NOVA IMS Information Management School

Data Mining

NOVA-IMS

02/11/2021

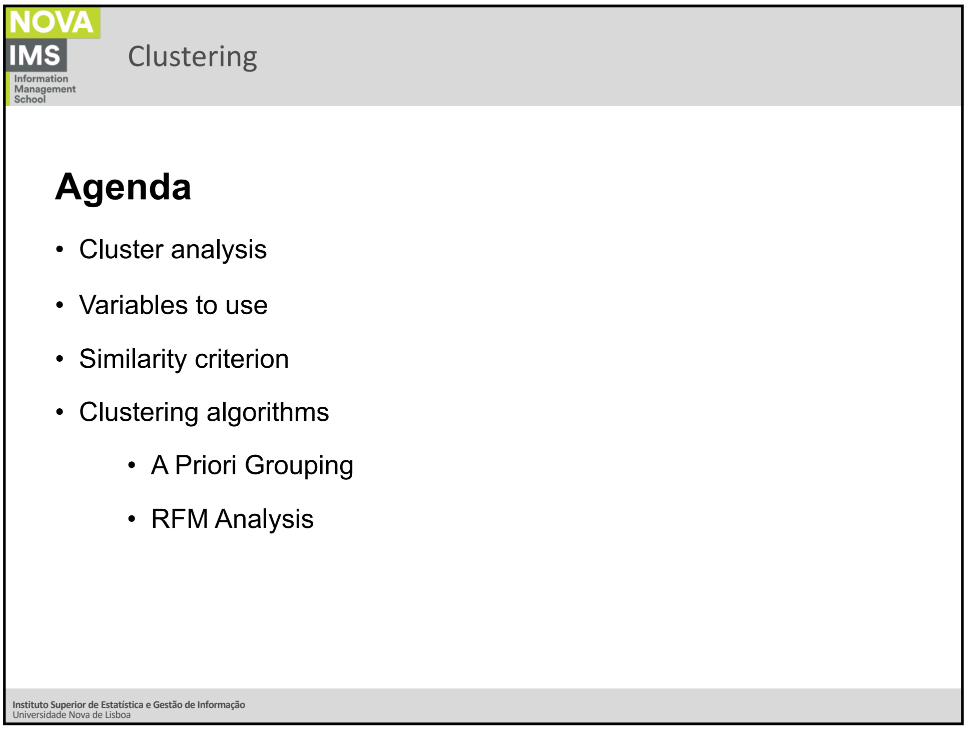
Fernando Lucas Bação

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

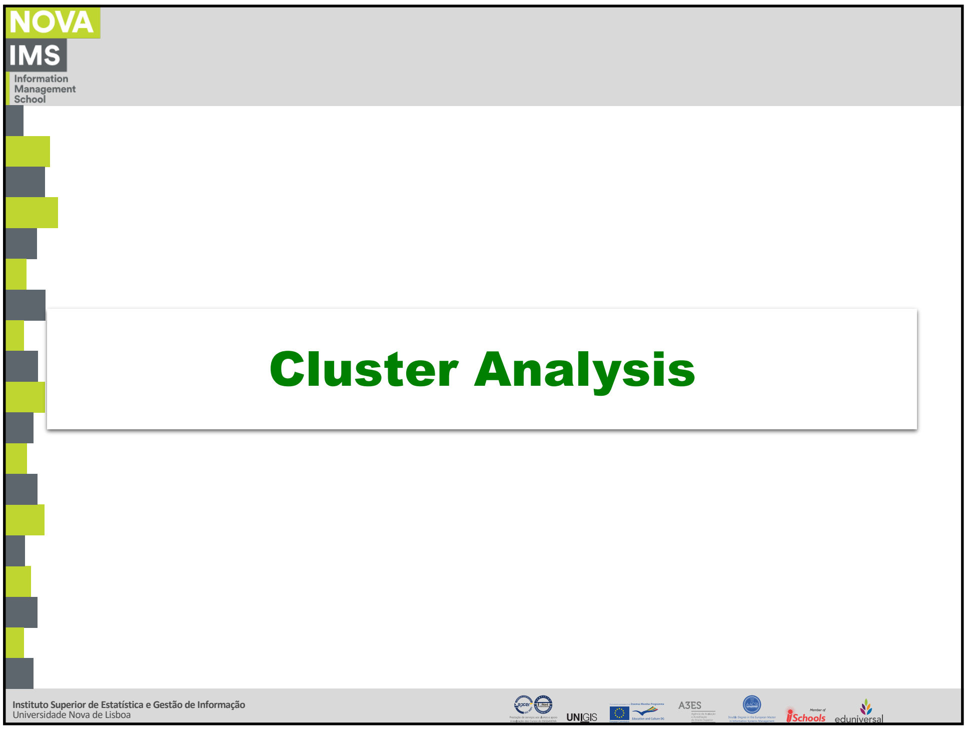
Clustering

Agenda

- Cluster analysis
- Variables to use
- Similarity criterion
- Clustering algorithms
  - A Priori Grouping
  - RFM Analysis

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

2

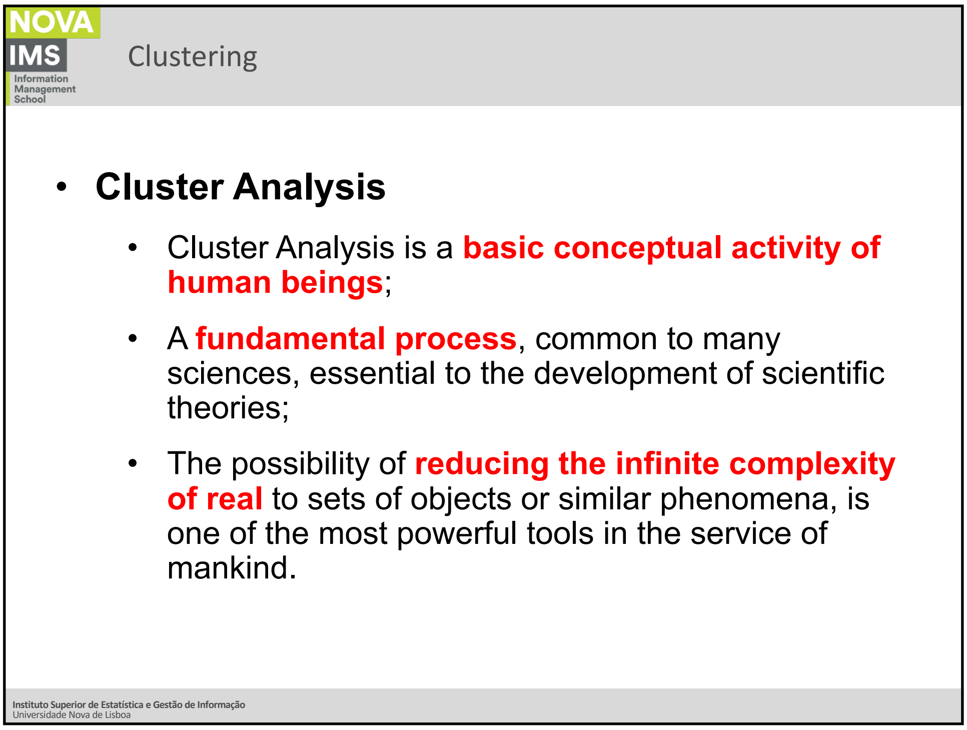


NOVA IMS Information Management School

Cluster Analysis

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

3



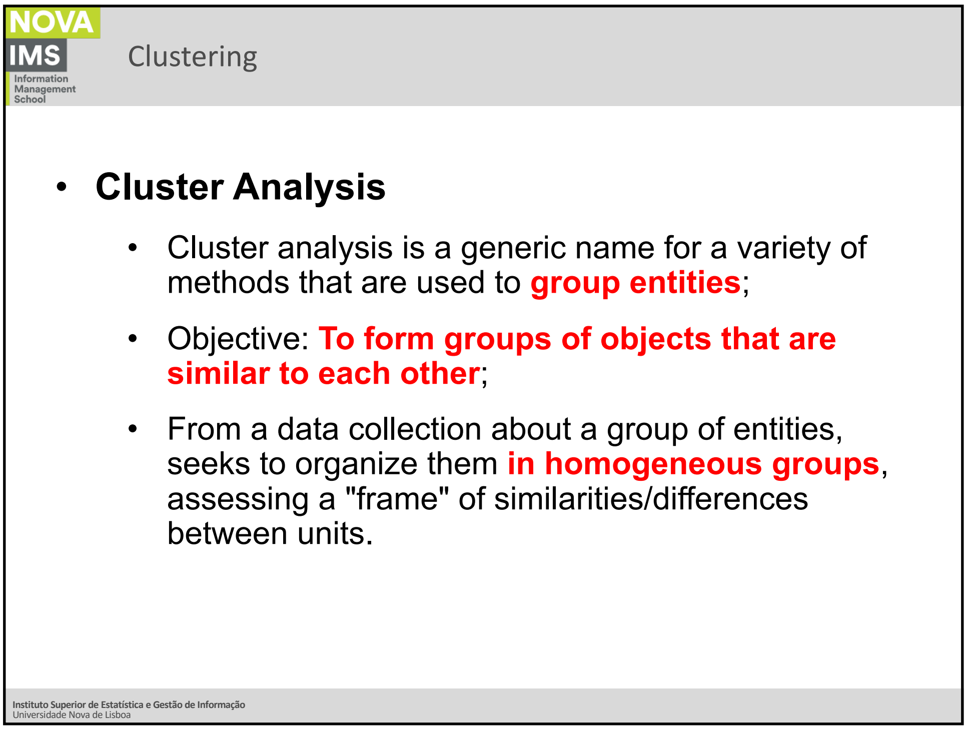
NOVA IMS Information Management School

Clustering

- **Cluster Analysis**
  - Cluster Analysis is a **basic conceptual activity of human beings**;
  - A **fundamental process**, common to many sciences, essential to the development of scientific theories;
  - The possibility of **reducing the infinite complexity of real** to sets of objects or similar phenomena, is one of the most powerful tools in the service of mankind.

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

4



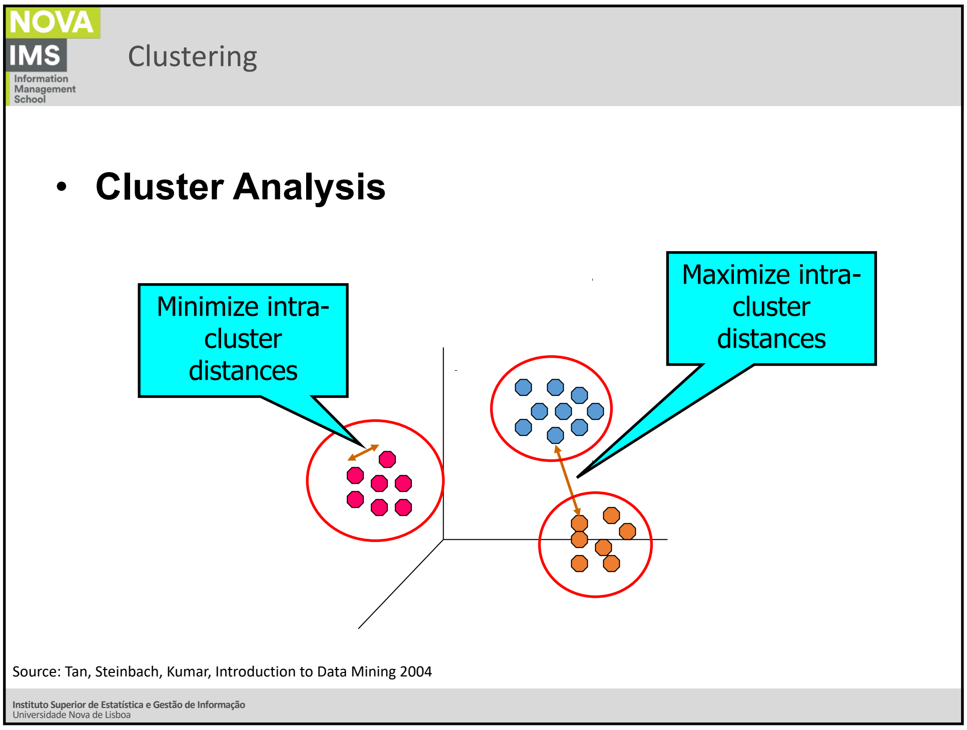
NOVA IMS Information Management School

Clustering

- **Cluster Analysis**
  - Cluster analysis is a generic name for a variety of methods that are used to **group entities**;
  - Objective: **To form groups of objects that are similar to each other**;
  - From a data collection about a group of entities, seeks to organize them **in homogeneous groups**, assessing a "frame" of similarities/differences between units.

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

5



NOVA IMS Information Management School

Clustering

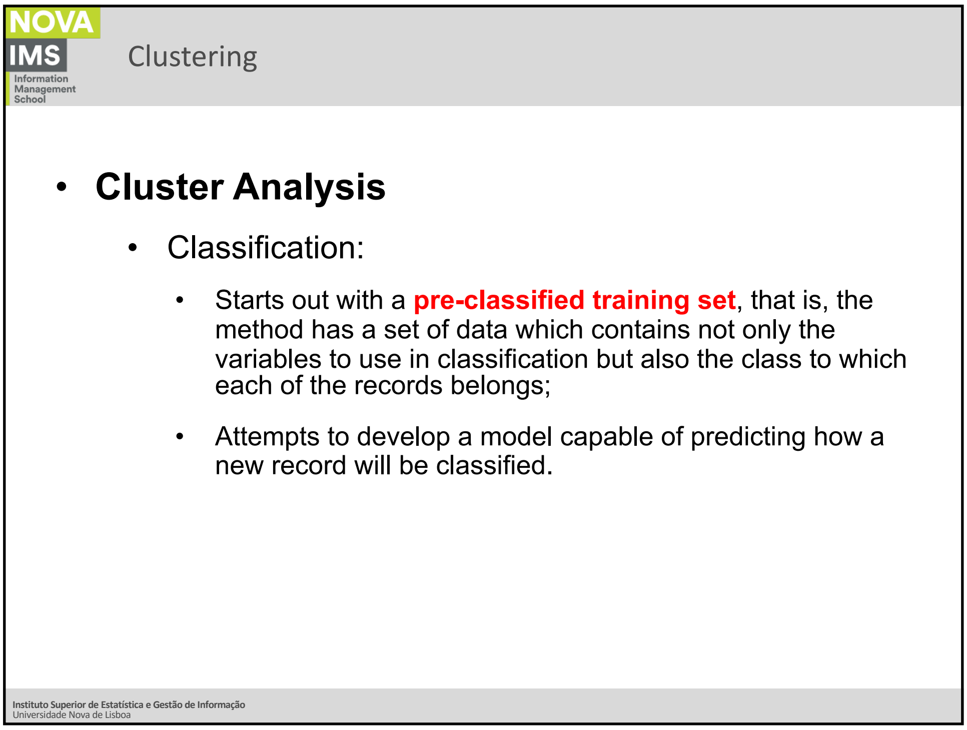
- **Cluster Analysis**

Minimize intra-cluster distances
Maximize intra-cluster distances

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

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

6

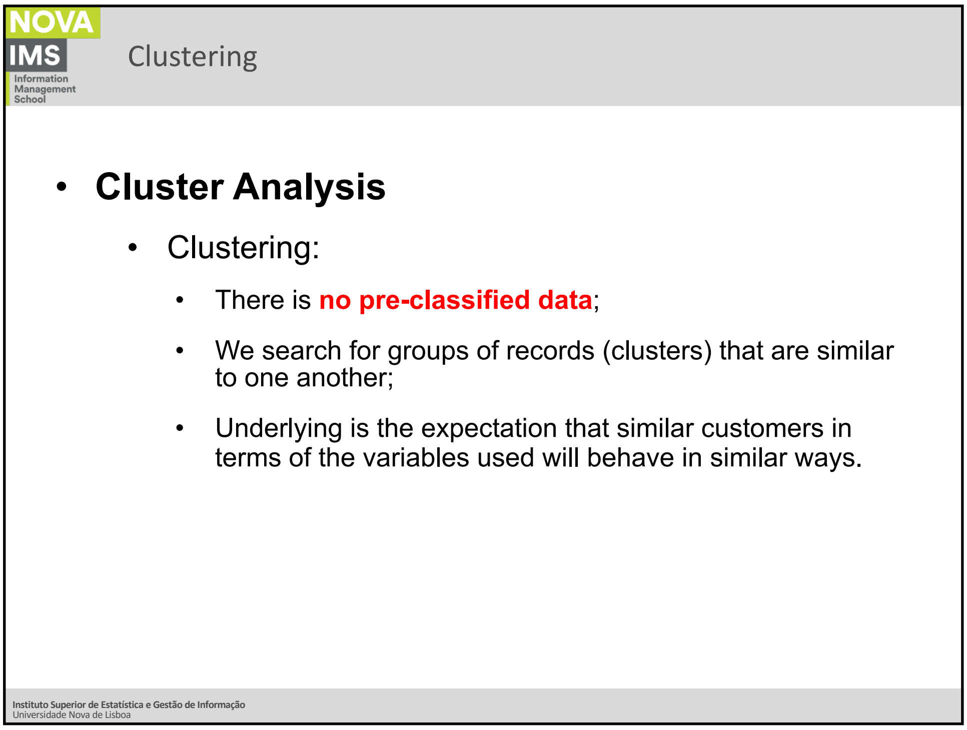


Clustering

- **Cluster Analysis**
  - Classification:
    - Starts out with a **pre-classified training set**, that is, the method has a set of data which contains not only the variables to use in classification but also the class to which each of the records belongs;
    - Attempts to develop a model capable of predicting how a new record will be classified.

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

7

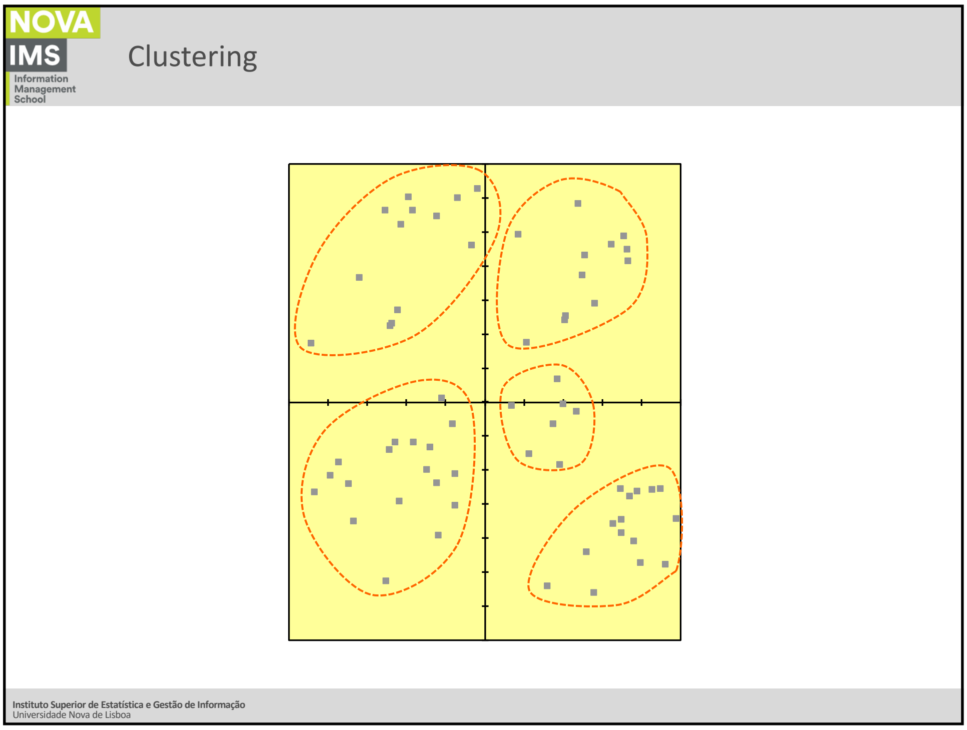


Clustering

- **Cluster Analysis**
  - Clustering:
    - There is **no pre-classified data**;
    - We search for groups of records (clusters) that are similar to one another;
    - Underlying is the expectation that similar customers in terms of the variables used will behave in similar ways.

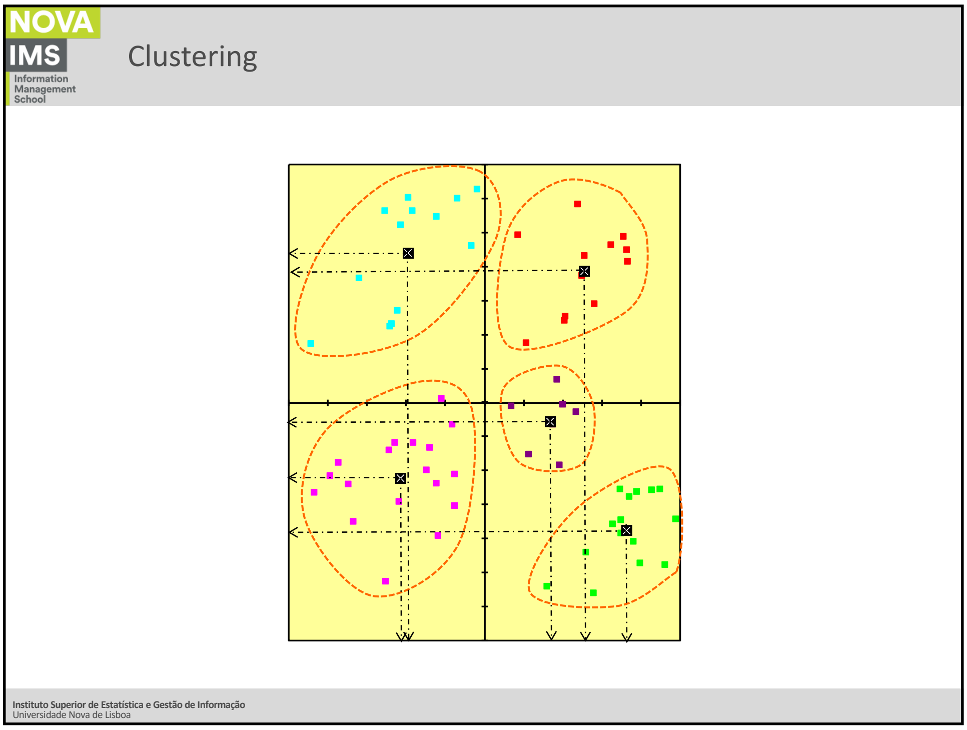
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

8



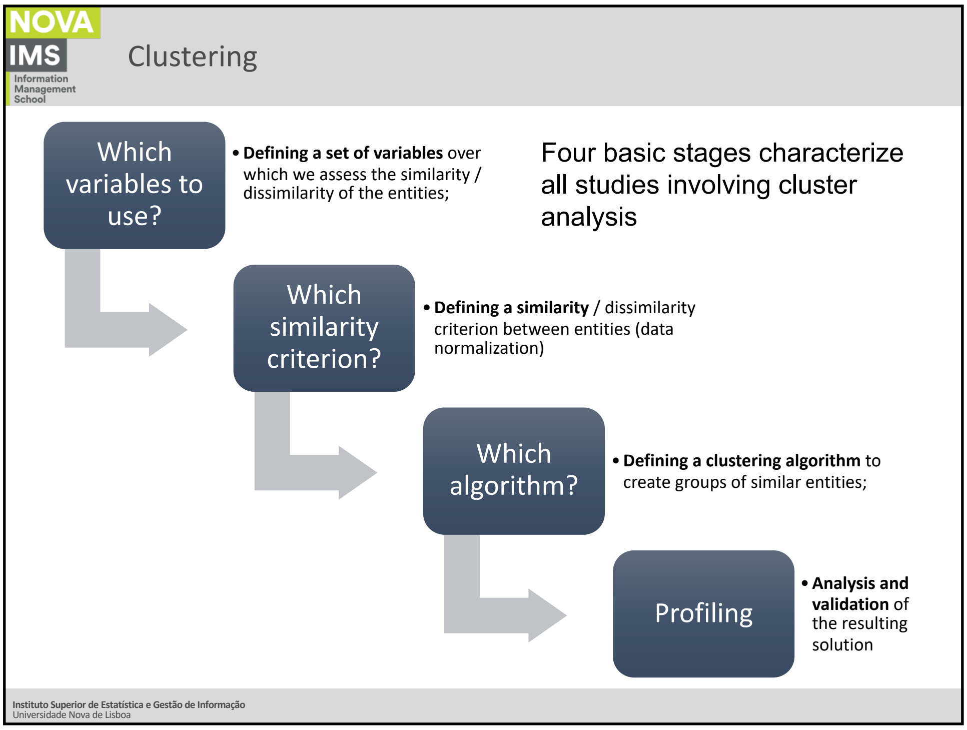
Clustering

9



Clustering

10



NOVA IMS Information Management School

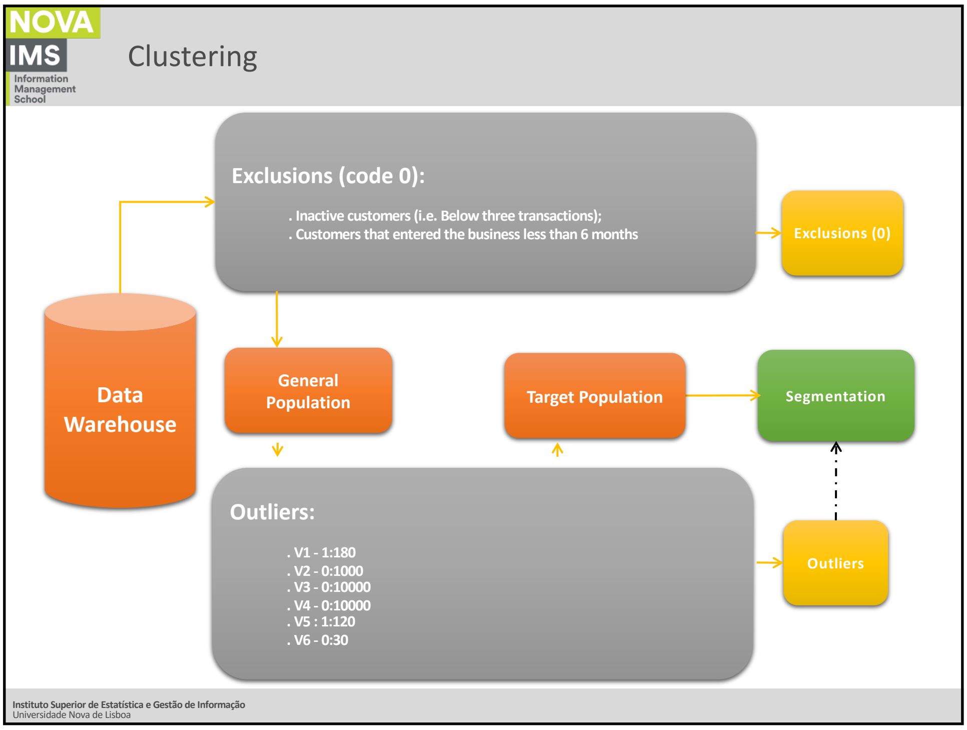
Clustering

Four basic stages characterize all studies involving cluster analysis

- **Defining a set of variables** over which we assess the similarity / dissimilarity of the entities;
- **Defining a similarity** / dissimilarity criterion between entities (data normalization)
- **Defining a clustering algorithm** to create groups of similar entities;
- **Analysis and validation** of the resulting solution

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

11



NOVA IMS Information Management School

Clustering

Exclusions (code 0):

- . Inactive customers (i.e. Below three transactions);
- . Customers that entered the business less than 6 months

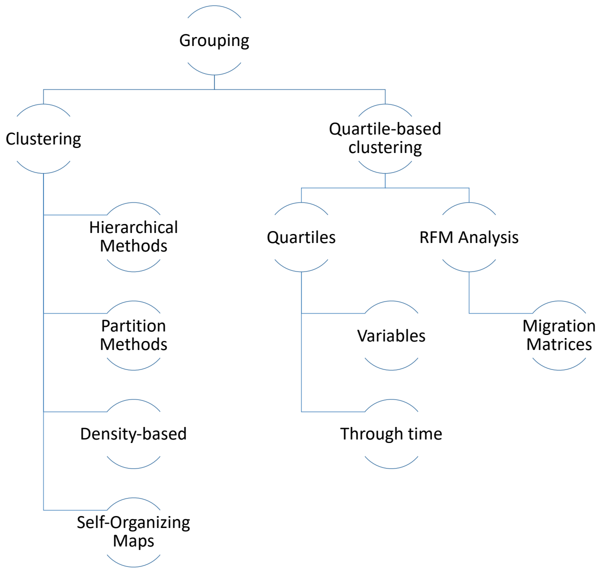
Outliers:

- . V1 - 1:180
- . V2 - 0:1000
- . V3 - 0:10000
- . V4 - 0:10000
- . V5 : 1:120
- . V6 - 0:30

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

12

Clustering



Grouping

Clustering

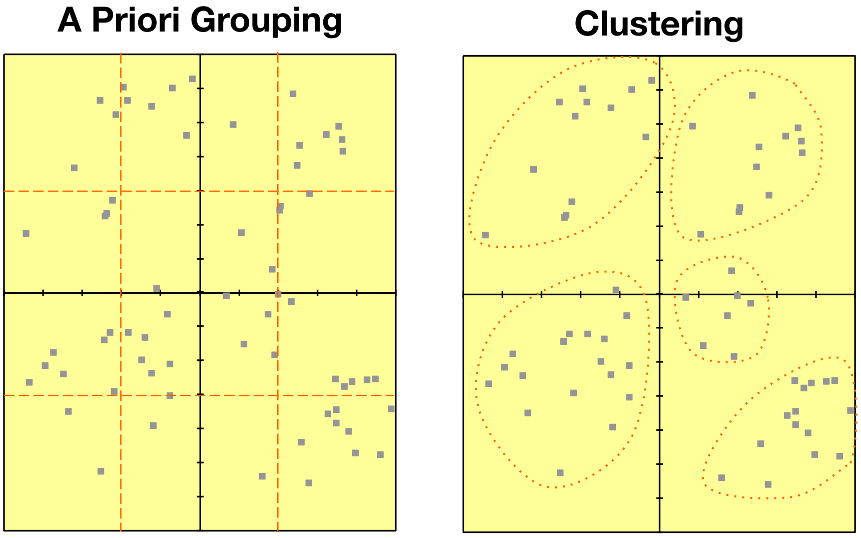
- Hierarchical Methods
- Partition Methods
- Density-based
- Self-Organizing Maps

Quartile-based clustering

- Quartiles
  - Variables
  - Through time
- RFM Analysis
  - Migration Matrices

13

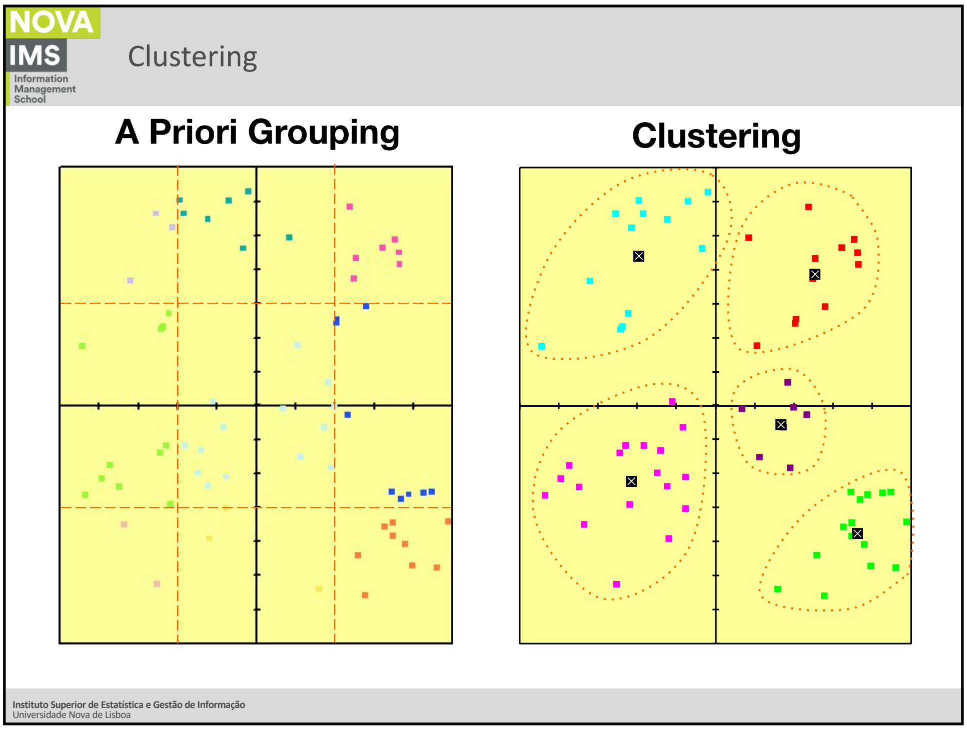
Clustering



A Priori Grouping

Clustering

14



NOVA IMS Information Management School

Clustering

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

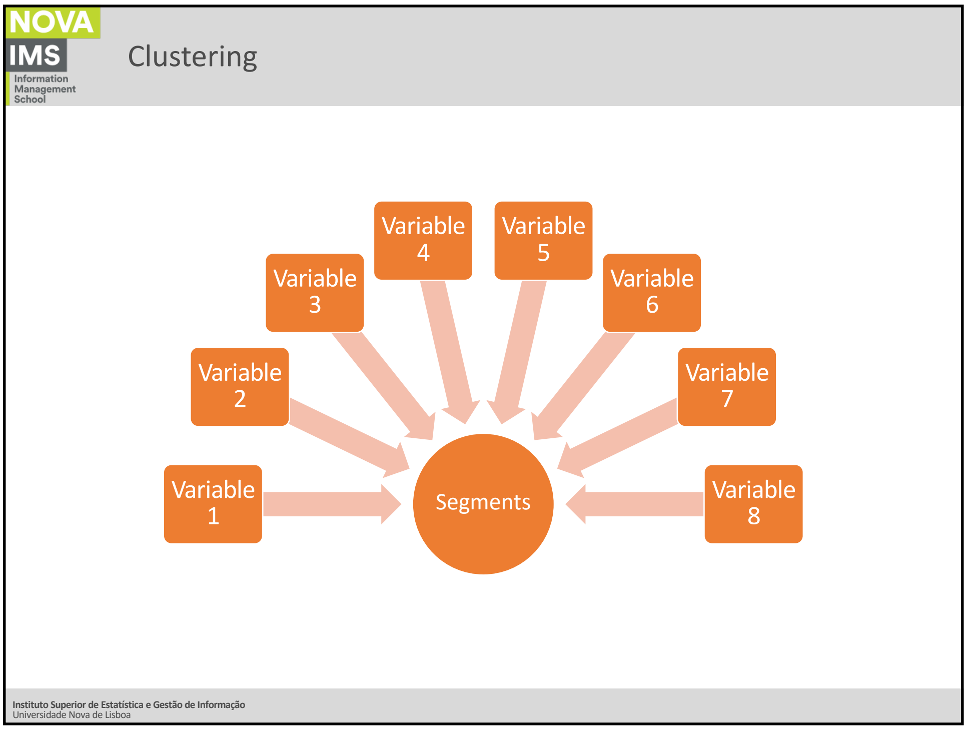
15

NOVA IMS Information Management School

Decide on the Variables to Use

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

16



Clustering

Variable 1

Variable 2

Variable 3

Variable 4

Variable 5

Variable 6

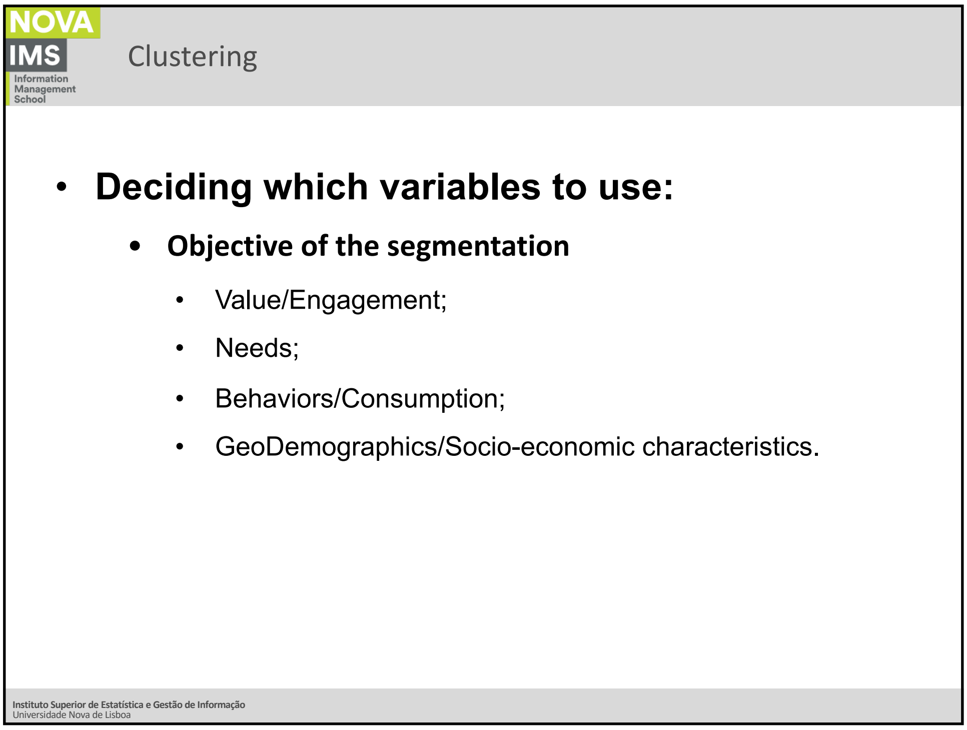
Variable 7

Variable 8

Segments

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

17

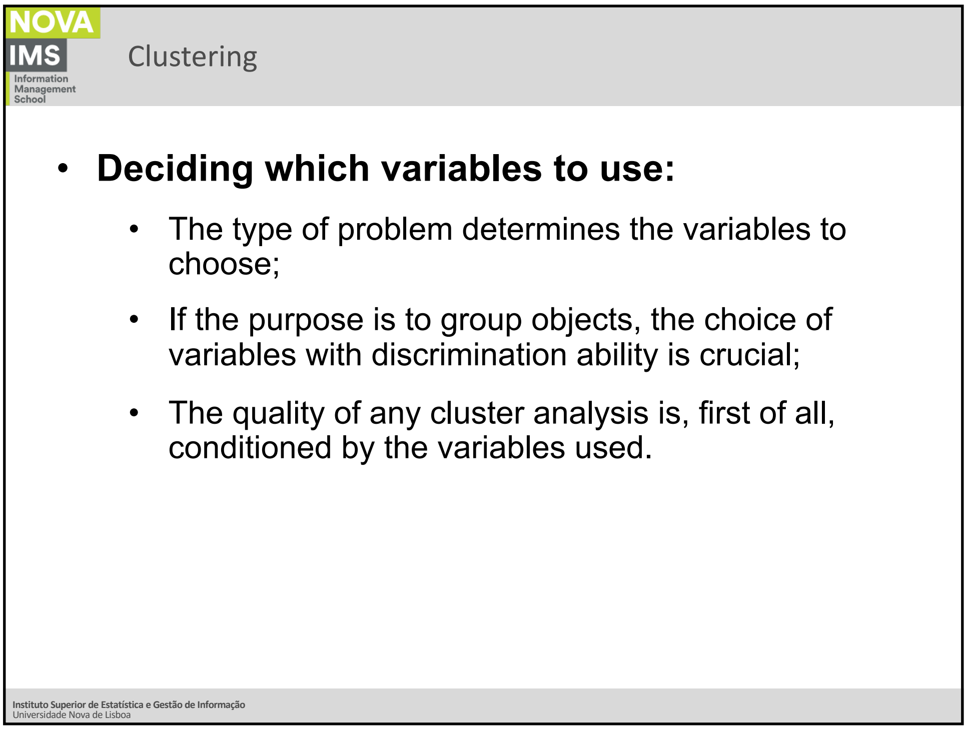


Clustering

- **Deciding which variables to use:**
  - **Objective of the segmentation**
    - Value/Engagement;
    - Needs;
    - Behaviors/Consumption;
    - GeoDemographics/Socio-economic characteristics.

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

18

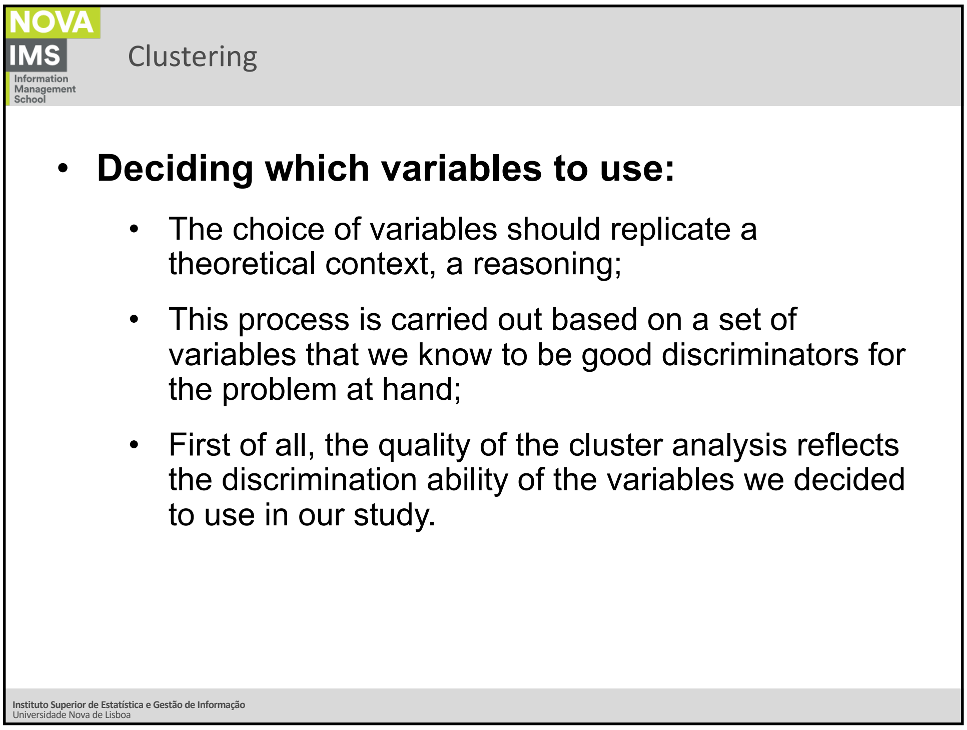


Clustering

- **Deciding which variables to use:**
  - The type of problem determines the variables to choose;
  - If the purpose is to group objects, the choice of variables with discrimination ability is crucial;
  - The quality of any cluster analysis is, first of all, conditioned by the variables used.

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

19

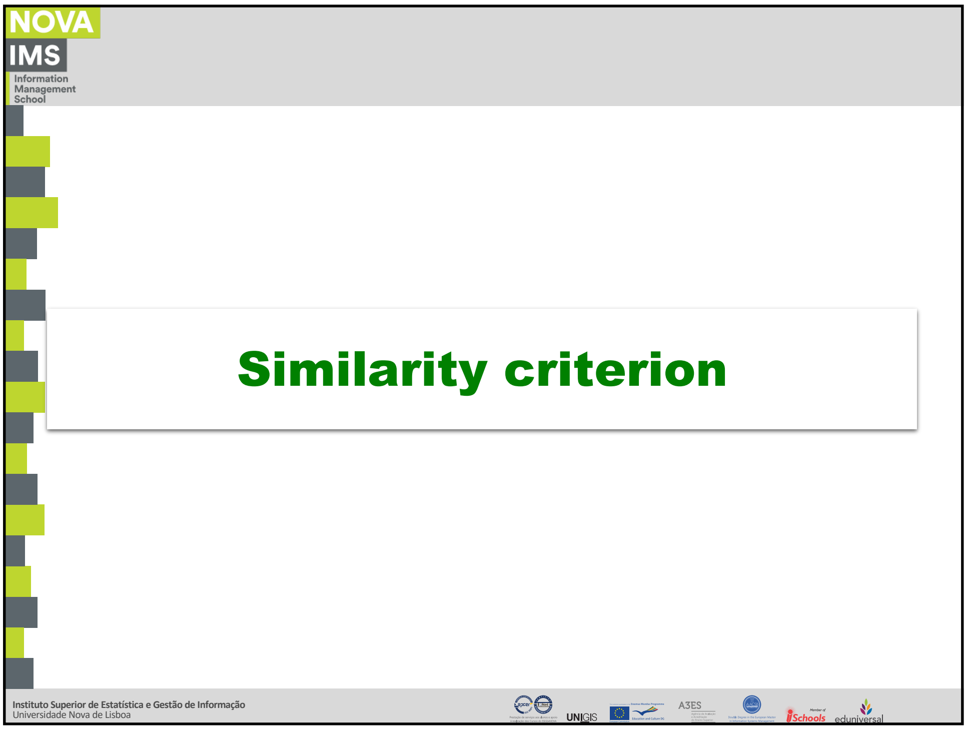


Clustering

- **Deciding which variables to use:**
  - The choice of variables should replicate a theoretical context, a reasoning;
  - This process is carried out based on a set of variables that we know to be good discriminators for the problem at hand;
  - First of all, the quality of the cluster analysis reflects the discrimination ability of the variables we decided to use in our study.

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

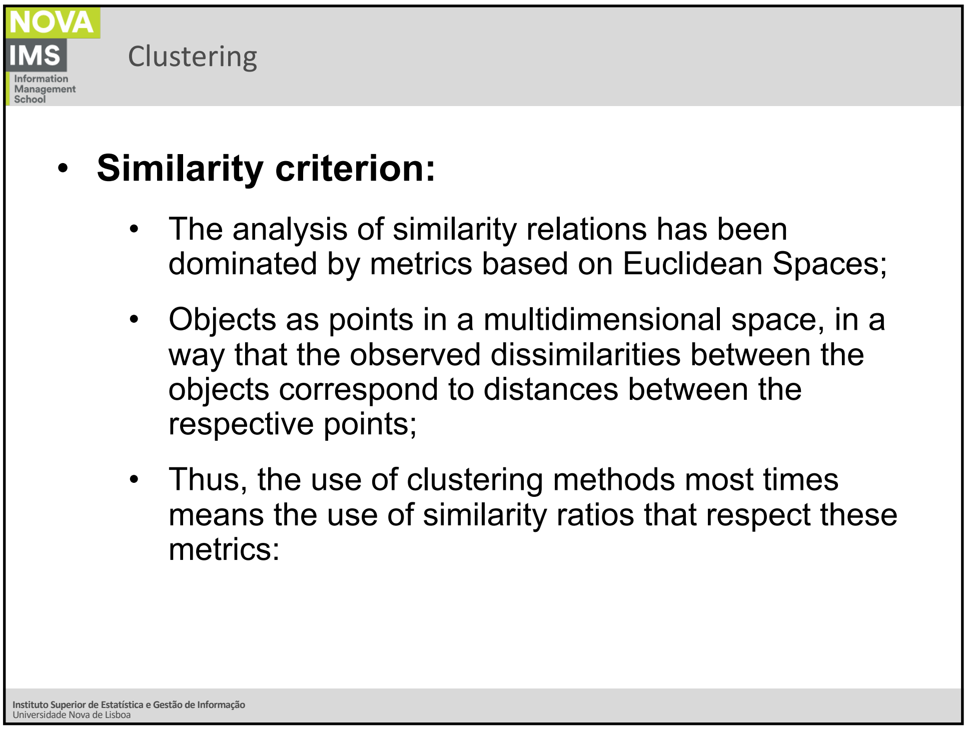
20



Similarity criterion

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

21



Clustering

- **Similarity criterion:**
  - The analysis of similarity relations has been dominated by metrics based on Euclidean Spaces;
  - Objects as points in a multidimensional space, in a way that the observed dissimilarities between the objects correspond to distances between the respective points;
  - Thus, the use of clustering methods most times means the use of similarity ratios that respect these metrics:

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

22

NOVA IMS Information Management School

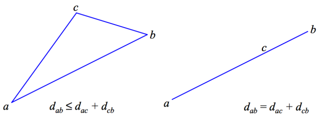
Clustering

- **Similarity criterion:**

- In mathematics, a true measure of distance, called a *metric*, obeys three properties. These metric axioms are as follows, where d_{ab} denotes the distance between objects *a* and *b*:

  1. $d_{ab} = d_{ba}$ *(measure is symmetric)*
  2. $d_{ab} \geq 0$ and = 0 if and only if $a = b$ *(distances are always positive except when the objects are identical)*
  3. $d_{ab} \leq d_{ac} + d_{ca}$ *(triangle inequality)*

Exhibit 5.1 Illustration of the triangle inequality for distances in Euclidean space.



NOVA IMS Information Management School

Clustering

- **Similarity criterion:**

- Euclidian distance: the distance between two elements (*i,j*) is the square root of the sum of the squares of the differences between *i* and *j* values for all variables (*v=1, 2,...., p*):

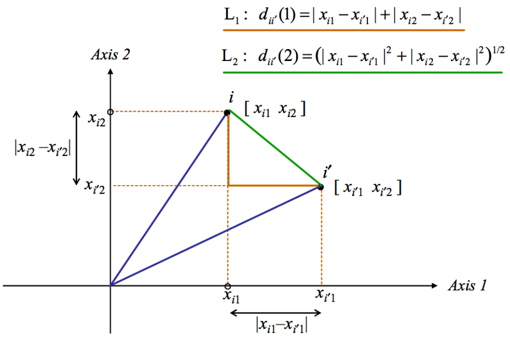
$$d_{ij} = \sqrt{\sum_{v=1}^{p} \left(X_{iv} - X_{jv}\right)^2} \, euclidean\, also\, known\, as\ L_2$$

$$d_{ij} = \sum_{v=1}^{p} \left|X_{iv} - X_{jv}\right| City\, Block\, or\, L_1$$

Clustering

- **Similarity criterion:**

- Euclidian distance: the distance between two elements (*i,j*) is the square root of the sum of the squares of the differences between *i* and *j* values for all variables (*v=1, 2,...., p*):



$$L_1:\ d_{ii'}(1) = |x_{i1} - x_{i'1}| + |x_{i2} - x_{i'2}|$$

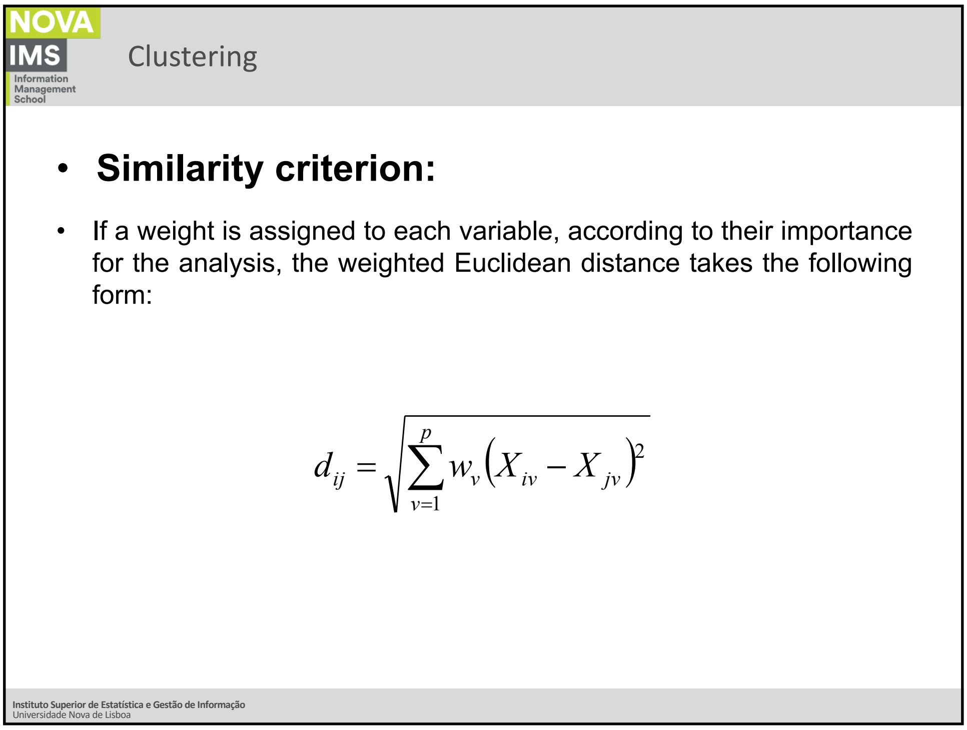
$$L_2:\ d_{ii'}(2) = (|x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2)^{1/2}$$

Clustering

- **Similarity criterion:**

- Minkowski distance: is defined from the absolute distance, and can be considered as a generalization of both the Euclidean distance and the Manhattan distance. It coincides with Euclidean distance when *r=2* and with Manhattan distance when *r=1*:

$$d_{ij} = \left(\sum_{v=1}^{p} \left| X_{iv} - X_{jv} \right|^{r} \right)^{1/r}$$



NOVA IMS Information Management School

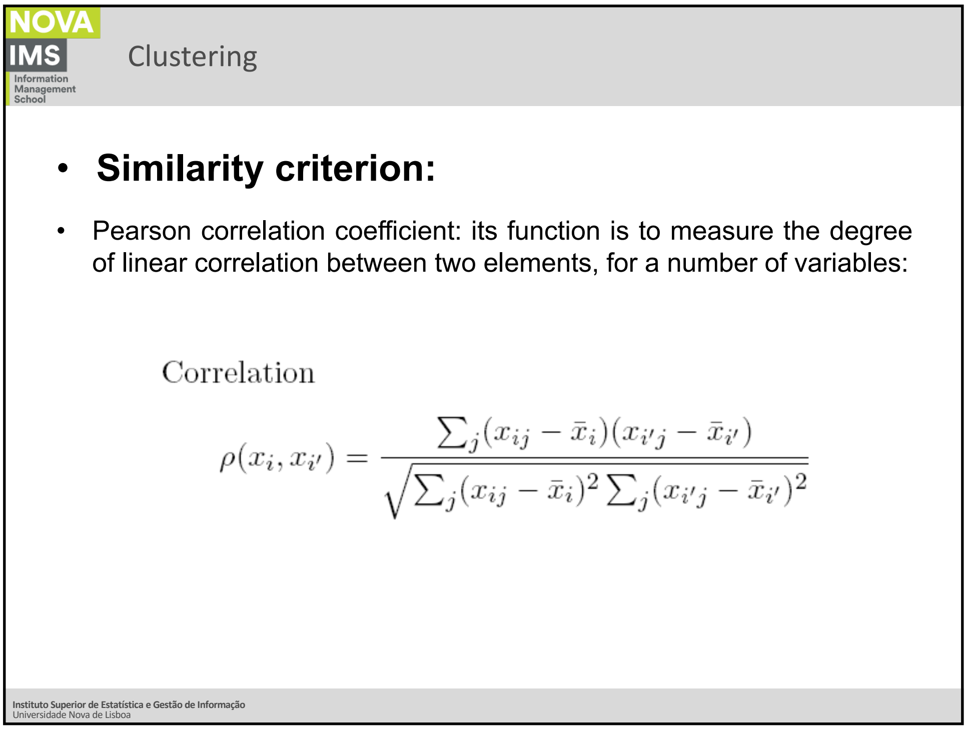
Clustering

- **Similarity criterion:**
  - If a weight is assigned to each variable, according to their importance for the analysis, the weighted Euclidean distance takes the following form:

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

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

27



NOVA IMS Information Management School

Clustering

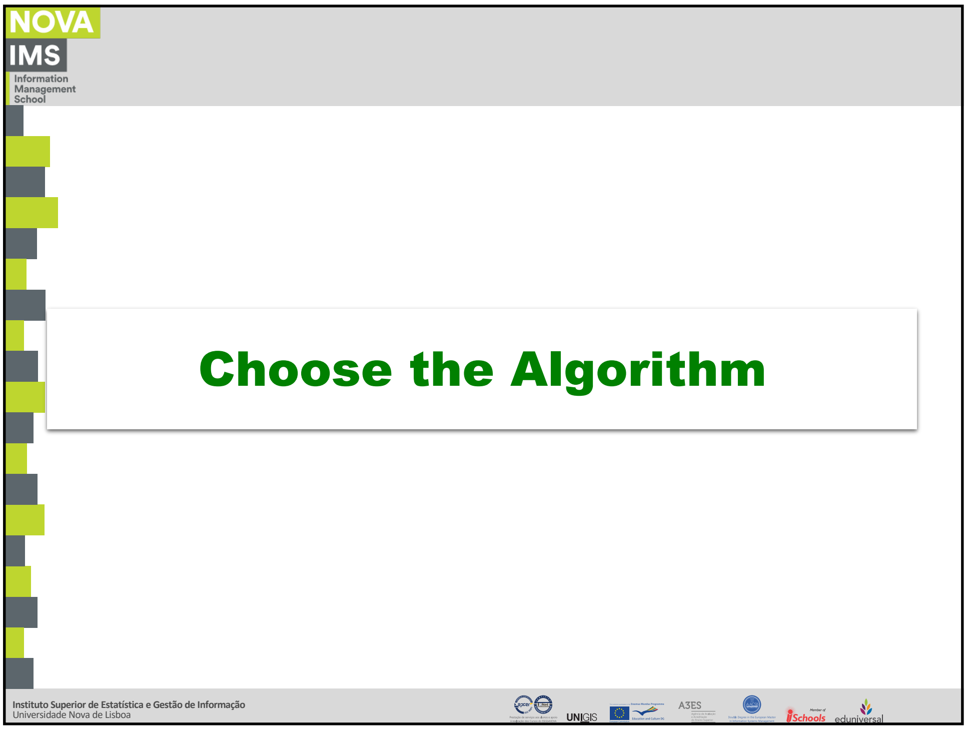
- **Similarity criterion:**
  - Pearson correlation coefficient: its function is to measure the degree of linear correlation between two elements, for a number of variables:

Correlation

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$$

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

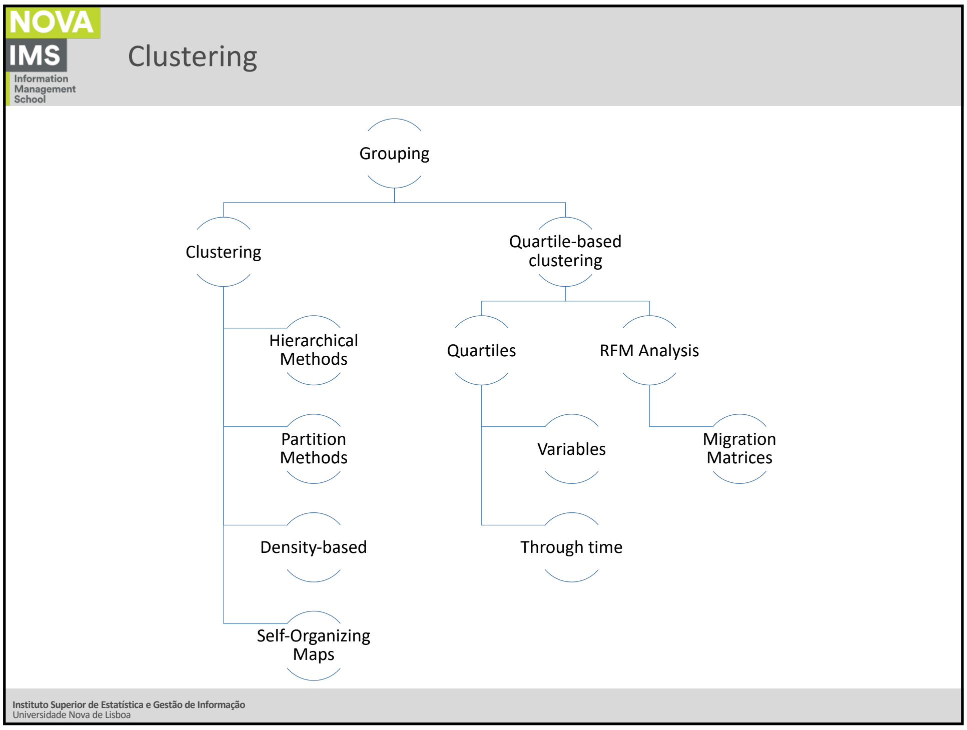
28



NOVA IMS Information Management School

Choose the Algorithm

29

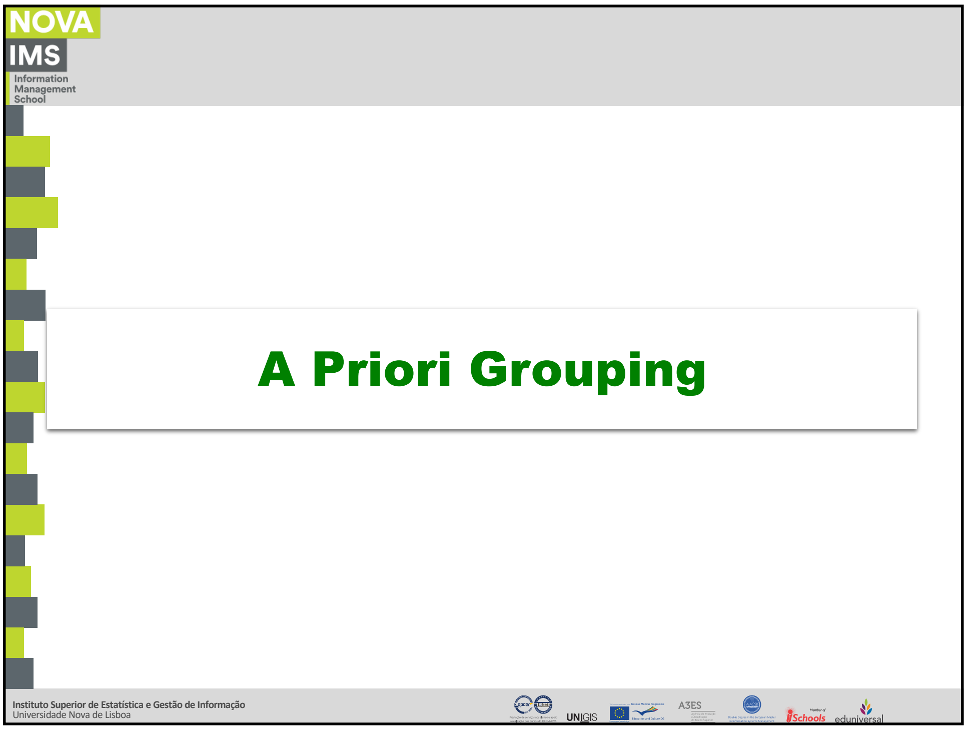


NOVA IMS Information Management School

Clustering

- Grouping
  - Clustering
    - Hierarchical Methods
    - Partition Methods
    - Density-based
    - Self-Organizing Maps
  - Quartile-based clustering
    - Quartiles
      - Variables
      - Through time
    - RFM Analysis
      - Migration Matrices

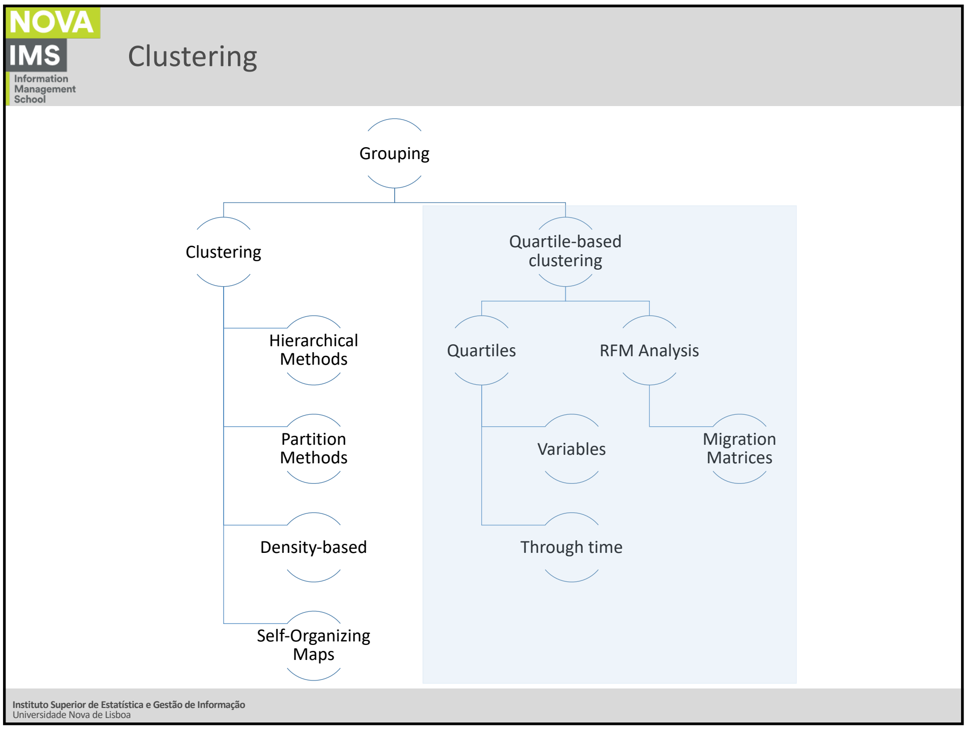
30



NOVA IMS Information Management School

A Priori Grouping

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

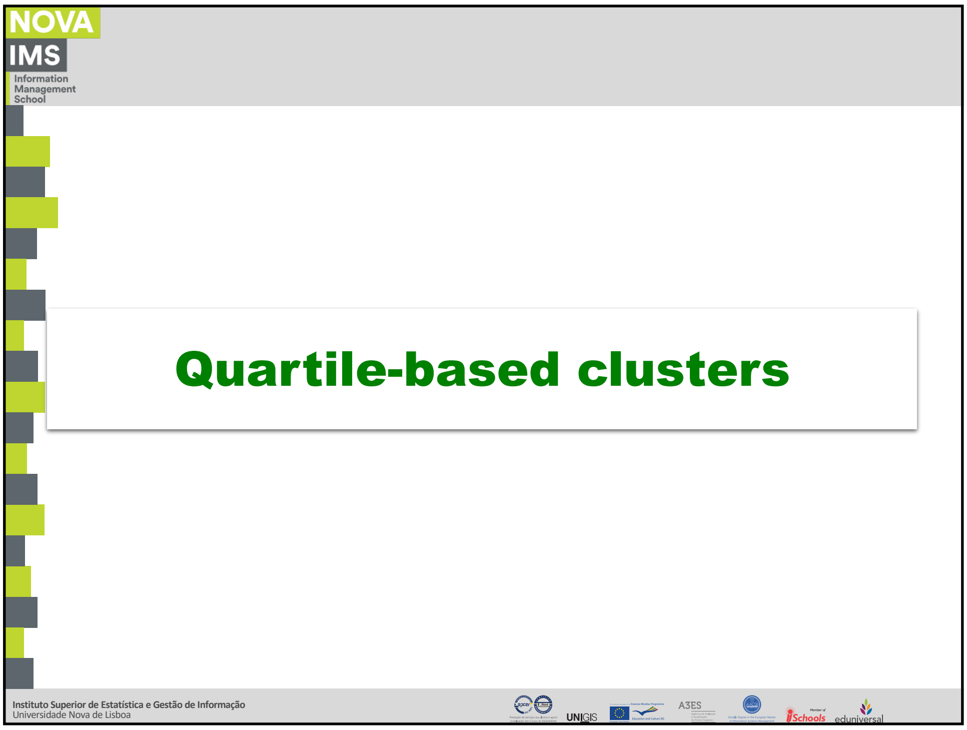


NOVA IMS Information Management School

Clustering

- Grouping
  - Clustering
    - Hierarchical Methods
    - Partition Methods
    - Density-based
    - Self-Organizing Maps
  - Quartile-based clustering
    - Quartiles
      - Variables
      - Through time
    - RFM Analysis
      - Migration Matrices

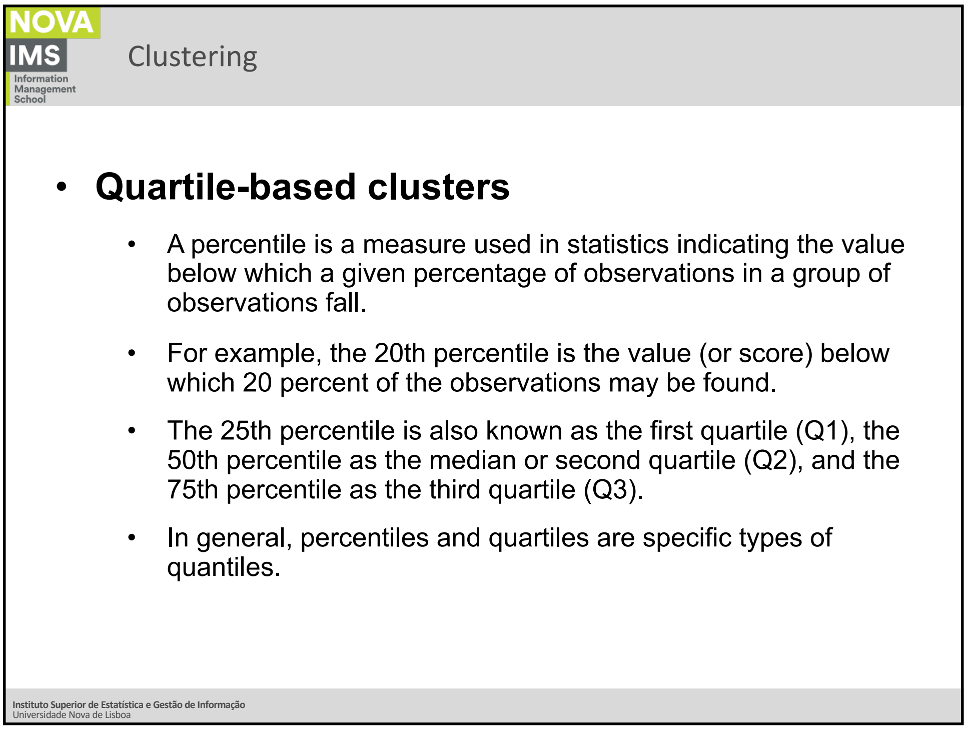
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa



Quartile-based clusters

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

33

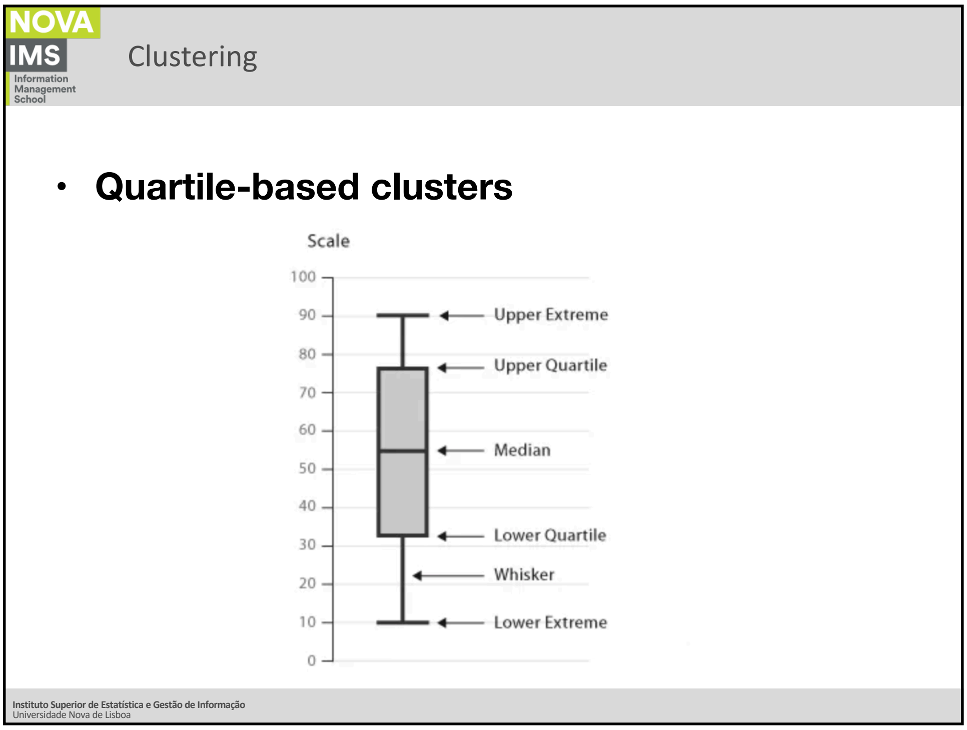


Clustering

- **Quartile-based clusters**
  - A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
  - For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found.
  - The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3).
  - In general, percentiles and quartiles are specific types of quantiles.

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

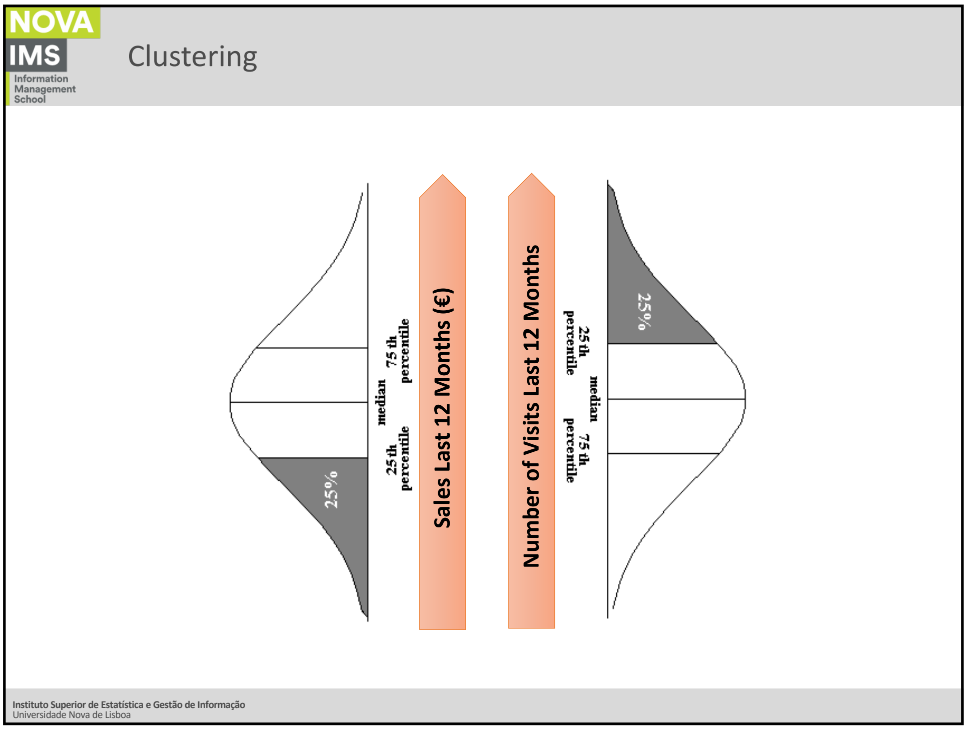
34



NOVA IMS Information Management School

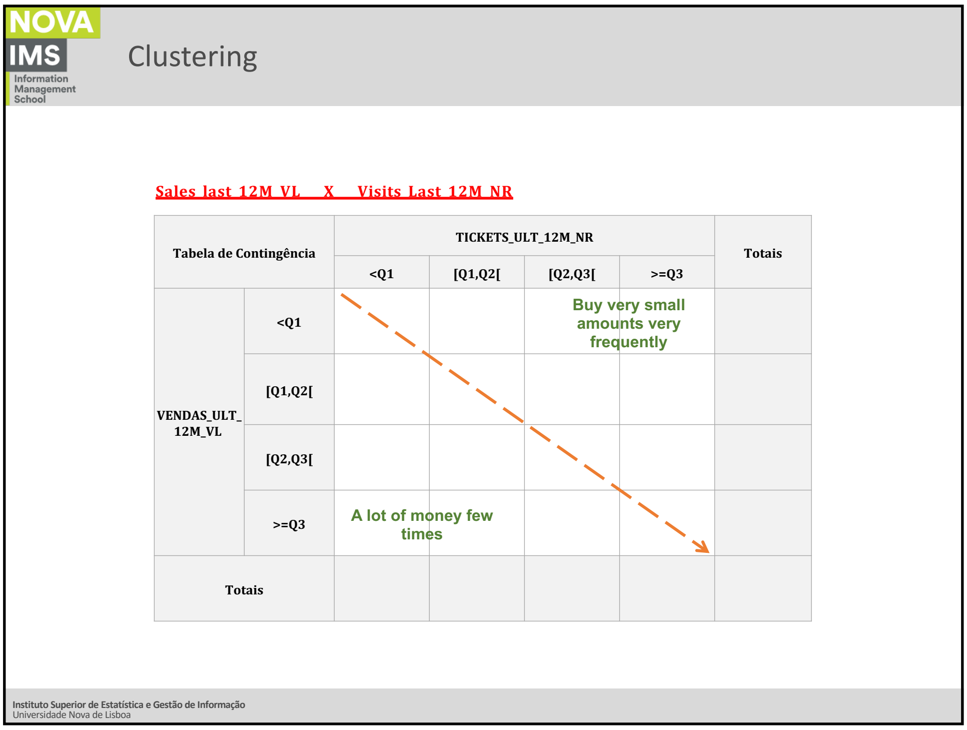
Clustering

- **Quartile-based clusters**



NOVA IMS Information Management School

Clustering



NOVA IMS Information Management School

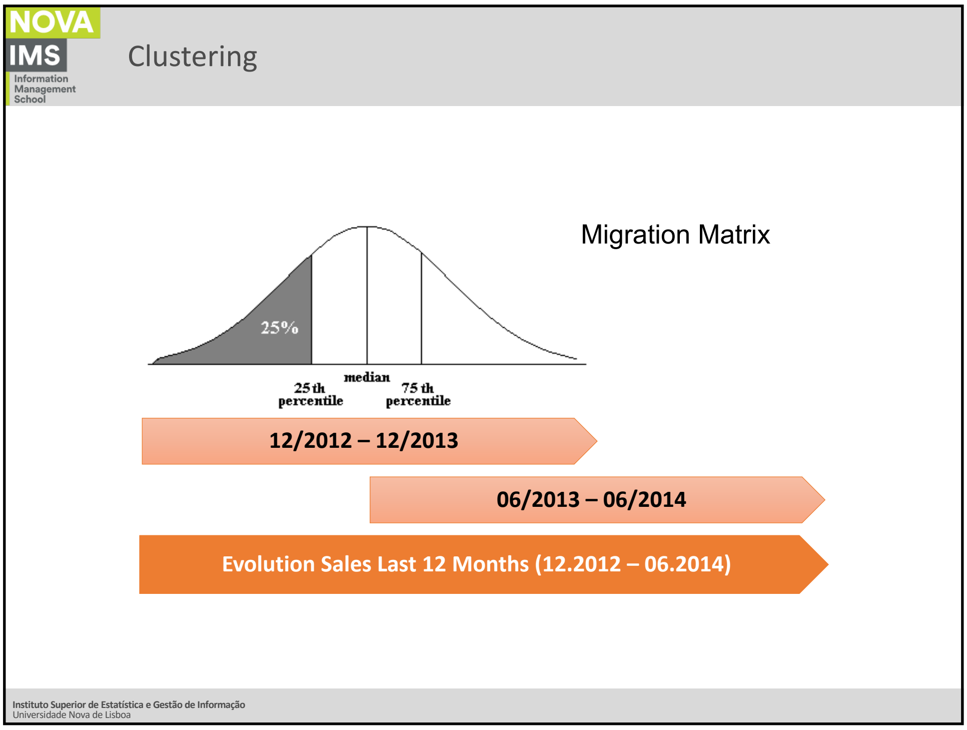
Clustering

Sales last 12M VL   X   Visits Last 12M NR

Tabela de Contingência		TICKETS_ULT_12M_NR				Totais
		<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
VENDAS_ULT_12M_VL	<Q1			Buy very small amounts very frequently		
	[Q1,Q2[					
	[Q2,Q3[					
	>=Q3	A lot of money few times				
Totais						

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

37



NOVA IMS Information Management School

Clustering

Migration Matrix

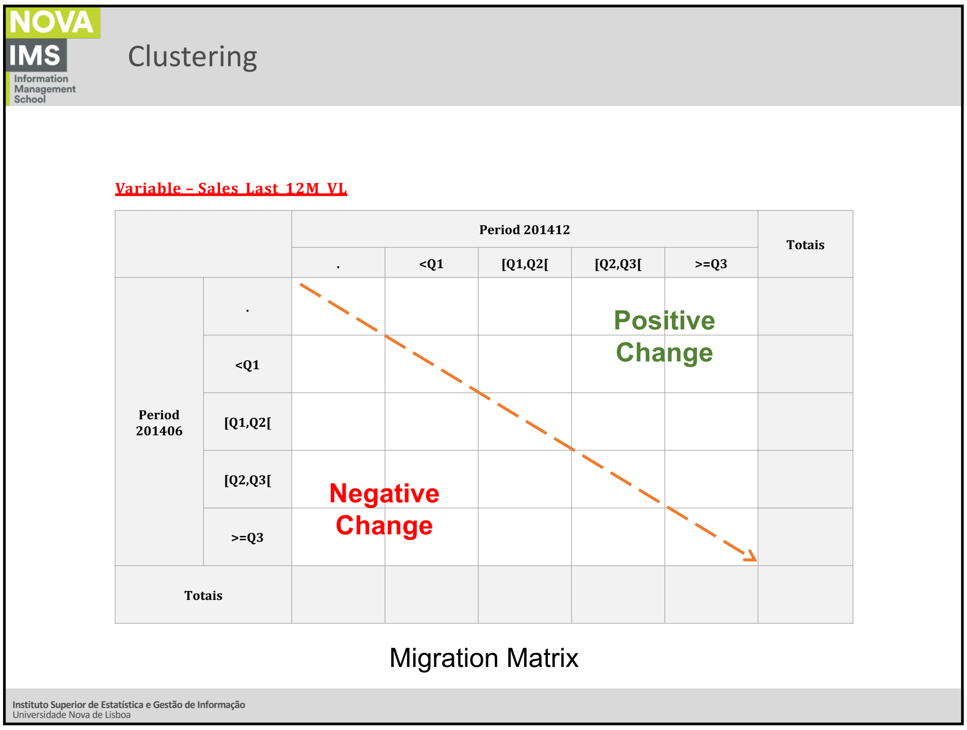
12/2012 – 12/2013

06/2013 – 06/2014

Evolution Sales Last 12 Months (12.2012 – 06.2014)

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

38

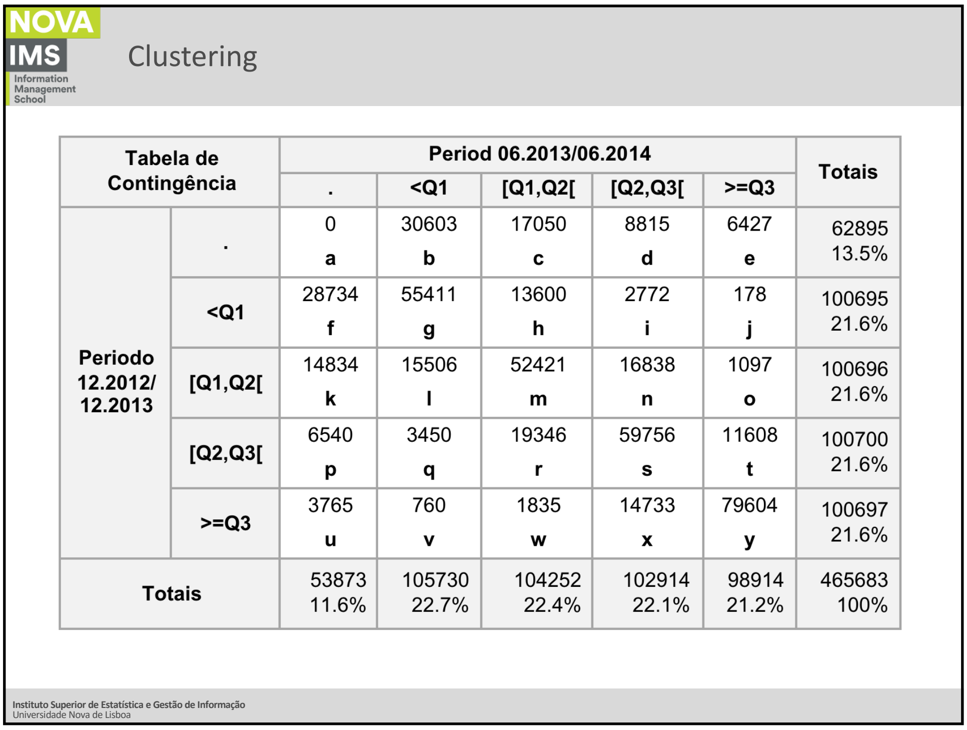


Clustering

Variable – Sales Last 12M VL

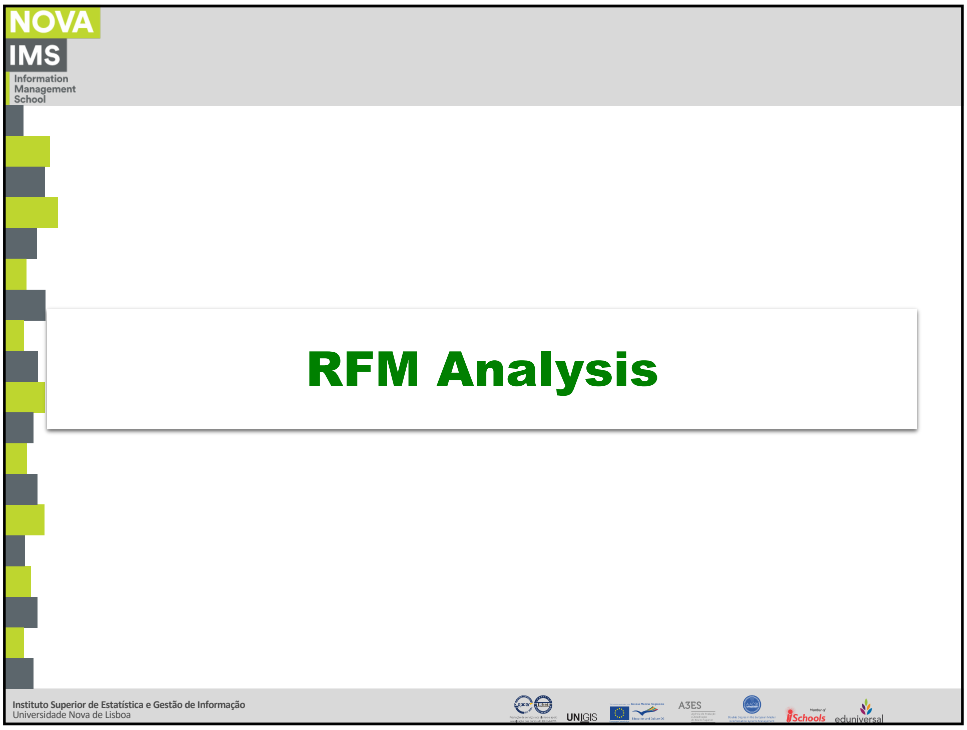
		Period 201412					Totais
		.	<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
Period 201406	.						
	<Q1						
	[Q1,Q2[						
	[Q2,Q3[						
	>=Q3						
Totais							

Migration Matrix



Clustering

Tabela de Contingência		Period 06.2013/06.2014					Totais
		.	<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
Periodo 12.2012/ 12.2013	.	0 a	30603 b	17050 c	8815 d	6427 e	62895 13.5%
	<Q1	28734 f	55411 g	13600 h	2772 i	178 j	100695 21.6%
	[Q1,Q2[	14834 k	15506 l	52421 m	16838 n	1097 o	100696 21.6%
	[Q2,Q3[	6540 p	3450 q	19346 r	59756 s	11608 t	100700 21.6%
	>=Q3	3765 u	760 v	1835 w	14733 x	79604 y	100697 21.6%
Totais		53873 11.6%	105730 22.7%	104252 22.4%	102914 22.1%	98914 21.2%	465683 100%

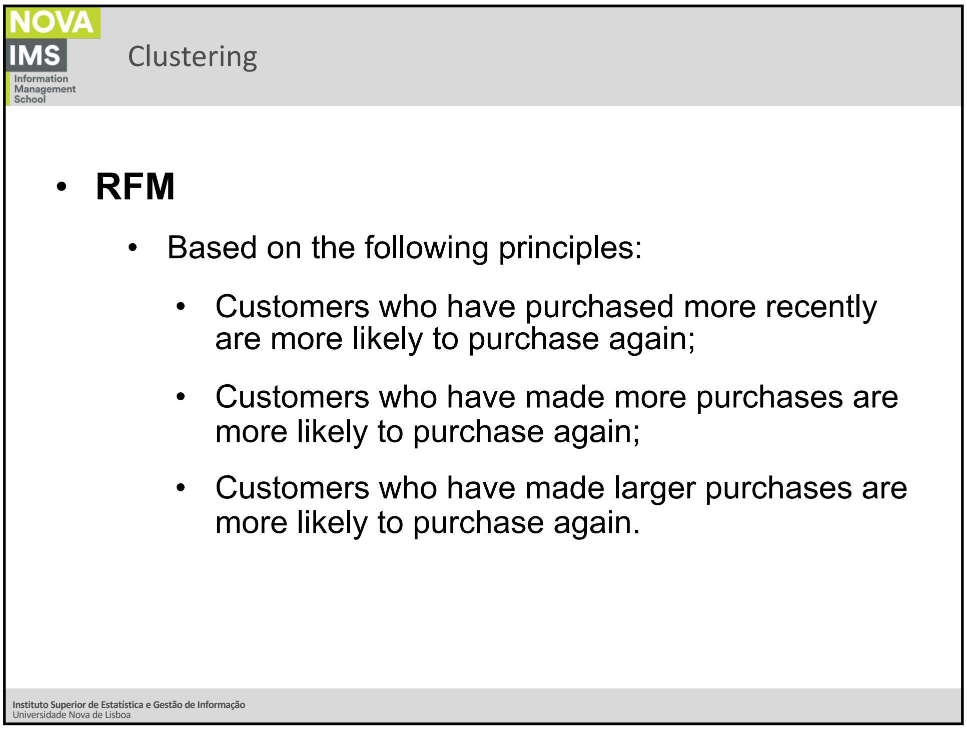


NOVA IMS Information Management School

RFM Analysis

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

41



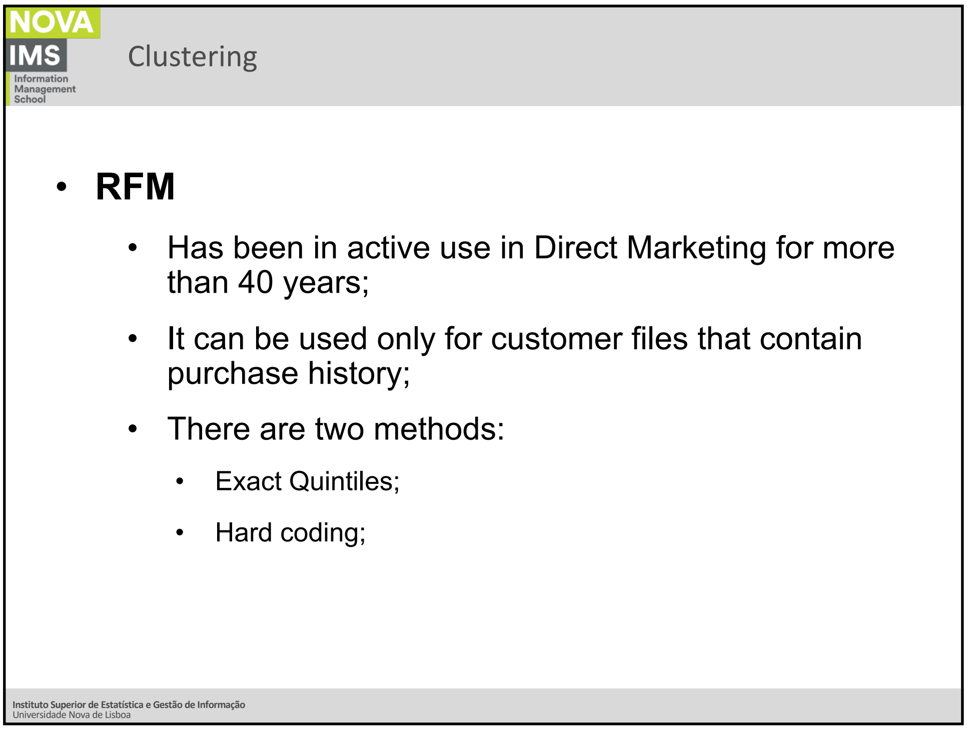
NOVA IMS Information Management School

Clustering

- **RFM**
  - Based on the following principles:
    - Customers who have purchased more recently are more likely to purchase again;
    - Customers who have made more purchases are more likely to purchase again;
    - Customers who have made larger purchases are more likely to purchase again.

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

42



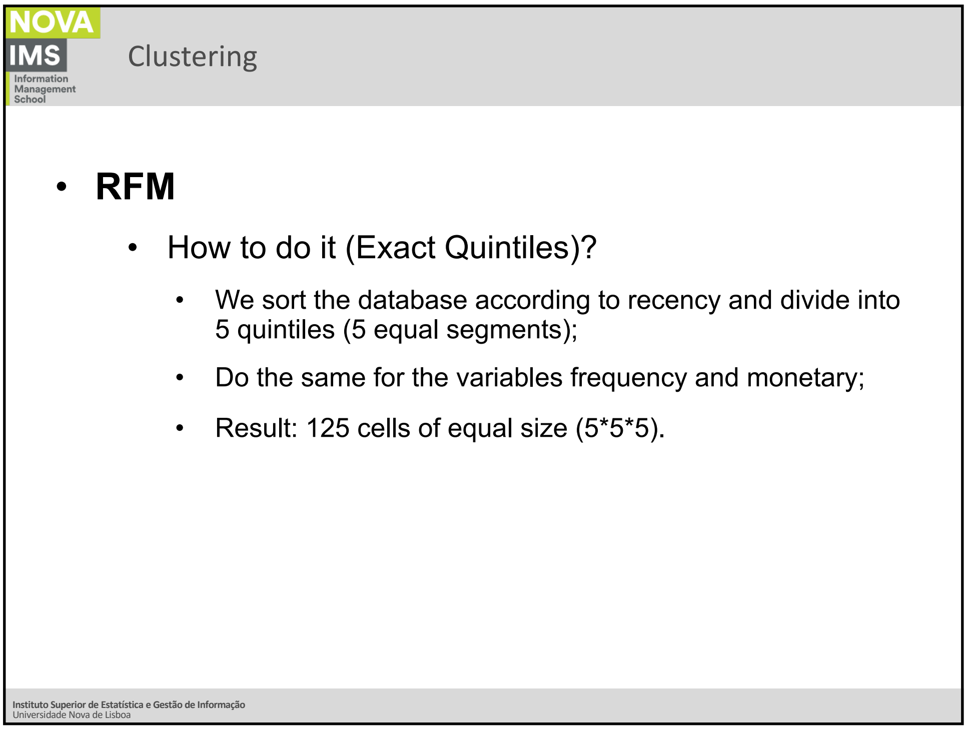
NOVA IMS Information Management School

Clustering

- **RFM**
  - Has been in active use in Direct Marketing for more than 40 years;
  - It can be used only for customer files that contain purchase history;
  - There are two methods:
    - Exact Quintiles;
    - Hard coding;

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

43



NOVA IMS Information Management School

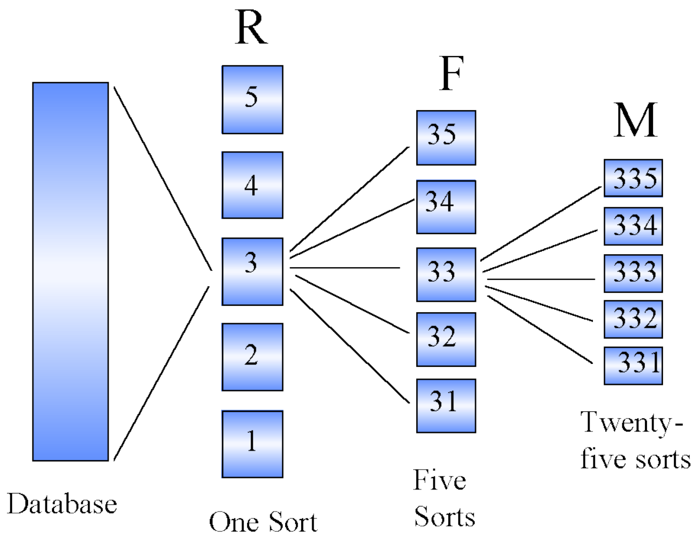
Clustering

- **RFM**
  - How to do it (Exact Quintiles)?
    - We sort the database according to recency and divide into 5 quintiles (5 equal segments);
    - Do the same for the variables frequency and monetary;
    - Result: 125 cells of equal size (5*5*5).

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

44

Clustering



R

F

M

Database

5
4
3
2
1

One Sort

35
34
33
32
31

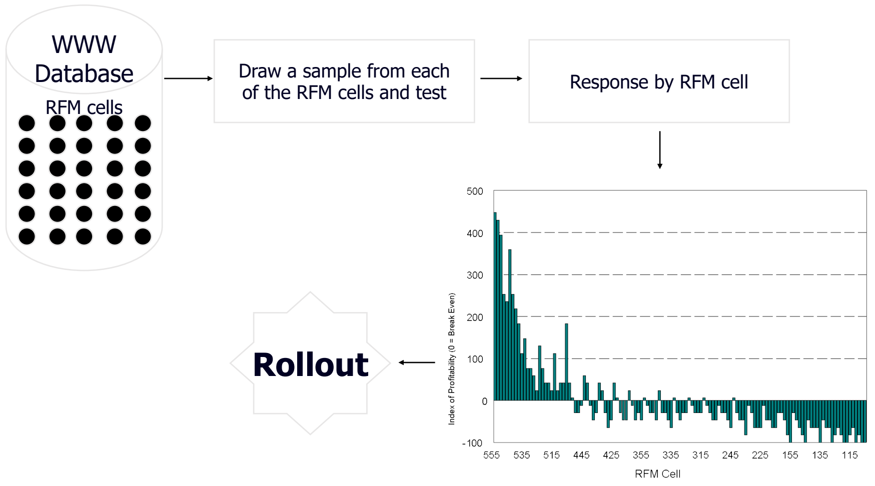
Five Sorts

335
334
333
332
331

Twenty-five sorts

Clustering

Typical use



WWW Database

RFM cells

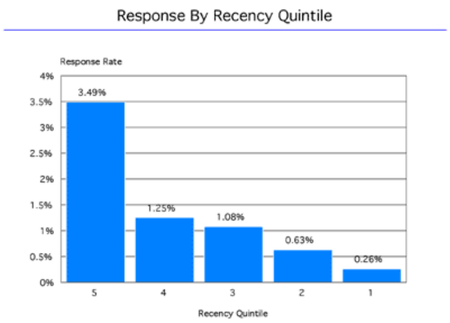
Draw a sample from each of the RFM cells and test

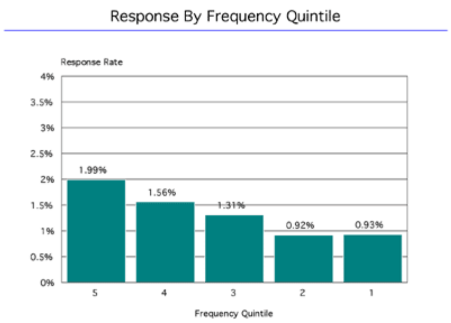
Response by RFM cell

Rollout

NOVA IMS Information Management School

Clustering

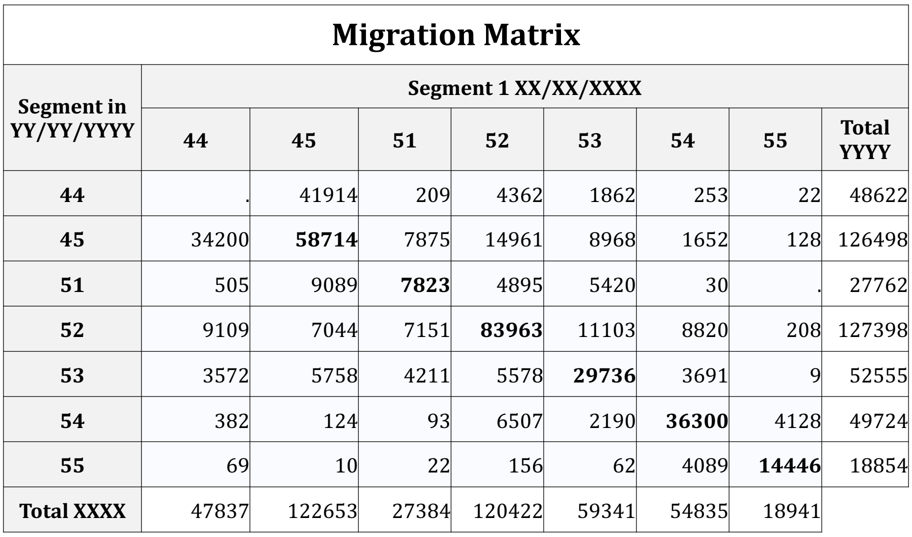




47

NOVA IMS Information Management School

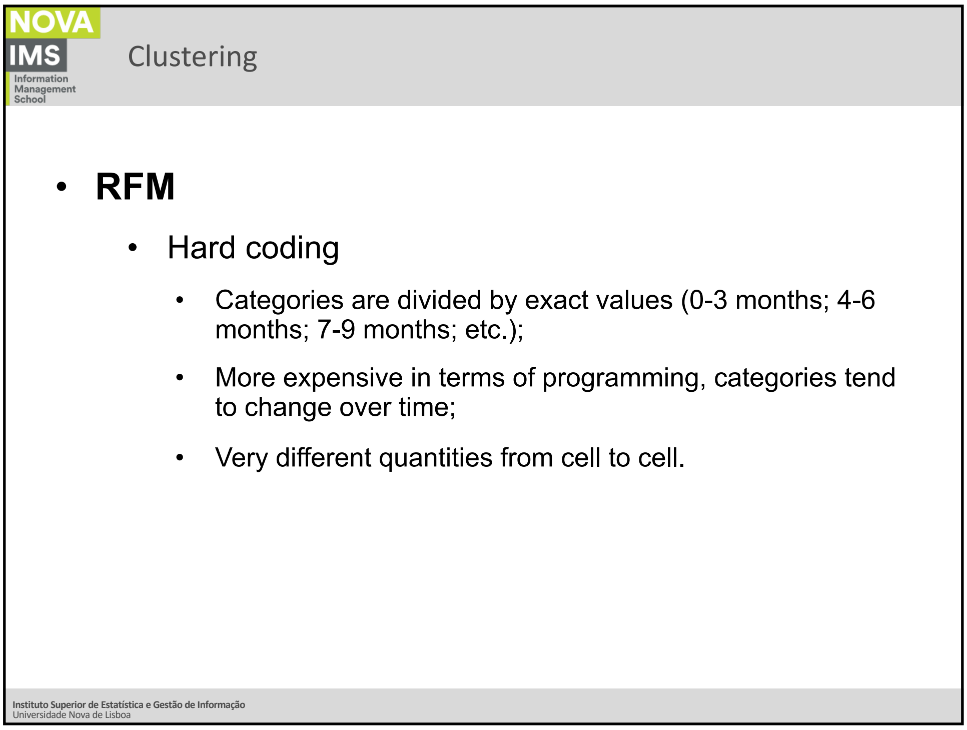
Clustering



Migration Matrix

Segment in YY/YY/YYYY	Segment 1 XX/XX/XXXX							
	44	45	51	52	53	54	55	Total YYYY
44	.	41914	209	4362	1862	253	22	48622
45	34200	**58714**	7875	14961	8968	1652	128	126498
51	505	9089	**7823**	4895	5420	30	.	27762
52	9109	7044	7151	**83963**	11103	8820	208	127398
53	3572	5758	4211	5578	**29736**	3691	9	52555
54	382	124	93	6507	2190	**36300**	4128	49724
55	69	10	22	156	62	4089	**14446**	18854
Total XXXX	47837	122653	27384	120422	59341	54835	18941	

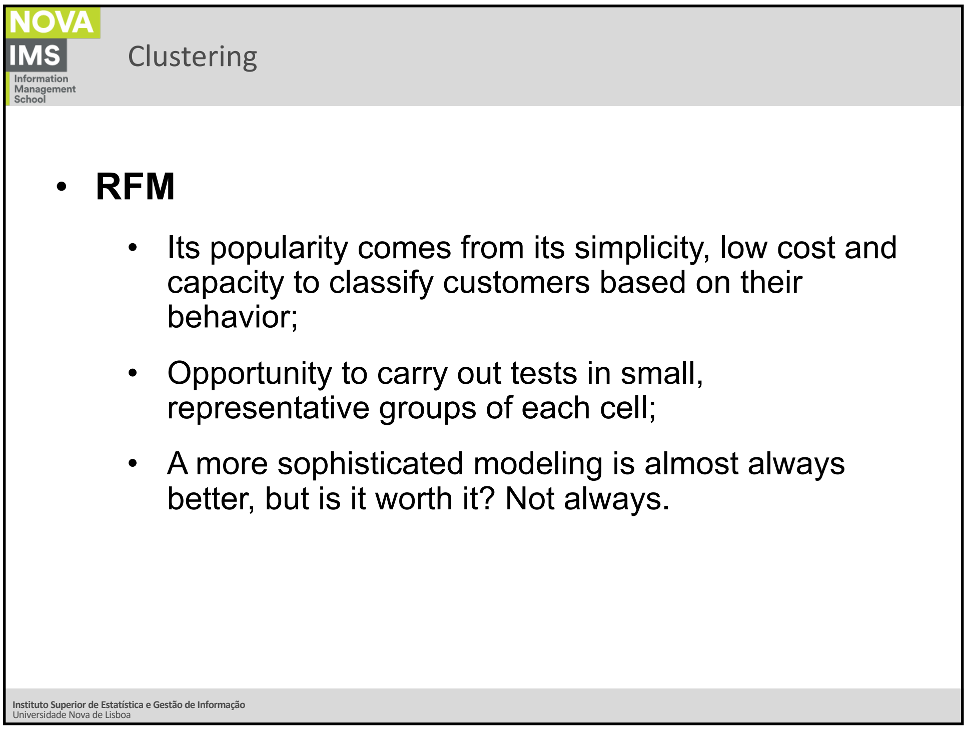
48



Clustering

- **RFM**
  - Hard coding
    - Categories are divided by exact values (0-3 months; 4-6 months; 7-9 months; etc.);
    - More expensive in terms of programming, categories tend to change over time;
    - Very different quantities from cell to cell.

49



Clustering

- **RFM**
  - Its popularity comes from its simplicity, low cost and capacity to classify customers based on their behavior;
  - Opportunity to carry out tests in small, representative groups of each cell;
  - A more sophisticated modeling is almost always better, but is it worth it? Not always.

50

Questions?

51

51