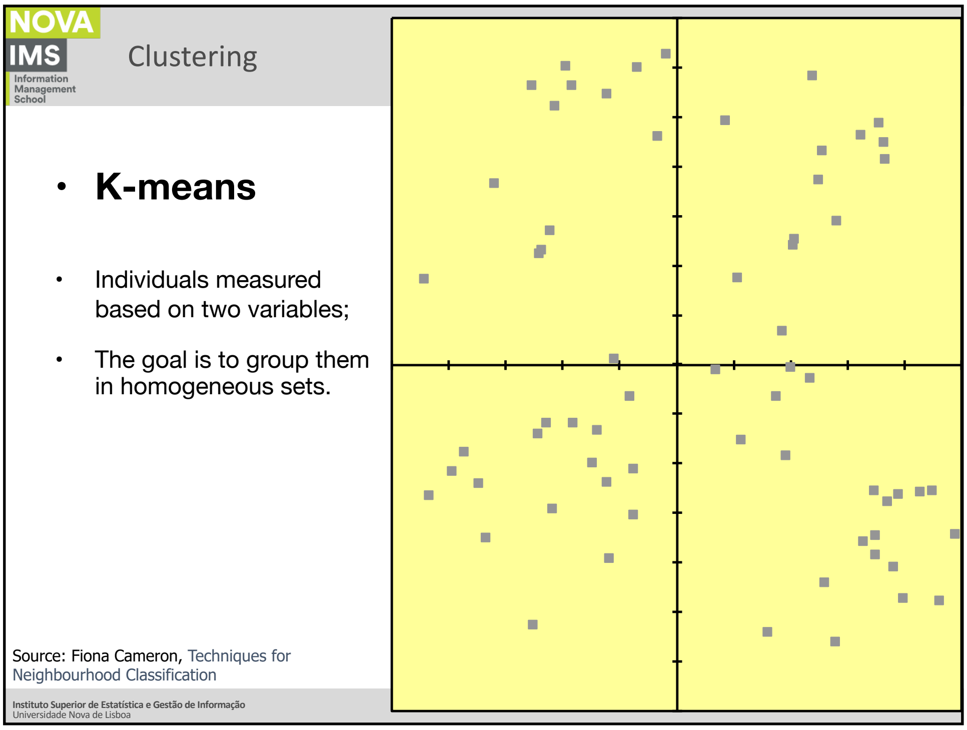
25



## K-Means Algorithm in figures

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

26



NOVA IMS Information Management School

## Clustering

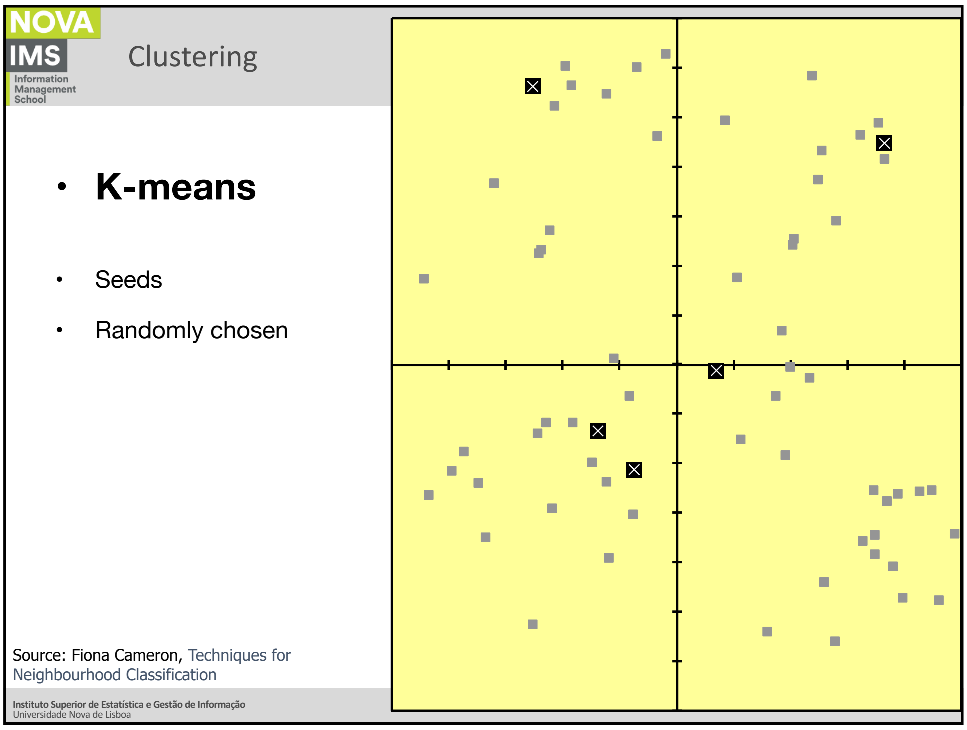
- **K-means**

- Individuals measured based on two variables;
- The goal is to group them in homogeneous sets.

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

27



NOVA IMS Information Management School

## Clustering

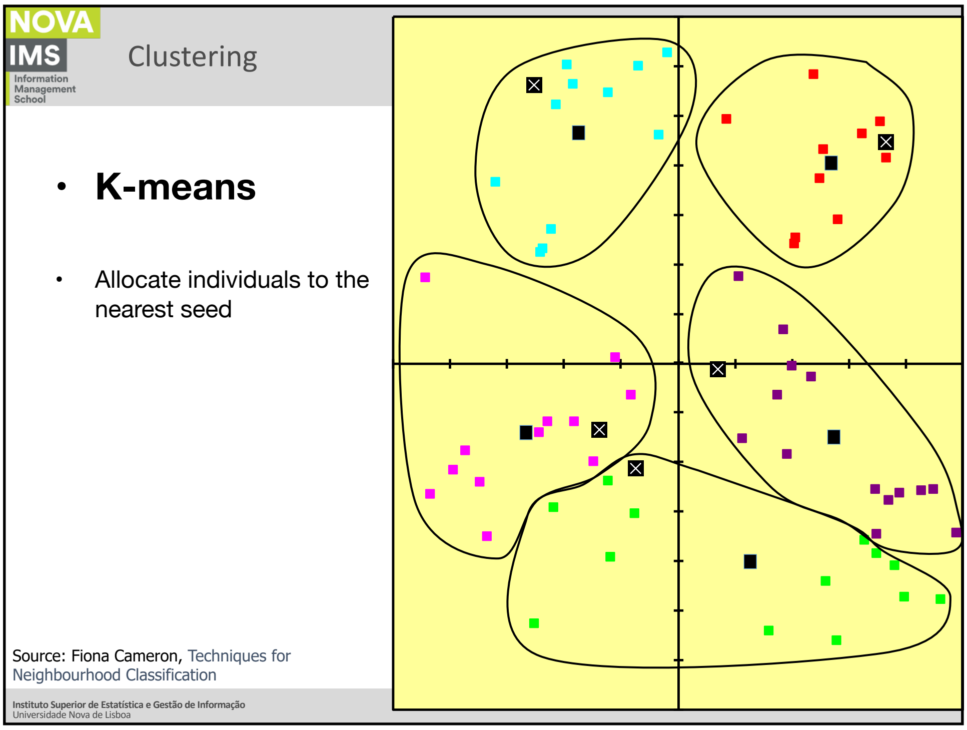
- **K-means**

- Seeds
- Randomly chosen

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

28



NOVA IMS Information Management School

## Clustering

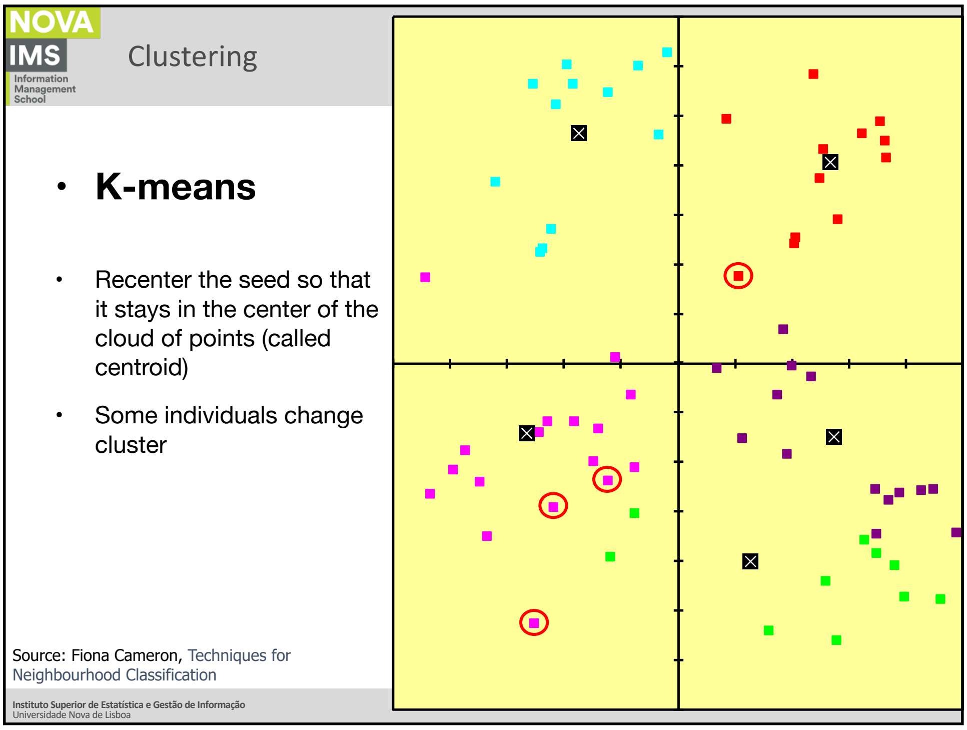
- **K-means**

- Allocate individuals to the nearest seed

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

29



NOVA IMS Information Management School

## Clustering

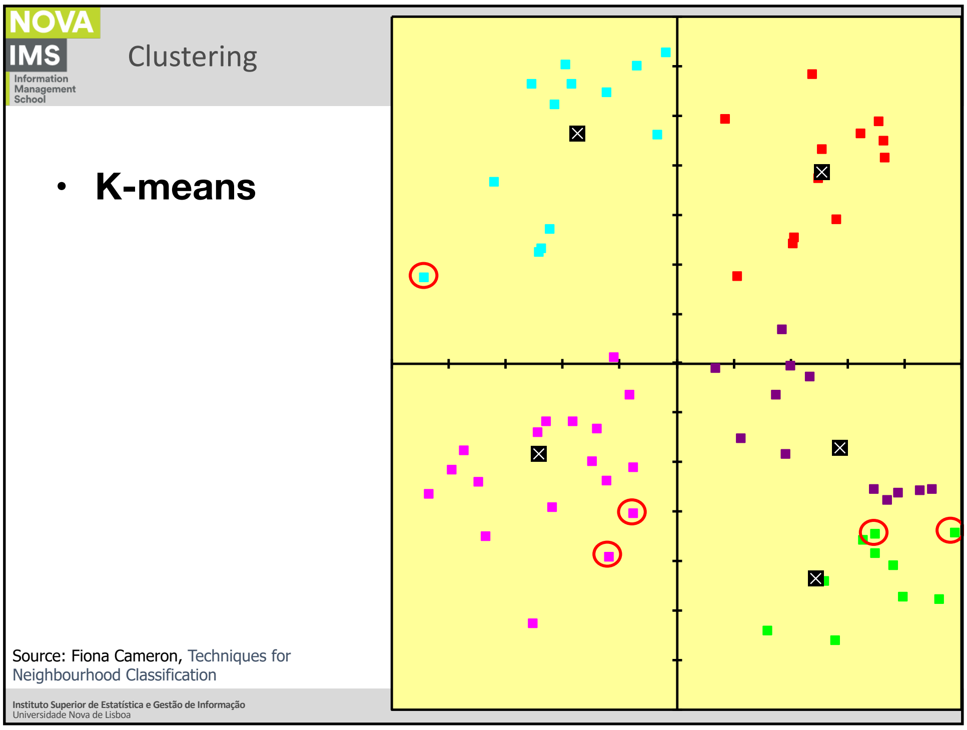
- **K-means**

- Recenter the seed so that it stays in the center of the cloud of points (called centroid)
- Some individuals change cluster

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

30



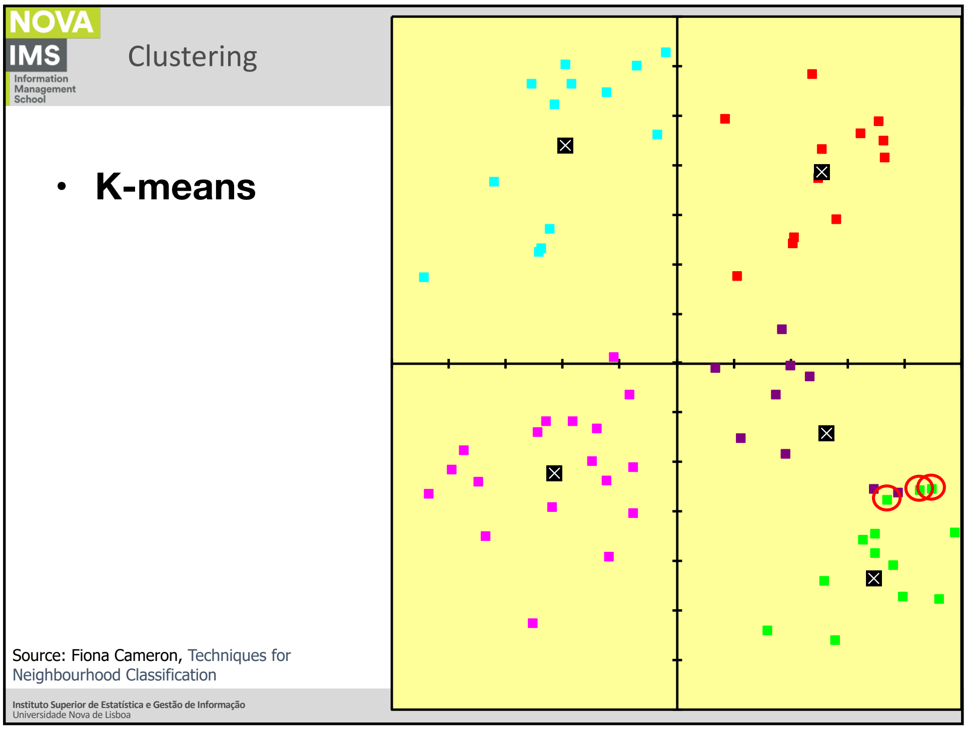
## Clustering

- **K-means**

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

31



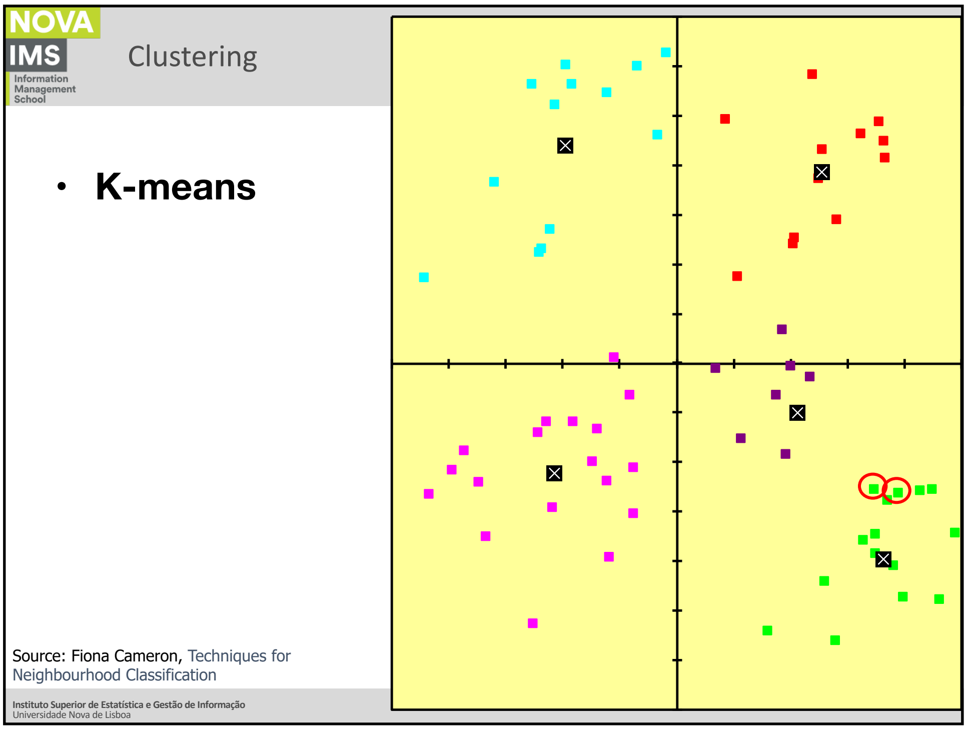
## Clustering

- **K-means**

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

32



NOVA IMS Information Management School

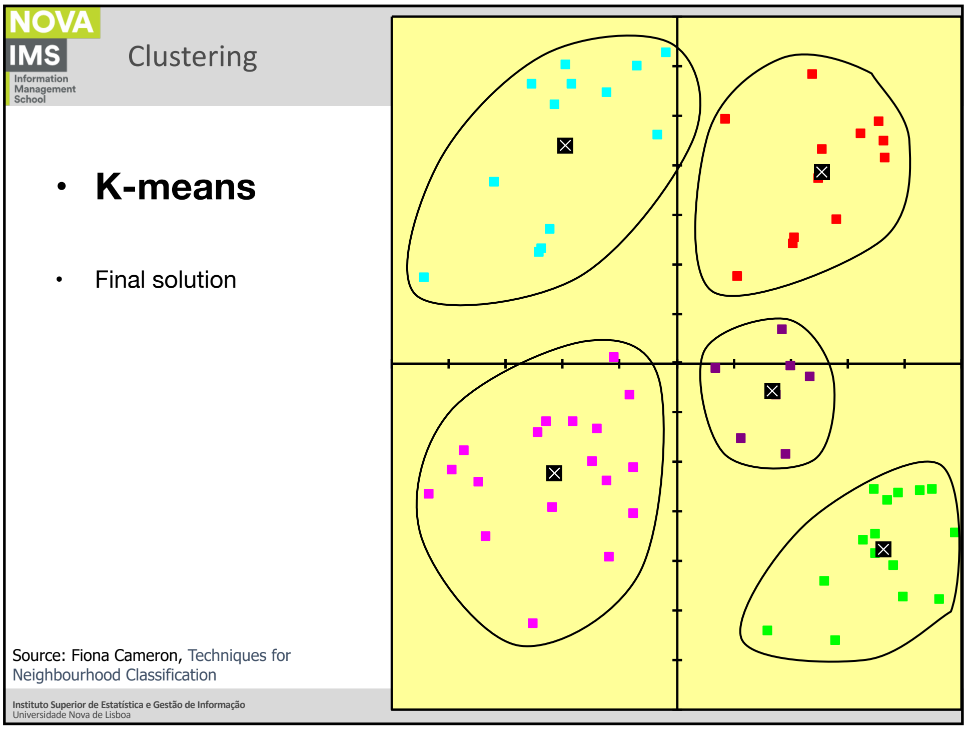
## Clustering

- **K-means**

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

33



NOVA IMS Information Management School

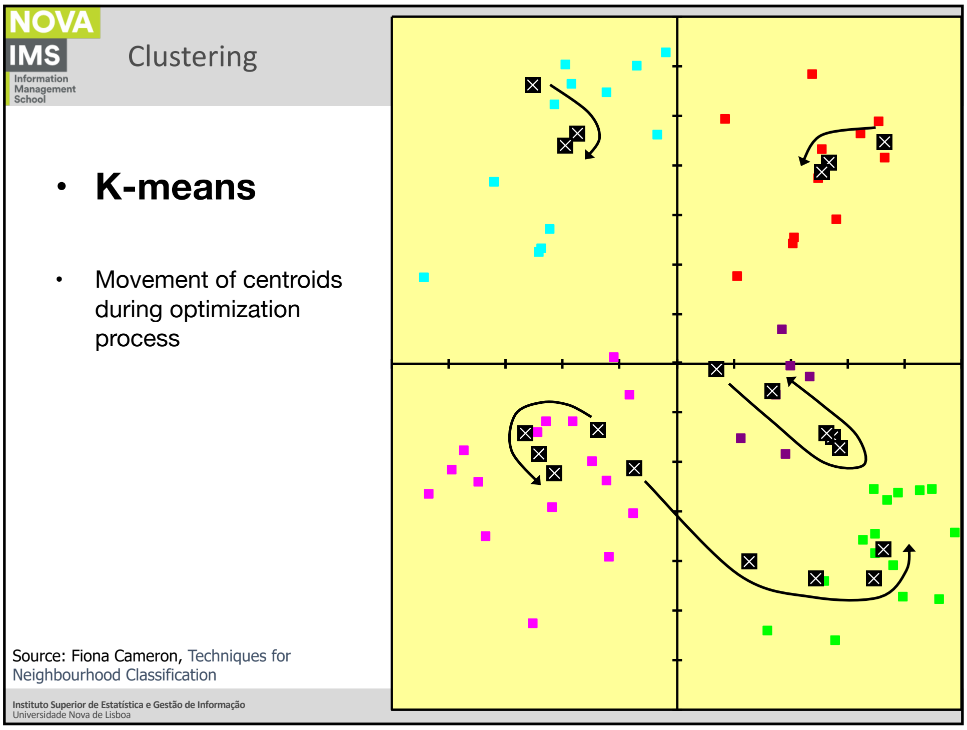
## Clustering

- **K-means**
- Final solution

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

34



NOVA IMS Information Management School

## Clustering

- **K-means**
  - Movement of centroids during optimization process

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

35

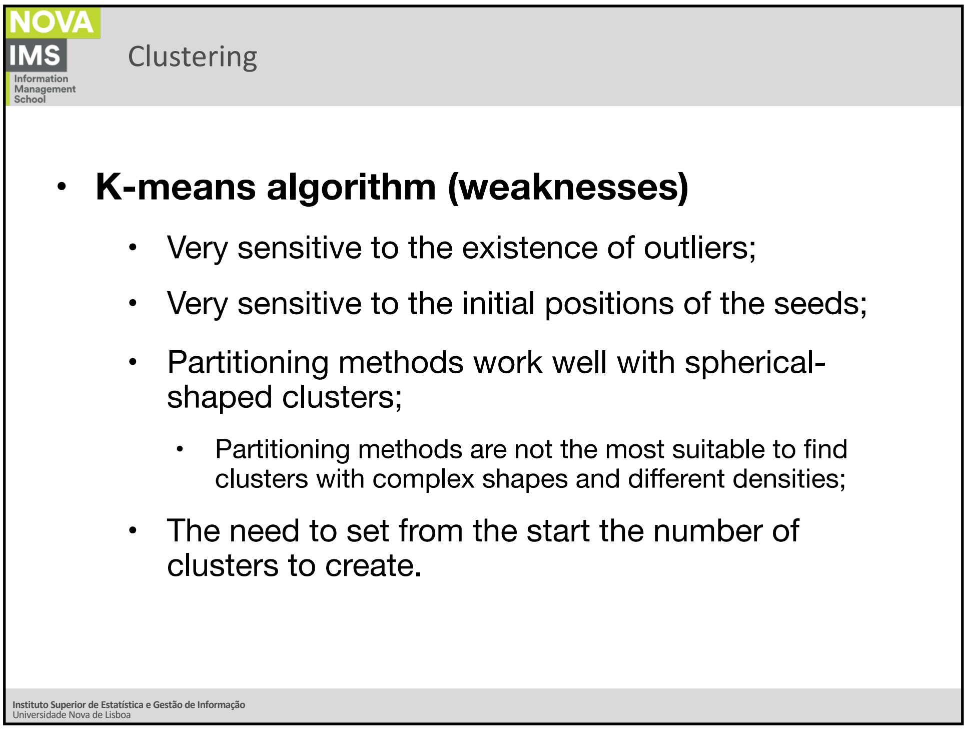
NOVA IMS Information Management School

## Clustering

- **K-means algorithm (strengths)**
  - Simple: easy to understand and to implement
  - Efficient: Time complexity O(tkn),
    - where n is the number of data points,
    - k is the number of clusters, and
    - t is the number of iterations.
  - Since both k and t are small, k-means is considered a linear algorithm.
  - K-means is the most popular clustering algorithm.
  - Note that: it terminates at a local optimum if SSE is used. The global optimum is hard to find due to complexity.

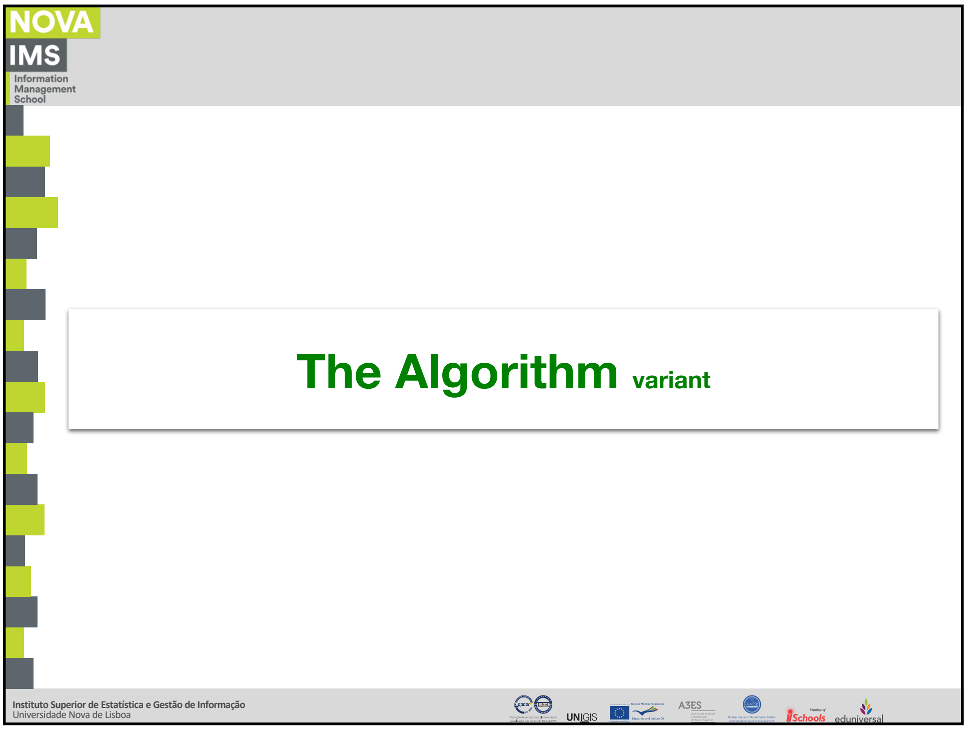
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

36



## Clustering

- **K-means algorithm (weaknesses)**
  - Very sensitive to the existence of outliers;
  - Very sensitive to the initial positions of the seeds;
  - Partitioning methods work well with spherical-shaped clusters;
    - Partitioning methods are not the most suitable to find clusters with complex shapes and different densities;
  - The need to set from the start the number of clusters to create.



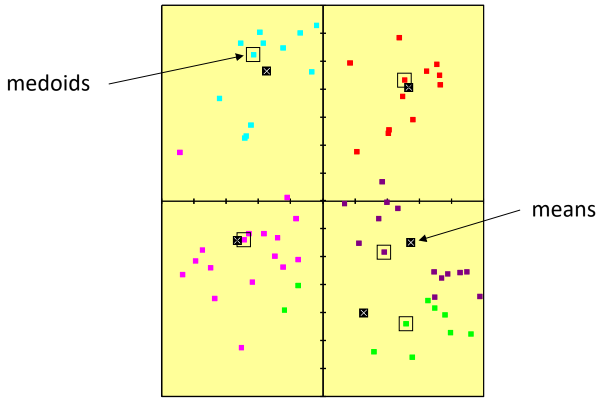
## The Algorithm variant

## Clustering

- **K-means and k-medoids algorithms**
  - Most algorithms adopt one of two very popular heuristics:
    - k-means algorithm, where each cluster is represented by the average of the values of the points in a cluster;
    - k-medoids algorithm, where each cluster is represented by one of the points located near the center of the cluster.

## Clustering

- **K-means and k-medoids algorithms**

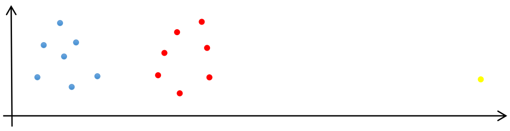


Source: Fiona Cameron, Techniques for Neighbourhood Classification

NOVA IMS Information Management School

Clustering

- **K-means and k-medoids algorithms**



Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

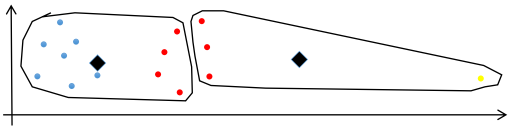
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

41

NOVA IMS Information Management School

Clustering

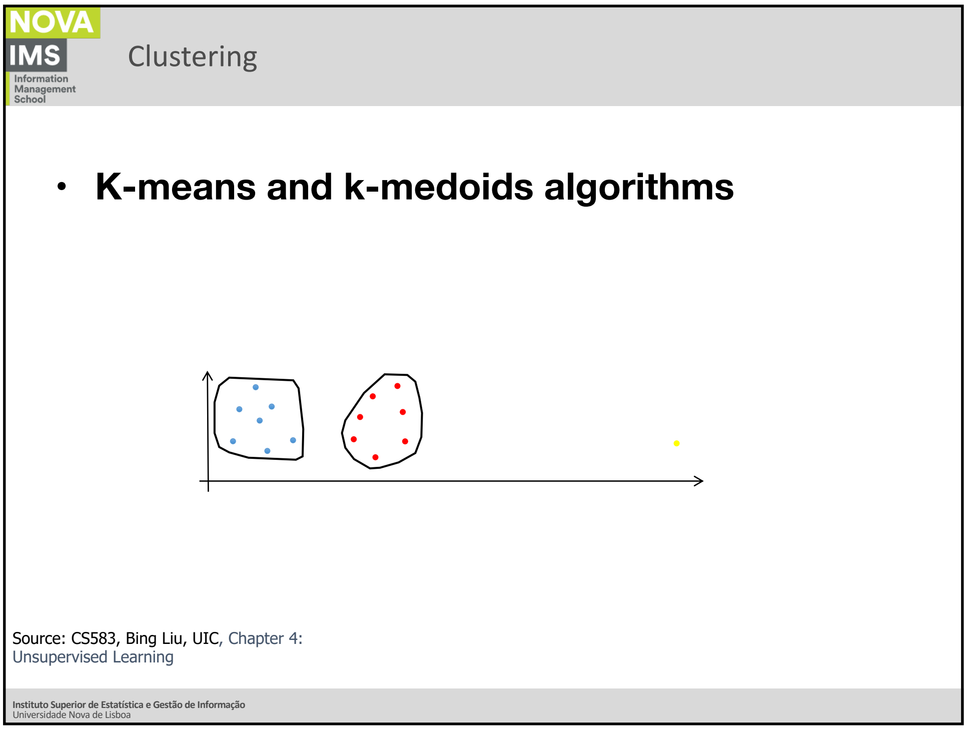
- **K-means and k-medoids algorithms**



Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

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

42



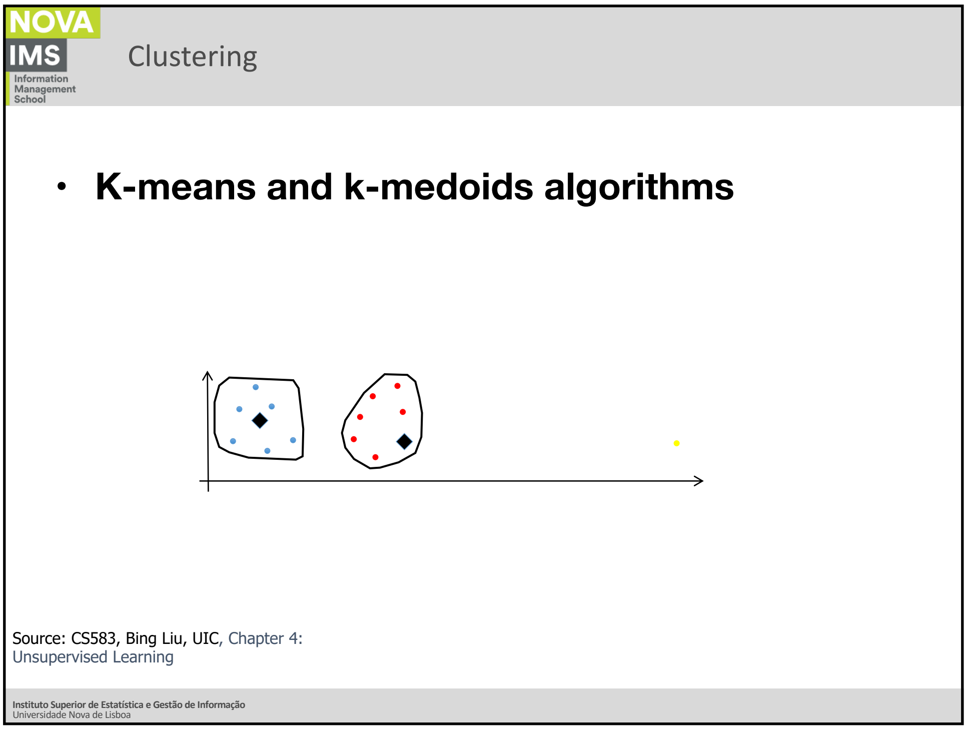
Clustering

- **K-means and k-medoids algorithms**

Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

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

43



Clustering

- **K-means and k-medoids algorithms**

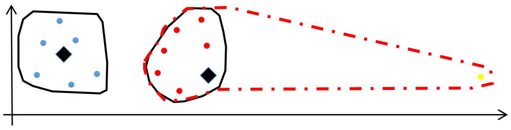
Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

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

44

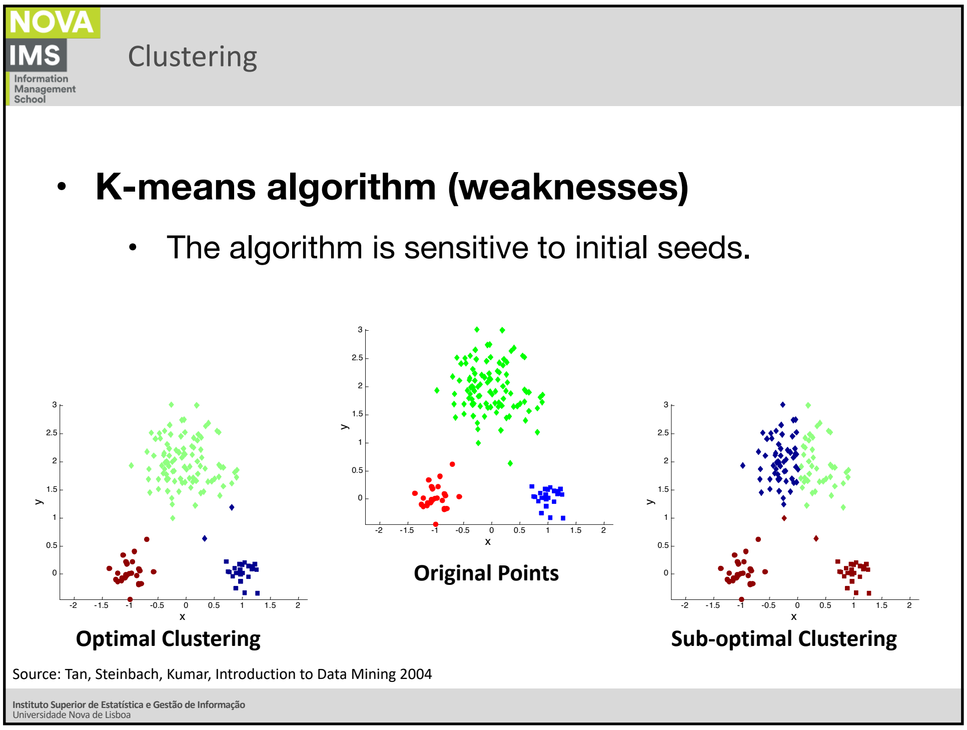
Clustering

- **K-means and k-medoids algorithms**



Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

## The initialization problem



NOVA IMS Information Management School

Clustering

- **K-means algorithm (weaknesses)**
  - The algorithm is sensitive to initial seeds.

Original Points

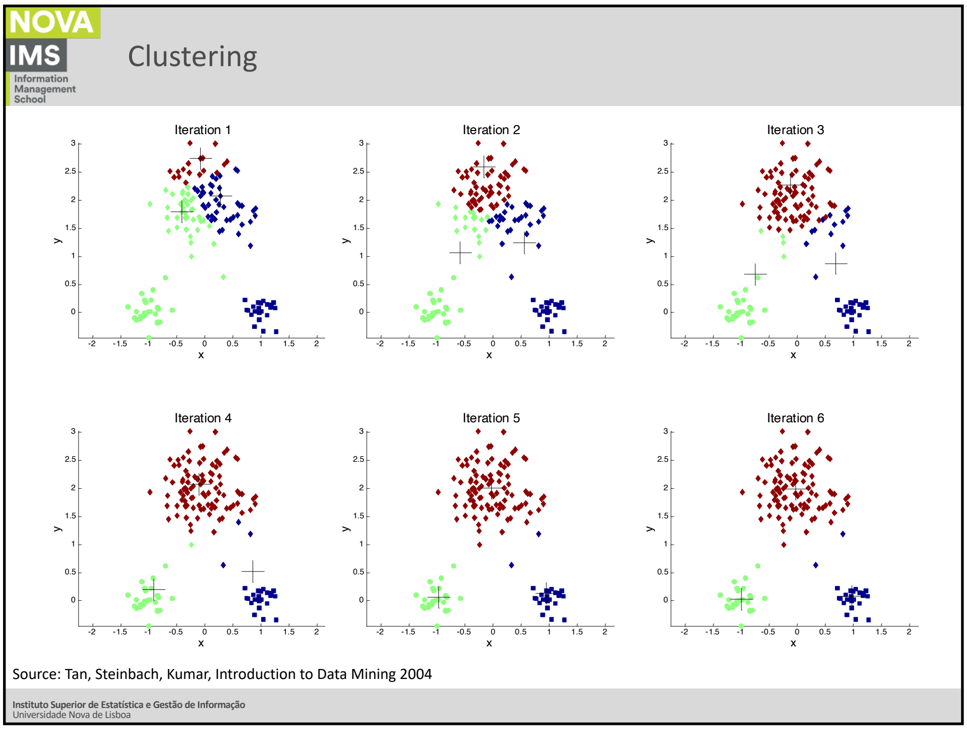
Optimal Clustering

Sub-optimal Clustering

Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

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

47



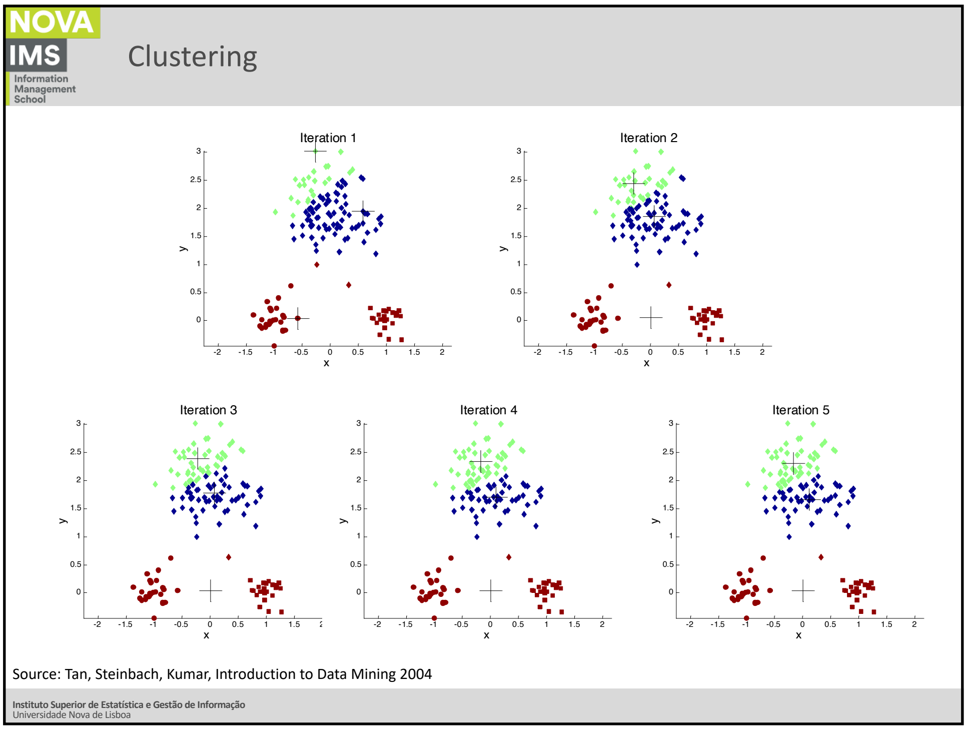
NOVA IMS Information Management School

Clustering

Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

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

48

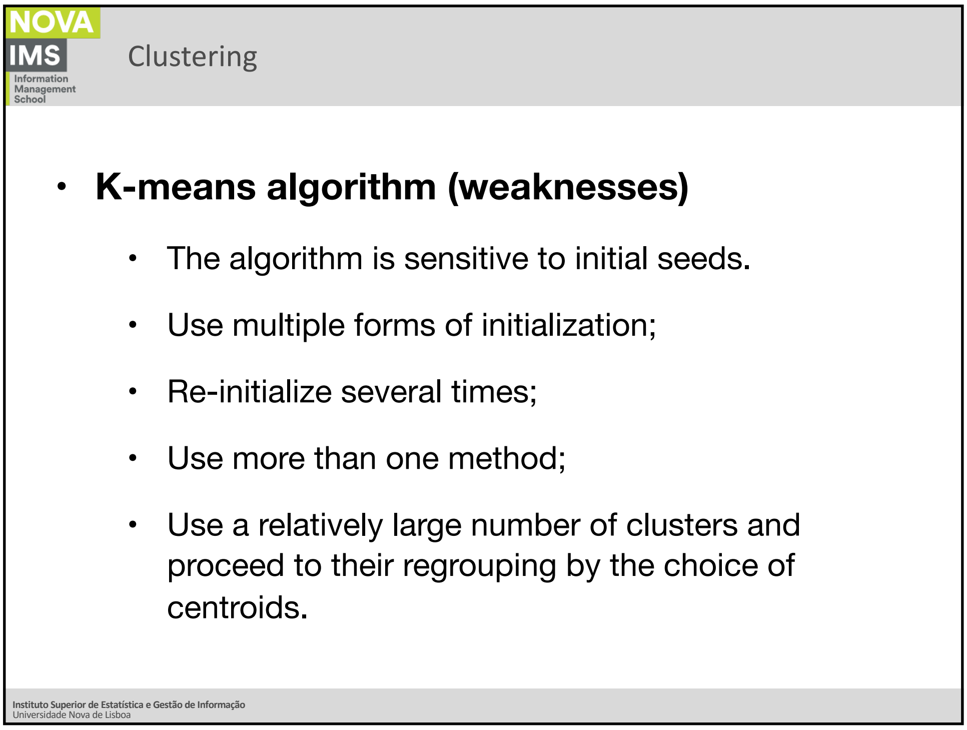


## Clustering

Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

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

49

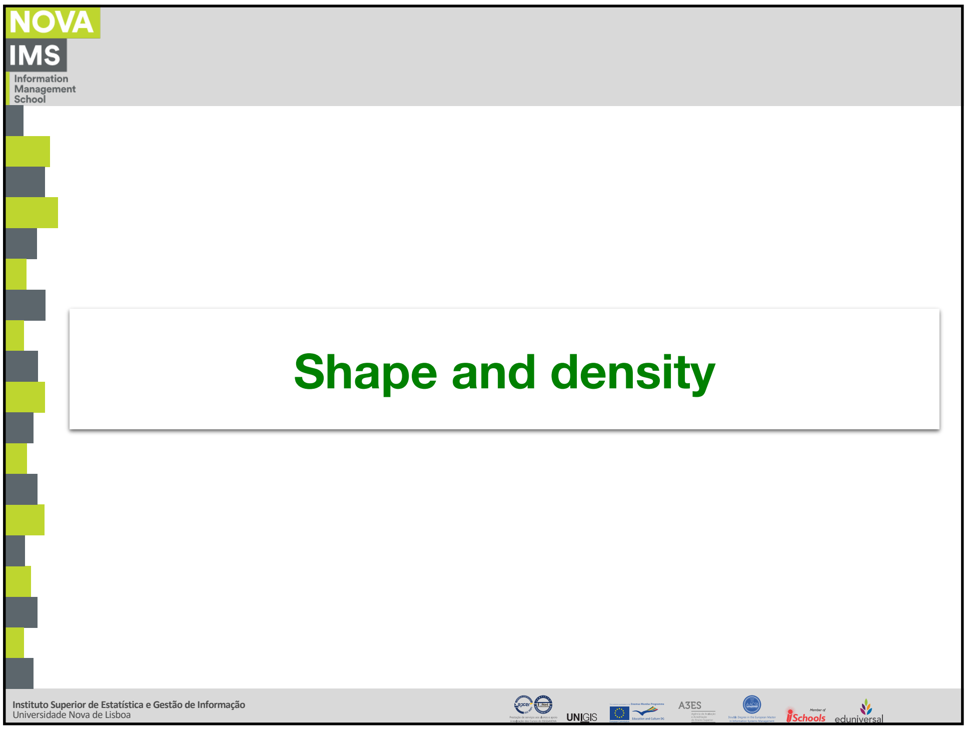

NOVA
IMS
Information Management School

## Clustering

- **K-means algorithm (weaknesses)**
  - The algorithm is sensitive to initial seeds.
  - Use multiple forms of initialization;
  - Re-initialize several times;
  - Use more than one method;
  - Use a relatively large number of clusters and proceed to their regrouping by the choice of centroids.

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

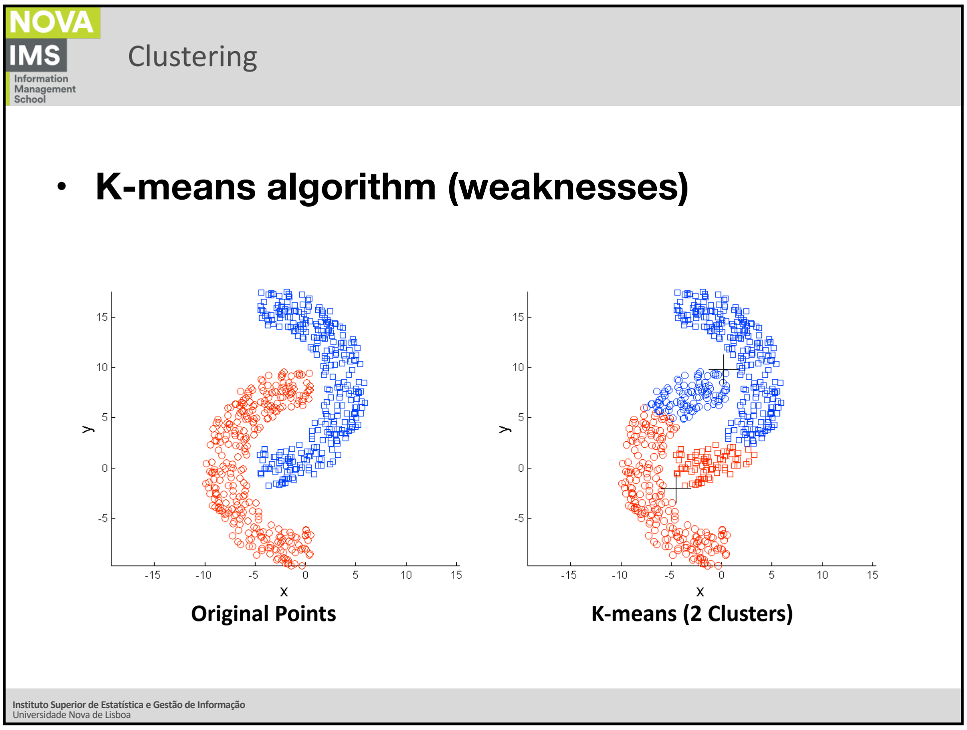
50



NOVA IMS Information Management School

# Shape and density

51



NOVA IMS Information Management School

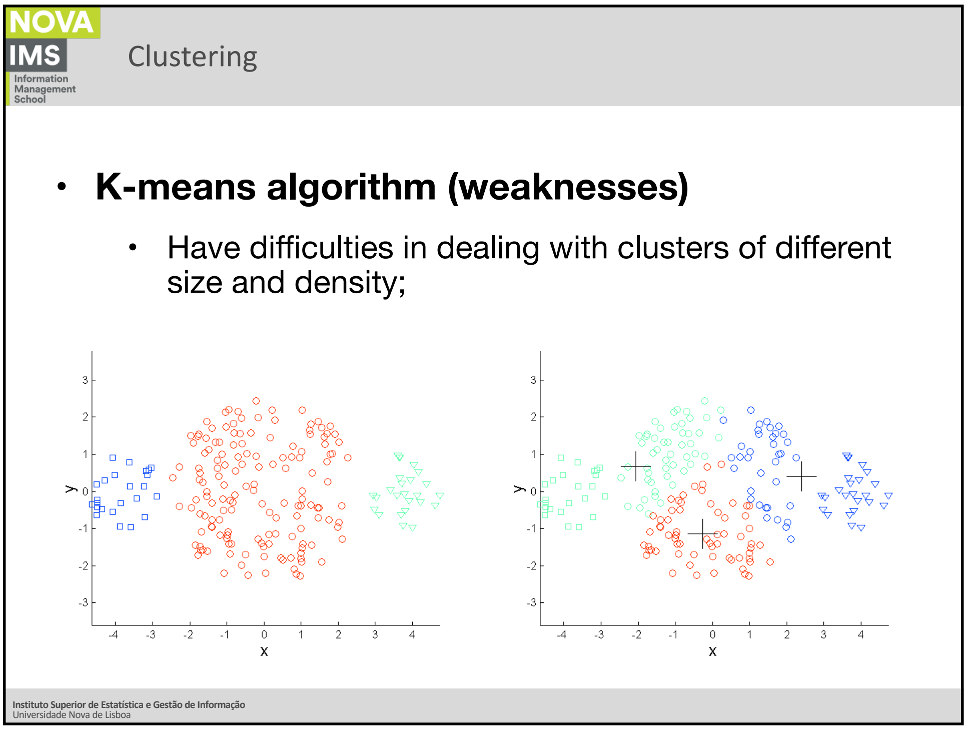
Clustering

- **K-means algorithm (weaknesses)**

**Original Points**

**K-means (2 Clusters)**

52

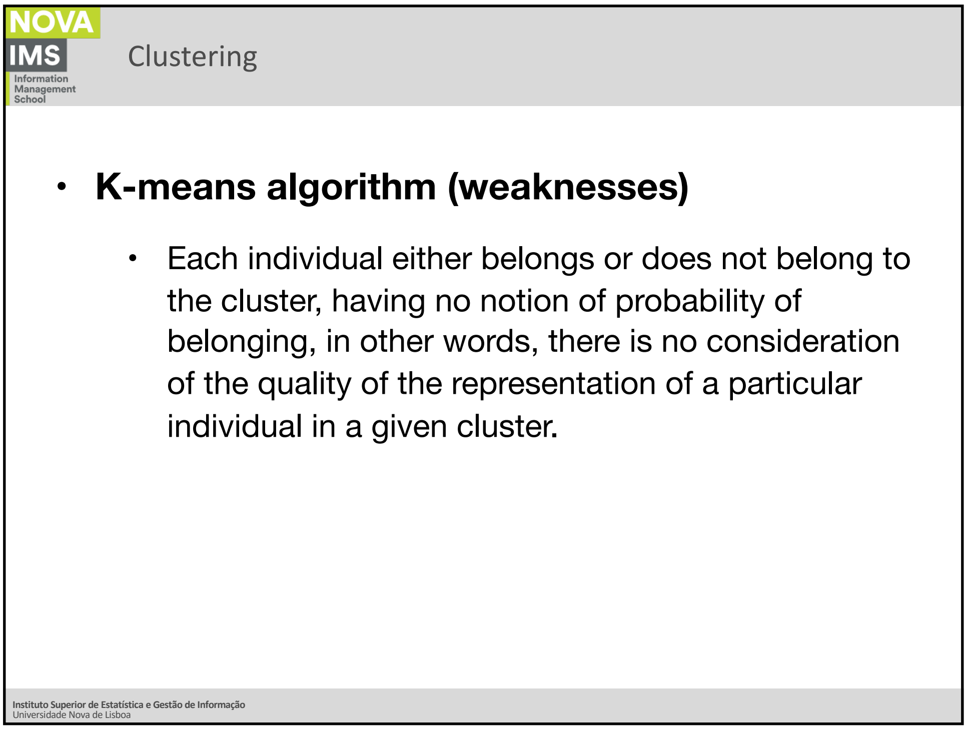


Clustering

- **K-means algorithm (weaknesses)**
  - Have difficulties in dealing with clusters of different size and density;

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

53



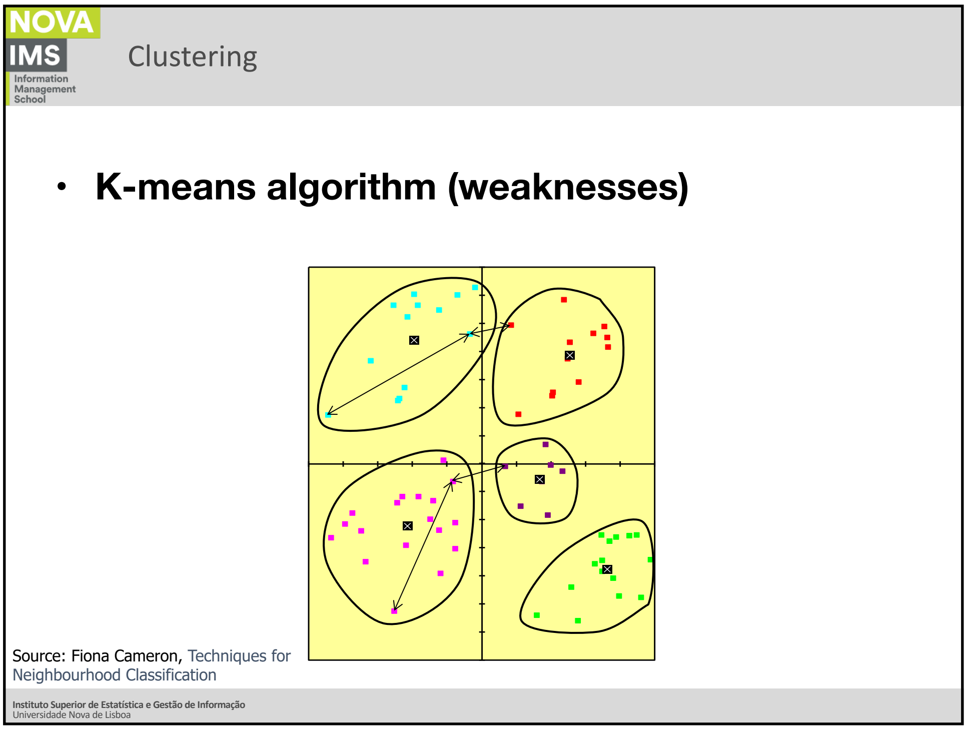
NOVA
IMS
Information Management School

Clustering

- **K-means algorithm (weaknesses)**
  - Each individual either belongs or does not belong to the cluster, having no notion of probability of belonging, in other words, there is no consideration of the quality of the representation of a particular individual in a given cluster.

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

54



NOVA IMS Information Management School

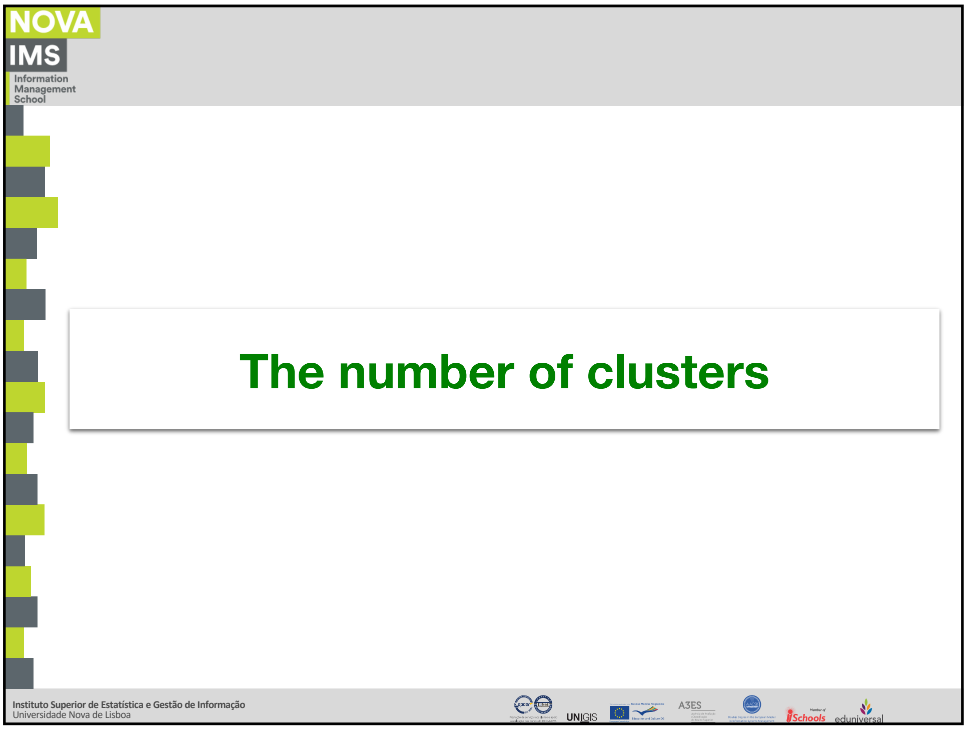
Clustering

- **K-means algorithm (weaknesses)**

Source: Fiona Cameron, Techniques for Neighbourhood Classification

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

55



NOVA IMS Information Management School

## The number of clusters

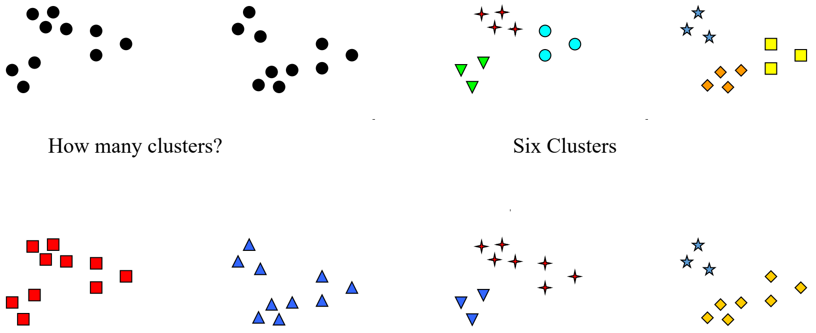
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

56

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**



How many clusters?

Six Clusters

Two Clusters

Four Clusters

Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

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

57

NOVA IMS Information Management School

Clustering

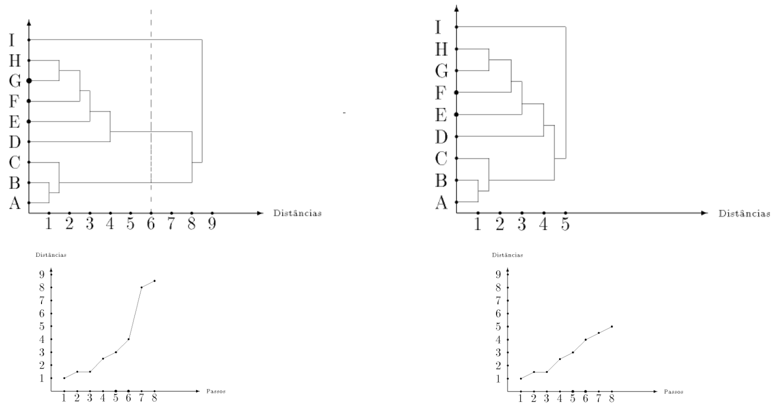
- **K-means algorithm the number of clusters**
  - This is always a difficult problem to solve, and there are no recipes to fix this.
  - One way to minimize the problem is to create various classifications with different K, and choose the best.
  - Use a hierarchical method in order to choose the number of clusters based on the dendrogram.

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

58

## Clustering

- **K-means algorithm the number of clusters**



## Clustering

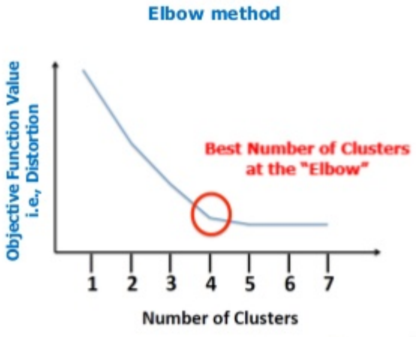
- **K-means algorithm the number of clusters**
  - The choice should be guided by three fundamental criteria:
    - intra-cluster variance,
    - evaluation of the profile of the cluster (subjective),
    - operational considerations.

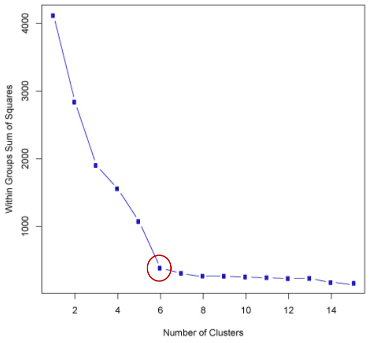
## Clustering

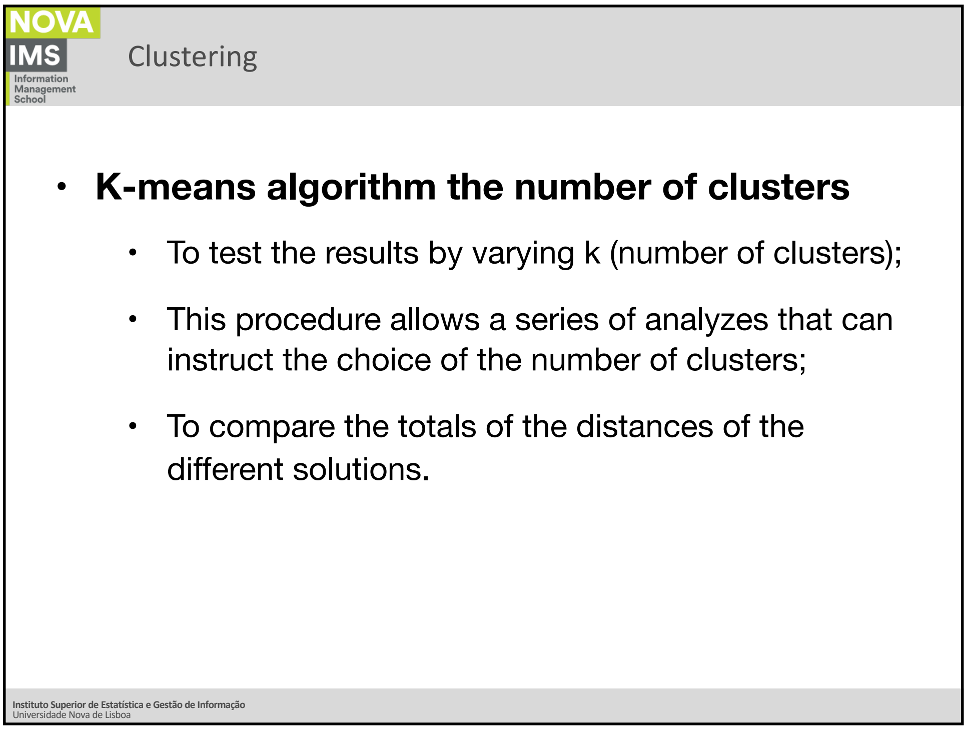
- **K-means algorithm the number of clusters**
  - Regarding the first criterion, the analysis is simple and not too subjective, since we know that the lower the intra-cluster variability the greater the cohesion of the cluster, a highly desirable feature in this type of analysis. However, as k increases, variability decreases;
  - Regarding the second criterion, the question is not as simple in the sense that it requires much more subjective assessments, which relate to the interpretation of the obtained clusters;
  - The third criterion is relatively simple in the sense that these issues are imposed on the analyst.

## Clustering

- **K-means algorithm the number of clusters**



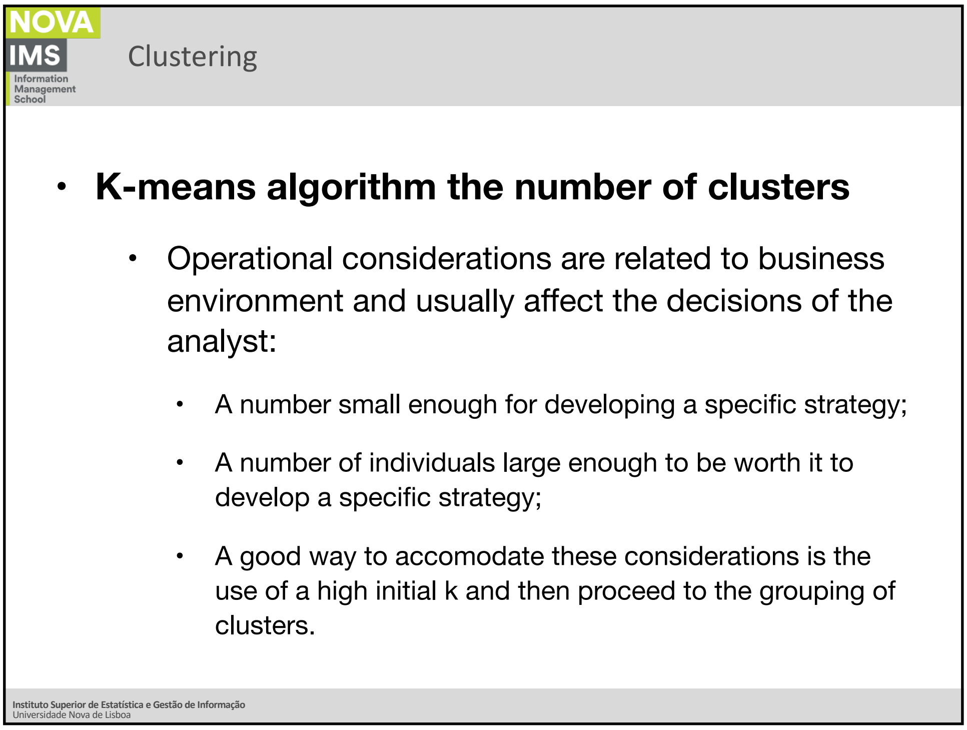




## Clustering

- **K-means algorithm the number of clusters**
  - To test the results by varying k (number of clusters);
  - This procedure allows a series of analyzes that can instruct the choice of the number of clusters;
  - To compare the totals of the distances of the different solutions.

63



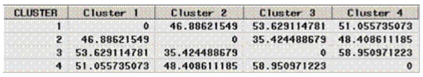
## Clustering

- **K-means algorithm the number of clusters**
  - Operational considerations are related to business environment and usually affect the decisions of the analyst:
    - A number small enough for developing a specific strategy;
    - A number of individuals large enough to be worth it to develop a specific strategy;
    - A good way to accomodate these considerations is the use of a high initial k and then proceed to the grouping of clusters.

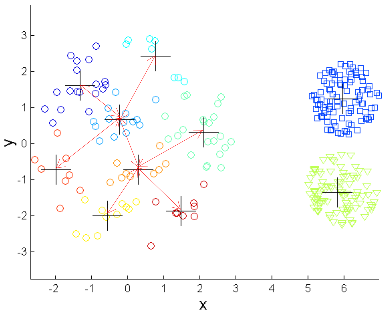
64

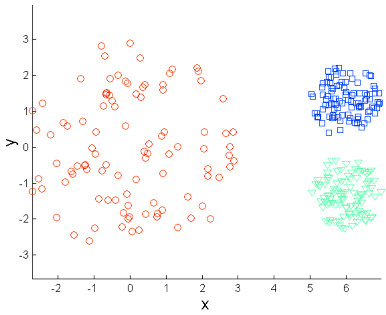
# Clustering

- **K-means algorithm the number of clusters**



| CLUSTER | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| 1 | 0 | 46.88621549 | 53.629114781 | 51.055735073 |
| 2 | 46.88621549 | 0 | 35.424488679 | 48.408611185 |
| 3 | 53.629114781 | 35.424488679 | 0 | 58.950971223 |
| 4 | 51.055735073 | 48.408611185 | 58.950971223 | 0 |





# Clustering

- **K-means algorithm the number of clusters**
  - This evaluation is carried out by comparing the mean values for each variable in each cluster with the mean values of the population for each variable;
  - In this case, it is particularly relevant to take into account the most important differences within the different clusters and the mean population.
  - That is why profiling is so important.