UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Profiling

Tomàs Aluja

Barcelona; January 11th, 2017

Outline

1. Introduction to Data Mining
2. Profiling
3. Application: Profiling of granted credits

INTRODUCTION TO DATA MINING

My data, my treasure

Paradigm: *Any stored data from any process always contain information about the generating mechanism(s) (**statistical regularity**).*

Data are routinely stored (and most will never be analyzed)

Data is a treasure for organizations (be aware of the data quality)

Any transactional process con be enhanced by Analytics

How? *Detecting and reporting what is interesting*

Goal: **To reveal the information** (model, patterns, associations, trends, clusters, …) hidden in the data

SQL queries are NOT ENOUGH.

Data Mining consist on transforming Data into Information (usable) (=Knowledge)

Data Mining = (exploratory) Statistics and Machine Learning

Some classic definitions …

Statistics: Methodology for extracting information from data and expressing the amount of uncertainity in decisions we make"
(C.R. Rao, Statistics and Truth, 1989)

Machine Learning: Data mining is the process of identifying valid, novel, potentially useful and understandable patterns in data
(U. Fayyad 1997, KDDB)

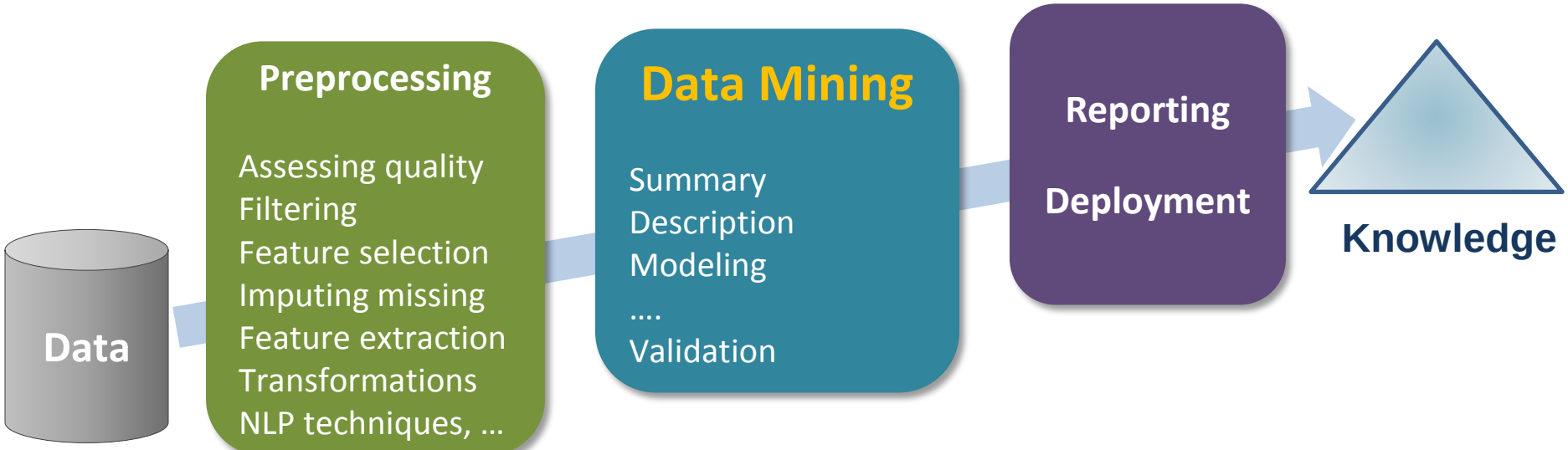
Data Mining is the natural continuation of the exploratory approach of Statistics
(J. Tukey 1962, J.P. Benzecry, 1965)
"le modèle doit suivre les données et non l'inverse"


"It is a capital mistake to theorize before one has data. Insensibly, one begins to twist facts to suit theories, instead of theories to suit facts."
A Scandal in Bohemia (1891) Arthur Conan Doyle

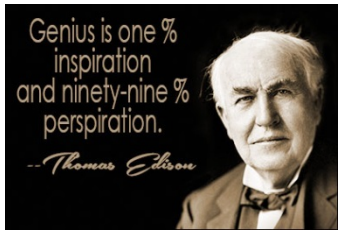
"Without data you are just another person with an opinion."
American Statistician (1988) Edward Deming

Knowledge Discovery in Data Bases

KDDB encompasses the whole process (from Data to Knowledge), whereas DM refers to the technical phase of applying statistical modeling or learning algorithms, but in practice DM is used indistinguishable of the whole process as well.

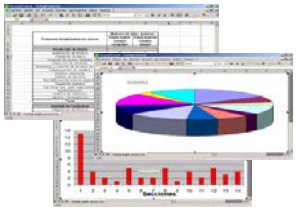


Numeric
Web
Text
Signals
Images



Data Mining
Web Mining
Text Mining
Multimedia Mining

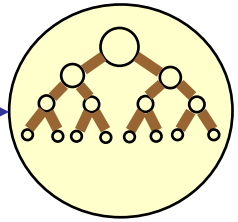
The virtuous cycle of Data Mining (the marketing campaign case)

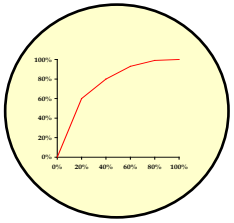


Data from last campaign

Corporate DW

Decision tree





Marketing campaign

Select if .. then ..

Reporting

Target definition

DB enrichment

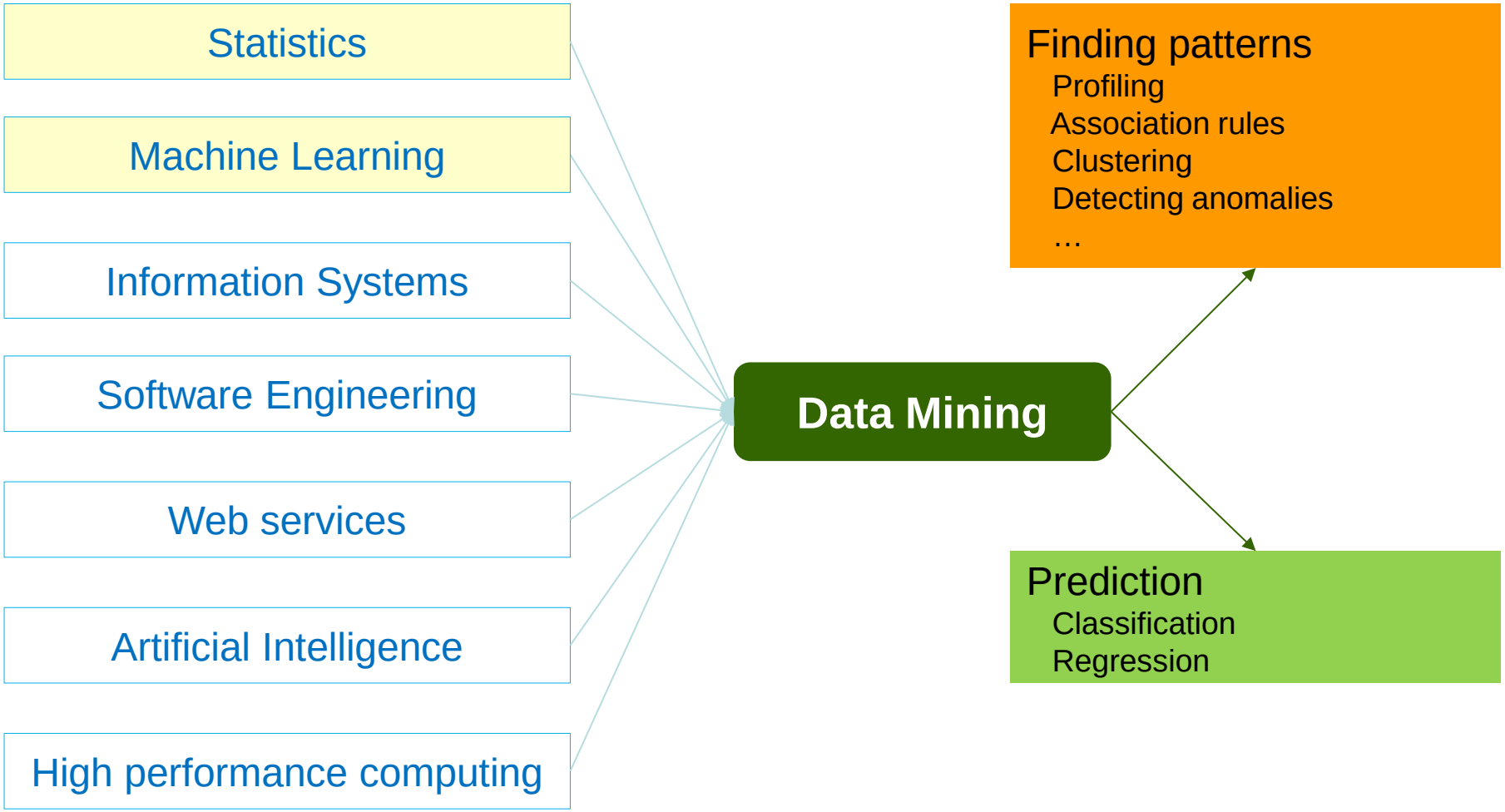
Target

Control

Random

New outcome

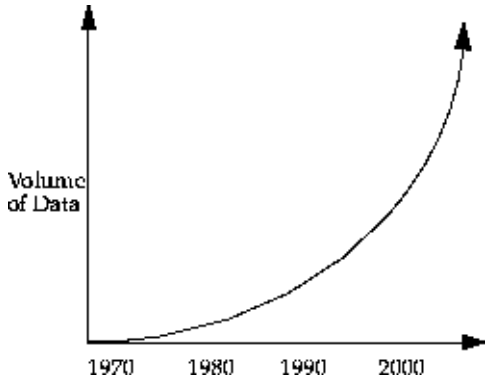
Technologies & goals of Data Mining



The new paradigm: The data flood

> "We are drowning in information but starved for knowledge."
>
> –John Naisbitt, "Megatrends" (1982)

- Exponential increase of data generation and storage
  - Bank, telecom, other business transactions ...
  - Scientific data: genomics, astronomy, health, ...
  - Web, text, and e-commerce
  - Social networks, ...
- Increase in data formats
  - Structured tables: Relational tables
  - Non structured tables, log files, ...
  - Textual data, image data, ...
- Real time
  - Streaming data, sensors,
  - On time personalized assessment, ...



The Big Data movement

The Big Data challenge

"Essentially, all models are wrong, but some are useful" (Empirical model building 1987)
G. E. P. Box 1919-2013

"*All models are wrong and increasingly you can suceed without them*"
Petar Norvig, Google research director
2008 O'Reilly Emerging Technology Conference. San Diego CA.

Google

"Petabytes allows to say: Correlations is enough"
"Correlations supersedes causation, and science can advance even without coherent models, unified theories , or really any mechanism explanation at all"
Chris Anderson
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete
wired.com 2008


WIRED MAGAZINE: 16.07

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

"All models are wrong, but some are useful."

The opportunities of Big Data

Association doesn't preclude causality

- Causality implies than a variable causes another one.
  In survey data (non experimental data) it is very difficult to ascertain (What was the cause of the colza epidemic 1981?)
- Association means that two variables covary together (like height and weight of persons).

Predictive analytics versus finding patterns

- The objective of predictive analytics is to explain the behavior of the response variable by means of a function (model) of its explanatory variables. This is a difficult task. In the context of Big Data modeling is even more difficult since it increases the danger of including into the model *significant noise variables* and *redundant variables*.

- However, we can find easily plenty of *significant associations of the response with explanatory variables*. Some of them would reflect noise, but since there are plenty of them, we will obtain knowledge by accumulation of significant associations (people who like a certain wine are of middle age, high education level and also like …). Strong partisans of Big Data claim that this is enough and we don't need models.

http://www.talent.upc.edu/cat/professionals/premsa/codi/331100/big-data-management-analytics/

The risks of Big Data

Nate Silver,
The signal and the noise
2015

- Data is a treasure for organizations. The acquisition of relevant information is a key factor of competitive advantage. The ability to learn from data provides better adaptation and taking faster better decisions.
- Explosion of information is nowadays seen as a universal solution "*a cure-all solution*". Although the amount of information is increasing exponentially, the amount of useful information certainly isn't. Most it is just redundant or noise.
- With more data to mine, we have more hypothesis to test, but a relatively the same amount of true signals to detect.
- Modeling (predictive analytics) is no an easy task. It is hard to distinguish the signal from the noise. Only succeeded predictions are reported, but not failures (except the famous ones: no chemical weapons in Irak, 2008 real estate bubble, ….). There is information overload. This leads to the problem of false positives; finding patterns in random noise.
- The huge amount of data would not obviate the need for a theoretical framework and scientific method.
- Numbers have no way to speak for themselves. We speak for them. Data tell us what we are ready to understand.
- Data driven predictions can succeed, but quite often fail. Data is useless without a context.

"Nobody believes a theory, except the one that has formulated it.
Everybody believes a figure, except the one who has calculated it"

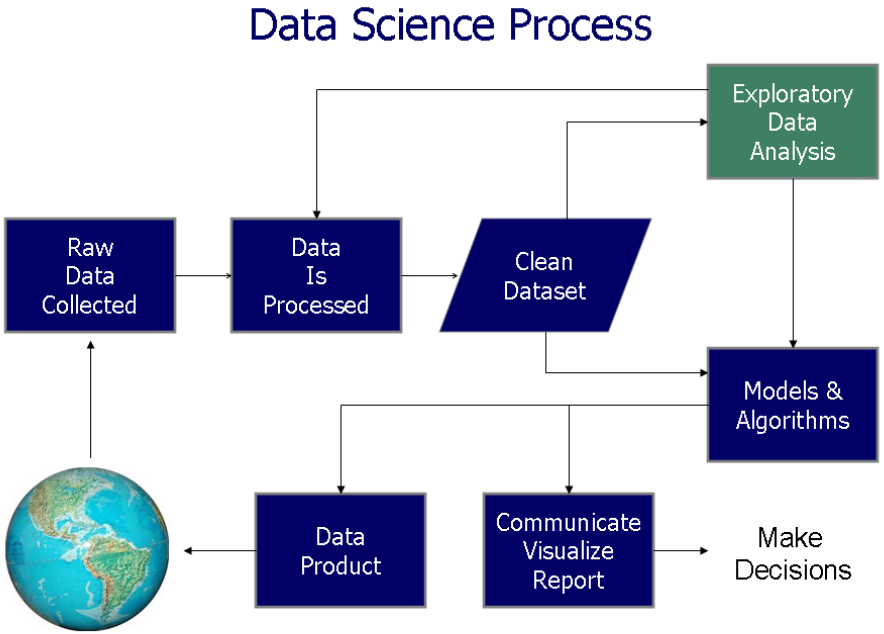
Lord William Beveridge, 1940

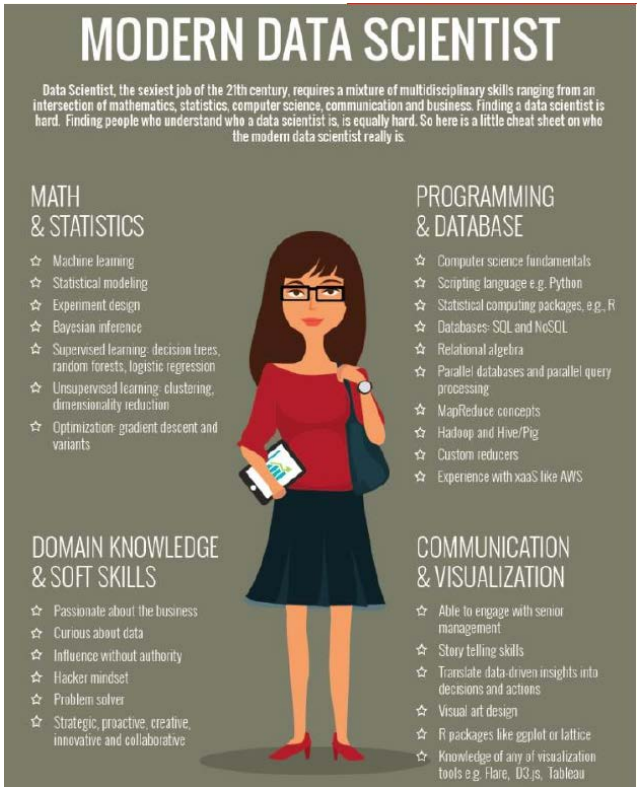
The new concept: Data Science

From Wikipedia

Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining and predictive analytics, as well as Knowledge Discovery in Databases (KDD).

At center stage of data science is the explosion of new data generated from smart devices, web, mobile and social media.

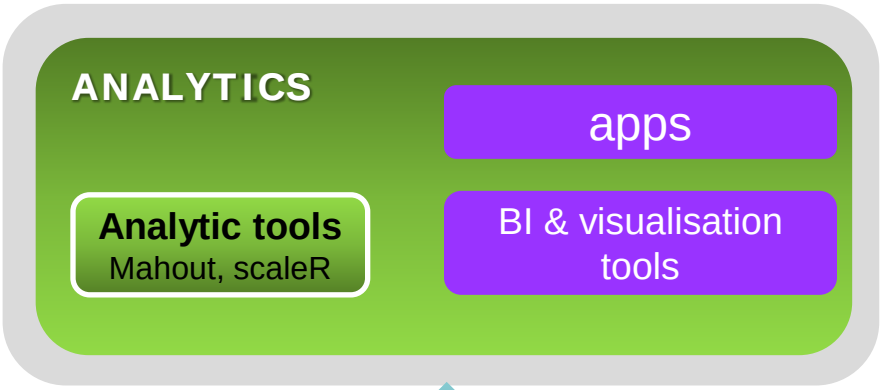




Data Science

Big Data Management

Big Data Analytics



Data scientist

Analytics emphasizes the visual representation of results and the interaction with the user

Analytics = Data Mining + Visual Interaction

PROFILING

Data is stored in tables

Data is multivariate

Rows of table

Represent individuals or instances, Also called records, …, forming the data under study.

Characterized by a predetermined set of attributes

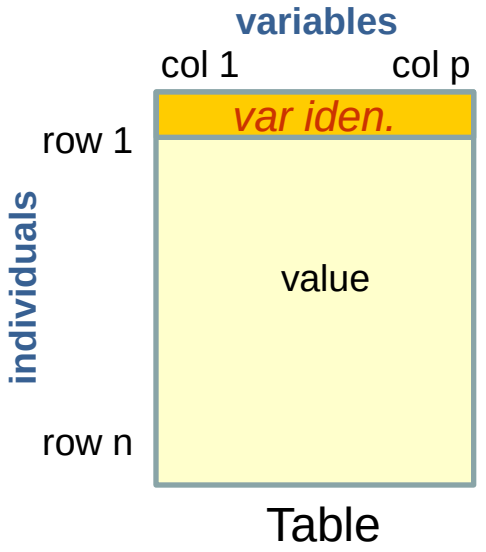
Columns of a table

Each instance is described by a predefined set of features, its variables or "attributes". A variable is a measure of individuals which can take different values (according a probabilistic function).

Possible attribute types ("levels of measurement"): binary, nominal, ordinal, interval, ratio, textual, …

Restriction: Same variables measured in all individuals and in the same order, but different formats are possible (fixed, csv, …), forming a Table.

First rows usually contain the labels of variables



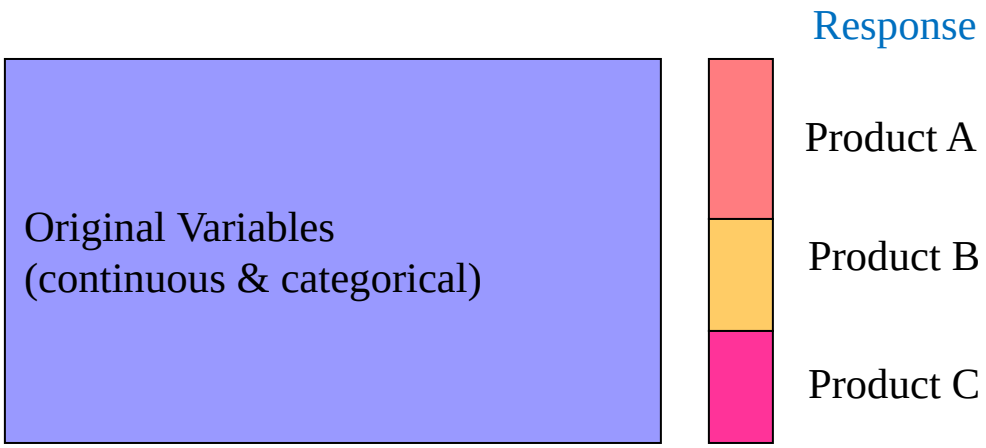
We use all available data. Sampling betrays the spirit of DM, where few individuals may keep the most precious information and only can be justified for alleviating the computational cost

Profiling : what is this?

Profiling. Automatic detection of the specific characteristics of a group of individuals.
What do buyers of my product have different from the others?
Concept characterization

The group of individuals of interest is taken as a modality of a categorical variable (i.e. variable: *which product do you buy?* (being product A one our product of interest).

The remaining variables, either continuous or categorical, are taken as explanatory.



Knowledge by accumulation of evidences of association (not by modeling)

Automatic profiling of groups of individuals

Problem:

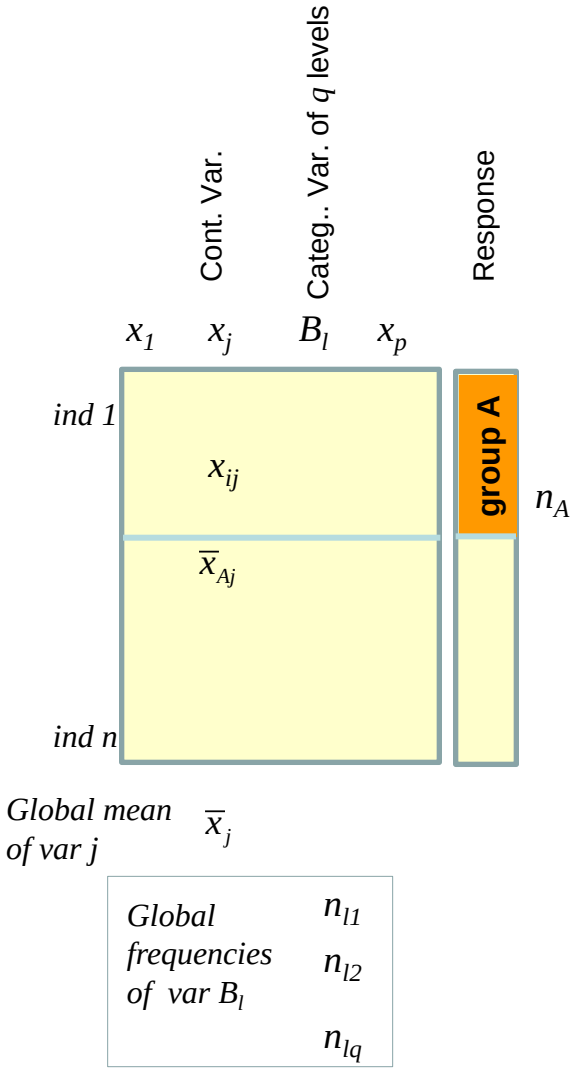
For every group of individuals we want to detect its significant characteristics:

1. Which means (of continuous variables) deviate from what is expected.
2. Which proportions of modalities (of categorical variables) deviate from what is expected.

Tool: Hypothesis test

For each group of individuals, rank the modalities of the categorical explanatory variables according their p-value (ascending). Likewise, rank the continuous variables according their p-value

Select the most significant by a threshold (0.05, 0.01, ..) defined a priori (what matters is the ordering, actual significance depends on the number of individuals)



Review: hypothesis test

We want to test whether Group k of n_k individuals is different from the population with all n individuals.

In our case:

H_0 : Individuals of group k are taken at random

H_1 : Individuals of group k are NOT taken at random

To test a hypothesis we need:

Null hypothesis	$\boldsymbol{H_0}$: denial of the hypothesis (like an evil advocate). = Group k is equal to the population
Alternative hypothesis	$\boldsymbol{H_1}$: the hypothesis we want to validate
Test statistic:	It depends on the problem.
Reference distribution:	Distribution of the *test statistic* if the H_0 is true.
Significance threshold:	Risk that we are ready to incur to reject H_0 when it is true (significance depend on the number of individuals, thus classical statistical thresholds need to be adapted to the n).

Profiling from continuous variables

for every continuous variable:

groups	means	counts
1	$\overline{x}_1$	n_1
⋮	⋮	⋮
K	$\overline{x}_K$	n_K
Global	$\overline{x}$	n
	s^2	

Comparison with a nominal value problem

$$H_0 : \mu_k = \mu \quad k = 1, \ldots, K$$

Test statistic: Difference between the mean in group *k* and the global mean, relativized respect to its random fluctuation.

$$t = \frac{\overline{x}_k - \overline{x}}{\sqrt{\left(1 - \frac{n_k}{n}\right)\frac{s^2}{n_k}}} \sim t_{n-1}$$

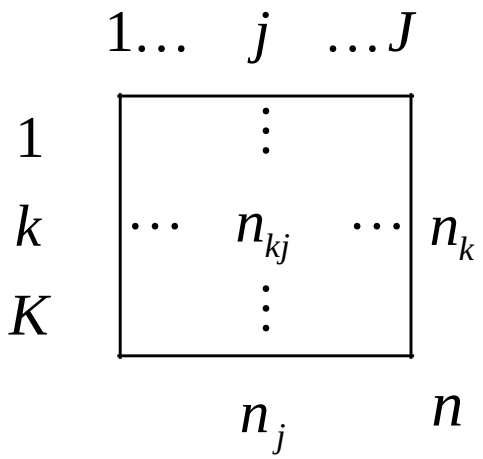
Student's *t*

Rank the continuous variables by p.value (ascending)

Profiling from categorical variables

modalities of the explanatory categorical variable:

modalities of the response categorical variable:



Comparison with a proportion problem

$$H_0 : p_{j|k} = p_j \quad k = 1,\ldots,K;\ j = 1,\ldots,J$$

Assumption of normality of proportions:

$$\frac{n_{kj}}{n_k} \sim N\left(p_j = \frac{n_j}{n}, \left(1 - \frac{n_k}{n}\right)\frac{p_j(1-p_j)}{n_k} \right)$$

Test statistic: Difference between proportion of modality *j* in group *k* and proportion of modality *j* in whole data

$$z = \frac{\frac{n_{kj}}{n_k} - \frac{n_j}{n}}{\sqrt{\left(1 - \frac{n_k}{n}\right)\left(\frac{p_j(1-p_j)}{n_k}\right)}} \sim N(0,1)$$

Rank the levels of the categorical explanatory variables by p.value (ascending)

Profiling in R

The variable to profile can be either categorical (buyers or non buyers, nationality, …) or continuous (revenue, level of satisfaction, …)

- Profiling of a categorical variable **– `catdes` function in `FactoMineR`**

  For every category (group) of the categorical variables find:

  – Which explanatory continuous variables are related with this group (comparing the mean of the continuous variable in the group with the overall mean)

  – Which modalities of the explanatory categorical variables are related with the group (comparing the proportion of the modality in the group with the overall proportion of the modality).

- Profiling of a continuous variable **– `condes` function in `FactoMineR`**

  – Finding which modalities of the explanatory variables are related with the continuous variable (comparing the mean of the continuous variable in the modality with the overall mean of the continuous variable).

APPLICATION: PROFILING OF GRANTED CREDITS

Application of profiling: to whom a bank is granting credits

The dataset

```
> sapply(credsco,levels)
Dictamen            "positiu" "negatiu"
Antig_feina
Vivenda             "lloguer" "escriptura" "contr_privat" "ignora_cont" "pares" "altres viv"
Plaç
Edat
Estat_civil         "solter" "casat" "vidu" "separat" "divorciat"
Registres           "reg_no" "reg_si"
Tipus_feina         "fixe" "temporal" "autonom" "altres sit"
Despeses
Ingressos
Patrimoni
Carrecs_pat
Import_sol
Preu_finan
Rati_fin
Estalvi
```

Application fo profiling: to whom a bank is granting credits

```
> library(FactoMineR)
> desc_Dict <- catdes(data, num.var=1)
```

```
$quanti$positiu
               v.test Mean in category Overall mean sd in category   Overall sd      p.value
Antig_feina  17.376225         9.319062     7.986753       8.486593     8.173388 1.249107e-67
Estalvi      12.955564         4.386752     3.928830       3.882315     3.767794 2.185239e-38
Ingressos     9.469047       150.470000   142.929951      85.385049    84.882954 2.824083e-21
Edat          6.359496        37.735625    37.080377      11.127476    10.983365 2.024164e-10
Patrimoni     6.299321      6142.187187  5451.422766   12548.029206 11689.307019 2.989516e-10
Plaç         -6.714938        45.515625    46.438707      15.200547    14.653817 1.881465e-11
Import_sol  -10.313197       993.012187  1038.918276     439.922117   474.492724 6.142358e-25
Rati_fin    -14.760093        69.786109    72.608246      20.932616    20.381757 2.649133e-49
```

```
$category$positiu
                       Cla/Mod  Mod/Cla    Global      p.value     v.test
Registres=reg_no       77.58761 89.25000 82.644814 1.632198e-70  17.753058
Tipus_feina=fixe       79.32264 69.53125 62.977099 1.801405e-46  14.313527
Vivenda=escriptura     81.45161 53.65625 47.328244 2.061899e-42  13.648460
Estat_civil=casat      74.42936 75.40625 72.788505 6.133385e-10   6.186964
Tipus_feina=altres sit 60.23392  3.21875  3.839246 8.650900e-04  -3.331082
Tipus_feina=autonom    67.41463 21.59375 23.013022 3.727057e-04  -3.558692
Estat_civil=solter     66.42784 20.28125 21.935339 2.572048e-05  -4.208383
Estat_civil=separat    50.76923  2.06250  2.918725 2.709754e-07  -5.142591
Vivenda=altres viv     54.37500  5.43750  7.184553 4.971211e-12  -6.906396
Vivenda=lloguer        60.06160 18.28125 21.867984 1.723938e-19  -9.029541
Tipus_feina=temporal   39.95585  5.65625 10.170633 5.156680e-51 -15.023432
Registres=reg_si       44.50194 10.75000 17.355186 1.632198e-70 -17.753058
```

Guide for interpreting the results of `catdes` function of R

Glosario para la descripción

A partir de las variables continuas

`Overall mean`: es la media global de la variable

`Mean in category`: es la media de la variable en la clase (=”grupo”) considerado

`v.test`: es ell valor del estadístico N(0,1) al comparar la “Mean in Category” con la “Overall mean”.

`p.value`: es el p.valor obtenido en la anterior comparación.

A partir de las variables categóricas

`Global`: es el porcentaje global de la modalidad (= categoría)

`Mod/Cla`: es el porcentaje de la modalidad en la clase (= cluster) considerado

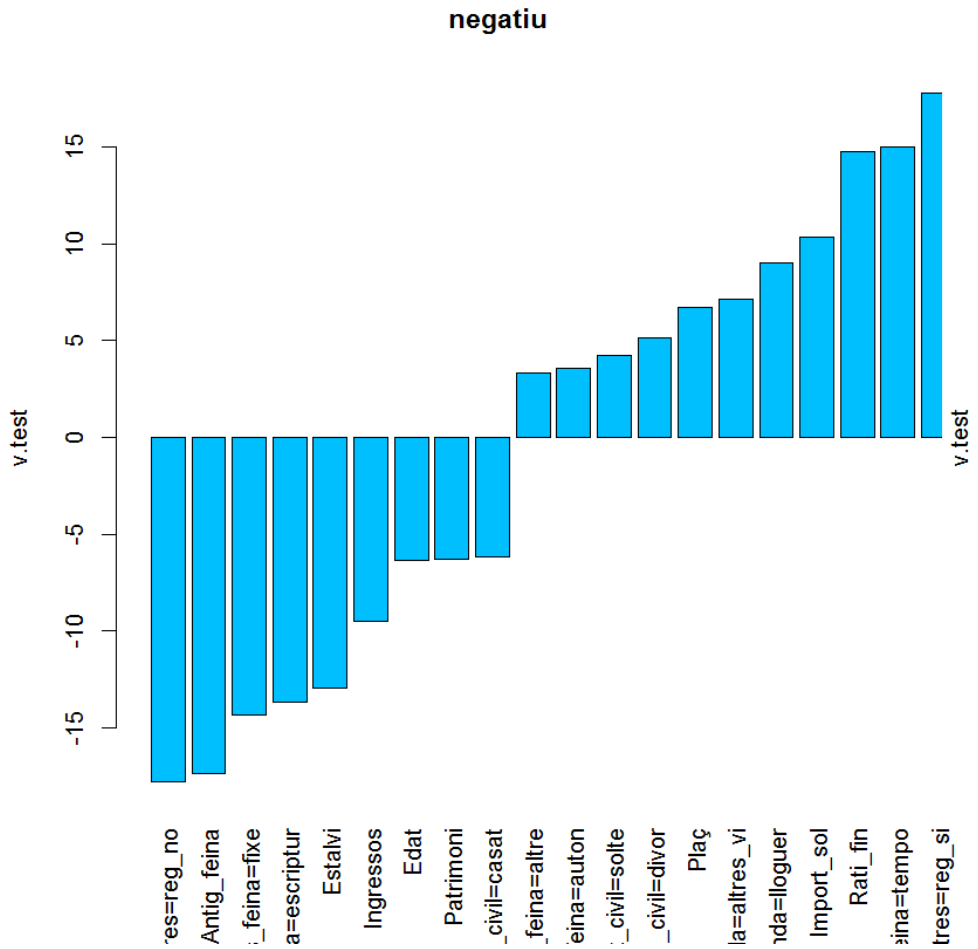
`v.test`: es el valor del estadístico al comparar la proporción Global con la proporción en la clase (= Mod/Cla)

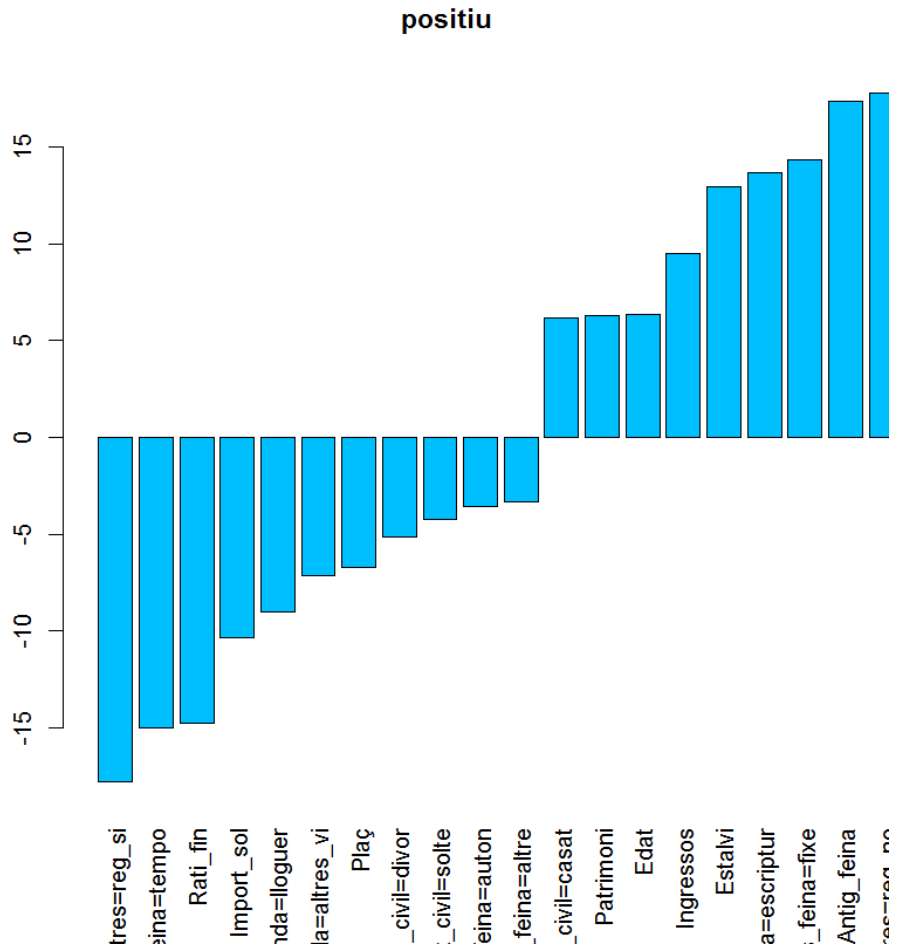
`p.value`: es el p.valor obtenido en la comparación de ambas proporciones

`Cla/Mod`: es el porcentaje de una modalidad en una clase, respecto del total de la modalidad. Da la especifidad de una clase respecto de una modalidad.

Plotting the catdes output

```
> plot(desc_Dict,numchar=17)
```





Example: Profiling of work candidates for a big company

	Variables
1	gusto por el trabajo
2	capacidad por el trabajo
3	voluntad y perseverancia
4	ambición
5	sentido de la competitividad
6	sentido de la eficacia
7	autoridad natural
8	capacidad para dirigir
9	capacidad para la persuasión
10	capacidad para la negociación
11	capacidad para asumir riesgos
12	espiritu de inciativa
13	capacidad para innovar
14	sentido de la organización y el método
15	capacidad adapatación nuevas técnicas
16	capacidad trabajar en el extranjero
17	presentación
18	respeto usos y modos sociales
19	tacto y delicadeza
20	facilidad de contacto
21	optimismo y alegria de vivir
22	capacidad para trabajar en equipo
23	tolerancia
24	respeto por la jerarquia
25	sensibilidad opinión de los demás
26	capacidad para escuchar
27	capacidad para hablar en público
28	capacidad adaptación situaciones nuevas
29	discreción
30	equilibrio personal
31	estabilidad de comportamiento
32	espontaneidad
33	resistencia a la frustración
34	independencia
35	confianza en si mismo
36	sentido de la realidad

Example: Profiling of work candidates for a big company

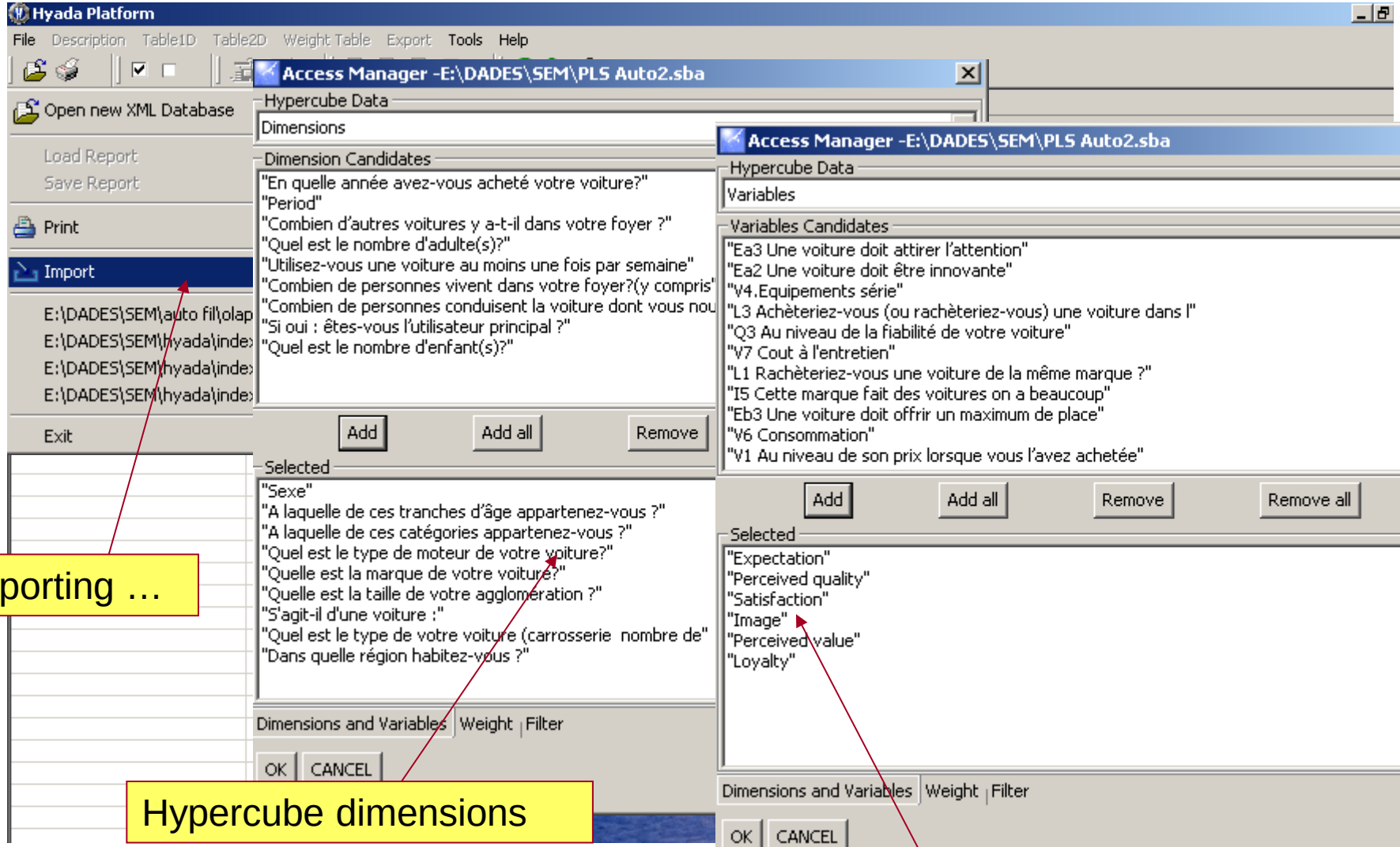
V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.	LIBELLE	IDEN
		CLASSE 1 / 4		(POIDS = 143.00		EFFECTIF = 143)		a01a
7.94	0.000	13.90	10.73	1.83	5.24	7.	autoridad natural	auto
6.09	0.000	13.46	10.58	1.95	6.21	27.	capacidad hablar en público	publ
5.91	0.000	13.01	10.72	1.50	5.09	35.	confianza en si mismo	conf
5.32	0.000	13.22	11.35	1.53	4.60	5.	sentido de la competitividad	comp
5.24	0.000	13.14	10.88	1.74	5.65	13.	capacidad para innovar	inno
		CLASSE 2 / 4		(POIDS = 307.00		EFFECTIF = 307)		a02a
-2.74	0.003	12.17	12.83	1.73	5.30	29.	discreción	disc
		CLASSE 3 / 4		(POIDS = 170.00		EFFECTIF = 170)		a03a
4.67	0.000	13.91	12.75	1.56	3.62	22.	capacidad para trabajar en equipo	equi
4.37	0.000	14.26	12.89	1.61	4.57	25.	sensibilidad opinión de los demás	dema
4.27	0.000	13.39	12.18	1.46	4.15	18.	respeto por los usos y modos sociales	resp
4.14	0.000	14.01	13.09	1.29	3.22	19.	tacto y delicadez	tact
4.02	0.000	13.83	12.58	1.46	4.55	24.	respeto de la jerarquia	jera
		CLASSE 4 / 4		(POIDS = 219.00		EFFECTIF = 219)		a04a
-7.07	0.000	10.10	11.50	1.71	3.33	17.	presentación	pres
-7.24	0.000	10.10	12.31	1.79	5.12	20.	facilidad de contacto	cont
-7.33	0.000	10.20	12.49	2.26	5.22	2.	capacidad por el trabajo	ctra
-7.42	0.000	10.07	12.34	2.15	5.10	1.	gusto por el trabajo	gtra
-7.56	0.000	10.17	11.92	1.81	3.87	14.	sentido de la organización y del método	orga

Enhancing the profiling by introducing the two-way interactions of categorical predictors

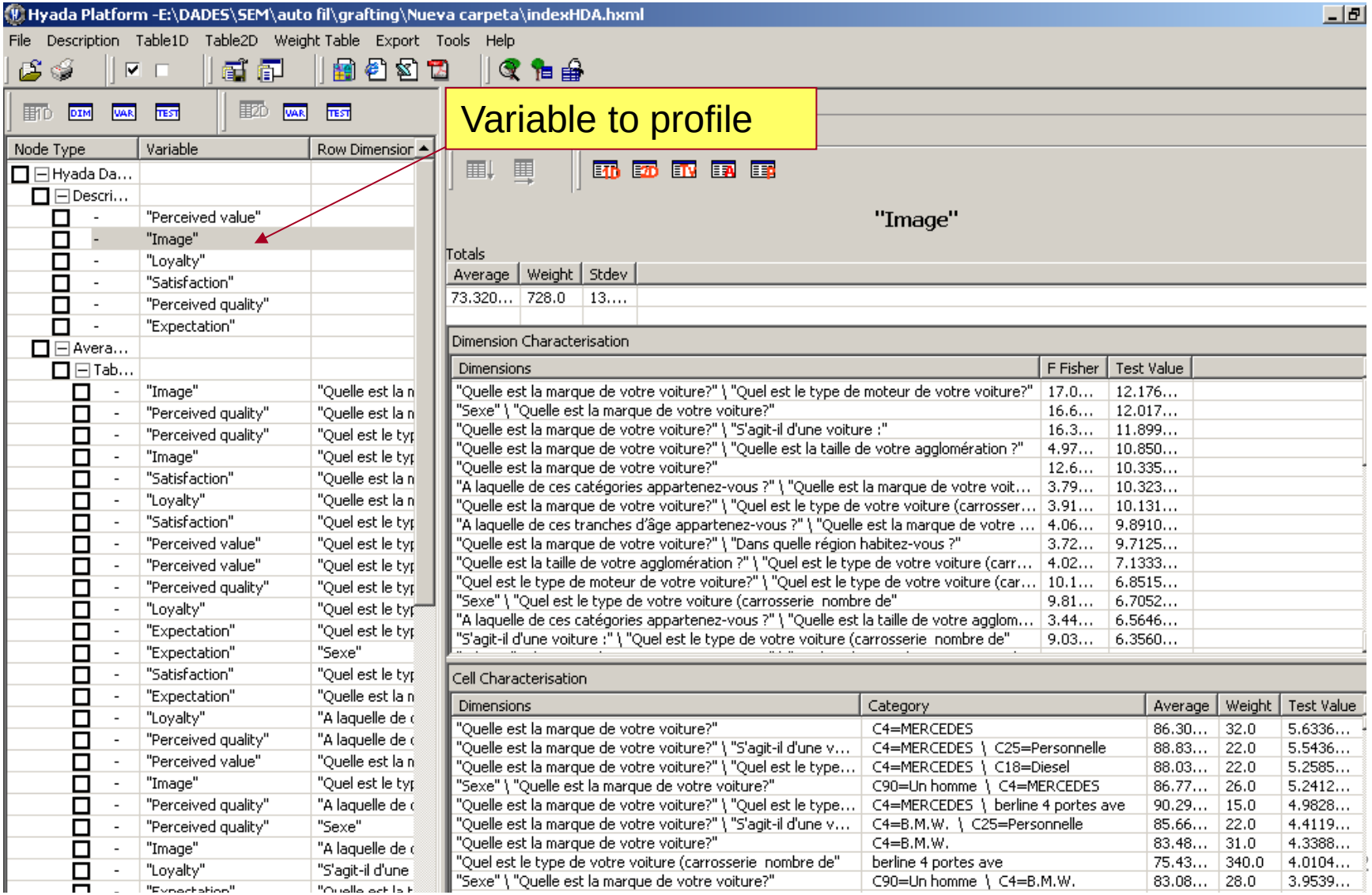
Hyada: Hypercube data analysis

Automatic profiling up to two dimensions

Hypercube definition



Automatic profiling up to two dimensions



SIMPLE PARALLEL COMPUTATION IN R

Apply functions in R

For and Apply in R

We have several ways to do loops:

```
x <- 1:10
for (i in 1:length(x)) { y <- x^2 }        #Classic For-loop
y <- sapply(x, function(input) input^2)    #Simple Apply
```

In this case, both lines calculate the square of the elements of the input vector. It's recommended to use apply instead of for-loops as in several situations they are faster for this kind of use-case.

Types of Apply function

- *apply: Apply that works over matrices. Given a selection parameter you can choose between passing to the function the rows or the columns of the matrix;* `apply(X, 2, mean, trim = .2)`
- *sapply: Apply a function over a vector element-wise;* `sapply(X, mean , trim = .2)`
- *lapply. Apply a function element-wise and return the results in a list;* `lapply(X,mean,trim=.2)`
- *….*

This set of functions are parallelizable in a trivial way.

Parallelism with Parallel package functions

Types of cluster

This library uses an abstraction object called cluster that is used to communicate with processes in the local machine or other machines seamlessly. We have available:

- *PSOCK clusters: Each process listen with sockets (used by default as they work in all OS)*
- *FORK clusters: Father-child process relation.*

The main difference between these two cluster types is that in socket cluster the environmental variables are not shared (but can be copied with clusterExport function). There are more kinds (MPI clusters), but this two suffice for basic parallelism.

Defining a simple cluster

To do so we can do the following:

```
cores <- detectCores(logical=TRUE)
cl    <- makeCluster(cores)
```

By doing this we will create a cluster object that will enable us to use all the threads available in our CPU. If logical=TRUE is set it will return the number of thread, e.g. Intel's i5-2410M has 2 cores but implements 4 threads with hyper-threading technology.

From Apply to Parallel Apply

```
library(parallel)                          #loading the Parallel package

m <- matrix(1:50, nrow=10)                 #Define a matrix
l <- data.frame(a=1:50,b=101:150,c=201:250)

#Monothread version
apply(m, 1, mean) #Average by rows
apply(m, 2, mean) #Average by columns
sapply(l,log) #log of data in Matrix form
lapply(l,log) #log of data separated in lists

#Parallel Version
cl <- makeCluster(detectCores(logical=TRUE))

parApply(cl, m, 1, mean) #Average by rows
parApply(cl, m, 2, mean) #Average by columns
sapply(l,log) #log of data in Matrix form
lapply(l,log) #log of data separated in lists

stopCluster(cl) #Do it always after computation to free resources!
```

Parallelization of Profiling

```
library(nnet)                                   # for using class.ind
library(parallel)                               # Parallelism

d <- read.csv(file="credsco2.txt", sep=" ")
binclass <- class.ind(d$Dictamen)       # Binarize data
d$Dictamen <- NULL                      # Remove Dictamen

cl <- makeCluster(detectCores(logical=TRUE))

clusterExport(cl,"d")       # Export the data to the cluster to make it available

a <- parApply(cl, binclass, 2,
       function(x) {
          library(FactoMineR)           # catdes imported in the worker nodes
          data <- cbind(as.factor(x),d) # Bind the current column with the dataset
          catdes(data,1)                # Perform catdes
       }
     )

stopCluster(cl)                         #Stop the cluster to free resources

#Each profiling will be available with the class name (e.g, negatiu vs the others)
a$negatiu
a$positiu
```

```
> head(binclass)
     negatiu positiu
[1,]       0       1
[2,]       0       1
[3,]       1       0
[4,]       0       1
```