Scuola di Scienze Matematiche, Fisiche e Naturali
Università degli Studi di Firenze

Mining and **Analysis** of **courses** and **students data** about the **Computer Science Degree**

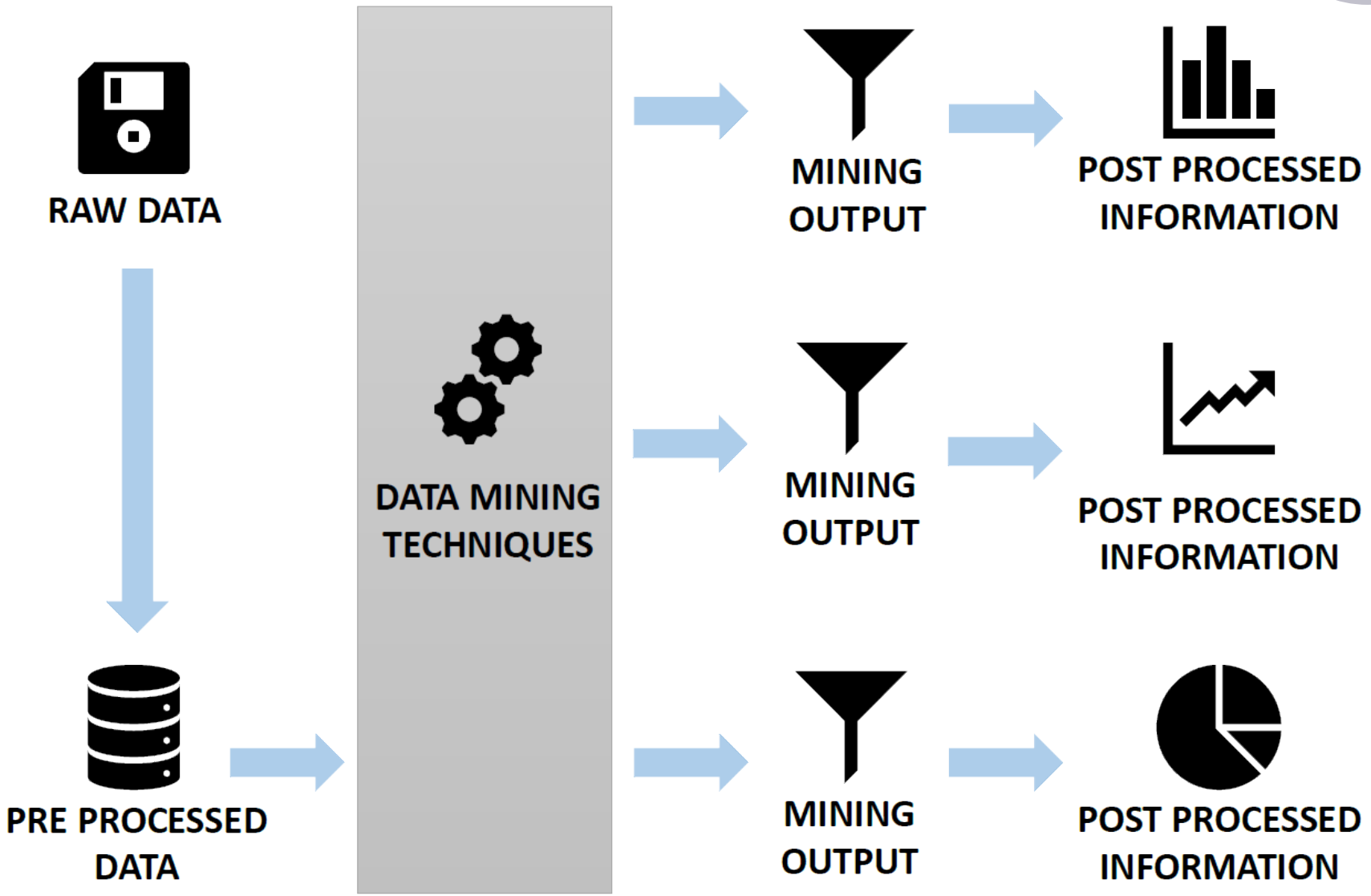
Analisi con tecniche di **Data Mining** su dati relativi a **corsi** e **studenti** del **C. d. L. Triennale in Informatica**

Simone Cipriani
prof. **Donatella Merlini**

12th October 2018

Introduction

A general view about the Data Mining process



Introduction

The choice of appropriate technologies: picking the right tools

- Data Processing: mongoDB

  New tech, noSQL paradigm, good skill to learn!

Introduction

The choice of appropriate technologies: picking the right tools

- Data Processing: mongoDB

  New tech, noSQL paradigm, good skill to learn!

- Data Mining Algorithms: WEKA The University of Waikato

  Open source data mining software/framework;

Introduction

The choice of appropriate technologies: picking the right tools

- Data Processing: mongoDB

  New tech, noSQL paradigm, good skill to learn!

- Data Mining Algorithms: WEKA The University of Waikato

  Open source data mining software/framework;

- Visualization Techniques: R X

  R is powerful, but spreadsheets are easier to use.

Raw Data

What is the nature of the raw material? How can we use it?

- **Data about students**: *anonymous students records* about academic career results, in a certain time span:

A.A.	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
	Coorte 2010						
		Coorte 2011					
			Coorte 2012				
				Coorte 2013			

Raw Data

What is the nature of the raw material? How can we use it?

- **Data about students**: *anonymous students records* about academic career results, in a certain time span:

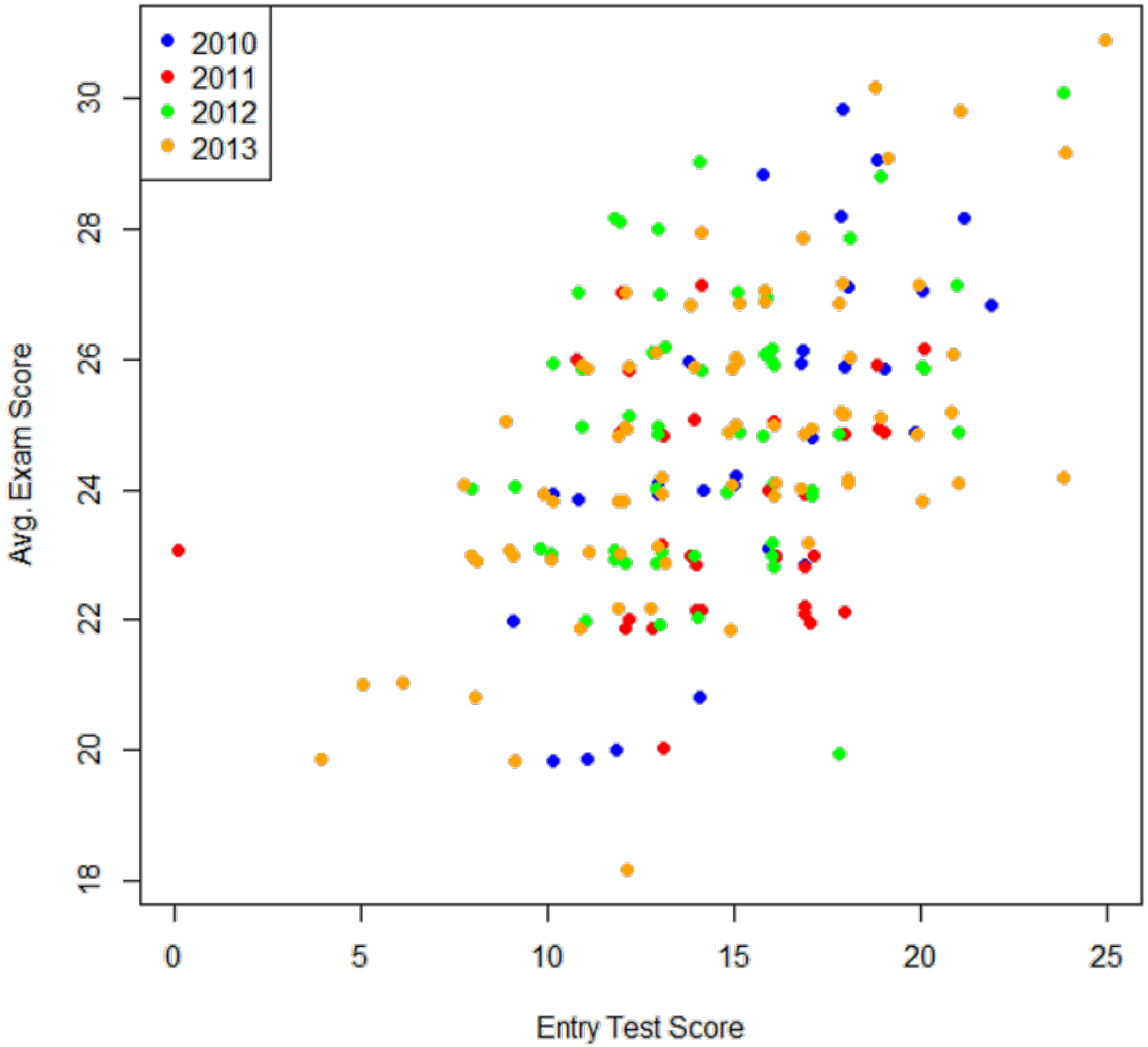
A.A.	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
	Coorte 2010						
		Coorte 2011					
			Coorte 2012				
				Coorte 2013			

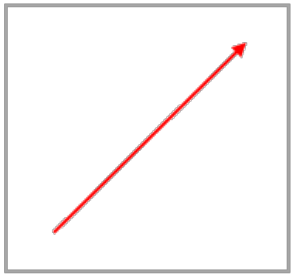
- **Data about courses**: teachings evaluations questionaries compiled by the students, in an *aggregate form*, about those Academical Years:

`2010-2011, 2011-2012, ..., 2016-2017`

Students Data Understanding

Example of visualization technique: interpreting a scatter plot

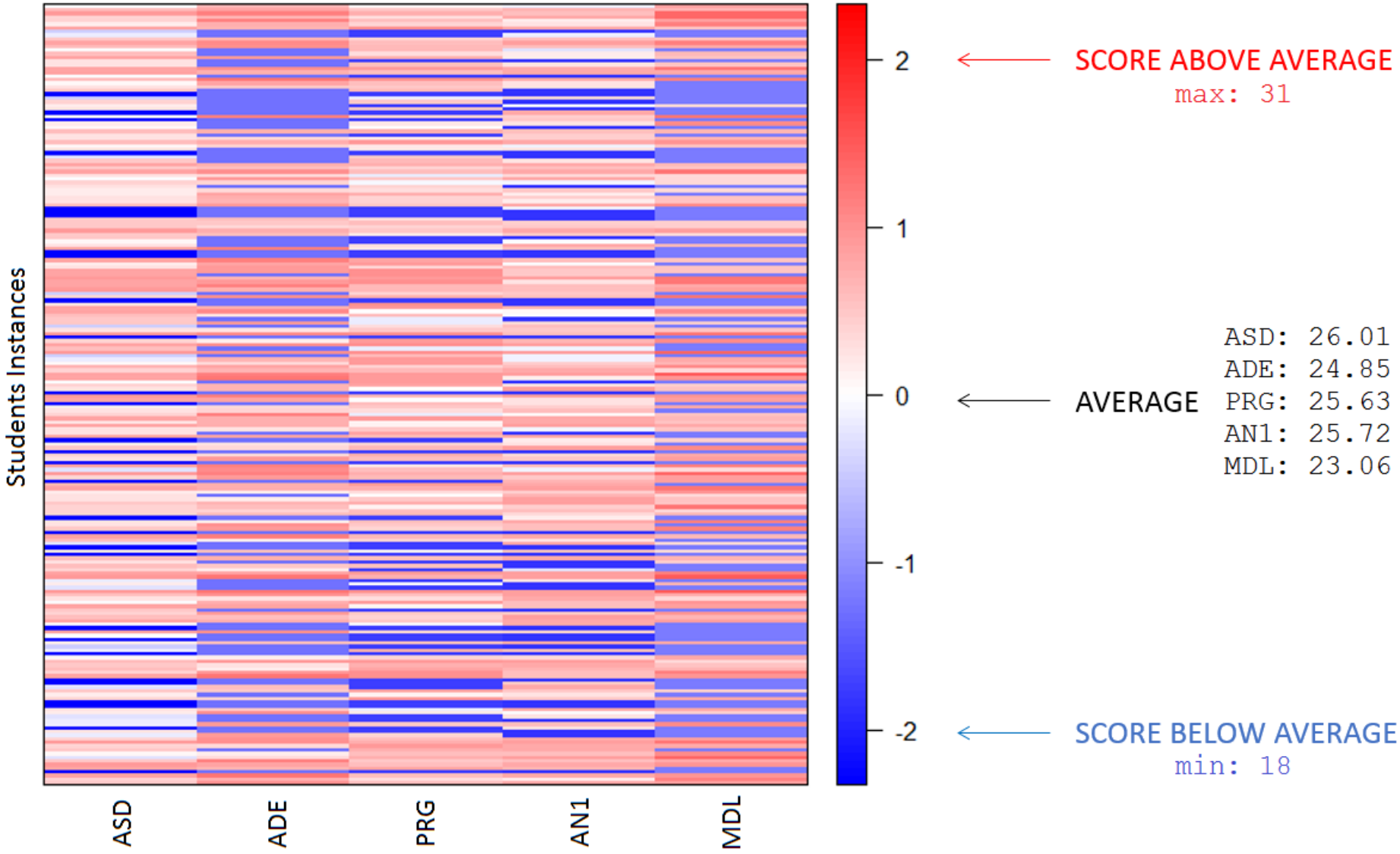




Direct correlation?

Students Data Understanding

Interpreting a std. dev. matrix about first year exam scores



Preprocessing

How can we combine all the available data in a single set?

Raw data sets has to be cleaned, and *disaggregated data* can be aggregated to match the appropriate *primary key*.

Then, they can be joined — for example, like this:

KEYS			STUDENTS DATA						COURSES DATA			
Anno Acc...	Hash Docente/i	Insegnamento	Produtt...	Produtt...	Produtt...	Produtt...	Produtt...	Produtt...	Valutazi...	Valutazi...	Valutazi...	Valutazi...
2015-2016	f6af79bac8bb	INTERPRETI E COM...	45	20	0.36	82.22	26.16	2.8	44	8.25	1.58	94.02
2015-2016	3bbe80b19b92	INFORMATICA TEO...	13	38	0.54	38.46	23.08	3.27	22	6.88	2.52	71.26
2015-2016	7acf3f684e24	RETI DI CALCOLAT...	35	46	0.89	91.43	27.6	2.77	35	7.96	1.97	90.35
2015-2016	fc2f8485971a	CALCOLO NUMERI...	27	0	0	55.56	24.04	3.89	48	6.09	2.83	63.29
2014-2015	f6af79bac8bb	INTERPRETI E COM...	39	33	0.56	82.05	25.92	2.88	28	7.68	1.89	87.43
2014-2015	3bbe80b19b92	INFORMATICA TEO...	17	94	1.41	58.82	23.76	3.81	19	6.21	2.33	69.66
2014-2015	ebf305b5bf29	PROGRAMMAZION...	49	53	1.2	67.35	25.92	3.84	48	6.31	2.4	71.75

Before attemping any data mining technique, data may still need discretization, normalization, etc.

Each analysis needs its own *specific preprocessing*.

Cluster Analysis

How can we find semantic groups among data instances?

How can we obtain a clustering on the joined data set?

- K-Means — Others algorithms as been tried, but K-Means gave the best results;

Cluster Analysis

How can we find semantic groups among data instances?

How can we obtain a clustering on the joined data set?

- K-Means — Others algorithms as been tried, but K-Means gave the best results;
- Euclidean distance metric — Straight line distance between two points in an Euclidean Space;

Cluster Analysis

How can we find semantic groups among data instances?

How can we obtain a clustering on the joined data set?

- K-Means — Others algorithms as been tried, but K-Means gave the best results;
- Euclidean distance metric — Straight line distance between two points in an Euclidean Space;
- Looking for 2 clusters — Deviding teaching courses instances in *good* ones and *not-so-good* ones;

Cluster Analysis

How can we find semantic groups among data instances?

How can we obtain a clustering on the joined data set?

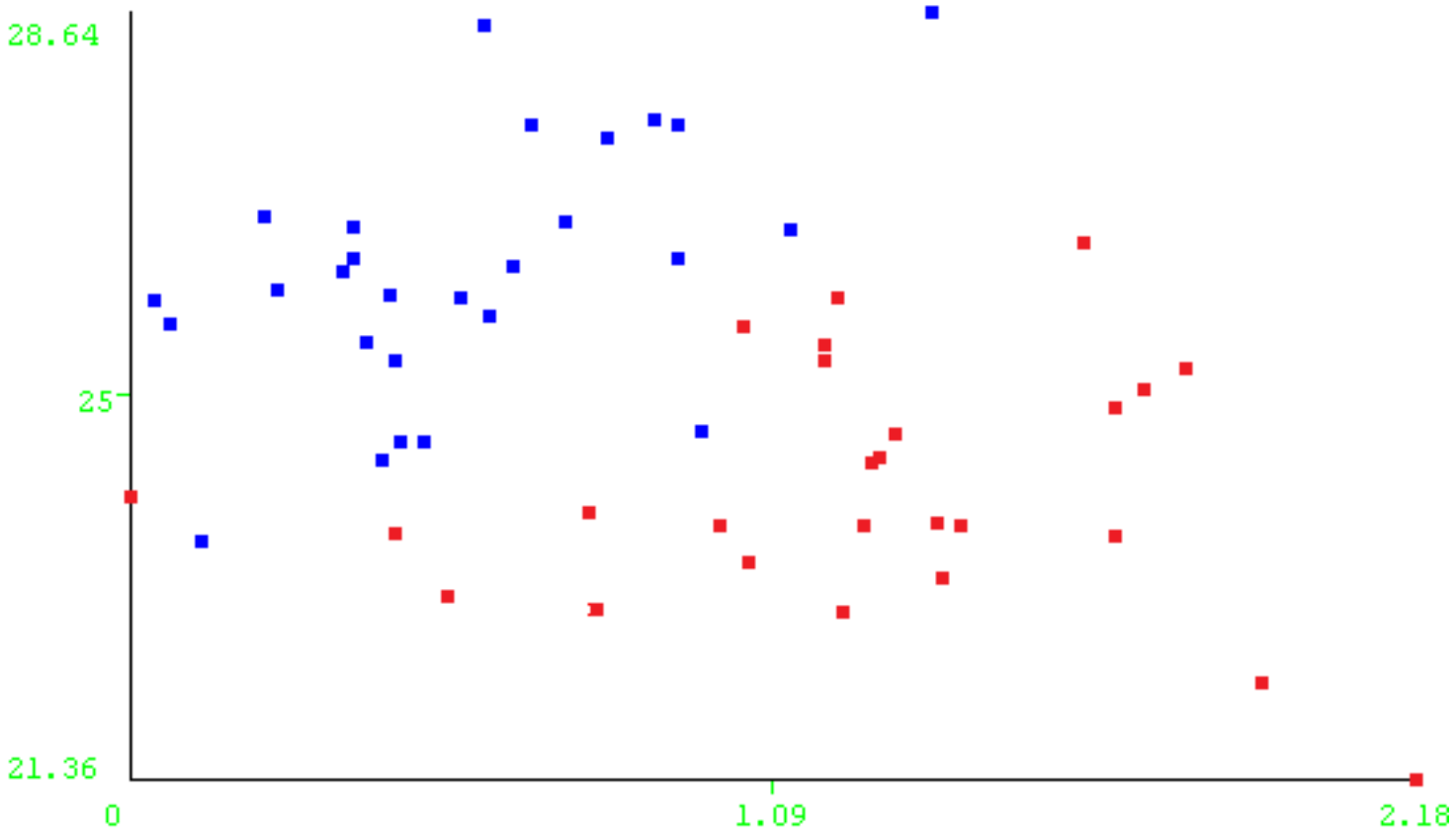
- K-Means — Others algorithms as been tried, but K-Means gave the best results;
- Euclidean distance metric — Straight line distance between two points in an Euclidean Space;
- Looking for 2 clusters — Deviding teaching courses instances in *good* ones and *not-so-good* ones;
- Considering 3 attributes — Each one express a fundamental aspect of the whole data set: *average exam score*, *average teaching evaluation* and *average delay*.

Cluster Analysis

How can we obtain a clustering on the joined data?

Section of the data space:

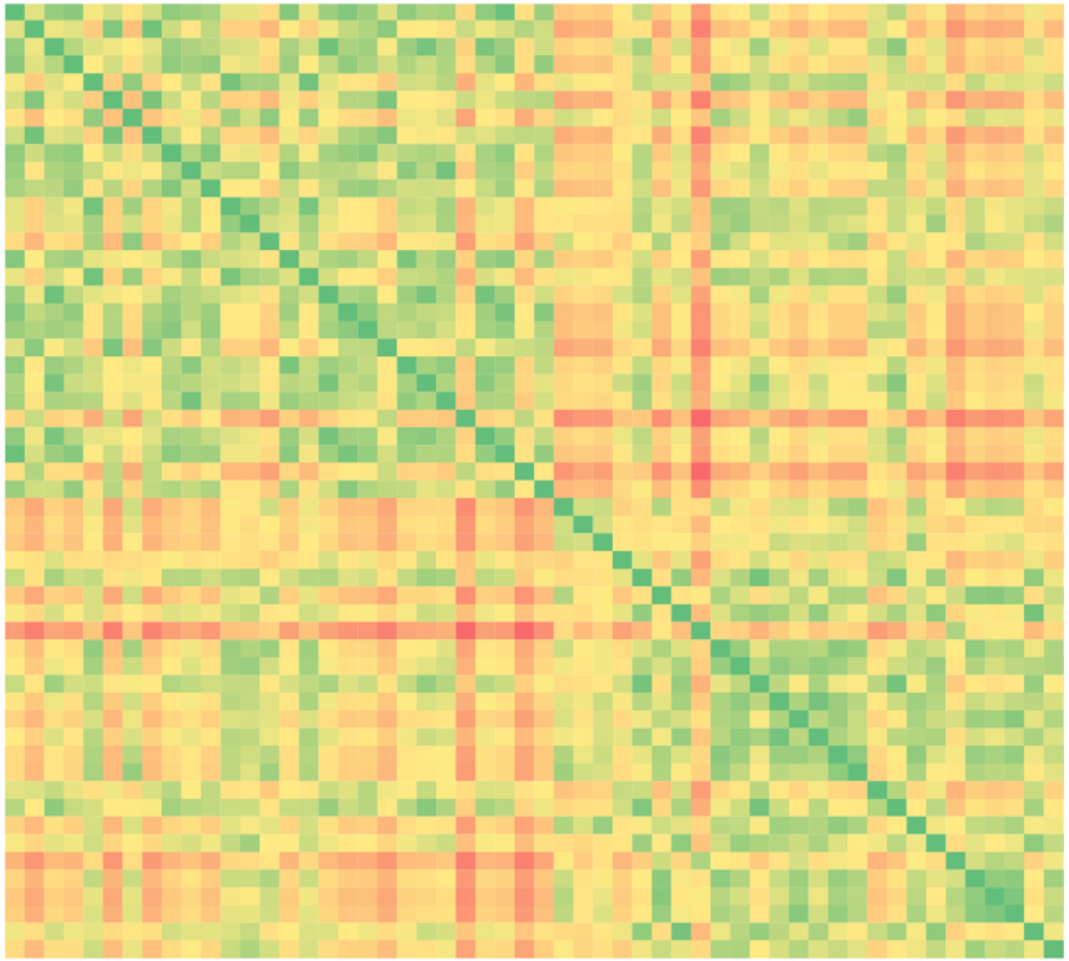
X axis: *average delay*
Y axis: *average exams mark*



Cluster 0: good courses — Cluster 1: bad courses

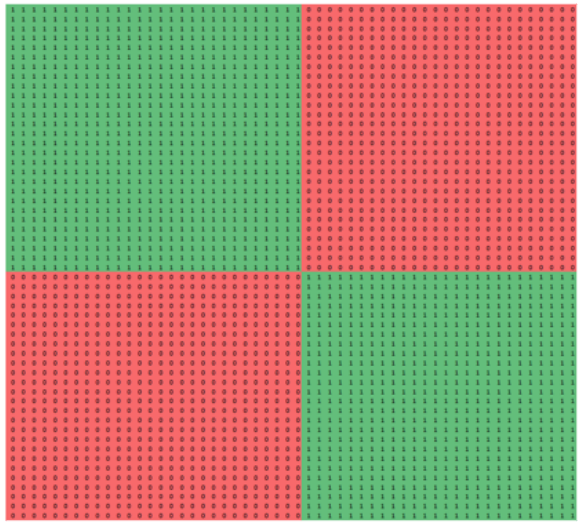
Cluster Analysis

How can we evaluate the obtained clustering?



Euclidean distance matrix
Instances sorted *by cluster*

LOW DISTANCE
BIG DISTANCE



Incidence matrix
Instances sorted *by cluster*

SAME CLUSTER
OTHER CLUSTER

Correlation between matrixes: **negative**

GOOD CLUSTERING

Cluster Analysis

Cluster's composition: analysis and interpretation

Cluster 0 composition:

2010-2011_ASD	2012-2013_ASD
2010-2011_PRG	2013-2014_CAN
2011-2012_BDS	2013-2014_ASD
2011-2012_SOP	2013-2014_BDS
2011-2012_ASD	2013-2014_PRG
2011-2012_MDP	2013-2014_AN1
2011-2012_ALG	2013-2014_MDP
2011-2012_AN1	2013-2014_REC
2012-2013_AN1	2014-2015_INC
2012-2013_ITE	2014-2015_MDP
2012-2013_CAN	2014-2015_BDS
2012-2013_PRG	2014-2015_REC
2012-2013_REC	2014-2015_CAN
2015-2016_REC	2015-2016_INC

Cluster 1 composition:

2010-2011_ADE	2013-2014_SOP
2010-2011_AN2	2013-2014_ADE
2010-2011_MDL	2013-2014_MDL
2011-2012_PRC	2013-2014_ALG
2011-2012_MDL	2013-2014_ITE
2011-2012_PRG	2013-2014_PRC
2011-2012_ADE	2014-2015_ITE
2012-2013_MDL	2014-2015_PRC
2012-2013_ALG	2014-2015_SOP
2012-2013_BDS	2014-2015_ALG
2012-2013_MDL	2015-2016_ITE
2012-2013_PRC	2015-2016_CAN
2012-2013_ADE	
2012-2013_SOP	

*Which courses have **all** their instances in a cluster?*

Only in **Cluster 0**:

ASD – Algoritmi e Strutture Dati
INC – Interpreti e Compilatori
REC – Reti di Calcolatori

Only in **Cluster 1**:

ADE – Architetture degli Elaboratori
MDL – Matematica Discreta e Logica
PRC – Programmazione Concorrente

Associative Rules Analysis

Looking for implications among the dataset's attributes

What do we need to perform an associative analysis?

- Apriori algorithm — It uses an heuristic technique to render the *candidate generation problem* computable;

Associative Rules Analysis

Looking for implications among the dataset's attributes

What do we need to perform an associative analysis?

- Apriori algorithm — It uses an heuristic technique to render the *candidate generation problem* computable;
- Confidence metric — A good one is lift, as it values a rule ability to predict cases comparing it against the odds of a random prediction.

Associative Rules Analysis

Looking for implications among the dataset's attributes

What do we need to perform an associative analysis?

- Apriori algorithm — It uses an heuristic technique to render the *candidate generation problem* computable;
- Confidence metric — A good one is lift, as it values a rule ability to predict cases comparing it against the odds of a random prediction.
- Discretization — We need *discrete attributes* in order to find logical implications between them;

Associative Rules Analysis

Looking for implications among the dataset's attributes

What do we need to perform an associative analysis?

- Apriori algorithm — It uses an heuristic technique to render the *candidate generation problem* computable;
- Confidence metric — A good one is lift, as it values a rule ability to predict cases comparing it against the odds of a random prediction.
- Discretization — We need *discrete attributes* in order to find logical implications between them;
- Focus on a few attributes — The same ones chosen for clustering will do: *exam score*, *delay* and teaching evaluation.

Associative Rules Analysis

Looking for implications among the dataset's attributes: results

The best mined rules are all *double implications!* $X \rightarrow Y$	***Confidence*** $\frac{supp(X\ U\ Y)}{supp(X)}$	***Lift*** $\frac{supp(X\ U\ Y)}{supp(X) \times supp(Y)}$
• Low Delay -> Good Evaluation • Good Evaluation -> Low Delay	0.59 0.48	1.51
• High Delay -> Low Marks • Low Marks -> High Delay	0.63 0.42	1.41
• Good Marks -> Good Evaluation • Good Evaluation -> Good Marks	0.52 0.67	1.33
• Low Marks -> Low Evaluation • Low Evaluation -> Low Marks	0.75 0.55	1.23
• High Delay -> Low Evaluation • Low Evaluation -> High Delay	0.75 0.36	1.23

Frequent Sequential Patterns

Looking for meaningful frequent patterns in exams sequences

Which exams are skipped the most by the students?

- Deeper preprocessing — We need to transform *students* data in a totally different way;

Frequent Sequential Patterns

Looking for meaningful frequent patterns in exams sequences

Which exams are skipped the most by the students?

- Deeper preprocessing — We need to transform *students* data in a totally different way;
- GSP algorithm — Apriori-based algorithm to extract *frequent sequential patterns* from preprocessed data;

Frequent Sequential Patterns

Looking for meaningful frequent patterns in exams sequences

Which exams are skipped the most by the students?

- Deeper preprocessing — We need to transform *students* data in a totally different way;
- GSP algorithm — Apriori-based algorithm to extract *frequent sequential patterns* from preprocessed data;
- Clever post processing — Which of those patterns are interesting? What do they mean?

Associative Rules Analysis

Looking for implications among the dataset's attributes

Example of a frequent, unusual pattern:

`3_2_CN, 3_2_ITE, 2_2_FIG`

...which stands for:

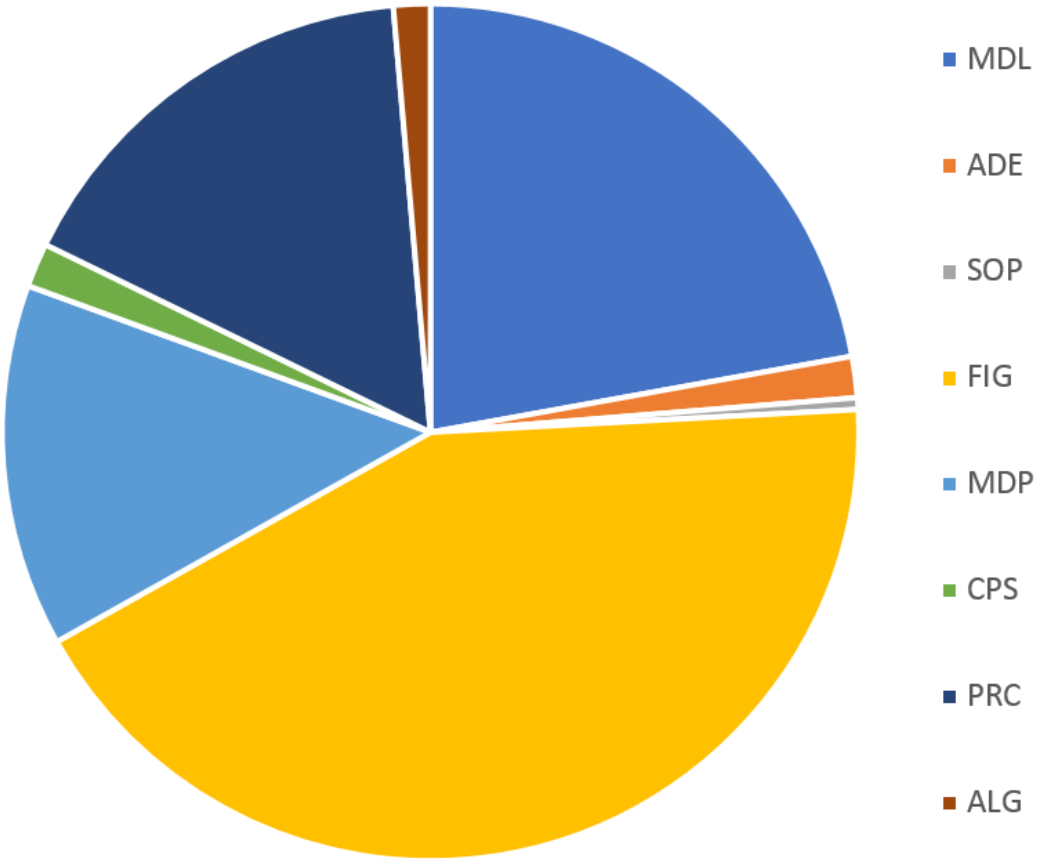
`Calcolo Numerico, Informatica Teorica, Fisica Generale`

`Fisica Generale` is out of place, for it is a 2nd year exam done after some 3rd year exams.

Associative Rules Analysis

Looking for implications among the dataset's attributes

Counting how many times an exams is out of place and making proportions with all out-of-place exams, we get...



That’s all!

…but we only scratched the surface…

Thank you for your time.