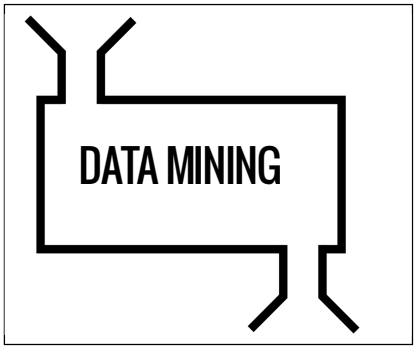
Madrid JUG - Minería de Datos sobre Weka (Data Mining)



9 de Mayo 2013

Jose María Gómez Hidalgo (@jmgomez)
Guillermo Santos García (@gsantosgo)

Universidad Pontificia Comillas

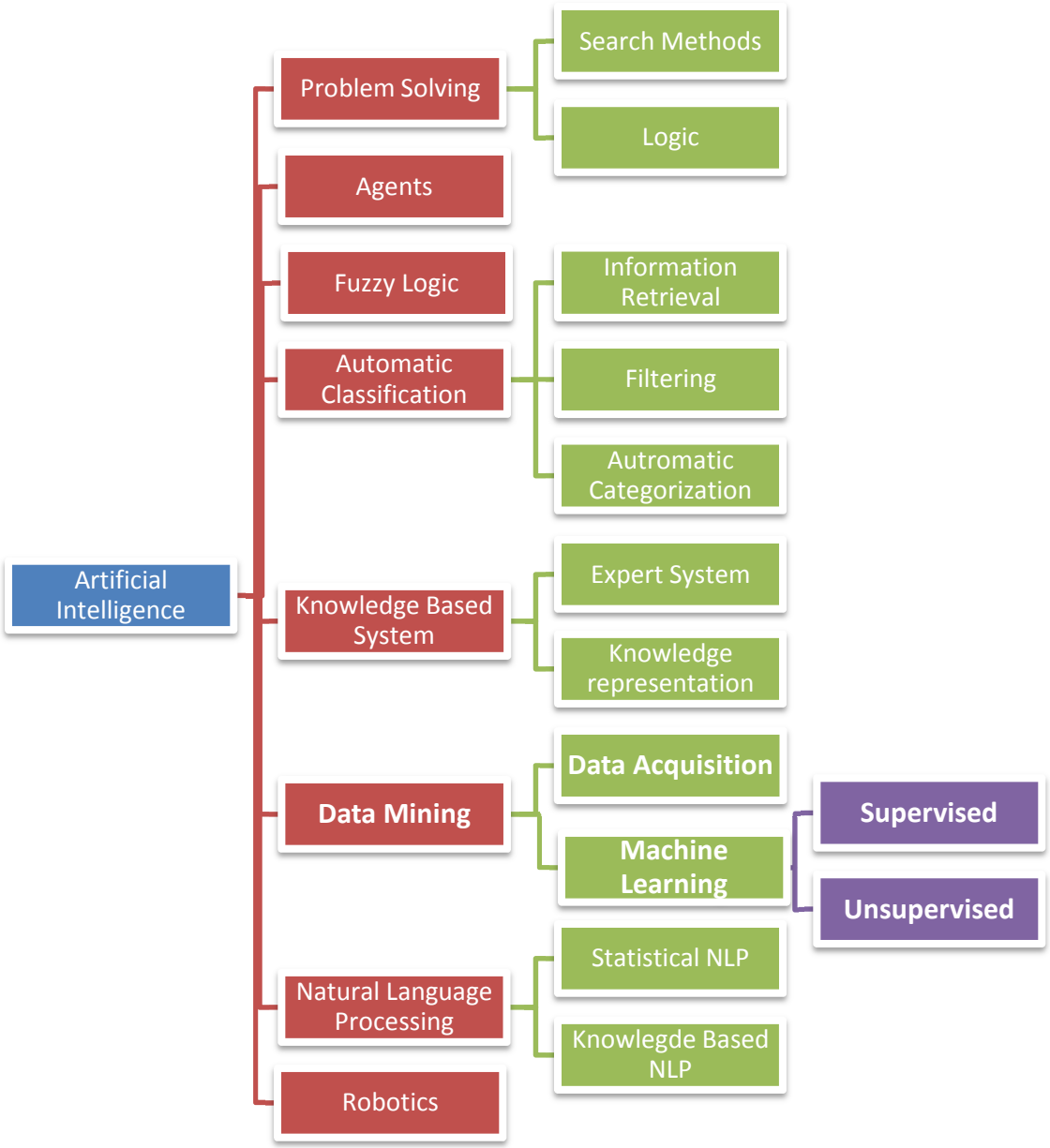
Cómo llegar | Escribir una reseña

Dirección: Calle Alberto Aguilera, 23, 28015 Madrid

Teléfono: 915 42 28 00

INDEX

1. Artificial Intelligence. Conceptual Map



Link: http://en.wikipedia.org/wiki/Artificial_intelligence

DATA MINING. LEARN FROM DATA

1.1 Knowledge Based System vs. Machine Learning System

Knowledge Based System (Expert System)

- Rules are codified manually (Represent knowledge)
- Experts (expert is a person with extensive knowledge about domain).
- Cost.

Expert Sytems (Credit Expert System)

If (Annual Income > 3 * Annual Debt) Then CREDIT = YES

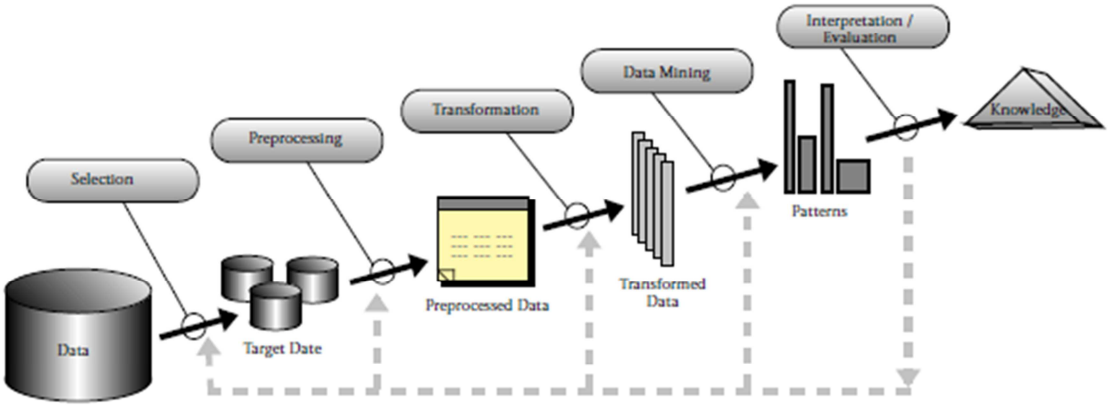
Annual Income	Annual Debt	Credit
42.000 €	15.000 €	NO
37.000 €	12.000 €	SI
80.000 €	40.500 €	NO
150.000 €	45.000€	SI

Machine Learning System

- The manual process is automated.
- There aren't experts.
- We take us advantage of data classified manually over years.
- Training phase and testing phase.
- At first, machine learning systems aren't as accurate as knowledge based systems, however they're can evolve and get better through time. (Ex. Spam Detection Spam)

2. Data Mining Process

KDD (Knowlegde Discovery in Databases)



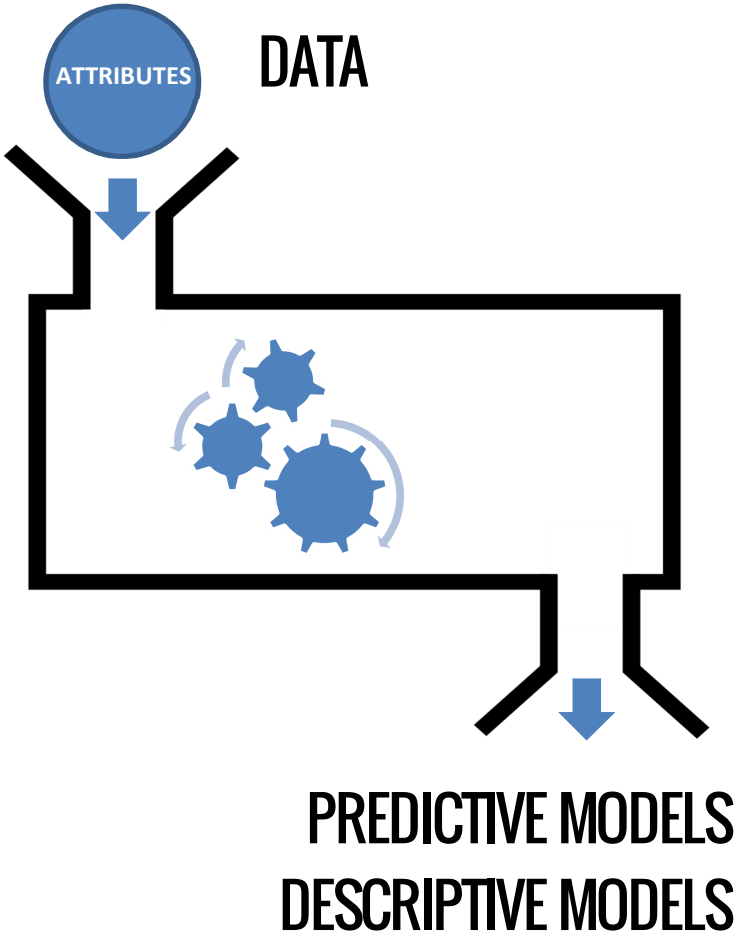
Source: From Data Mining to Knowledge Discovery in Databases (Fayyad. 1997)

1. Selection. The data relevant to select.
2. Preprocessing.
3. Transformation.
4. Data-Mininq. Building Models and Patterns. (MODELLING)
5. Interpretation/Evaluation . Evaluation and Results

The term *DATA-MINING* sometimes refers to the complete process KDD, and sometimes refers only to the phase of *MODELLING* (4). Here mainly are applied algorithms in the scope of Machine Learning.

2.1 Machine Learning

Aim. Building or creating programs capable of **generalizing behavior** from weakly structured information.



2.1.1 Supervised Machine Learning

Aim. Predict the value of a variable based on a number of input variables.

Regression Problem.

Classification Problem.

Result: **PREDICTIVE MODELS** or **CLASSIFIERS**.

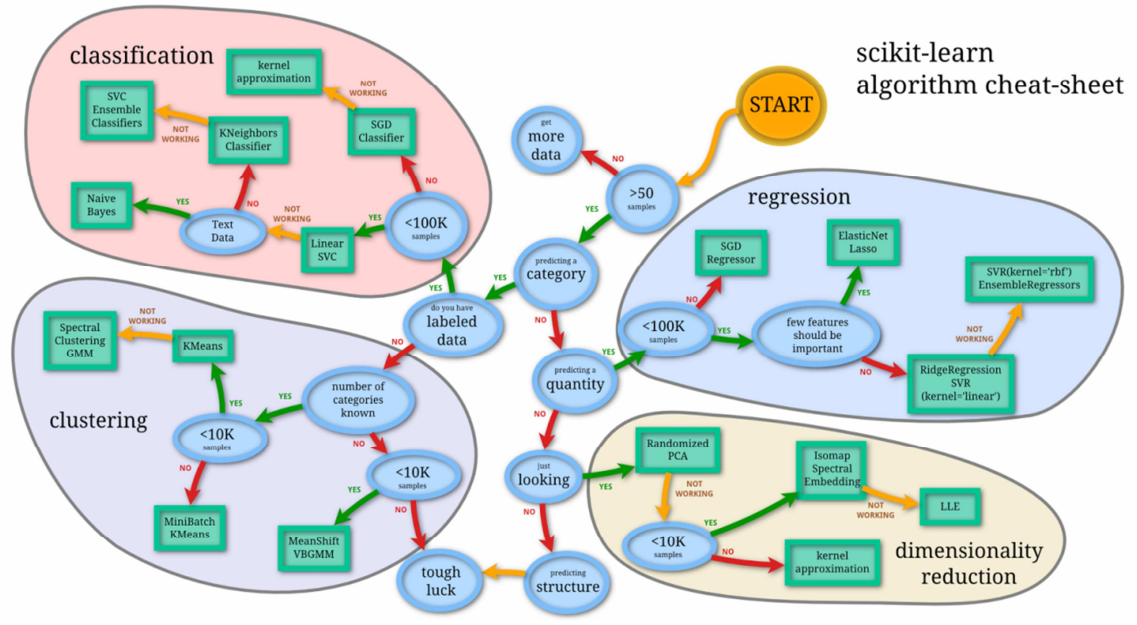
2.1.2 Unsupervised Machine Learning

Aim. Describe patterns or associations among a set of input measures.

Patterns or Associations

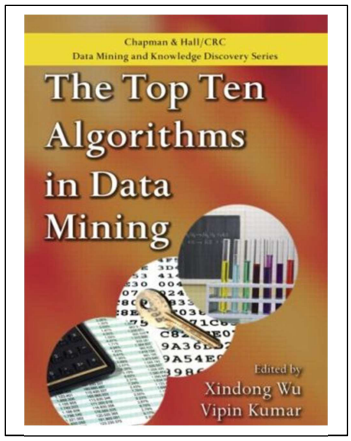
Clustering

Result: **DESCRIPTIVE MODELS**



2.1.3 The Top Ten Algorithms in Data Mining

IEEE International Conference on Data Mining (ICDM). http://www.cs.uvm.edu/~icdm/
The most influential algorithms used in the Data Mining Community.



1. C 4.5 (Decision Tree).
2. K-Means.
3. Support Vector Machine (SVM). The Best Generalization Ability
4. Apriori. To find frequent itemsets from a transaction dataset and derive association rules
5. EM (Expectation- Maximization) Pattern Recognition
6. PageRank. Link-based ranking algorithm, which also powers the Google search engine.
7. AdaBoost.
8. k-Nearest Neighbors (k-NN)
9. Naïve Bayes.
10. CART. Classification and Regression Trees

Source: http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf

3. Tools

3.1 WEKA (Waikato Environment for Knowledge Analysis)

http://www.cs.waikato.ac.nz/ml/weka/

- Data Mining Software in Java.
- Implemented in Java
- Multi-platform
- GUI (Limitations)
- GPL License.
- University of Waikato, New Zealand

3.2 R (#RStats)

http://www.r-project.org/

R is a language and environment for statistical computing and graphics.
- S Language (Bell Laboratories)
- Implemented in C/C++
- Highly extensible. R can be extended via packages.
- R Environment. Uses a command line interface. (NO GUI)
- RStudio. Graphical User Interfaces (GUI)
- GPL License.
- Created by University of Auckland, New Zealand and currently developed R Development Core Team

R for Linux | R for Mac OSX | R for Windows

RWeka

Links: How R grows
Books: Machine Learning for Hackers, The Elements of Statistical Learning: Data Mining, Inference and Prediction, OpenIntro Statistics
Enterprises: Revolution Analytics, Oracle R Enterprise, …

3.3 RapidMiner

http://rapid-i.com/content/view/181/190/lang,en/

- Open-Source Data Mining and Analysis System
- Implemented in Java
- Multi-platform
- Machine Learning library Weka fully integrated.
- Access to data sources: Excel, MySQL, Oracle
- ETL
- Reporting
- Data Analysis
- AGPL License
- Created by Dortmund University of Technology

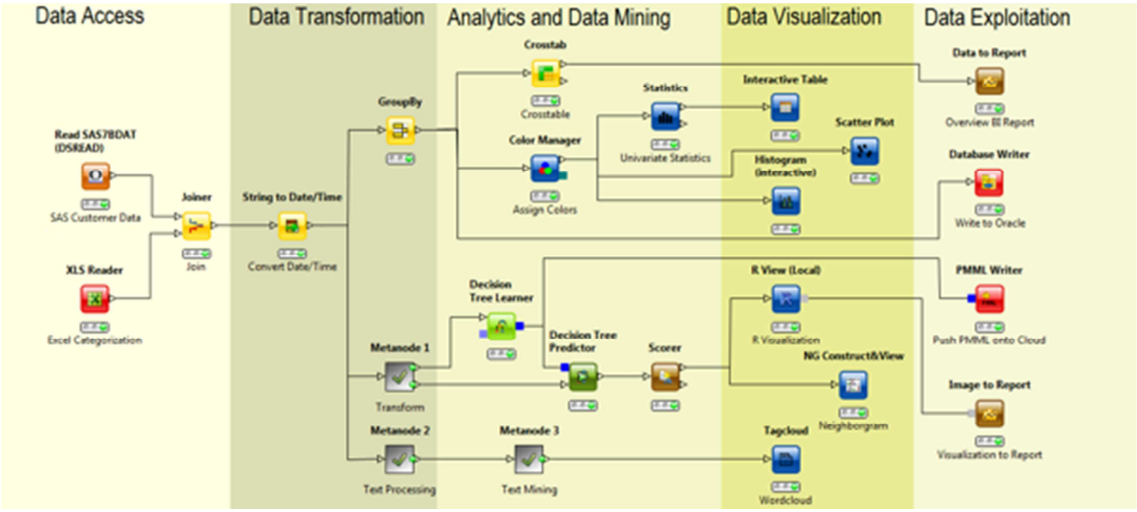
R Extension for RapidMiner

3.4 KNIME Desktop

http://www.knime.org/knime

- Data Analytics (Data access, data transformation, predictive analytics, visualization and reporting).
- Implemented in Java (Based in Eclipse Platform)
- Reporting
- ETL
- KNIME Extensions. Excel support, R integration, Weka
- GPL License
- Konstanz University, Germany

R Extension for Knime



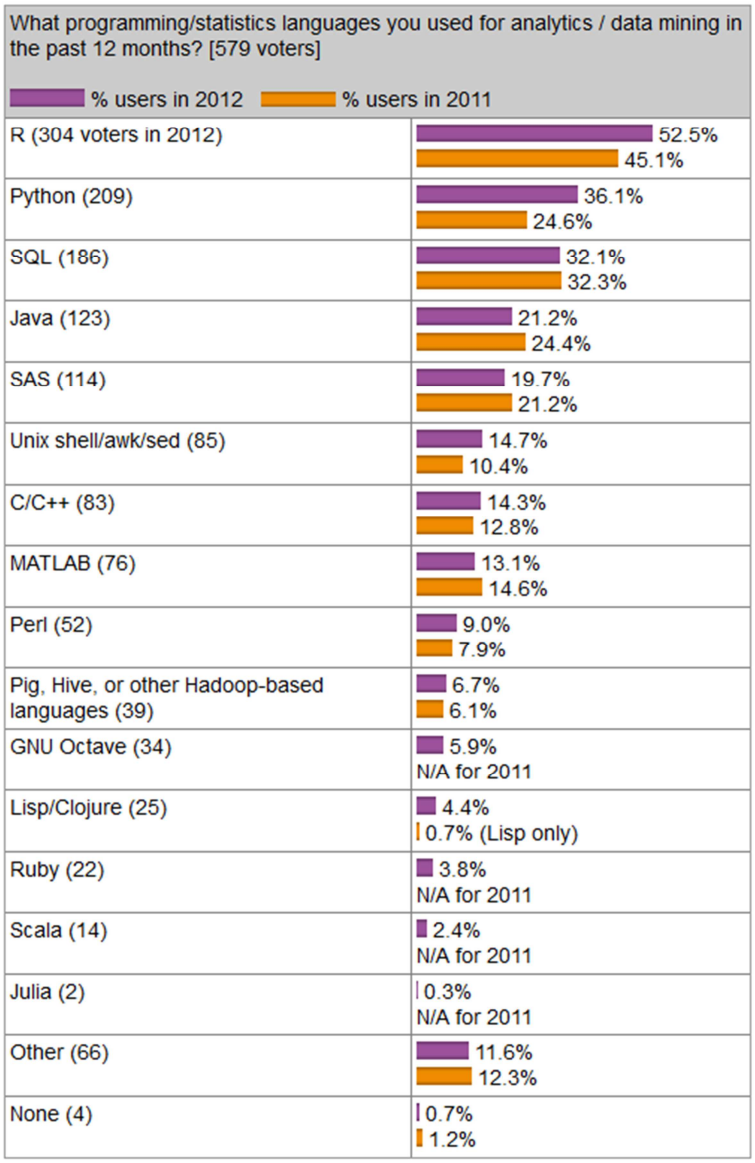
3.5 Orange

http://orange.biolab.si/

- A component-based data mining and machine learning software suite
- A visual programming front-end for explorative data analysis and visualization
- Multi-platform.
- Python
- GPL License
- University of Ljubljana, Slovenia

3.6 Polls

3.6.1 What programming/statistics languages you used for analytics / data mining in the past 12 months? [579 voters] (Aug 2012)



Source: http://www.kdnuggets.com/polls/2012/analytics-data-mining-programming-languages.html

3.6.2 What Analytics, Data mining, Big Data software you used in the past 12 months for a real project? (May 2012)

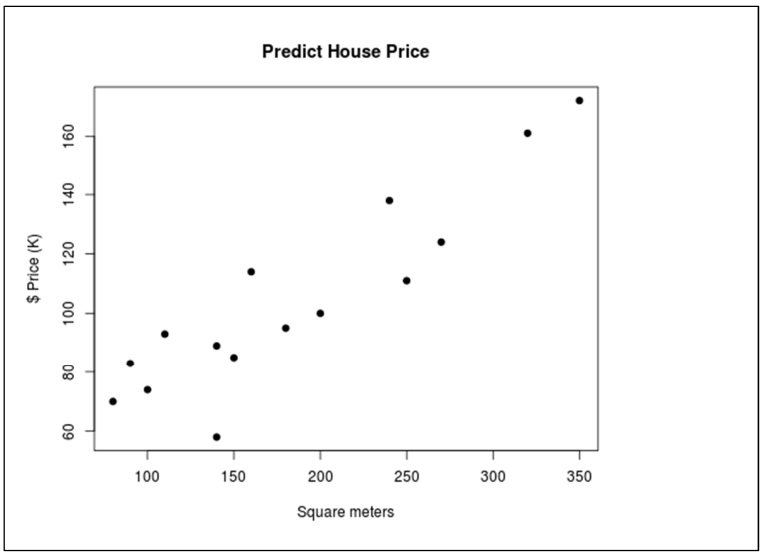
What Analytics, Data mining, Big Data software you used in the past 12 months for a real project (not just evaluation) [798 voters]	
Legend: Free/Open Source tools Commercial tools	% users in 2012 % users in 2011
R (245)	30.7% 23.3%
Excel (238)	29.8% 21.8%
Rapid-I RapidMiner (213)	26.7% 27.7%
KNIME (174)	21.8% 12.1%
Weka / Pentaho (118)	14.8% 11.8%
StatSoft Statistica (112)	14.0% 8.5%
SAS (101)	12.7% 13.6%
Rapid-I RapidAnalytics (83)	10.4% not asked in 2011
MATLAB (80)	10.0% 7.2%
IBM SPSS Statistics (62)	7.8% 7.2%
IBM SPSS Modeler (54)	6.8% 8.3%
SAS Enterprise Miner (46)	5.8% 7.1%
Orange (42)	5.3% 1.3%
Microsoft SQL Server (40)	5.0% 4.9%

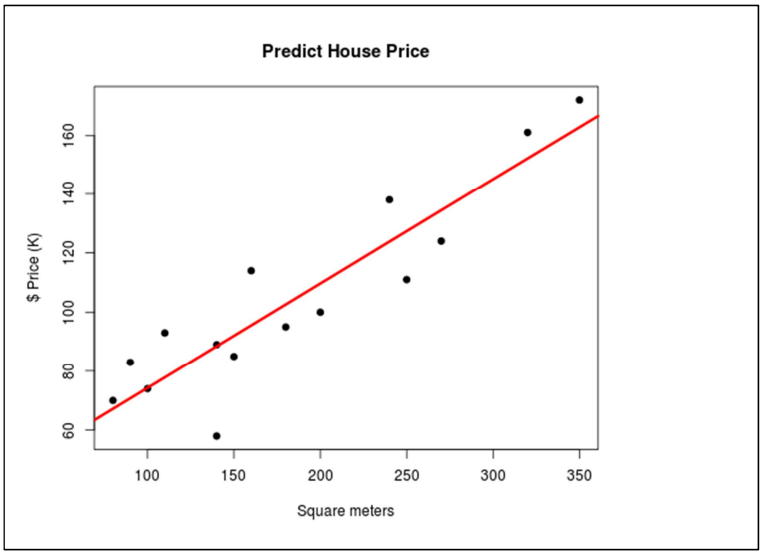
Source: http:/www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html

4. Examples

4.1 Predicting Price House

Size	Price (K)
80	70
90	83
100	74
110	93
140	89
140	58
150	85
160	114
180	95
200	100
240	138
250	111
270	124
320	161
350	172





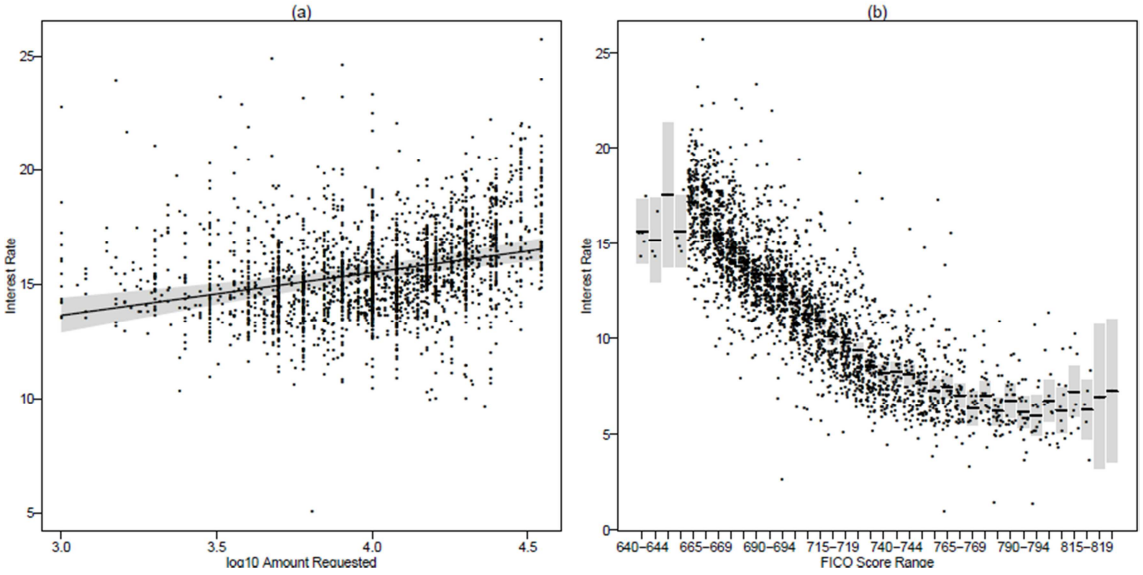
Regression Problem

Link: https://github.com/gsantosgo/RStats/blob/master/MadridJUG-DataMining/predictHousePrice.md

4.2 Lending Club

Peer to peer lending company.

What are the variables associated with the interest rate of a loan? Multivariate



Regression Problem

Links: http://www.lendingclub.com/
http://en.wikipedia.org/wiki/Lending_Club
https://github.com/gsantosgo/RStats/blob/master/MadridJUG-DataMining/loansLendingClub.md

4.3 Spam or Ham Email

Classification Problem

Links: https://github.com/gsantosgo/RStats/blob/master/MadridJUG-DataMining/spam.md

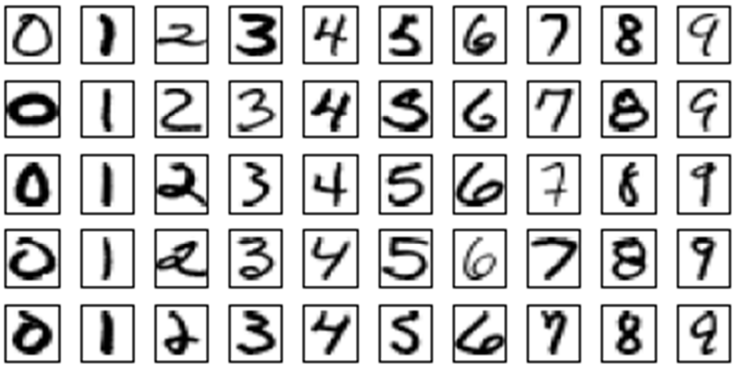
4.4 Handwritten Digit Recognition

Identification the numbers in a handwritten ZIP code, from a digitized image.

16x16

```
001 002 003 004 ... 015 016
017 018 019 020 ... 031 032
033 034 035 036 ... 037 038
 |   |   |   |  ...  |   |
209 210 211 212 ... 223 224
225 226 227 228 ... 239 240
241 242 243 244 ... 255 256
```

Each image is a 16 x 16 (256) 8-bit grayscale representation of a handwritten digit



Classification Problem

http://www.kaggle.com/c/digit-recognizer

Link: https://github.com/gsantosgo/RStats/blob/master/MadridJUG-DataMining/handwritten.md

4.5 Human Activity Recognition using Smartphones

We used data obtained from accelerometer and gyroscope sensor signals of the smartphones

3-axial linear acceleration
3-axial angular velocity
We can monitor acceleration, positions, rotation and angular motion.

Laying, Sitting, Standing, Walk, WalkDown, WalkUp

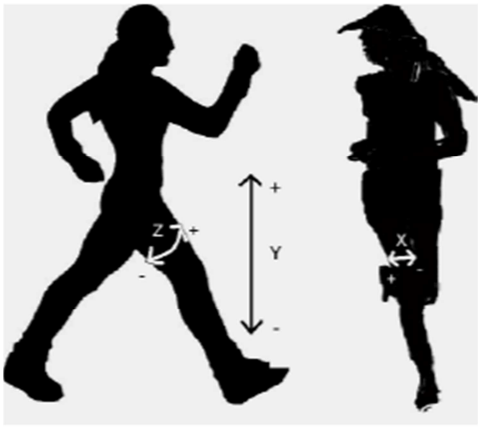
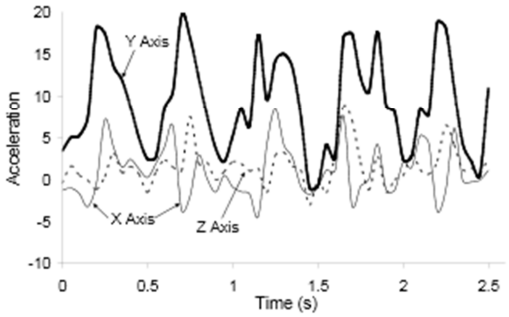
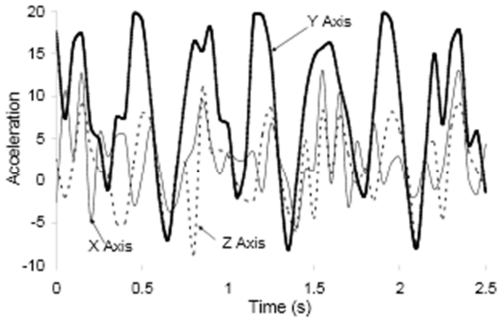


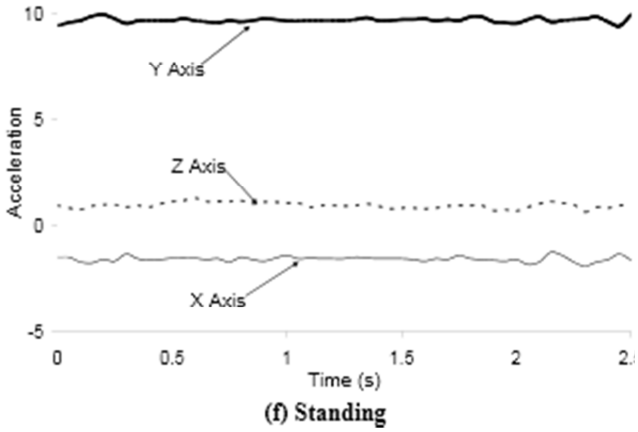
Figure 1: Axes of Motion Relative to User



(a) Walking



(b) Jogging



(f) Standing

Classification Problem

DataSet: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
Source: Activity Recognition using Cell Phone Accelerometers
http://www.cis.fordham.edu/wisdm/public_files/sensorKDD-2010.pdf

Link: https://github.com/gsantosgo/RStats/blob/master/MadridJUG-DataMining/handwritten.md

4.6 Inventory

A large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

Regression Problem

4.7 Image Classification

Computer Vision (C.V.)

Haralick texture features. Haralick described 14 statistics that can be calculated from the co-occurrence matrix with the intent of describing the texture of the image:

- Angular Second Moment
- Constrast
- Correlation

..

Source: https://github.com/gsantosgo/RStats/tree/master/MadridJUG-DataMining/data/faces.arff

Alessandra Ambrosio

Jessica Alba

Megan Fox

Classification Problem

Links: http://murphylab.web.cmu.edu/publications/boland/boland_node26.html

4.8 Clustering

Google News

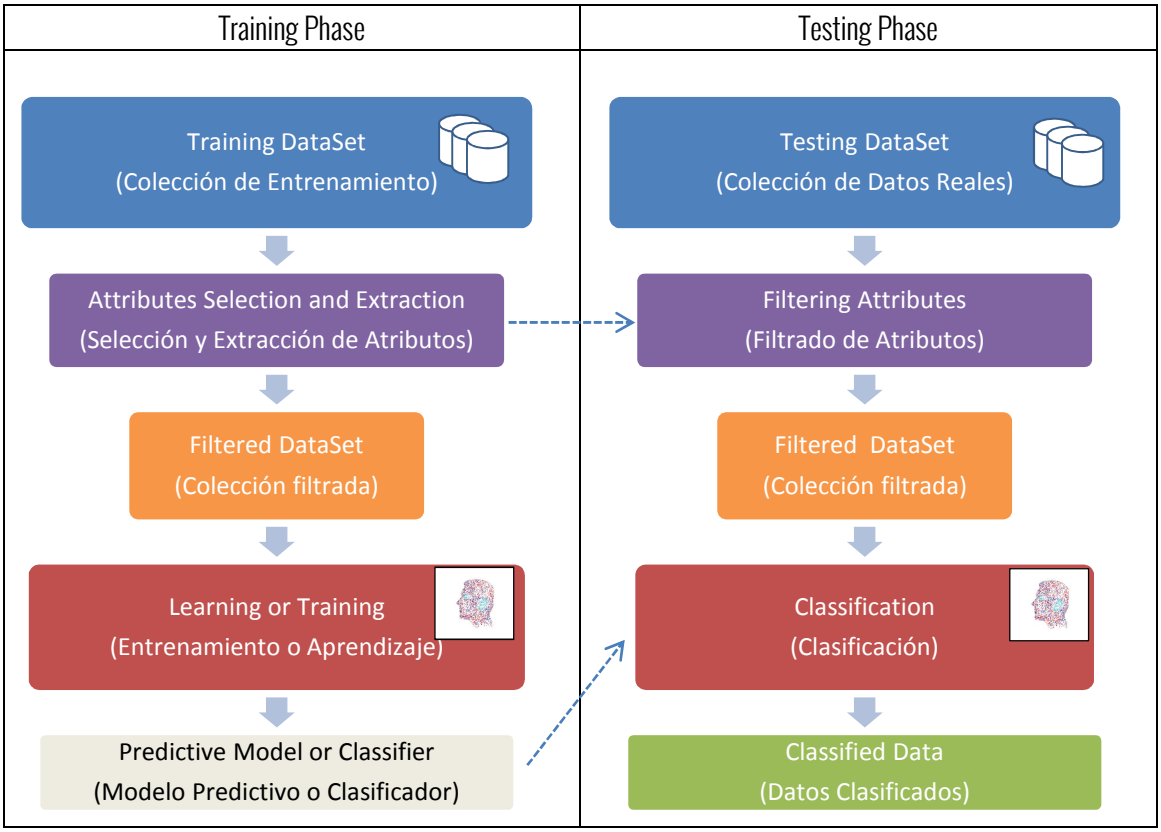
News Clustering

Clustering Problem

Source: http://news.google.es/

5. Supervised Machine Learning

Guide for Supervised Machine Learning



6. Evaluation

STATE OF ART

6.1 Random Subsampling

1. Use the training set.
2. Split it into training set (66.66 %) and testing set (33.33%). (RANDOM)
3. Build a model on the training set.
4. Evaluate on the test set.

6.2 Cross Validation (K-FOLD)

1. Use the training set.
2. Split it into training/test sets.
3. Build a model on the training set
4. Evaluate on the test set.
5. Repeat and average the estimated

Never Overlap!

K-FOLD

K = 1

K = 2

.....

K = 10

$$Total\ error\ estimate = \frac{1}{N}\sum_{i=1}^{N} e_i$$

Link: https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada

6.3 Confusion Matrix

- Accuracy (Precisión o Efectividad) . The rate of correct predictions
- Error rate. The rate of incorrect predictions.
- Performance (Eficiencia). The algorithm is quick or nor in the training phase or in the testing phase.

Actual/Real Class	Predicted Class		Total
	Yes	No	
Yes (1)	True Positive (TP)	False Negative (FN)	Total Positive Real (TPR)
No (0)	False Positive (FP)	True Negative (TN)	Total Negative Real (TNR)
Total	Total Positive Predicted (TPP)	Total Negative Predicted (TNP)	Total

Link: http://en.wikipedia.org/wiki/Confusion_matrix

A.1. ¿What is a DATASET?

Example: Dataset email50

```
##    spam num_char line_breaks format number
## 1     0   21.705          551      1  small
## 2     0    7.011          183      1    big
## 3     1    0.631           28      0   none
## 4     0    2.454           61      0  small
## 49    0    8.937          211      1  small
## 50    0   15.829          242      1  small
```

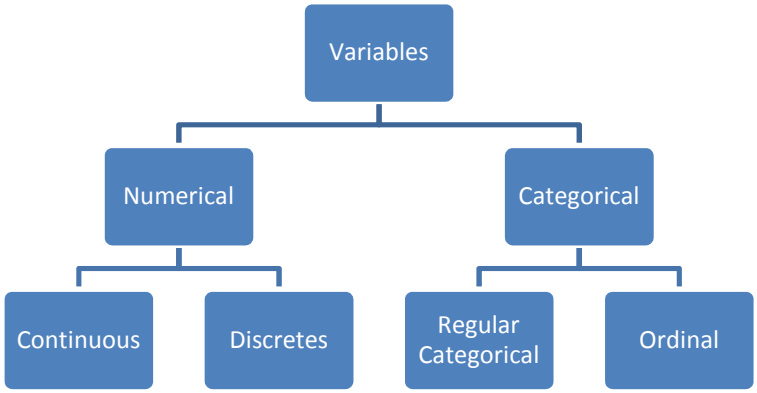
Row represents a **case**, a **unit of observation**, an **observational unit**, an **instance**. **OBSERVATIONS. EXAMPLE OR EXEMPLARY.**
Column represents an **attribute**, a **variable**, a **feature** (represent characteristics).
Special column. **the class**, **the class label** (two values or multi-valued)

For example: The email 4, which is not spam, contains 2454 characters, 61 line breaks, is written in Text format (0=text, 1=html), and contains only small numbers.

Variable	Description
spam	Specifies whether the message was spam
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
Format	Indicates if the email contained special formatting, such as bolding, tables or links, which would indicate the message is in HTML format
Number	Indicates whether the email contained no number, a small number (under 1 million) or a large number

Dataset represents a **data matrix**, **data frame**. Each row of a data matrix corresponds to unique case (example), and each column corresponds to a variable.

A.2 Types of variables



num_char and line_breaks CUANTITATIVE, NUMERICAL AND CONTINOUS VARIABLES.

spam CUANTITATIVE, NUMERICAL AND DISCRETE VARIABLE.

number indicates whether the email contained no number, a small number (under 1 million) or a large number. It takes values *none*, *small* and *big*. The different levels have a natural ordering. CUALITATIVE, CATEGORICAL VARIABLES AND ORDINAL VARIABLE.