

Introduction to Data Mining

José Hernández-Orallo

*Dpto. de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain*

jorallo@dsic.upv.es

Roma, 14-15th May 2009

Outline

- Motivation. BI: Old Needs, New Tools.

- Some DM Examples.

- Data Mining: Definition and Applications

- The KDD Process

- Data Mining Techniques

- Development and Implementation

Relevance

- **The volume and variety of the information which is stored in digital databases has grown spectacularly in the last decade.**
- **Most of this information is historical, i.e., represents transactions or situations which have happened.**
 - **Apart from its function as “state of the organisation”, and ultimately “memory of the organisation”,**

**the historical information is also
useful to predict future
information.**

Relevance

- Most *decisions* in companies, organisations and institutions are based on information from past experience, which are extracted from very different sources.
- **Collective decisions** use to entail most critical consequences, especially economical and, recently, must be based on **data volumes which overflow human capacity.**

The area of the (semi-)automatic knowledge extraction from databases has recently acquired an unusual scientific and economical significance.

Relevance

- The final user is not an expert in data analysis tools (statistics, machine learning, ...).
- The user cannot lose more time on analysing the data inefficiently:
 - industry: competitive advantages, more effective decisions.
 - science: data never analysed, data banks never related, etc.
 - personal: “information overload”...

The classical statistical packages are not easy to use and are not scalable to the size and type of data usual in databases.

Relation of DM and other disciplines

KDD arrives on the scene...

- *Knowledge Discovery from Databases.*

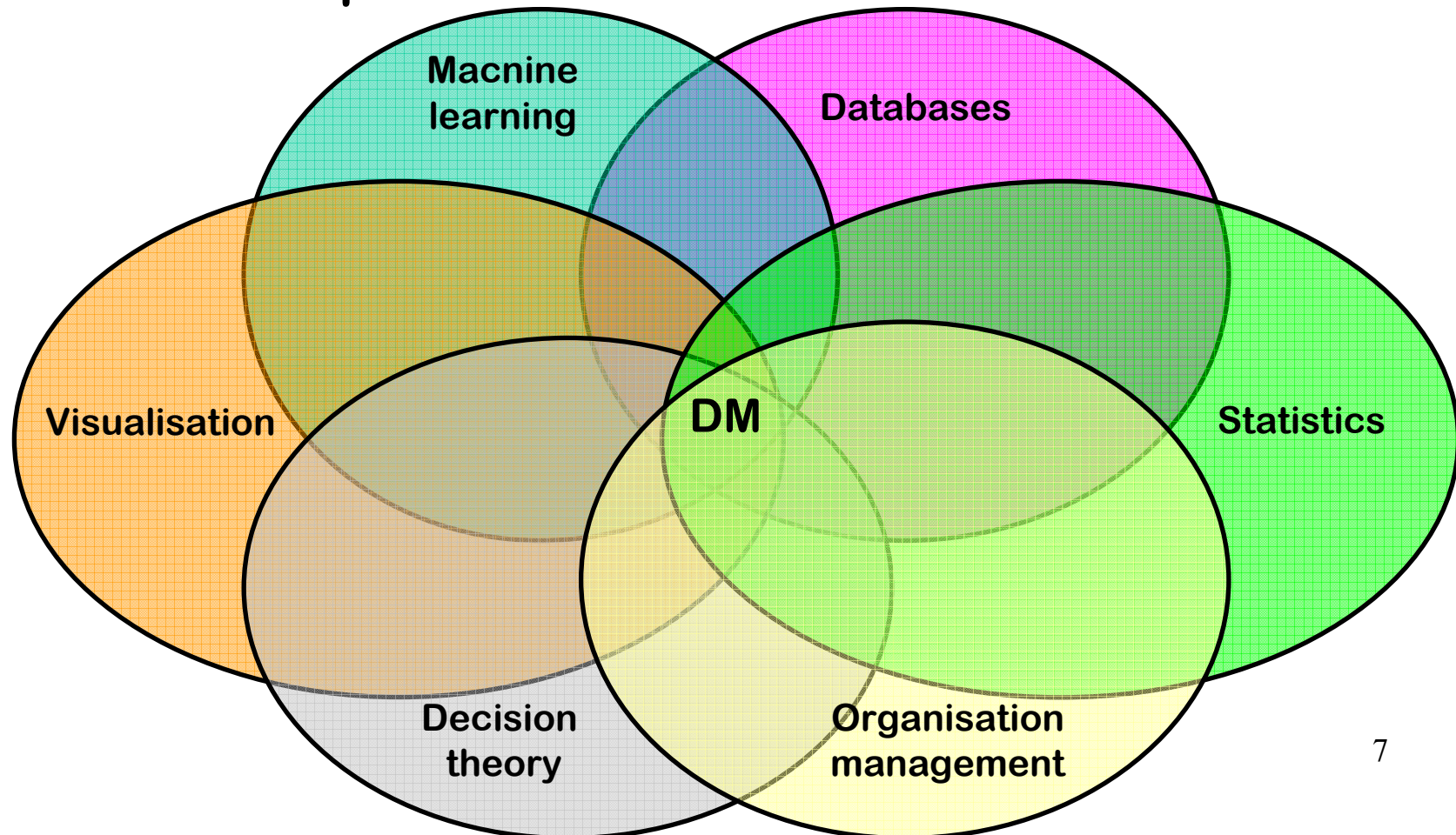
“non-trivial process of identifying valid, novel, potentially useful and ultimately comprehensible patterns from data”.

(Fayyad et al. 1996)

- The discipline integrates techniques from many other different disciplines, without prejudices.

Relation of DM and other disciplines

KDD appears as an interface between and is fed from several disciplines:



Typical Application Areas

KDD for decision making (Dilly 96)

- Retail/Marketing:
- Identify customer purchase patterns.
 - Find associations between customers and demographical features.
 - Predict the response to a *mailing campaign*.
 - Analyse shopping baskets.
- Bank:
- Detect patterns of unlawful credit card use.
 - Identify loyal clients.
 - Predict customers with risk of churn.
 - Determine the credit card spending by several groups.
 - Find correlations between financial indicators.
 - Identify stock market rules from historical data.
- Insurance / Private Health Care:
- Analyse medical procedures which are demanded together.
 - Predict which customers buy new insurance policies.
 - Identify behaviour patterns for customers with high risk.
 - Identify illegitimate behaviour.
- Transportation:
- Determine the schedule for store delivering.
 - Analyse load patterns.

Typical Application Areas

KDD for decision making

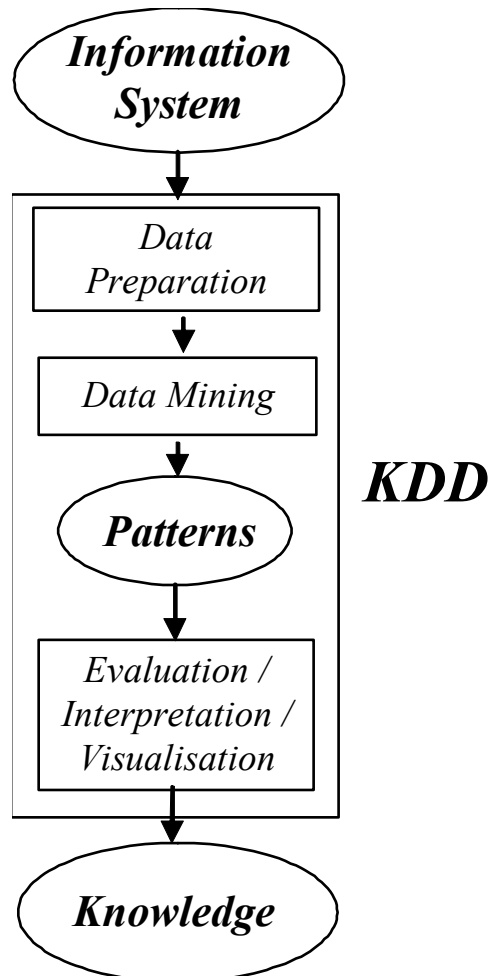
Medicine:

- Identify satisfactory medical therapies for several illnesses.
- Find symptom association and provide differential classifications for several pathologies.
- Study risk/health factors (genetic, precedents, habits, dietary, etc.) in different pathologies.
- Patient segmentation (clustering) for a more “intelligent” (specialised) attention according to each cluster/group.
- Temporal predictions in healthcare centres for a better use of resources, visits, wards and rooms.
- Epidemiological studies, throughput analysis of information campaigns, prevention, drug substitution, etc.

Outline

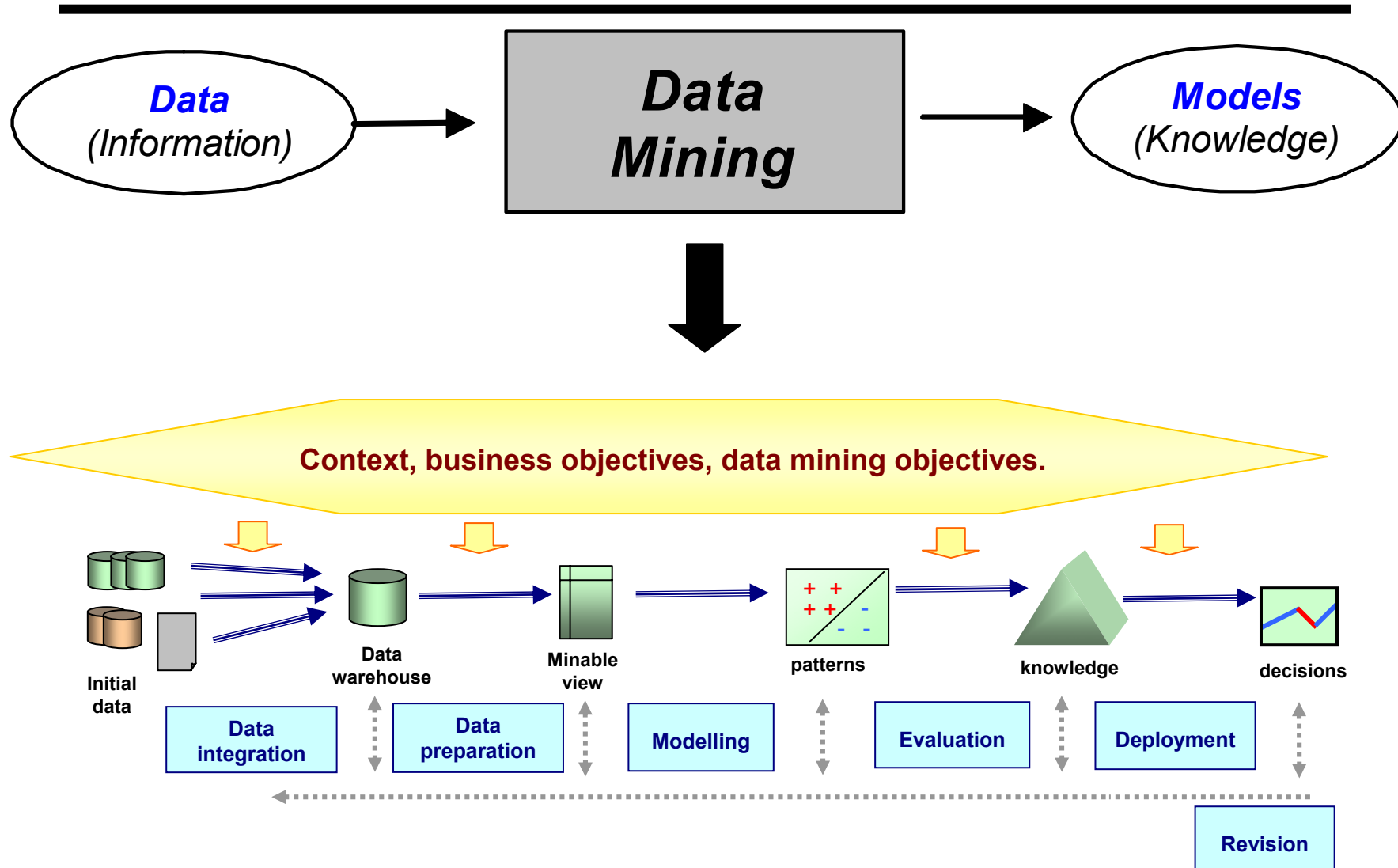
- Motivation. BI: Old Needs, New Tools.
- Some DM Examples.
- Data Mining: Definition and Applications
- The KDD Process
- Data Mining Techniques
- Development and Implementation

KDD Process. Stages



1. Determine the sources of information which can be useful and where to find them.
2. Design a common data repository (data warehouse) which can unify in an operative way all the gathered information.
3. Implement the data warehouse which allows for data "navigation" and prior visualisation, to determine which issues deserve analysis.
4. Data selection, cleansing and transformation of the data which will be analysed. Construction of the *minable views*.
5. Choose and apply the most appropriate data mining method(s).
6. Interpretation, transformation and representation of the extracted patterns.
7. Spread and use (deployment) of the new knowledge.
8. Monitoring and revision (in case).

KDD Process. Stages



Problem Definition

- An important issue is to establish the data mining objectives from the business objectives.



- Example:
 - Business Goal: “Reduce waiting queues”
↓
 - Refined business goal: “Assign more adjusted resources at cash desks according to customer flow”.
↓
 - Data mining objective: “Predict beforehand the customer flow for each supermarket at any period of the day”.

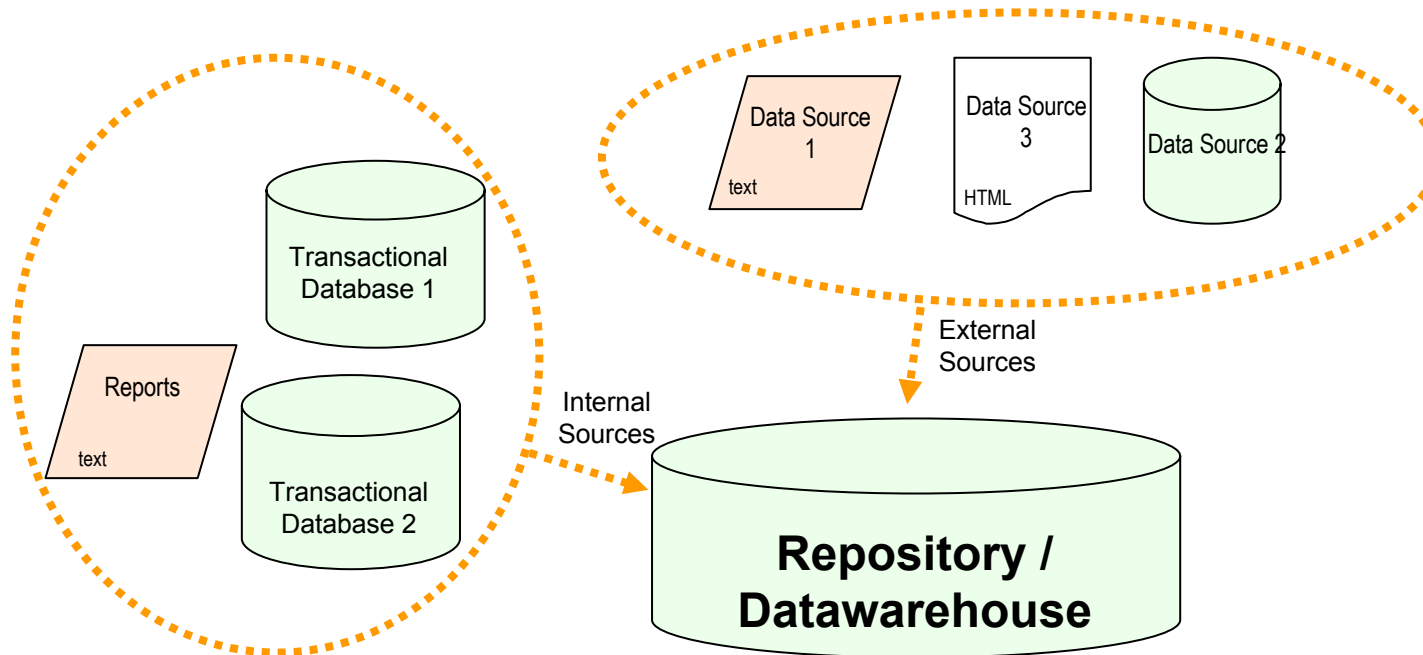
KDD stages: Data Integration

- The first KDD stages determine whether the subsequent stages are able to extract valid and useful knowledge from the original information.
- Generally, the information which requires analysis in the organisation may be found:
 - in databases and other highly diverse sources,
 - both internal and external.
 - many of these sources are those used for the transactional work.

The subsequent analysis will be much simpler if the source is unified, accessible (internal) and disconnected from the transactional work.

KDD stages: Data Integration

- Information gathering

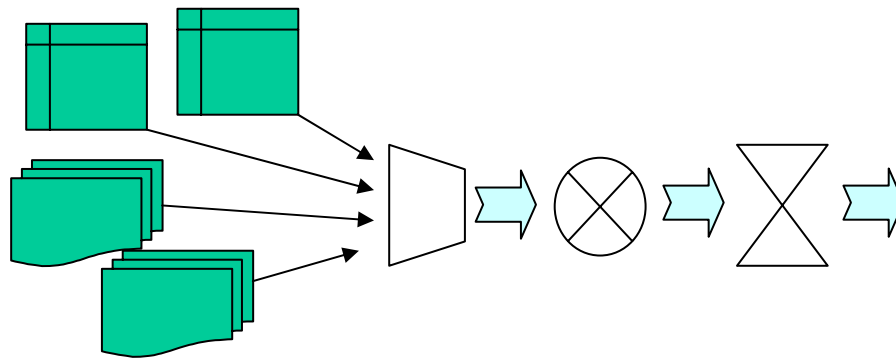


KDD stages: Data Integration

- **Information gathering:**
 - Granularity is usually higher than in DW
 - Information to be recorded into the database must be carefully planned beforehand.
 - If we now realise we need the age of the customer and we haven't recorded it, the problem has a difficult solution now.
 - External sources are very important:
 - Demographic data, sociological studies, general economical data, ...
 - Business data, competitors, ..
 - Calendars, weather, traffic, TV/sport schedule, catastrophes,
 - External information is frequently sold and bought.

KDD stages: Data Preparation

- After data integration:
 - The goal of “data preparation” is to obtain the “**MINABLE VIEW**”, from a set of data which may be inadequate, missing, wrong, irrelevant, scattered, etc.



MINABLE VIEW

Idc	D-credit (years)	C-credit (euros)	Salary (euros)	Own house	Default account	...	Good customer
101	15	60000	2200	Yes	2	...	no
102	2	30000	3500	Yes	0	...	yes
103	9	9000	1700	Yes	1	...	no
104	15	18000	1900	No	0	...	yes
105	10	24000	2100	No	0	...	no
...

Minable view: set of data which includes all and only the interest variables for the given problem in the adequate format

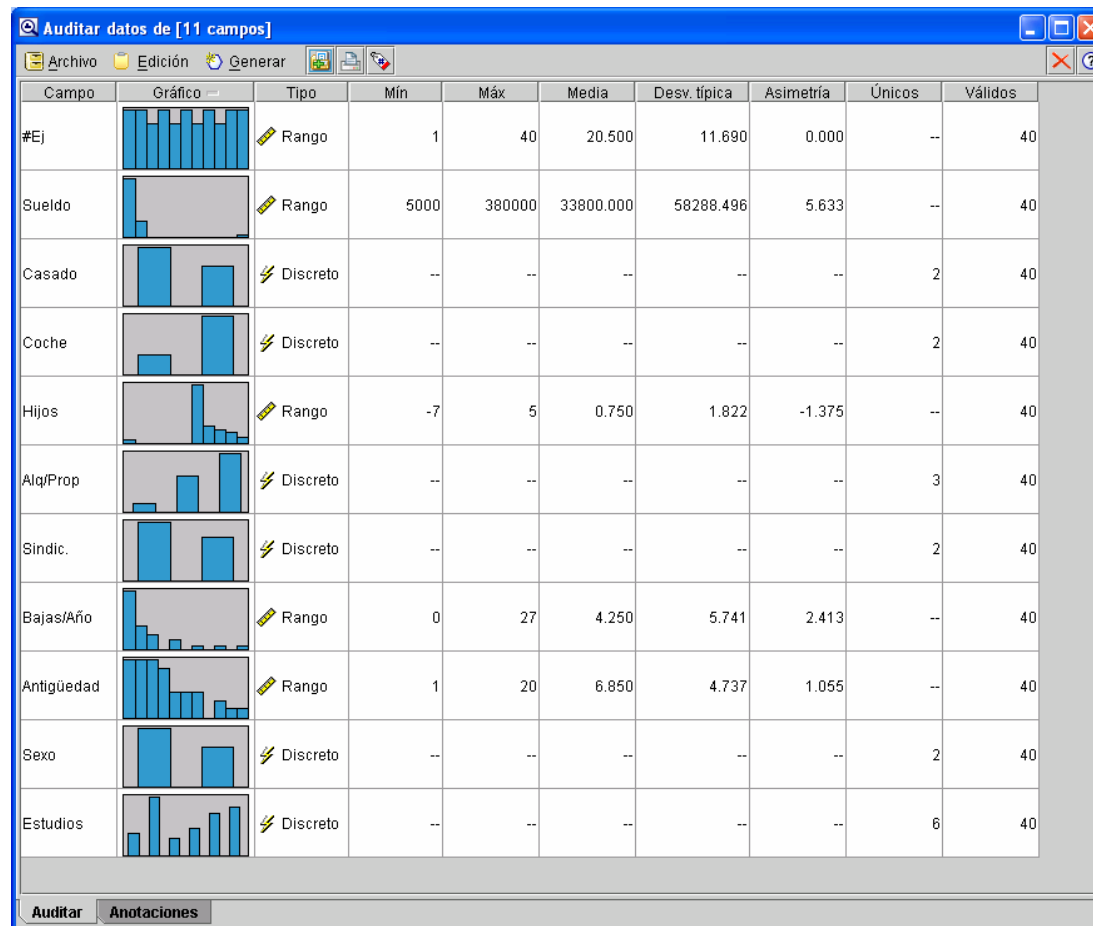
KDD stages: Data Preparation

- Data preparation includes:
 - Data comprehension
 - Data visualisation
 - Data cleansing
 - Data transformation
 - Data selection

This stage usually takes half of the time/effort from the overall KDD process.

KDD stages: Data Preparation

- The first step is to know and understand the data: a feature summary is very useful:



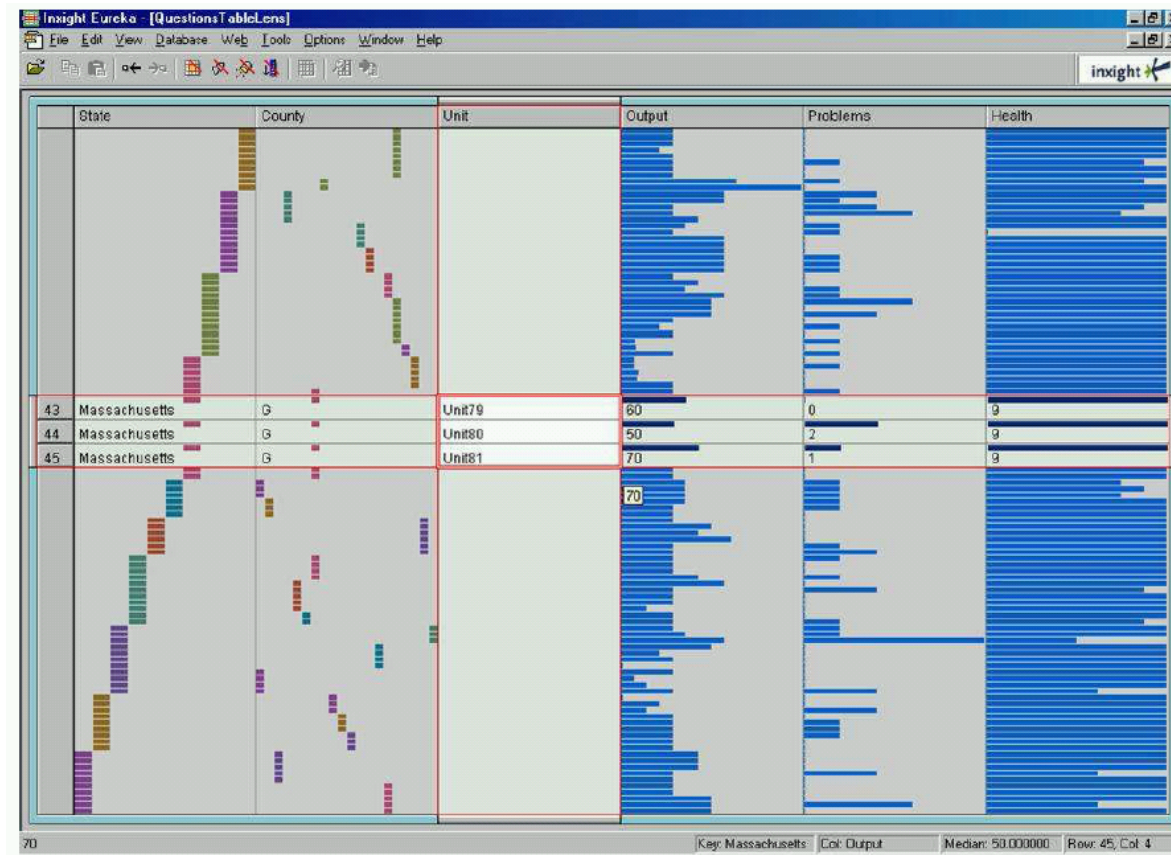
Auditar datos de [11 campos]

Campo	Gráfico	Tipo	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
#Ej		Rango	1	40	20.500	11.690	0.000	--	40
Sueldo		Rango	5000	380000	33800.000	58288.496	5.633	--	40
Casado		Discreto	--	--	--	--	--	2	40
Coche		Discreto	--	--	--	--	--	2	40
Hijos		Rango	-7	5	0.750	1.822	-1.375	--	40
Alq/Prop		Discreto	--	--	--	--	--	3	40
Sindic.		Discreto	--	--	--	--	--	2	40
Bajas/Año		Rango	0	27	4.250	5.741	2.413	--	40
Antigüedad		Rango	1	20	6.850	4.737	1.055	--	40
Sexo		Discreto	--	--	--	--	--	2	40
Estudios		Discreto	--	--	--	--	--	6	40

Auditar Anotaciones

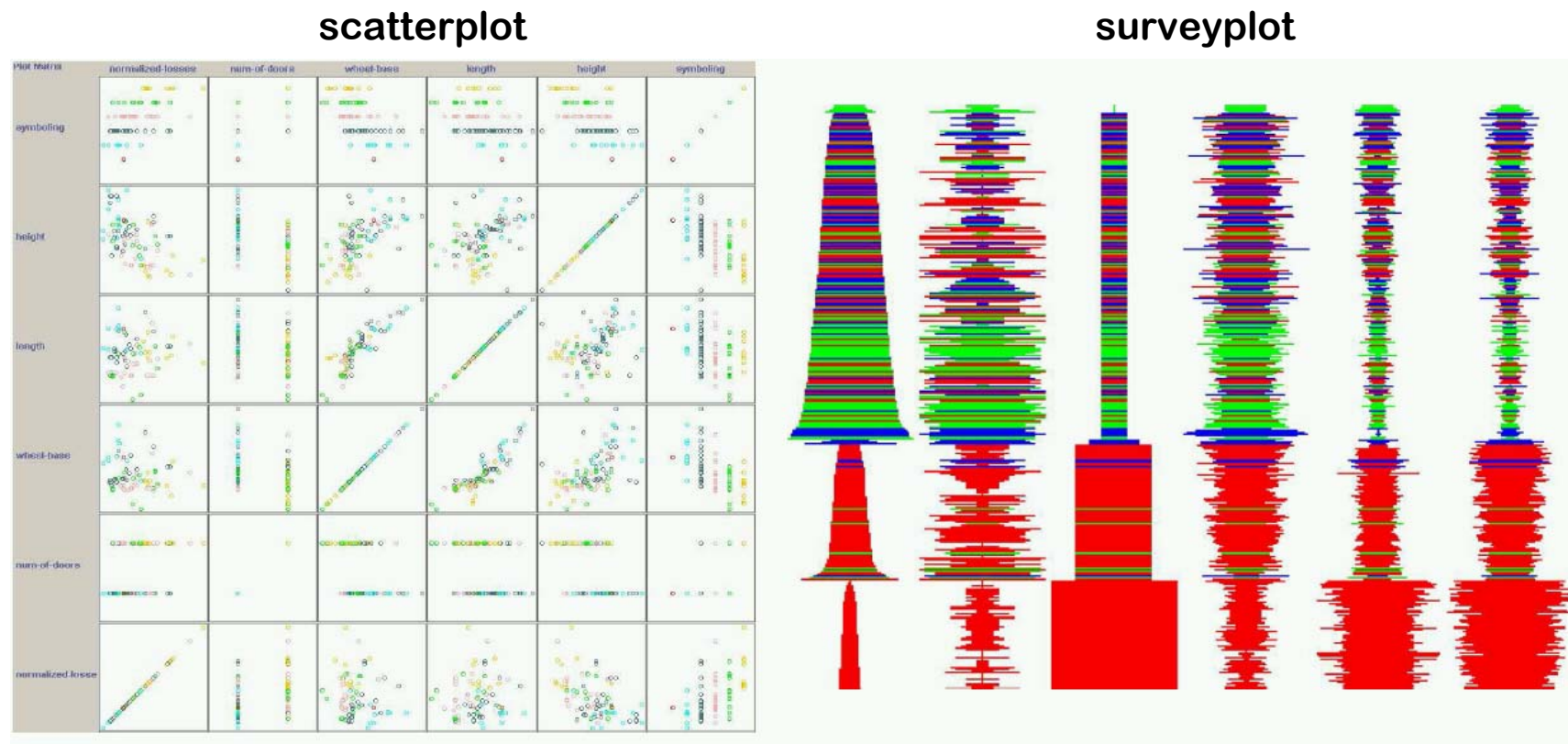
KDD stages: Data Preparation

- Visualisation techniques help understand the data.



KDD stages: Data Preparation

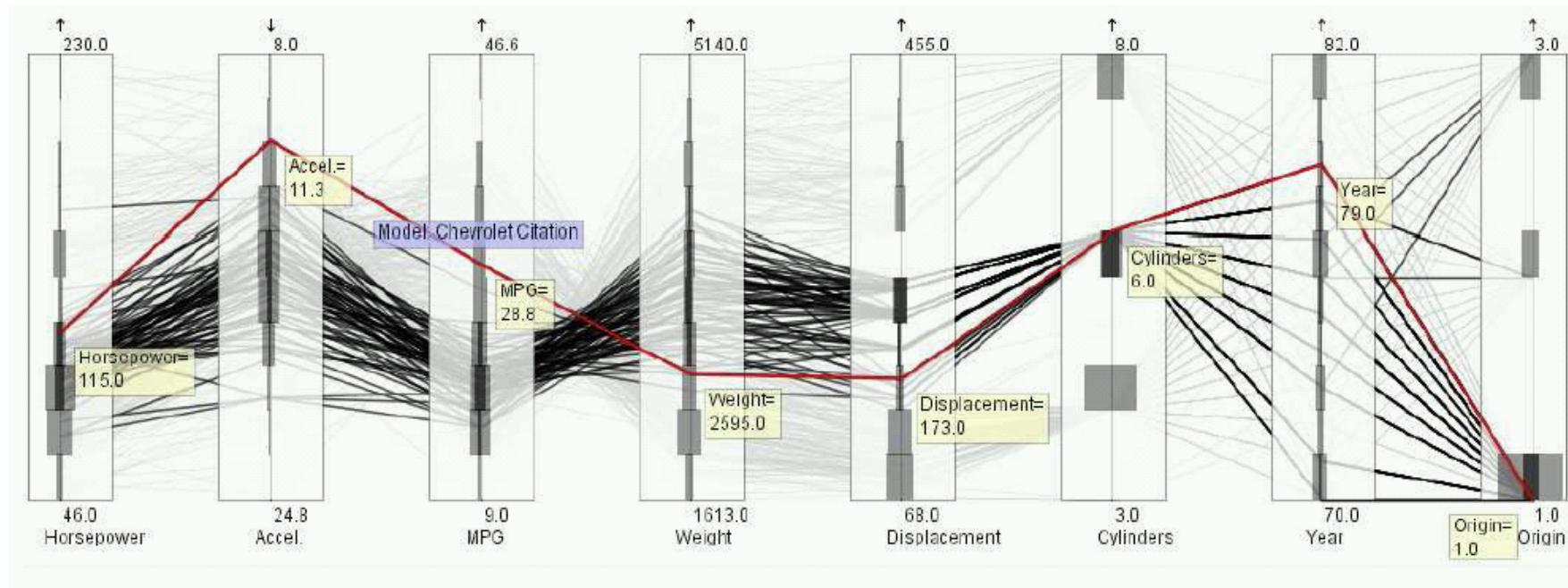
- Visualisation techniques help understand the data.



KDD stages: Data Preparation

- Visualisation techniques help understand the data.

Parallel coordinates



KDD stages: Data Preparation

- Visualisation techniques help understand the data.



© Francisco Javier Ferrer Troyano

KDD stages: Data Preparation

Data cleansing:

- **Possible actions against outliers or missing values:**
 - ignore.
 - filter (eliminate or substitute) the column.
 - filter the row.
 - replace the value by an average or predicted value.
 - segment the rows between correct data and the rest, and work separately.
 - discretise numerical attributes.
 - give up and modify the data quality policy for the next time.

KDD stages: Data Preparation

- Transformations and selections:
- Transformations:
 - Global transformation: e.g. exchange rows and columns.
 - Attribute creation or modification:
 - Discretisation and numerisation.
 - Normalisation.
 - Derived attributes.
 - **Attribute reduction.**
- Selections:
 - Vertical (over features / attributes):
 - **Feature selection.**
 - Horizontal (over instances):
 - Sampling.



The same goal:
data reduction

KDD stages: Data Preparation

Attribute creation:

- A good knowledge of the domain is the most important factor to create good derived attributes:

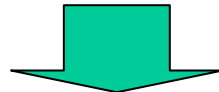
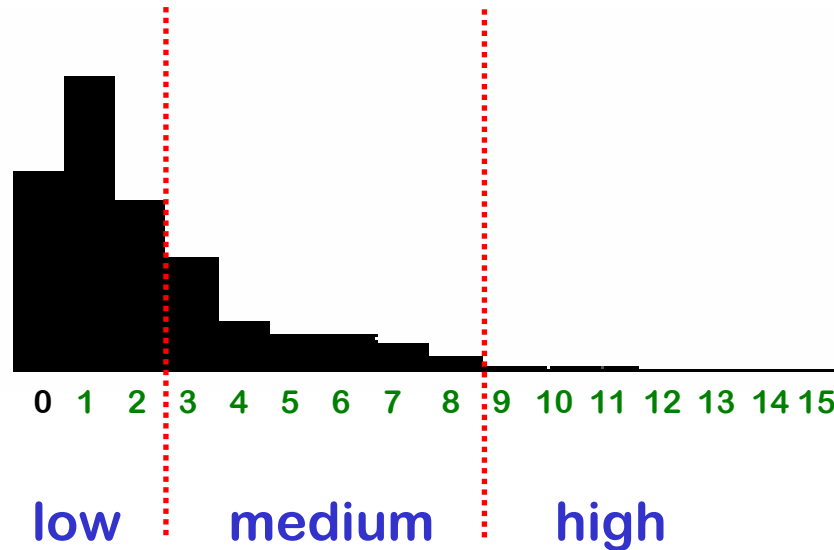
Examples:

- $\text{height}^2/\text{weight}$ (obesity index)
- $\text{debt}/\text{earnings}$
- $\text{passengers} * \text{miles}$
- $\text{credit limit} - \text{balance}$
- $\text{population} / \text{area}$
- $\text{minutes of use} / \text{number of telephone calls}$
- $\text{activation_date} - \text{application_date}$
- $\text{number of web pages visited} / \text{total amount purchased}$

KDD stages: Data Preparation

Discretisation :

Example: attribute “weektickets” (numerical, 1 ... 15).



New attribute “weekticketsNOM” (nominal: low, medium, high).

KDD stages: Data Preparation

Numerisation:

- Numerisation “1 to n” (or n-1):
 - **EXAMPLE:** Convert the field “card” with values: { “VISA”, “4B”, “Amer”, “Maestro” } into four binary fields.
- Numerisation “1 to 1”:
 - **EXAMPLE:** if we have four categories such as {child, young, adult, senior} we can create one attribute with values from 1 to 4.

KDD stages: Data Preparation

Attribute reduction by transformation:

- Well-known techniques such as:
 - *principal component analysis (PCA)*.
 - PCA transforms the m original attributes into a new set of attributes p where $p \leq m$.
 - It is a geometrical projection.
 - New attributes are independent from each other, and they are ordered by information relevance.

KDD stages: Data Preparation

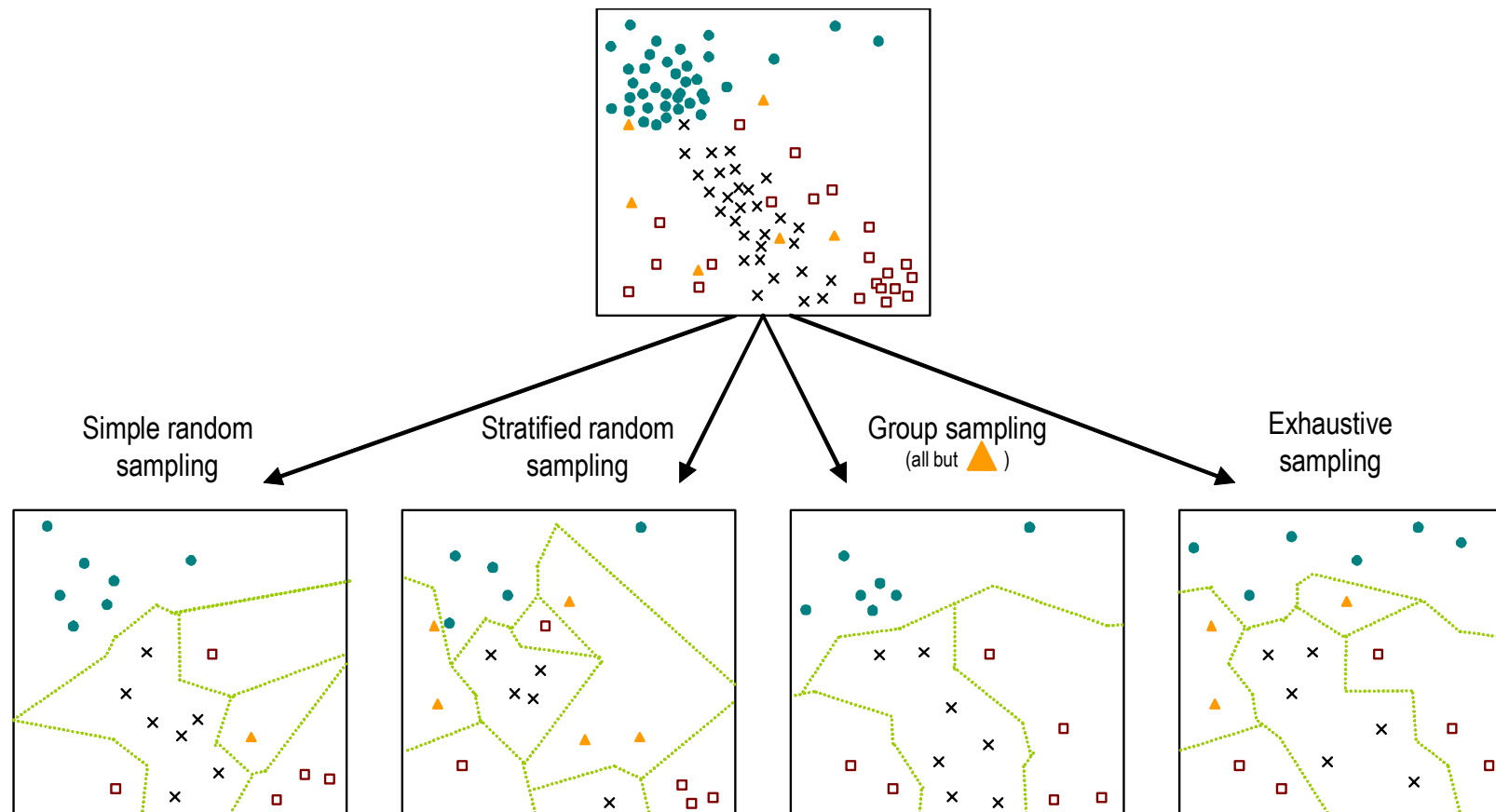
Feature selection:

Choice of p variables from m variables:

- **Filter methods:** selection is made independently from data mining models.
- **Wrapper methods:** selection is made using a data mining model.

KDD stages: Data Preparation

Sampling: reduce the number of rows/instances



KDD stages: Data Modelling

Pattern discovery:

- Once the data is integrated, clean, transformed and appropriately selected into a minable view, the “core” stage can start.
- The kind of knowledge (task) determines the possible *techniques*.
- Substages:
 - Define the task and type the variables.
 - Choose a technique/algorithm.
 - Gauge its parameters.
 - Apply/train the model.

KDD stages: Data Modelling

Typology / Tasks / Kinds of knowledge:

- **Associations:** An association between two attributes happens when the frequency of two specific values to happen together is relatively high.
 - **Example:** in a supermarket we analyse whether nappies and baby jars are bought together.
- **Dependencies:** A functional dependency (approximate or absolute) is a pattern in which it is established that one or more attributes determine the value of the other. But, alert! There are many void (non-interesting) dependencies (inverse causalities).
 - **Example:** if a patient has been allocated to the maternity ward implies that their gender is female.

KDD stages: Data Modelling

Kinds of knowledge (contd.):

- **Classification:** A classification can be seen as the clarification of a dependency, in which each dependent attribute can take a value from several classes, known in advance (supervised learning).
 - **Example:** we know (perhaps from a dependency study or factor analysis) that the attributes *age*, *myopic degree* and *astigmatism level* determine which patient may take a refractory surgical operation satisfactorily.
 - We can try to determine the exact rules which classify a case as positive or negative from these attributes.

KDD stages: Data Modelling

Kinds of knowledge (contd.):

- **Clustering / Segmentation:** clustering is the detection of groups of individuals. It's different from classification in that we do not know in advance the classes (or even its number) (unsupervised learning). The goal is to determine groups or clusters which are different from the other.
 - **Example: find types (clusters) of telephone calls or credit card purchases.**
 - These groups are useful to design different policies for each group or to analyse one group in more detail.
 - Also useful to summarise the data.

KDD stages: Data Modelling

Kinds of knowledge (contd.):

- **Trends / regression:** the goal is to predict the values of a continuous variable from the evolution of another continuous variable (generally time) or from other continuous or nominal variables.
 - **Example:** we need to know the number of future customers or patients, the incomes, calls, earnings, costs, etc. from previous results (day, weeks, months or years before).
- **Other types of knowledge:**
 - **Schema information:** (to discover alternative primary keys, integrity constraints, ...).
 - **General rules:** patterns which cannot be classified into the previous kinds. Other more complex models/patterns (dynamic, ...).

KDD stages: Data Modelling

Predictive Model Example:

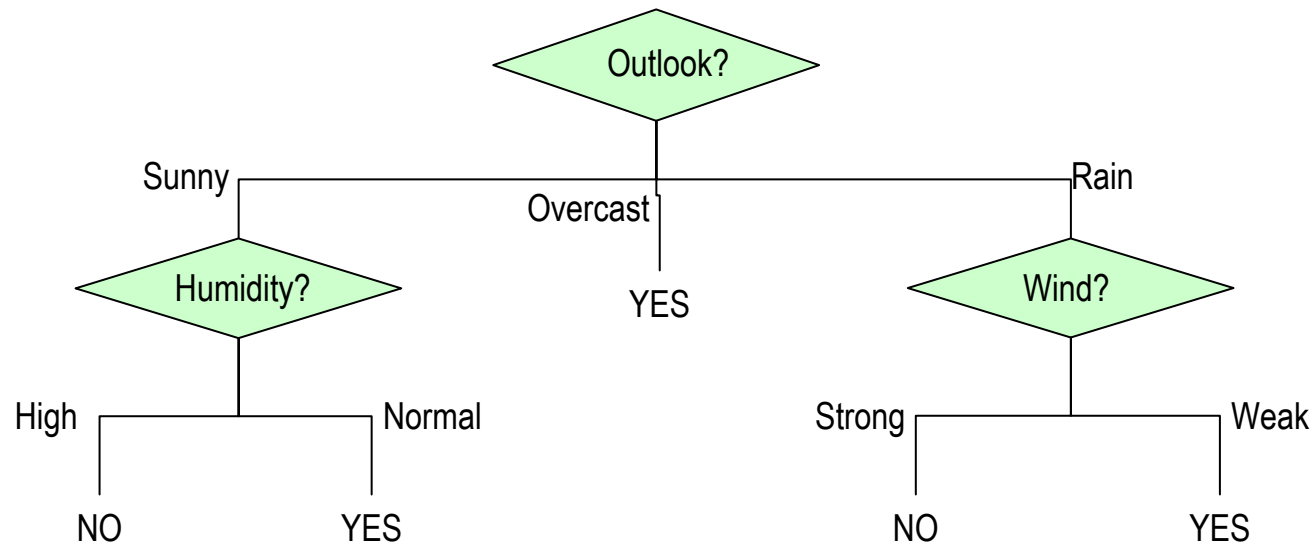
- Will it be nice to play tennis this afternoon?
- Previous experiences:

Example	Sky	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

KDD stages: Data Modelling

Predictive Model Example:

- We type “PlayTennis” as the class (output).
- We choose a decision tree and we gauge some parameters (pruning).
- We train the model and get this result:



- **Now we can use to predict the output for a new instance:**
(Outlook = sunny, Temperature = hot, Humidity = high, Wind = strong)
is NO.

KDD stages: Data Modelling

Descriptive Model Example

- We have the following table with employee data:

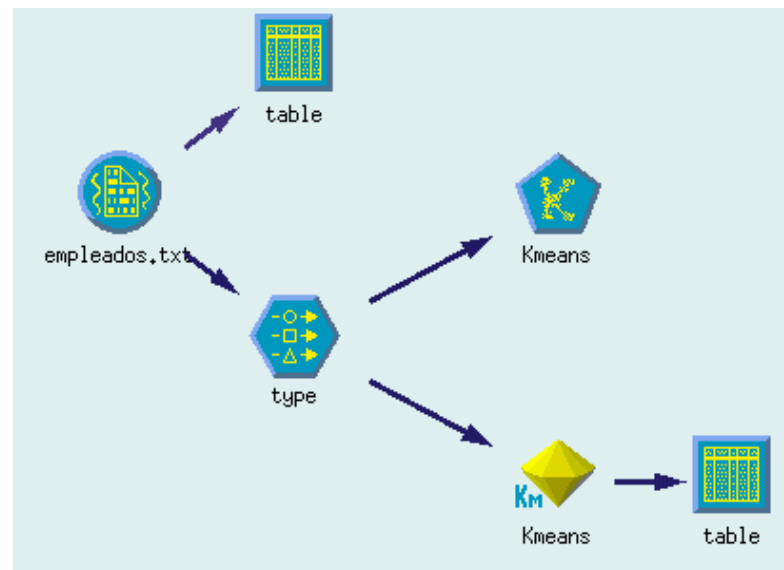
#	Salary	Marrd	Car	Chldrn	House	Union	offsick/Year	WorkYears	Gender
1	10000	Yes	No	0	Rent	No	7	15	M
2	20000	No	Yes	1	Rent	Yes	3	3	F
3	15000	Yes	Yes	2	Owner	Yes	5	10	M
4	30000	Yes	Yes	1	Rent	No	15	7	F
5	10000	Yes	Yes	0	Owner	Yes	1	6	M
6	40000	No	Yes	0	Rent	Yes	3	16	F
7	25000	No	No	0	Rent	Yes	0	8	M
8	20000	No	Yes	0	Owner	Yes	2	6	F
9	20000	Yes	Yes	3	Owner	No	7	5	M
10	30000	Yes	Yes	2	Owner	No	1	20	M
11	50000	No	No	0	Rent	No	2	12	F
12	8000	Yes	Yes	2	Owner	No	3	1	M
13	20000	No	No	0	Rent	No	27	5	F
14	10000	No	Yes	0	Rent	Yes	0	7	M
15	8000	No	Yes	0	Rent	No	3	2	M

We want to obtain representative subgroups.

KDD stages: Data Modelling

- **Descriptive Model Example:**
 - We import the data into a data mining package, we give types (nominal or numerical) to the data, we check for anomalous data, etc.
 - We apply the *k-means* algorithm to find clusters. We indicate the algorithm to find the three most significant groups.

The data mining process is as follows:



KDD stages: Data Modelling

Descriptive Model Example:

- After executing the algorithm we get a model, which shows three groups

cluster 1	cluster 2	cluster 3
5 examples	4 examples	6 examples
Salary : 226000 Married : No -> 0.8 Yes -> 0.2 Car : No -> 0.8 Yes -> 0.2 Children : 0 House : Rent -> 1.0 Union : No -> 0.8 Yes -> 0.2 Offsick/Year : 8 WorkYear : 8 Gender : M -> 0.6	Salary : 225000 Married : No -> 1.0 Car : Yes -> 1.0 Children : 0 House : Rent -> 0.75 Owner -> 0.25 Union : Yes -> 1.0 Offsick/Year : 2 WorkYear : 8 Gender : M -> 0.25 F -> 0.75	Salary : 188333 Married : Yes -> 1.0 Car : Yes -> 1.0 Children : 2 House : Rent -> 0.17 Owner -> 0.83 Union : No -> 0.67 Yes -> 0.33 Offsick/Year : 5 WorkYear : 8 Gender : M -> 0.83 F -> 0.17

How do we interpret these results?

KDD stages: Data Modelling

Descriptive Model Example:

- After executing the algorithm we get a model, which shows three groups

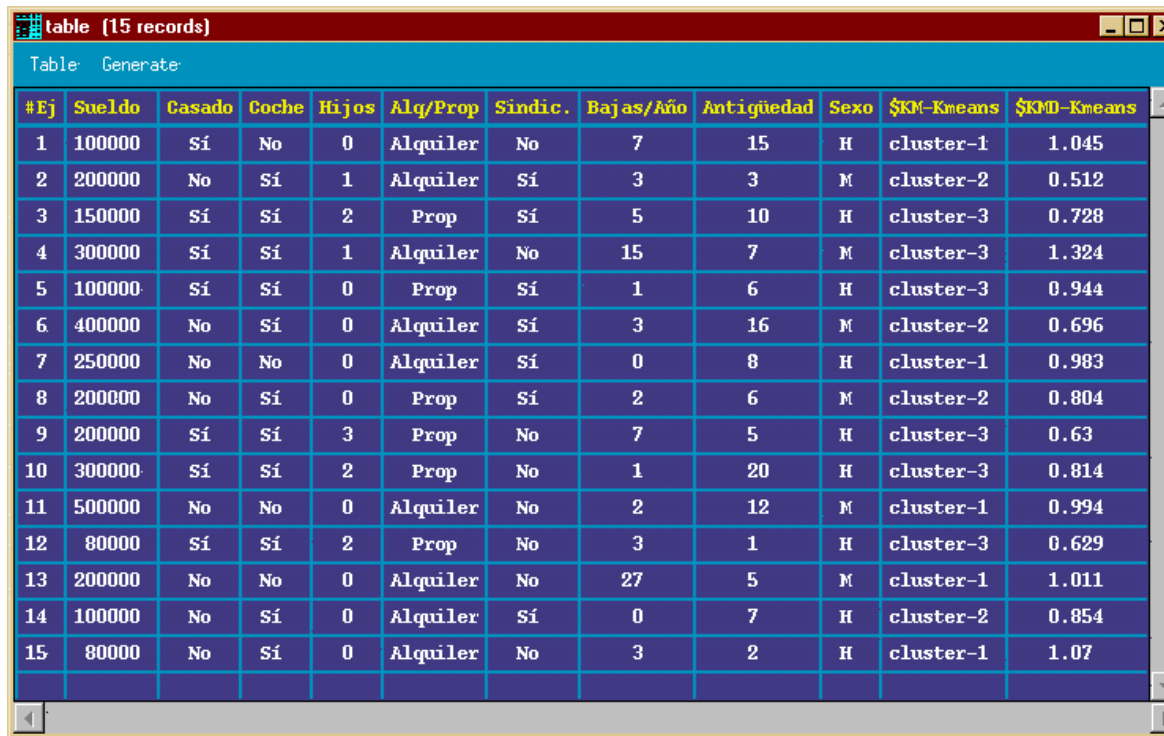


table (15 records)

Table Generate

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	\$KM-Kmeans	\$KMD-Kmeans
1	100000	Sí	No	0	Alquiler	No	7	15	H	cluster-1	1.045
2	200000	No	Sí	1	Alquiler	Sí	3	3	M	cluster-2	0.512
3	150000	Sí	Sí	2	Prop	Sí	5	10	H	cluster-3	0.728
4	300000	Sí	Sí	1	Alquiler	No	15	7	M	cluster-3	1.324
5	100000	Sí	Sí	0	Prop	Sí	1	6	H	cluster-3	0.944
6	400000	No	Sí	0	Alquiler	Sí	3	16	M	cluster-2	0.696
7	250000	No	No	0	Alquiler	Sí	0	8	H	cluster-1	0.983
8	200000	No	Sí	0	Prop	Sí	2	6	M	cluster-2	0.804
9	200000	Sí	Sí	3	Prop	No	7	5	H	cluster-3	0.63
10	300000	Sí	Sí	2	Prop	No	1	20	H	cluster-3	0.814
11	500000	No	No	0	Alquiler	No	2	12	M	cluster-1	0.994
12	80000	Sí	Sí	2	Prop	No	3	1	H	cluster-3	0.629
13	200000	No	No	0	Alquiler	No	27	5	M	cluster-1	1.011
14	100000	No	Sí	0	Alquiler	Sí	0	7	H	cluster-2	0.854
15	80000	No	Sí	0	Alquiler	No	3	2	H	cluster-1	1.07

And assign new employees to the discovered clusters.

KDD stages: Data Modelling

We will give more details about data mining techniques after finishing with the overview of all the KDD stages.

KDD stages: Model Evaluation

- How are models discarded or validated?
- How can we choose among several models?
- How does the number of examples affect?
- How does noise affect?
- How well will my model behave in the future?

KDD stages: Model Evaluation

- **Model evaluation depends on the type of DM task:**
 - **Predictive: simpler and more general evaluation**
 - **Descriptive: the evaluation depends on the used technique.**

KDD stages: Model Evaluation

- **Evaluation of predictive models:**

Which measure can we use to compare the correct/actual results (target “f”) with the estimated ones (model “h”) ?

- **Classification:**

- %Accuracy or, inversely, %Error
- Recall & precision.
- Area under the ROC curve.
- ...

- **Regression:**

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE).
- ...

KDD stages: Model Evaluation

- **Evaluation of predictive models:**
 - Given a set S of n data, the error is defined:
 - **Classification: Error**

$$error_S(h) = \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$

Where $\delta(a,b)=0$ if $a=b$ and 1 otherwise.

Predicted class ($h(x)$)	Actual class ($f(x)$)	Error
Buys	Buys	No
Doesn't buy	Buys	Si
Buys	Doesn't buy	Si
Buys	Buys	No
Doesn't buy	Doesn't buy	No
Doesn't buy	Buys	Si
Doesn't buy	Doesn't buy	No
Buys	Buys	No
Buys	Buys	No
Doesn't buy	Doesn't buy	No

Misclassifications / Total



Error = 3/10 = 0.3

KDD stages: Model Evaluation

- **Evaluation of predicted models:**
 - Given a set S of n data, the error is defined:
 - Regression: Mean Squared Error

$$error_S(h) = \frac{1}{n} \sum_{x \in S} (f(x) - h(x))^2$$

Predicted value (h(x))	Actual Value (f(x))	Error	Error ²
100 mill. €	102 mill. €	2	4
102 mill. €	110 mill. €	8	64
105 mill. €	95 mill. €	10	100
95 mill. €	75 mill. €	20	400
101 mill. €	103 mill. €	2	4
105 mill. €	110 mill. €	5	25
105 mill. €	98 mill. €	7	49
40 mill. €	32 mill. €	8	64
220 mill. €	215 mill. €	5	25
100 mill. €	103 mill. €	3	9



Error = 744/10 = 74,4

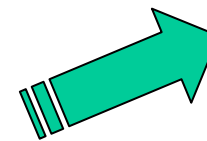
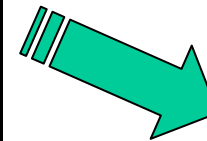
KDD stages: Model Evaluation

- **Evaluation with costs.**

- We can have cost functions for both classification and regression. In classification these are usually cost matrices:

COST		<i>actual</i>		
		low	medium	high
<i>predicted</i>	low	0€	5€	2€
	medium	200€	-2000€	10€
	high	10€	1€	-15€

ERROR		<i>actual</i>		
		low	medium	high
<i>predicted</i>	low	20	0	13
	medium	5	15	4
	high	4	7	60



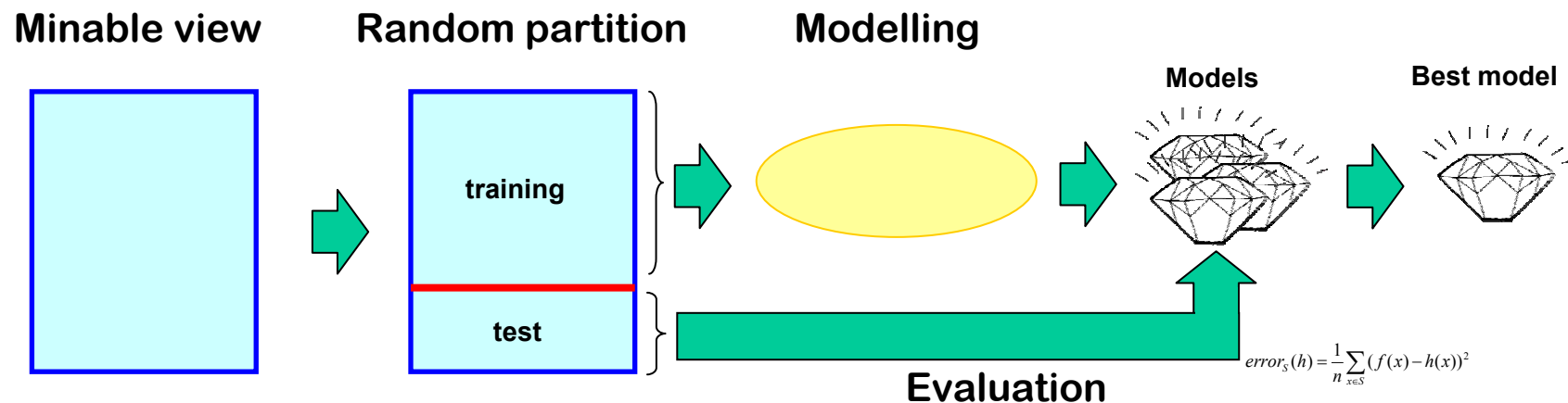
Total cost:

-29787€

We will chose the model with lowest cost, not with lowest error

KDD stages: Model Evaluation

- Evaluation of predicted models.
 - DATA PARTITION:
 - Split the data into:
 - Training set.
 - Models are trained with these data.
 - Test set.
 - Models are evaluated with these data.



KDD stages: Model Evaluation

- Evaluation of predicted models.
 - More elaborate partitions:
 - Cross-validations: Data is randomly split into n folds of equal size. All the combinations are trained/tested to get a better error estimate.
 - Bootstrap: n samples with repetition are made over the original data.

👍 **Very useful with few data**

👎 **Slow**

KDD stages: Model Evaluation

- **Evaluation of Descriptive Models:**
 - Association rules:
 - Simple evaluation:
 - *support*
 - *confidence*
 - It can be ordered using a combination of both indicators.
 - It is not usual to use a test set. The measures are estimated on the whole dataset.

KDD stages: Model Evaluation

- **Evaluation of Descriptive Models:**

- Clustering: much more complex

Concept of error is more difficult to define

- With distance-based techniques, one can use:
 - distance between cluster edges
 - distance between cluster centres (or centroids)
 - cluster radius and density (standard deviation).
 - Another option is to do several groupings with several techniques and to compare the groups (confusion matrix)

KDD stages: Model Evaluation

- **Other evaluation criteria:**
 - Comprehensibility.
 - Simplicity.
 - Interest.
 - Applicability.
 - ...
- **Before deployment the model, a pilot experience should be done.**

KDD stages: Model Deployment

- **Model deployment is sometimes trivial but in other occasions it requires a complex interpretation and implementation process:**
 - **Interpreting and understanding the model. It requires to contrast it with the previous knowledge in the organisation.**
 - **The model may require implementation (e.g. real-time detection of fraud credit card usage).**
 - **The model may have many users and requires to be spread and communicated: the model may require a comprehensible representation to be distributed in the organisation (e.g. beers and frozen products are bought together ⇒ place them in distant shelves).**

KDD Stages: Monitoring and Revision

- A model cannot work well indefinitely:
 - Reality and application context change.

- **Monitoring:**

- Detects whether a revision is necessary.
 - Periodical evaluation with new test data (fresh data).
 - Receptivity to users' comments.
 - Be aware of context change.

- **Revision:**

- Partial: part of the model is changed (e.g. obsolete rules) and part of the model is preserved.
- Total: the new model is drastically changed or completely re-trained.