PRESENTATION

BIG DATA INTELLIGENCE

METHODS AND TECHNOLOGIES

A.K.A. MACHINE LEARNING I

RICARDO ALER MUR (aler@inf.uc3m.es). 2.2B29

MASTER IN BIG DATA ANALYTICS

GOALS

1. To introduce the basics of **Machine Learning**: training, testing, models, hyper-parameter tuning, etc.
2. To show how to do Machine Learning in a **Big Data** Context
3. To apply the above goals with currently used **tools**

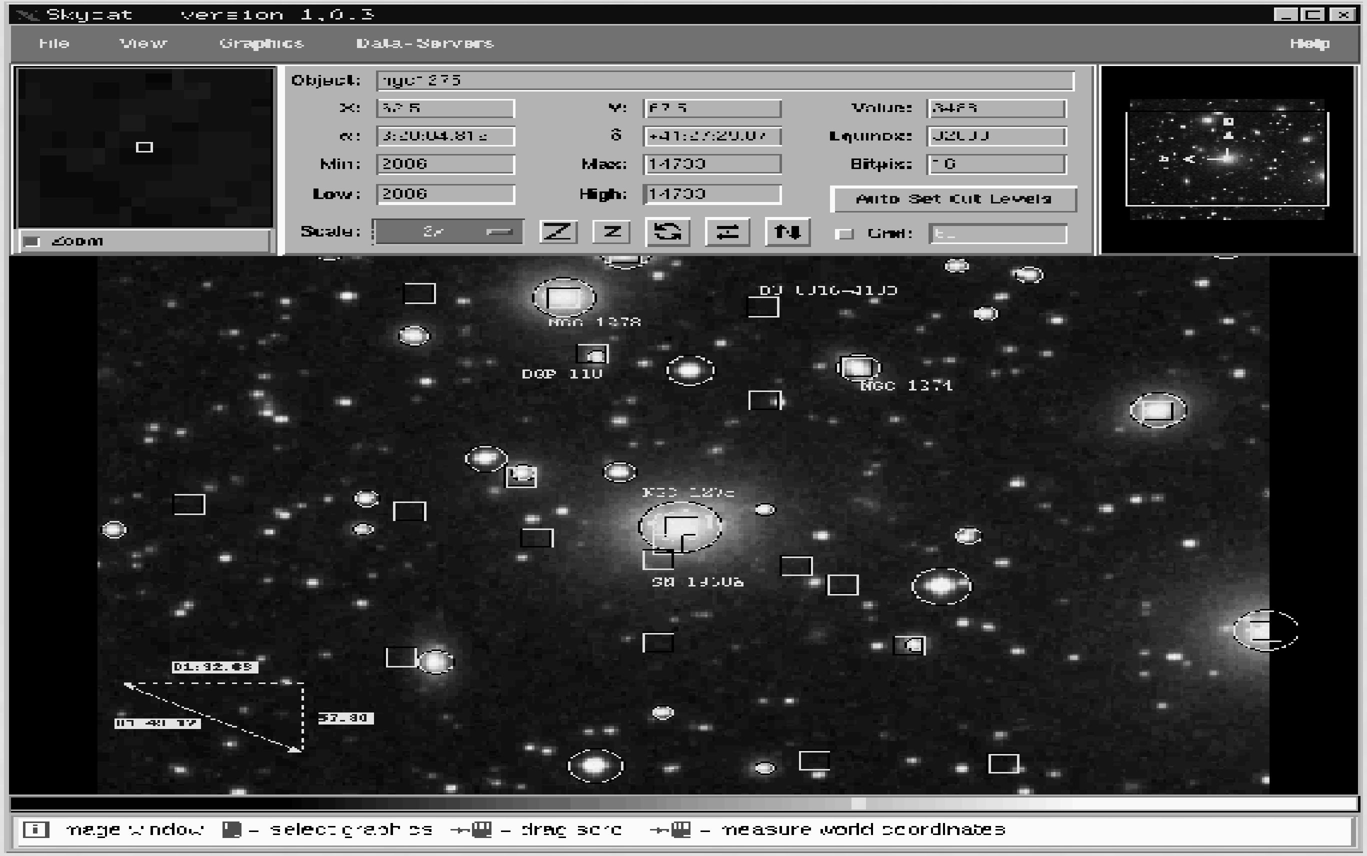
MACHINE LEARNING

- In general, it’s a subfield of **Artificial Intelligence** that tries to make computers and machines learn

- In practice, it tries to create models from data and thus is closely related to statistics. This is the point of view we will follows in this course

WHAT IS MACHINE LEARNING

- Example: Skycat: AUTOMATIC CLASSIFICATION OF OBJECTS IN THE SKY



Skycat version 1.0.3

Object: ngc 1275

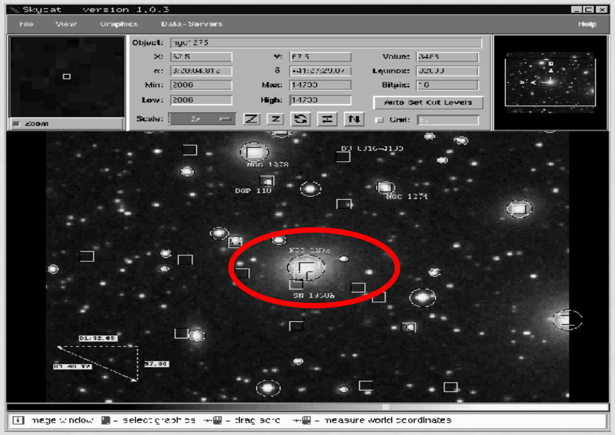
X: 32.5 Y: 67.5 Value: 3483

α: 3:20:04.81s δ: +41:27:29.07 Equinox: J2000

Min: 2006 Max: 14700 Bitpix: 16

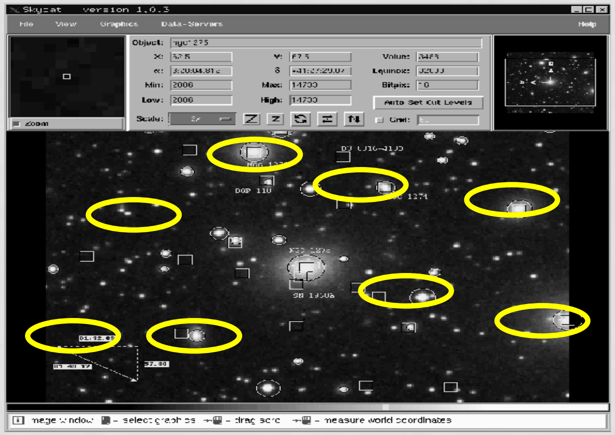
Low: 2006 High: 14700

Scale: 2x



?

Training data (labeled pictures of sky objects: galaxies, stars, nebulae, …)



Pictures in the catalog have been labeled by a human expert (astronomer)

ML Algorithm

Model

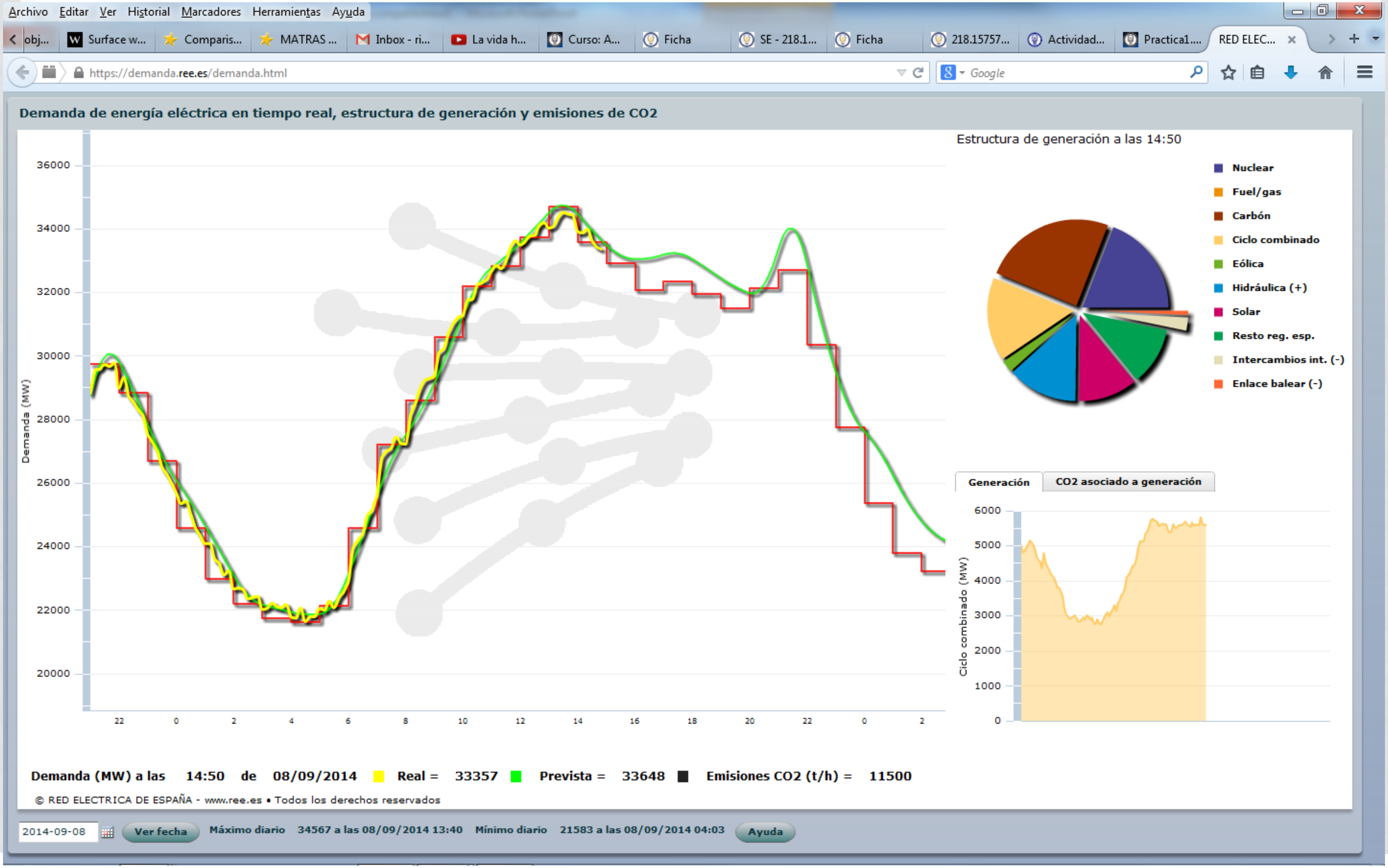
Spiral galaxy

APPLICATIONS

- Finances and banking
  - Credit card fraud detection
  - Credit default prediction
- Market analysis:
  - Market basket analysis
  - Market segmentation
- Insurance:
  - Expensive clients
- Education:
  - Prediction of school dropouts
- Industry:
  - Electric (energy) load forecasting
  - Solar / wind energy forecasting

https://demanda.ree.es/demanda.html

ELECTRIC LOAD FORECASTING



Demanda de energía eléctrica en tiempo real, estructura de generación y emisiones de CO2

Estructura de generación a las 14:50

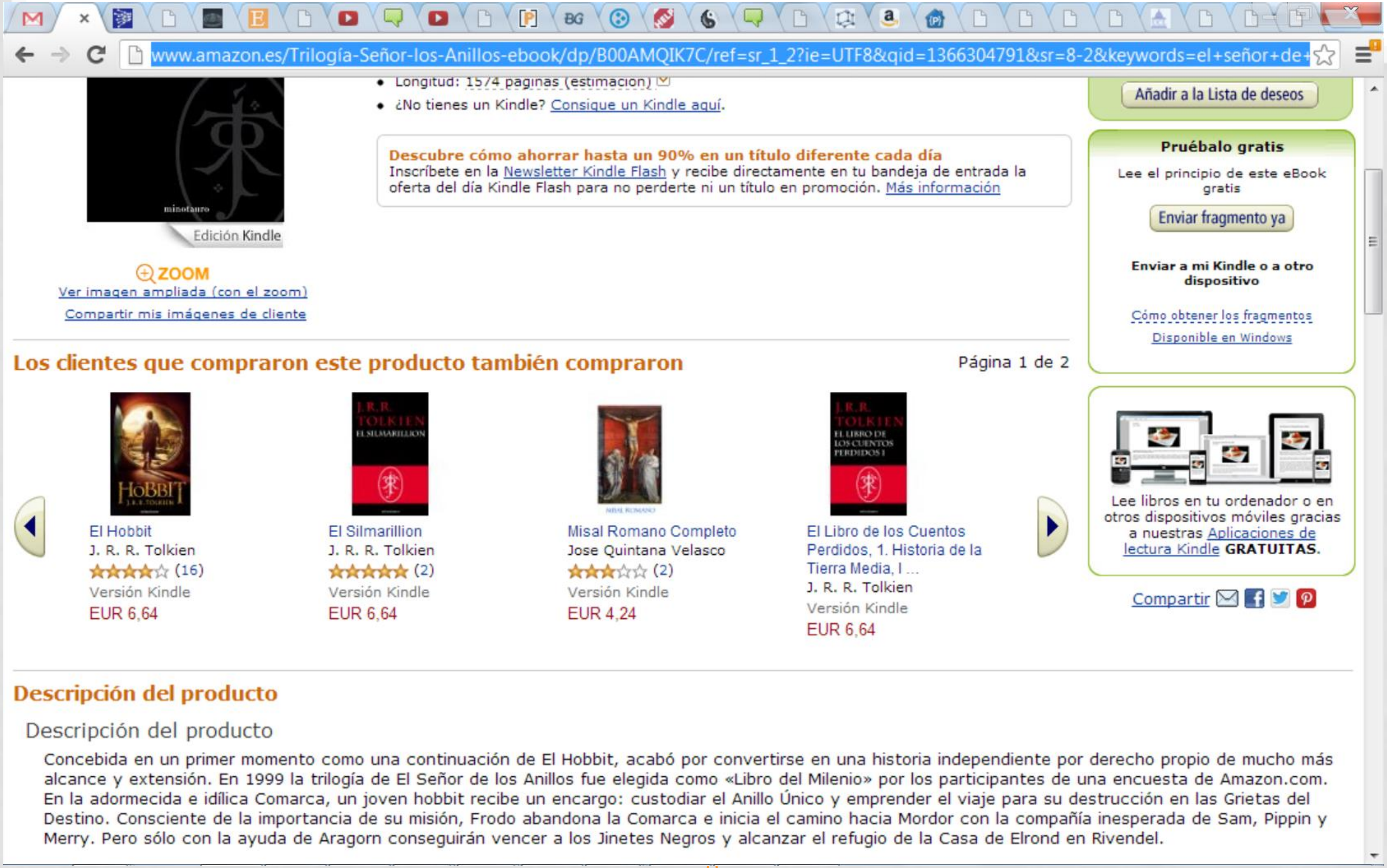
Demanda (MW) a las 14:50 de 08/09/2014 🟨 Real = 33357 🟩 Prevista = 33648 ■ Emisiones CO2 (t/h) = 11500

© RED ELECTRICA DE ESPAÑA - www.ree.es • Todos los derechos reservados

2014-09-08 Ver fecha Máximo diario 34567 a las 08/09/2014 13:40 Mínimo diario 21583 a las 08/09/2014 04:03 Ayuda

APPLICATIONS II

- Medicine:
  - Illness diagnosis
- Science:
  - Illness prediction from DNA analysis
  - Prediction if a new substance causes cancer
  - SKYCAT
- Internet:
  - Spam detection (SpamAssassin)
  - Web: book recommendation (amazon.com)



www.amazon.es/Trilogía-Señor-los-Anillos-ebook/dp/B00AMQIK7C/ref=sr_1_2?ie=UTF8&qid=1366304791&sr=8-2&keywords=el+señor+de+

- Longitud: 1574 páginas (estimación)
- ¿No tienes un Kindle? Consigue un Kindle aquí.

Descubre cómo ahorrar hasta un 90% en un título diferente cada día
Inscríbete en la Newsletter Kindle Flash y recibe directamente en tu bandeja de entrada la oferta del día Kindle Flash para no perderte ni un título en promoción. Más información

Ver imagen ampliada (con el zoom)
Compartir mis imágenes de cliente

Pruébalo gratis
Lee el principio de este eBook gratis

Enviar a mi Kindle o a otro dispositivo
Cómo obtener los fragmentos
Disponible en Windows

Los clientes que compraron este producto también compraron

Página 1 de 2

El Hobbit J. R. R. Tolkien ★★★★☆ (16) Versión Kindle EUR 6,64	El Silmarillion J. R. R. Tolkien ★★★★★ (2) Versión Kindle EUR 6,64	Misal Romano Completo Jose Quintana Velasco ★★★☆☆ (2) Versión Kindle EUR 4,24	El Libro de los Cuentos Perdidos, 1. Historia de la Tierra Media, I ... J. R. R. Tolkien Versión Kindle EUR 6,64

Lee libros en tu ordenador o en otros dispositivos móviles gracias a nuestras Aplicaciones de lectura Kindle **GRATUITAS**.

Compartir

Descripción del producto

Descripción del producto

Concebida en un primer momento como una continuación de El Hobbit, acabó por convertirse en una historia independiente por derecho propio de mucho más alcance y extensión. En 1999 la trilogía de El Señor de los Anillos fue elegida como «Libro del Milenio» por los participantes de una encuesta de Amazon.com. En la adormecida e idílica Comarca, un joven hobbit recibe un encargo: custodiar el Anillo Único y emprender el viaje para su destrucción en las Grietas del Destino. Consciente de la importancia de su misión, Frodo abandona la Comarca e inicia el camino hacia Mordor con la compañía inesperada de Sam, Pippin y Merry. Pero sólo con la ayuda de Aragorn conseguirán vencer a los Jinetes Negros y alcanzar el refugio de la Casa de Elrond en Rivendel.

SYLLABUS

1. Overview and introduction to Machine Learning: tasks and models.
2. Predictive models:
   - Decision trees, regression trees
   - K Nearest Neighbour (KNN)
   - Machine Learning pipeline: training, => ML algorithm => model => test / evaluation . Preprocessing, hyperparameter tuning, …
3. Ensemble methods: bagging, boosting, stacking
4. Preprocessing: selection of attributes and methods of dimensionality reduction
5. Machine learning software for Big Data:
   1. Python: scikit-learn, numpy
   2. Mapreduce
   3. Spark: pyspark, MLLIB
6. Other topics:
   1. Online learning
   2. Metaheuristics: genetic algorithms, genetic programming, …

SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models

2. Methods for training classification and regression models:
   - Nearest Neighbour (KNN)
   - Decision / regression trees & rules

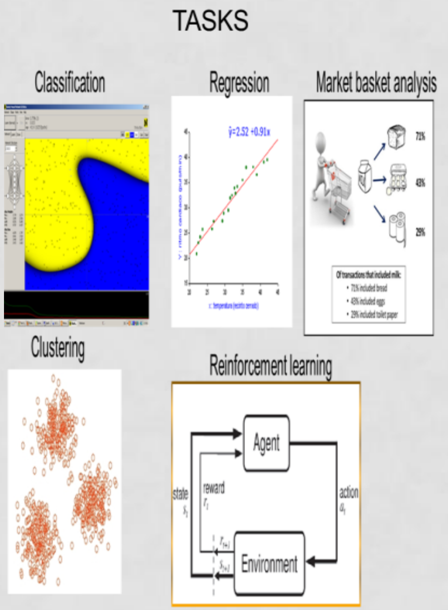
3. Methodology and the Machine Learning pipeline

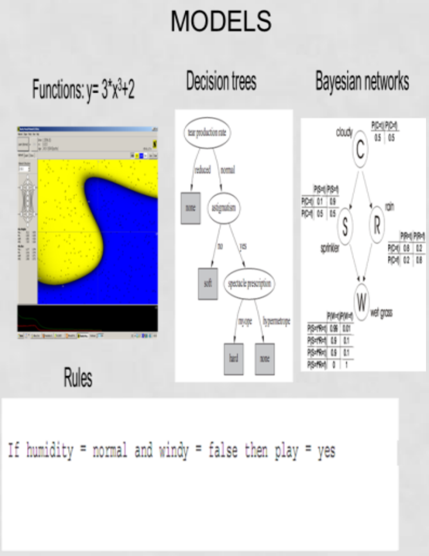
4. Methods for attributes (feature selection, transformation)

5. Methods based on ensembles of models: Bagging, boosting, stacking

6. Large Scale Machine Learning. Big Data
   - Map-reduce & Spark

7. Software tools:
   - Python + scikit-learn & Pyspark + MLIB





If humidity = normal and windy = false then play = yes

And many more: neural networks, nearest neighbor, ...

SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models

2. Methods for training classification and regression models:

- Nearest Neighbour (KNN)
- Decision / regression trees & rules

3. Methodology and the Machine Learning pipeline

4. Methods for attributes (feature selection, transformation)

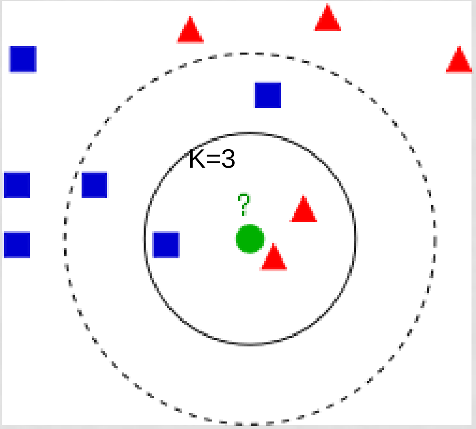
5. Methods based on ensembles of models: Bagging, boosting, stacking

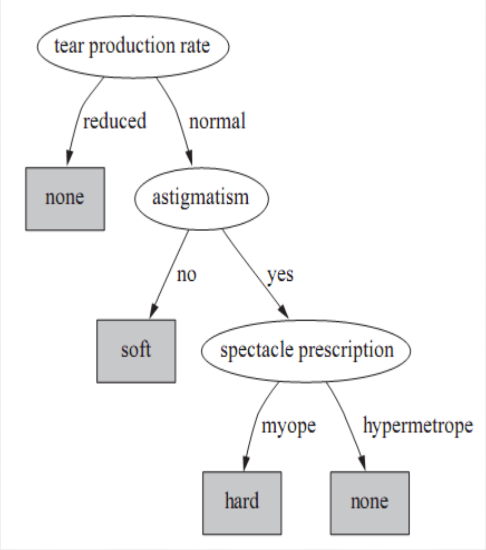
6. Large Scale Machine Learning. Big Data

- Map-reduce & Spark

7. Software tools:

- Python + scikit-learn & Pyspark + MLIB





SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models

2. Methods for training classification and regression models:
   - Nearest Neighbour (KNN)
   - Decision / regression trees & rules

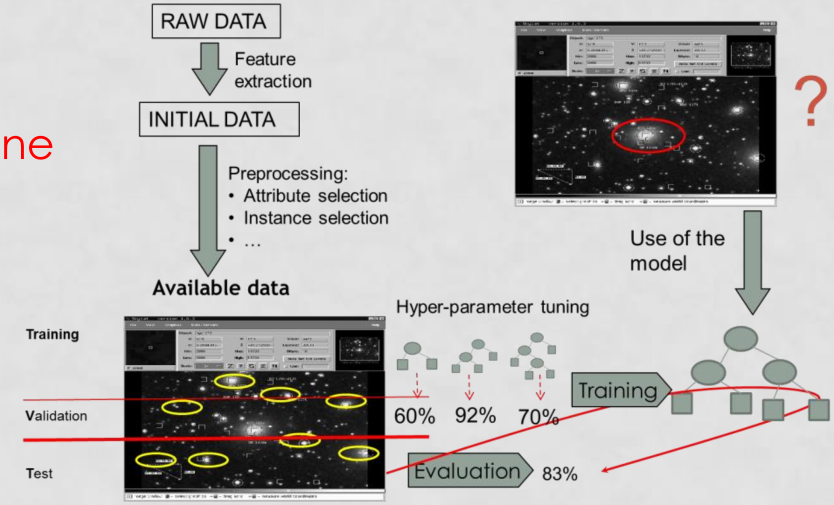
3. Methodology and the Machine Learning pipeline

4. Methods for attributes (feature selection, transformation)

5. Methods based on ensembles of models: Bagging, boosting, stacking

6. Large Scale Machine Learning. Big Data
   - Map-reduce & Spark

7. Software tools:
   - Python + scikit-learn & Pyspark + MLIB



RAW DATA

Feature extraction

INITIAL DATA

Preprocessing:
- Attribute selection
- Instance selection
- …

Available data

Training

Validation

Test

Hyper-parameter tuning

60% 92% 70%

Training

Evaluation 83%

Use of the model

?

SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models

2. Methods for training classification and regression models:

- Nearest Neighbour (KNN)
- Decision / regression trees & rules

3. Methodology and the Machine Learning pipeline

4. Methods for attributes (feature selection, transformation)

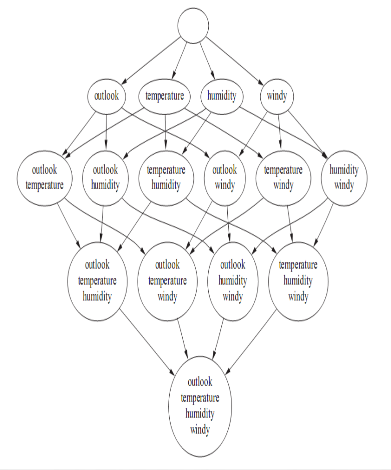
5. Methods based on ensembles of models: Bagging, boosting, stacking

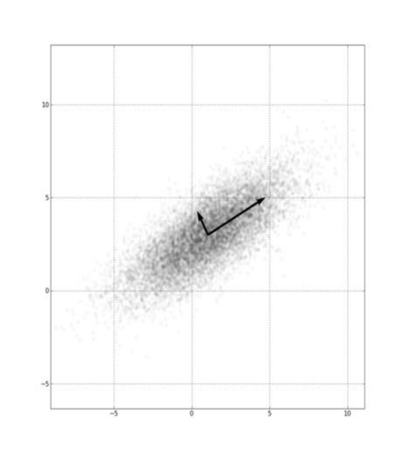
6. Large Scale Machine Learning. Big Data

- Map-reduce & Spark

7. Software tools:

- Python + scikit-learn & Pyspark + MLIB





SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models

2. Methods for training classification and regression models:

- Nearest Neighbour (KNN)
- Decision / regression trees & rules

3. Methodology and the Machine Learning pipeline

4. Methods for attributes (feature selection, transformation)

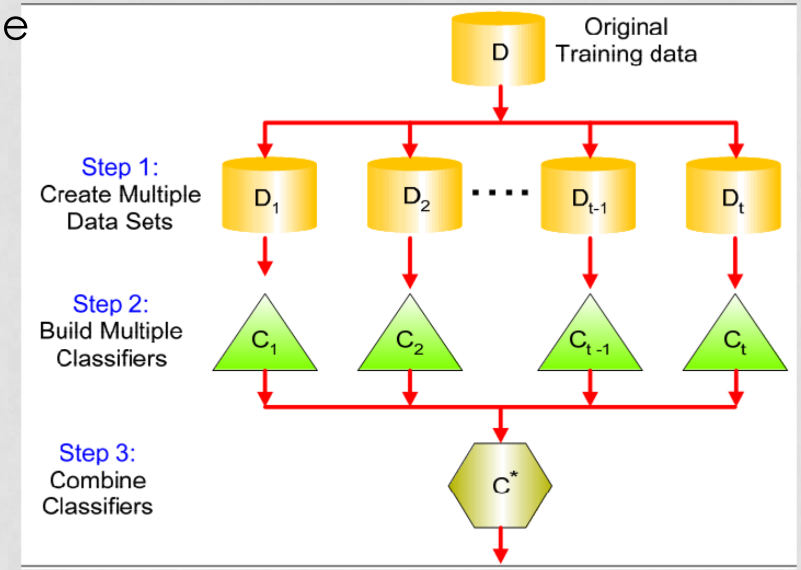
5. Methods based on ensembles of models: Bagging, boosting, stacking

6. Large Scale Machine Learning. Big Data

- Map-reduce & Spark

7. Software tools:

- Python + scikit-learn & Pyspark + MLIB



SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models

2. Methods for training classification and regression models:

   - Nearest Neighbour (KNN)
   - Decision / regression trees & rules

3. Methodology and the Machine Learning pipeline

4. Methods for attributes (feature selection, transformation)

5. Methods based on ensembles of models: Bagging, boosting, stacking

6. Large Scale Machine Learning. Big Data

   - Map-reduce & Spark

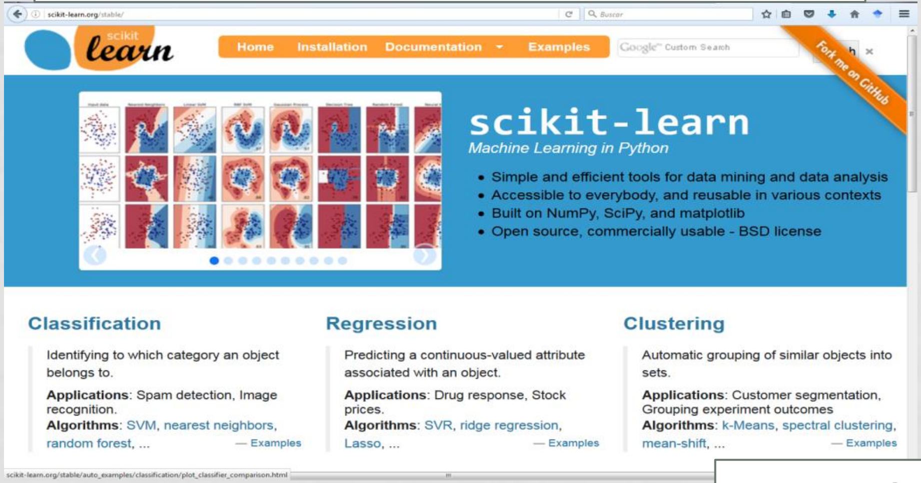
7. Software tools:

   - Python + scikit-learn & Pyspark + MLIB

SYLLABUS:

7. SOFWARE TOOLS

FOR MACHINE LEARNING BASICS
ANACONDA (python + scikit-learn)



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

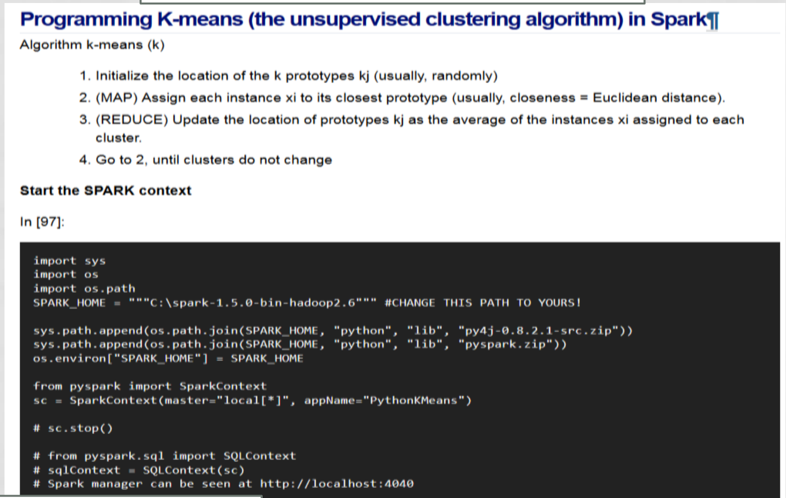
Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

PYSPARK AND MLIB



Programming K-means (the unsupervised clustering algorithm) in Spark¶

Algorithm k-means (k)

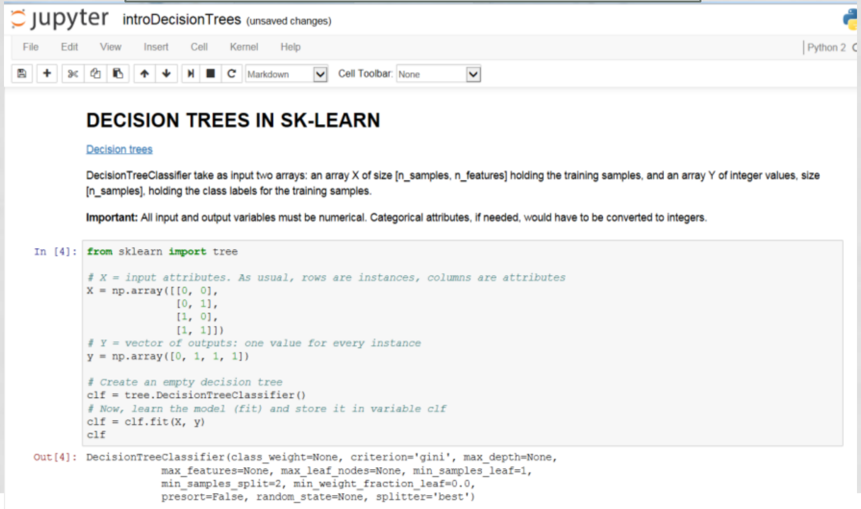
1. Initialize the location of the k prototypes kj (usually, randomly)
2. (MAP) Assign each instance xi to its closest prototype (usually, closeness = Euclidean distance).
3. (REDUCE) Update the location of prototypes kj as the average of the instances xi assigned to each cluster.
4. Go to 2, until clusters do not change

Start the SPARK context

In [97]:

```
import sys
import os
import os.path
SPARK_HOME = """C:\spark-1.5.0-bin-hadoop2.6""" #CHANGE THIS PATH TO YOURS!

sys.path.append(os.path.join(SPARK_HOME, "python", "lib", "py4j-0.8.2.1-src.zip"))
sys.path.append(os.path.join(SPARK_HOME, "python", "lib", "pyspark.zip"))
os.environ["SPARK_HOME"] = SPARK_HOME

from pyspark import SparkContext
sc = SparkContext(master="local[*]", appName="PythonKMeans")

# sc.stop()

# from pyspark.sql import SQLContext
# sqlContext = SQLContext(sc)
# Spark manager can be seen at http://localhost:4040
```

IPYTHON NOTEBOOKS



DECISION TREES IN SK-LEARN

Decision trees

DecisionTreeClassifier take as input two arrays: an array X of size [n_samples, n_features] holding the training samples, and an array Y of integer values, size [n_samples], holding the class labels for the training samples.

Important: All input and output variables must be numerical. Categorical attributes, if needed, would have to be converted to integers.

```
In [4]: from sklearn import tree

        # X = input attributes. As usual, rows are instances, columns are attributes
        X = np.array([[0, 0],
                      [0, 1],
                      [1, 0],
                      [1, 1]])
        # Y = vector of outputs: one value for every instance
        y = np.array([0, 1, 1, 1])

        # Create an empty decision tree
        clf = tree.DecisionTreeClassifier()
        # Now, learn the model (fit) and store it in variable clf
        clf = clf.fit(X, y)
        clf
```

```
Out[4]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                    max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
                    min_samples_split=2, min_weight_fraction_leaf=0.0,
                    presort=False, random_state=None, splitter='best')
```

GRADING

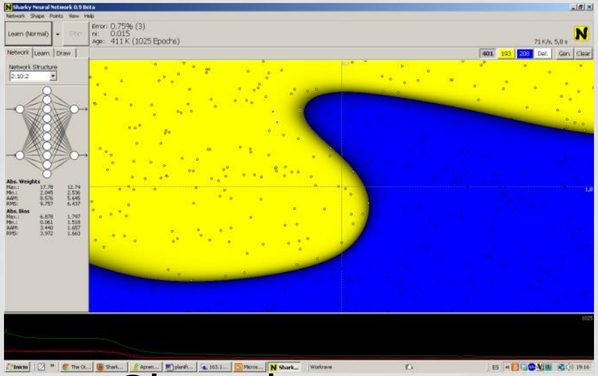
- A=30% FINAL EXAM
- B = 70% ASSIGNMENTS
  - A1: Programming
  - A2: Scikit-learn
  - A3: Pyspark / MLLIB

- Pass if A+B>=50%

TASKS AND ALGORITHMS

- What can be done?

Classification



Error: 0.75% (3)
ni: 0.015
Age: 411 K (1025 Epochs)
71 K/s, 5,8 s

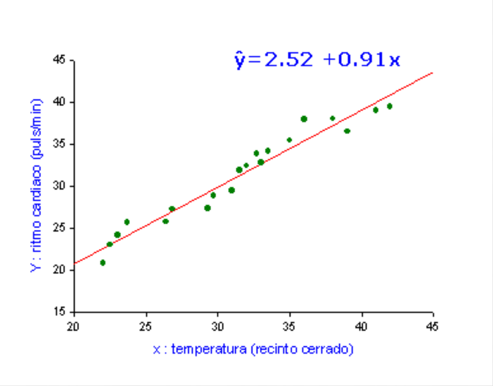
Abs. Weights

Max.:	17.78	12.74
Min.:	2.045	2.536
AAM:	8.576	5.645
RMS:	9.757	6.437

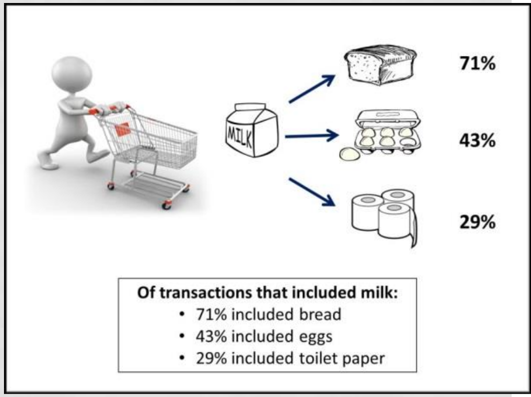
Abs. Bias

Max.:	6.878	1.797
Min.:	0.061	1.518
AAM:	3.440	1.657
RMS:	3.972	1.663

Regression

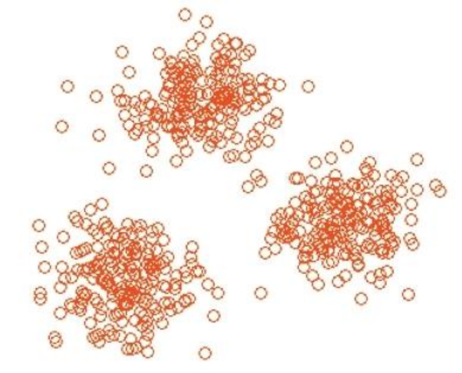


Market basket analysis

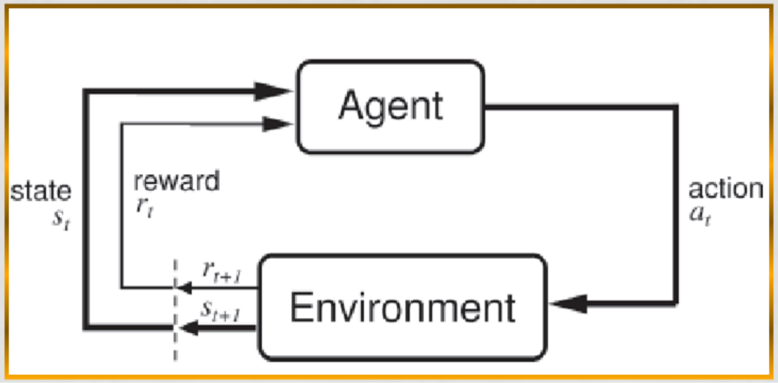


Of transactions that included milk:
- 71% included bread
- 43% included eggs
- 29% included toilet paper

Clustering



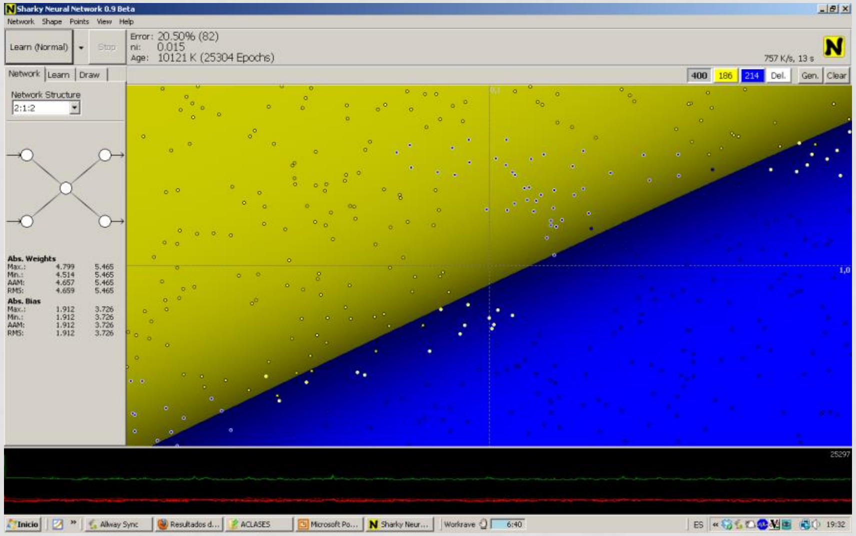
Reinforcement learning



MODELS

- What models can be obtained?

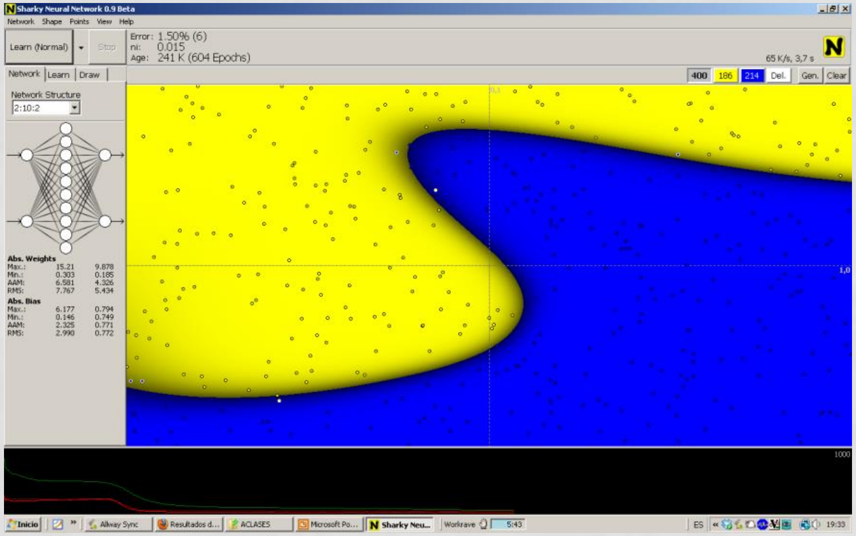
Linear



Error: 20.50% (82)
ni: 0.015
Age: 10121 K (25304 Epochs)

```
Abs. Weights
Max.:     4.799   5.465
Min.:     4.514   5.465
AAM:      4.657   5.465
RMS:      4.659   5.465
Abs. Bias
Max.:     1.912   3.726
Min.:     1.912   3.726
AAM:      1.912   3.726
RMS:      1.912   3.726
```

Non linear



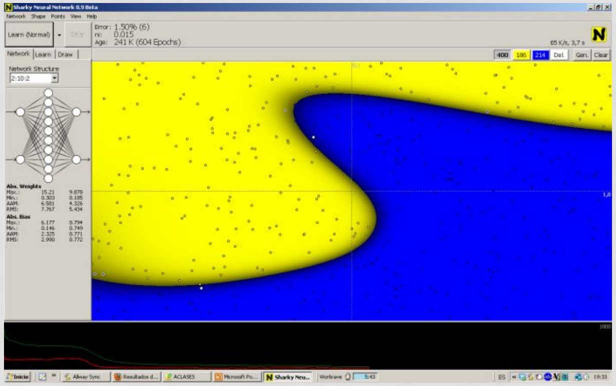
Error: 1.50% (6)
ni: 0.015
Age: 241 K (604 Epochs)

```
Abs. Weights
Max.:     15.21   9.878
Min.:     0.303   0.185
AAM:      6.581   4.326
RMS:      7.767   5.434
Abs. Bias
Max.:     6.177   0.794
Min.:     0.146   0.749
AAM:      2.325   0.771
RMS:      2.990   0.772
```

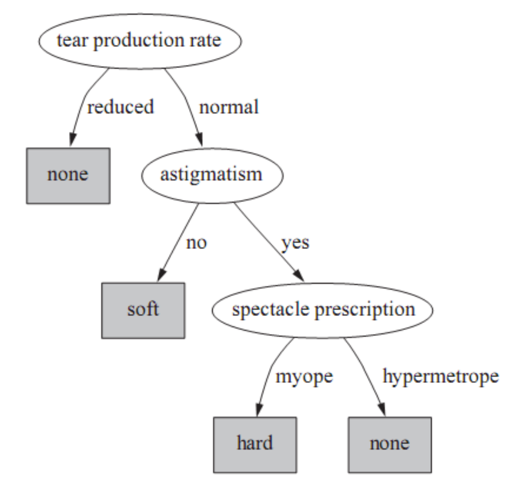
MODELS

- What models can be obtained?

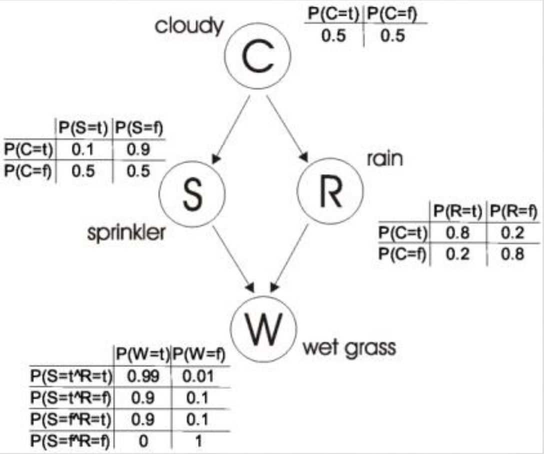
Functions: $y= 3*x^3+2$



Decision trees



Bayesian networks



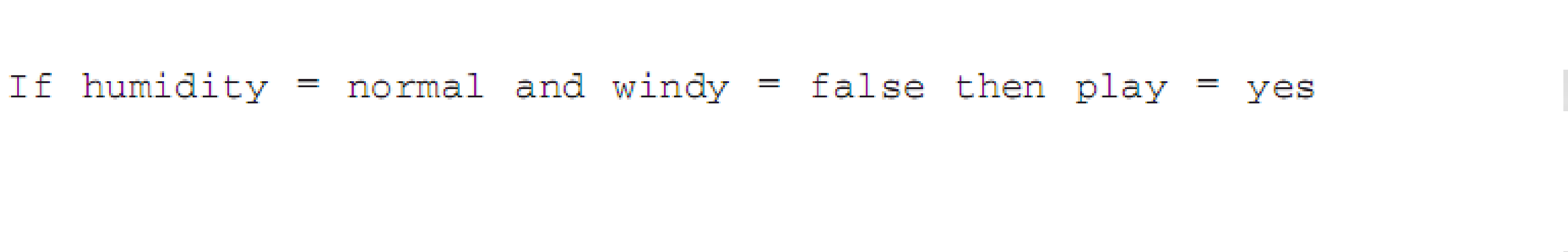
P(C=t)	P(C=f)
0.5	0.5

	P(S=t)	P(S=f)
P(C=t)	0.1	0.9
P(C=f)	0.5	0.5

	P(R=t)	P(R=f)
P(C=t)	0.8	0.2
P(C=f)	0.2	0.8

	P(W=t)	P(W=f)
P(S=t^R=t)	0.99	0.01
P(S=t^R=f)	0.9	0.1
P(S=f^R=t)	0.9	0.1
P(S=f^R=f)	0	1

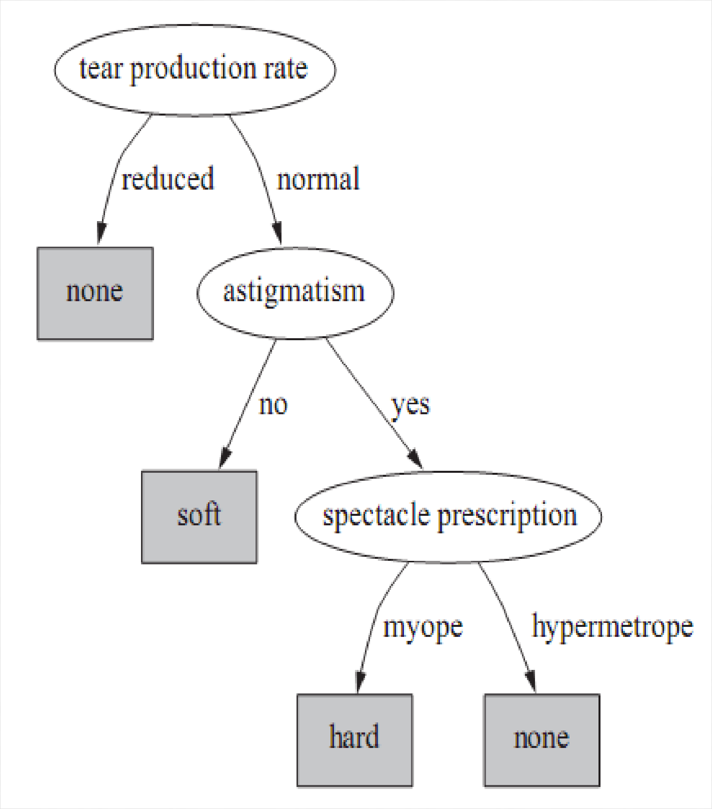
Rules

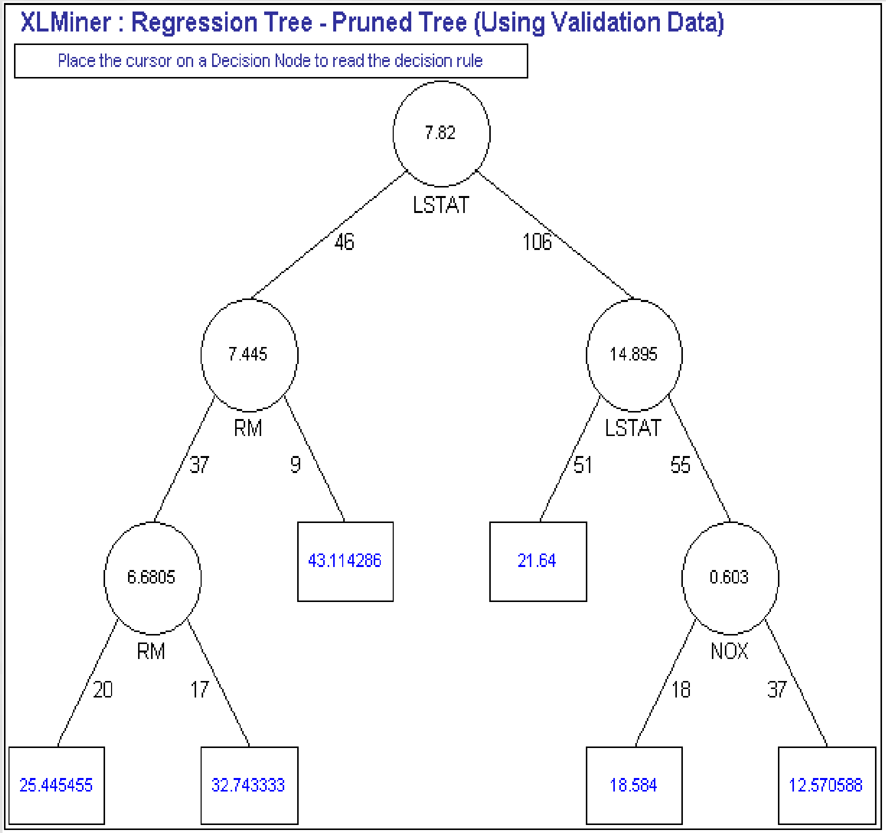


```
If humidity = normal and windy = false then play = yes
```

And many more: neural networks, nearest neighbor, …

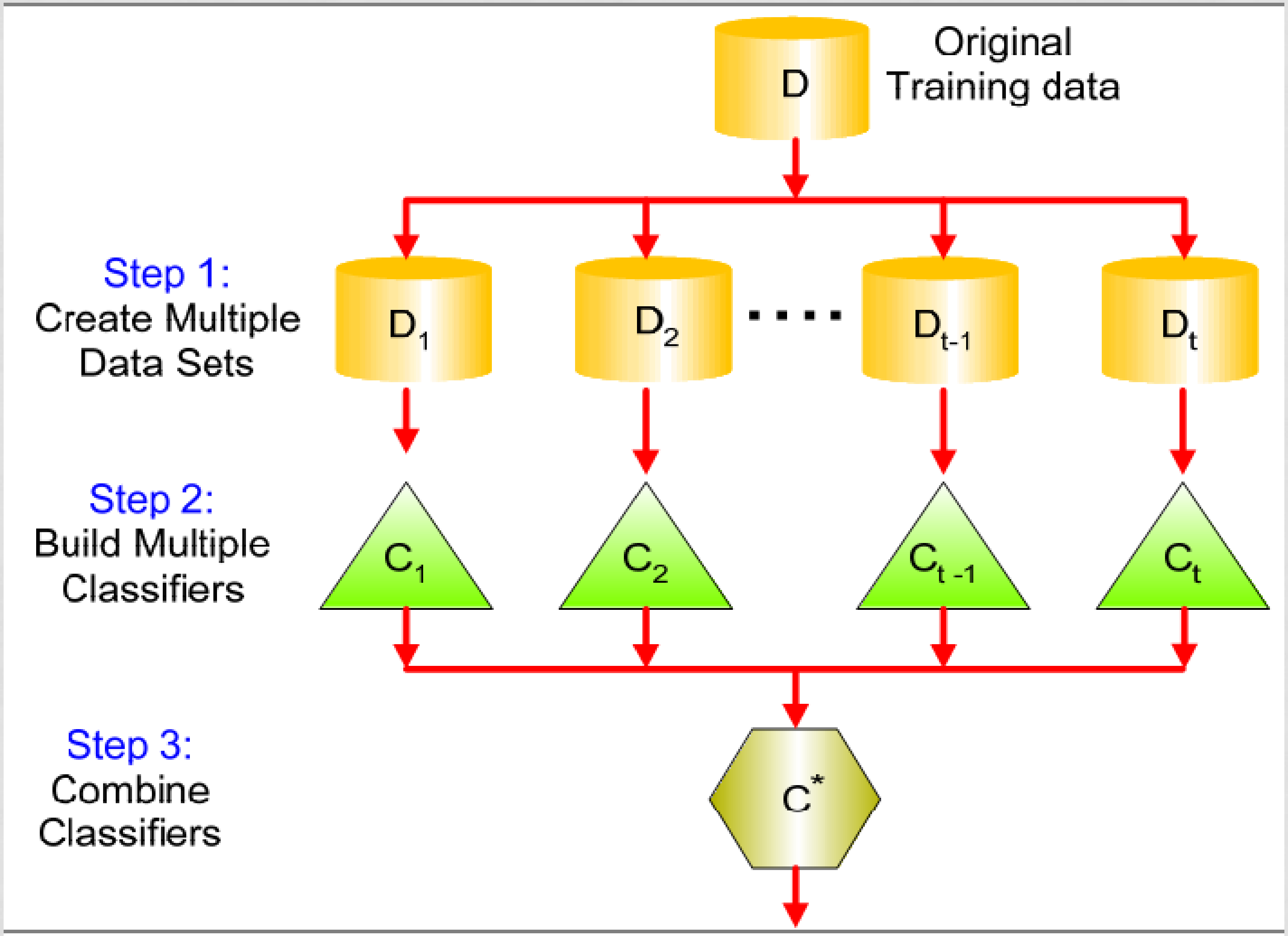
Decision trees and regression trees



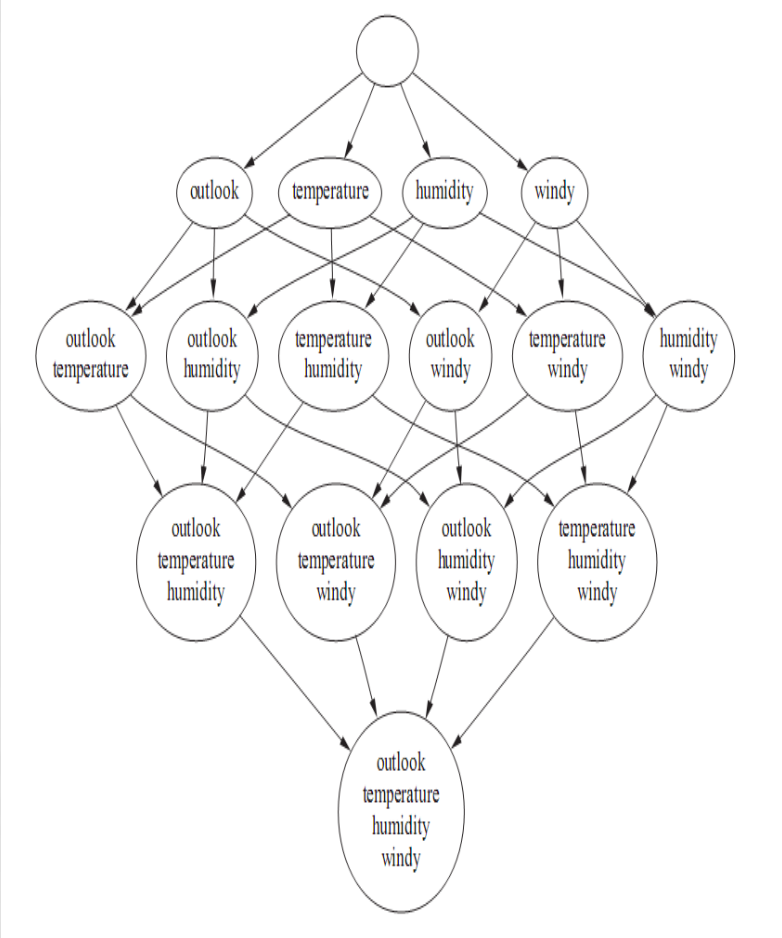

XLMiner : Regression Tree - Pruned Tree (Using Validation Data)

Place the cursor on a Decision Node to read the decision rule

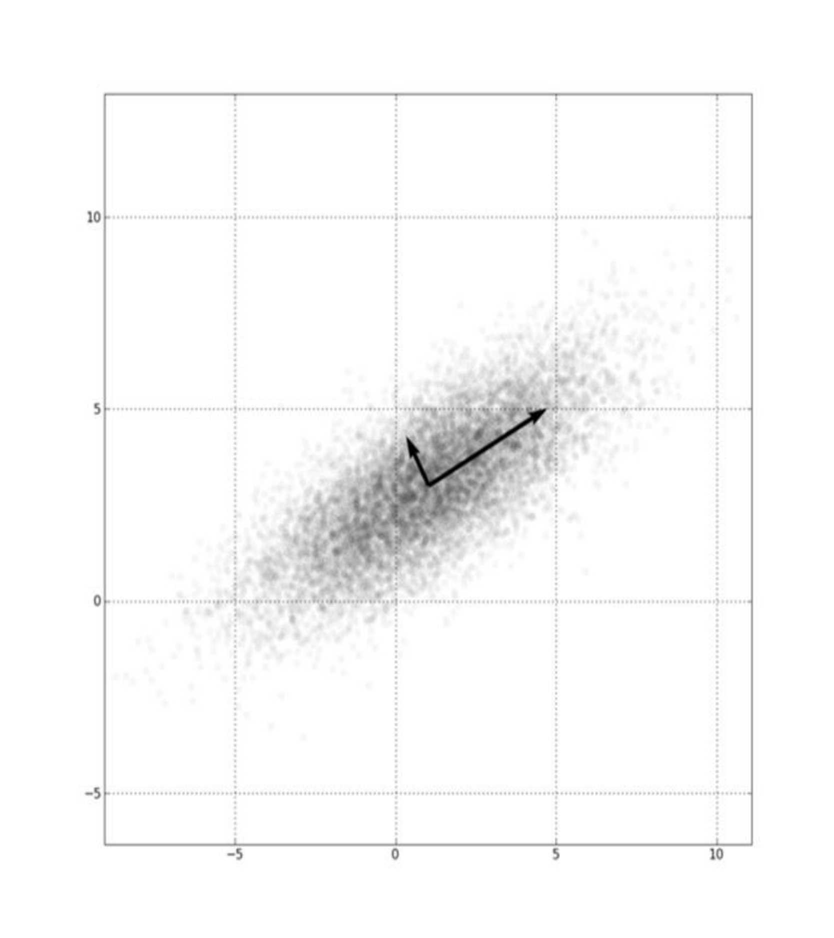
Ensembles of classifiers



ATTRIBUTE SELECTIÓN AND TRANSFORMATION

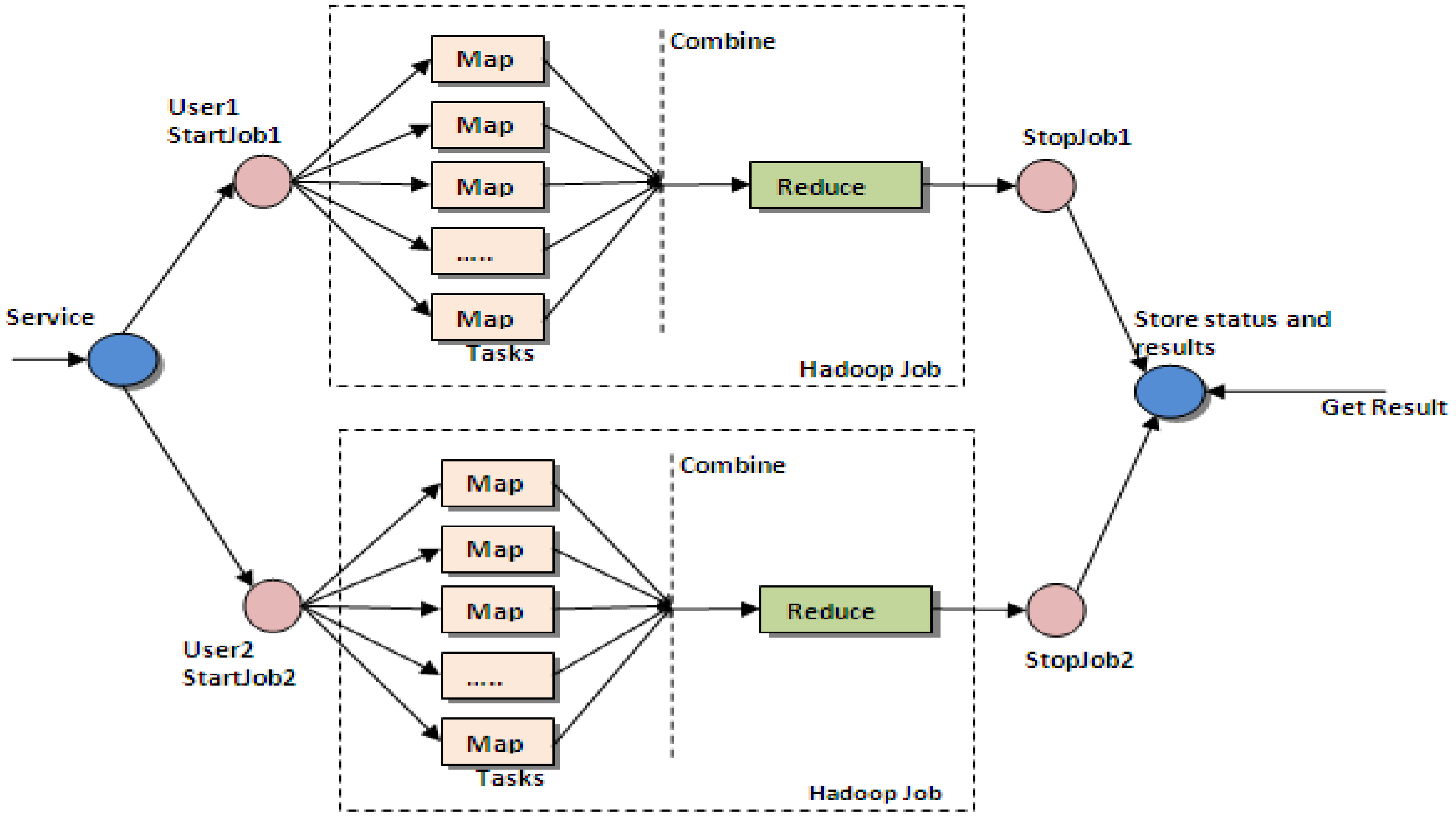


Attribute selection



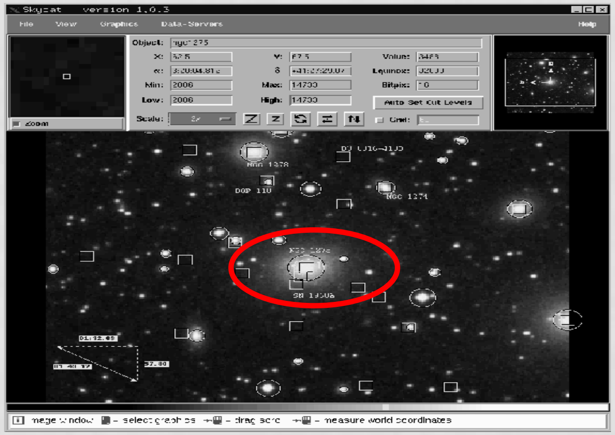
Principal Component Analysis and Random Projections

BIG DATA / MAP-REDUCE, SPARK (MLLIB)



TASKS / MODELS / ALGORITHMS

- What can be done? Tasks:
  - Supervised ML: classification, regression, …
  - Unsupervised ML: clustering, association, …
  - Semi-supervised ML
  - Reinforcement learning
- What kind of models can be learned?
  - Attribute-value:
    - Trees
    - Nearest neighbor
    - Functions: neural networks, support vector machines, …
    - Bayesian networks
    - Ensembles (bagging, boosting, stacking, …)
  - Relational
- How can models be learned? Algorithms:
  - Linear models: linear regression, simple perceptron, naive bayes, SVM with linear kernel, …
  - Neural networks: backpropagation, rprop, …
  - Decision trees: ID3, C4.5, C5.0, …
  - Nearest neighbour: IB1, …



Object: ngc1275

X: 32.5 Y: 67.5 Value: 3483

α: 3:20:04.812 δ: +41:27:29.07 Equinox: J2000

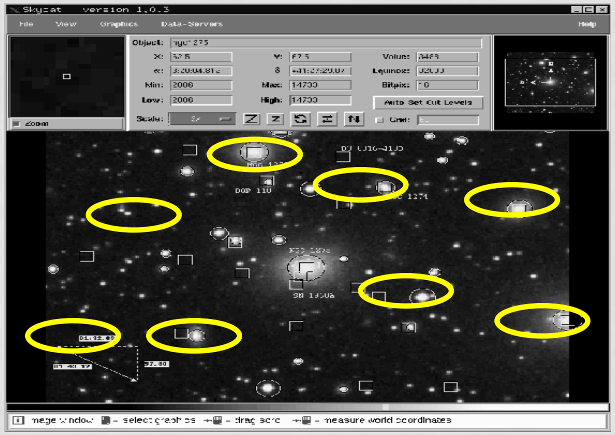
Min: 2006 Max: 14700 Bitpix: 16

Low: 2006 High: 14700

Scale: 2x □ Grid:

?

Training data (labeled pictures of sky objects: galaxies, stars, nebulae, …)



Object: ngc1275

X: 32.5 Y: 67.5 Value: 3483

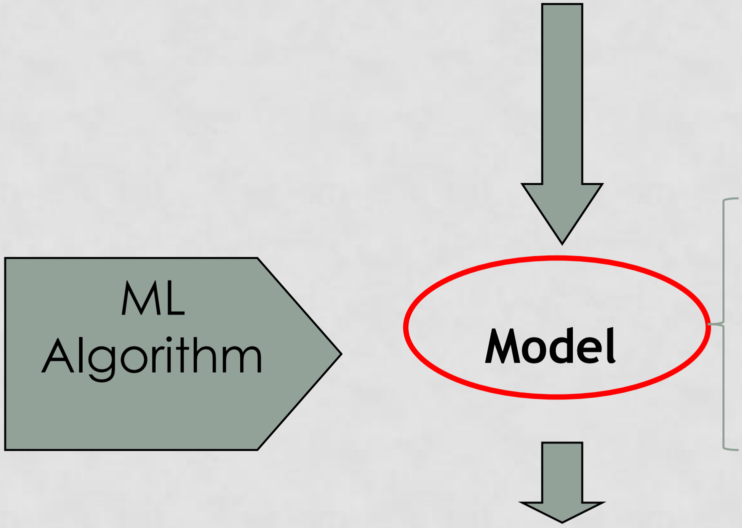
α: 3:20:04.812 δ: +41:27:29.07 Equinox: J2000

Min: 2006 Max: 14700 Bitpix: 16

Low: 2006 High: 14700

Scale: 2x □ Grid:

Pictures in the catalog have been labeled by a human expert (astronomer)



- Trees
- Nearest neighbor
- Functions: neural networks, support vector machines, …
- Bayesian networks
- Ensembles (bagging, boosting, stacking, …)

Spiral galaxy

TASKS

- **Inductive learning**(from instances)
  - **Supervised learning:**
    - **Classification:**
    - Regression
  - Semi-supervised learning
  - Unsupervised learning:
    - Clustering
    - Association
  - Reinforcement learning

CLASSIFICATION TASK. AN EXAMPLE:

- Bank credit approval:
  - An Internet bank owns a large data base with information about clients whose credits were approved or rejected
  - The banks requires a model to determine if a new customer will repay the loan or not
  - Instances (client records in the database):
    - Input attributes : credit time-length (years), amount, overdue accounts?, own house?
    - Class: yes/no
  - Rule-based model:
    - **IF** (overdue accounts > 0) **THEN** repay loan = no
    - **IF** (overdue accounts = 0) **AND** ((salary > 2500) **OR** (years > 10)) **THEN** repay loan = yes

CLASSIFICATION TASK. AN EXAMPLE:

test set

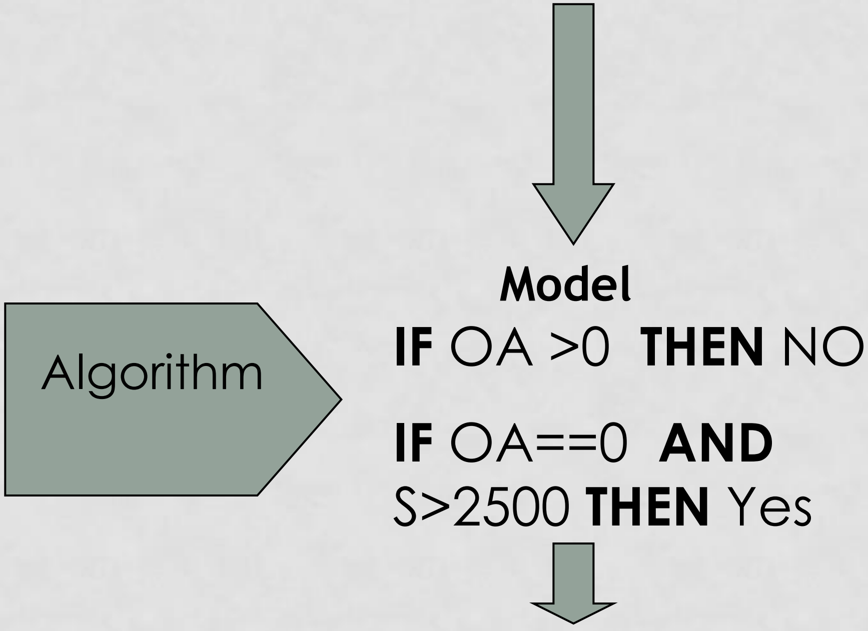
Years	Amount	Salary	Own house?	Overdue accounts?	Repay loan
10	50000	3000	Yes	0	??

T = training set (instances)

Years	Amount	Salary	Own house?	Overdue accounts?	Repay loan
15	60000	1900	Yes	2	No
2	30000	3500	Yes	0	Yes
9	9000	1700	Yes	1	No
15	18000	3000	No	0	Yes
10	24000	2100	No	0	No
	…	…	…	…	…

x (or input attributes)

y (class, or output attribute)



Model

IF OA >0 **THEN** NO

IF OA==0 **AND** S>2500 **THEN** Yes

Repay loan = yes

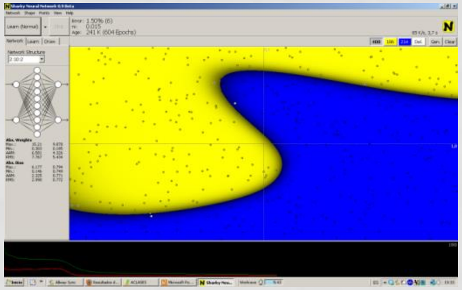
IMPORTANT: MODELS

- In the previous slide, the model built from training data is a set of rules:

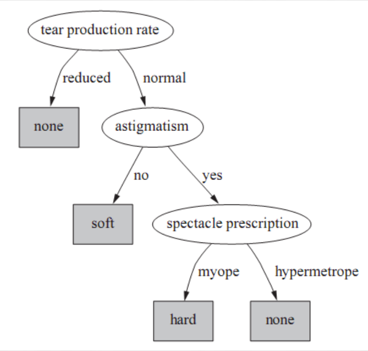
IF OA >0 **THEN** NO **ELSEIF** OA==0 **AND** S>2500 **THEN** Yes

- But there are many more that can be learned:

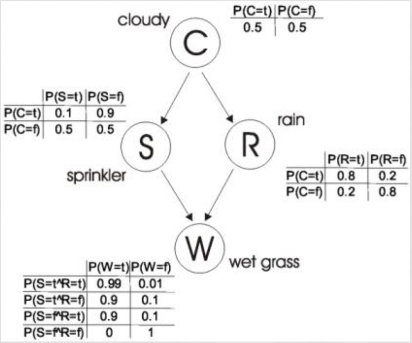
Functions: $y= 3*x^3+2$



Decision trees



Bayesian networks



P(C=t)	P(C=f)
0.5	0.5

	P(S=t)	P(S=f)
P(C=t)	0.1	0.9
P(C=f)	0.5	0.5

	P(R=t)	P(R=f)
P(C=t)	0.8	0.2
P(C=f)	0.2	0.8

	P(W=t)	P(W=f)
P(S=t^R=t)	0.99	0.01
P(S=t^R=f)	0.9	0.1
P(S=f^R=t)	0.9	0.1
P(S=f^R=f)	0	1

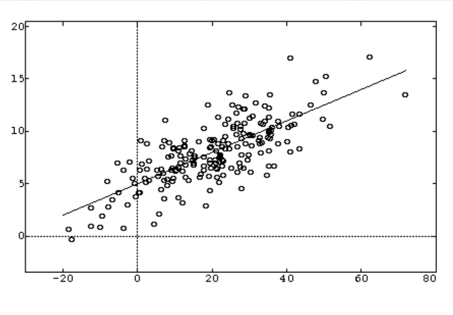
And many more: neural networks, nearest neighbor, support vector machines (SVMs).

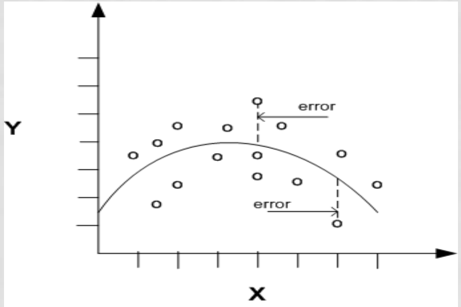
TASKS

- **Inductive learning**(from instances)
  - **Supervised learning:**
    - Classification
    - **Regression**
  - Semi-supervised learning
  - Unsupervised learning:
    - Clustering
    - Association
  - Reinforcement learning

REGRESSION

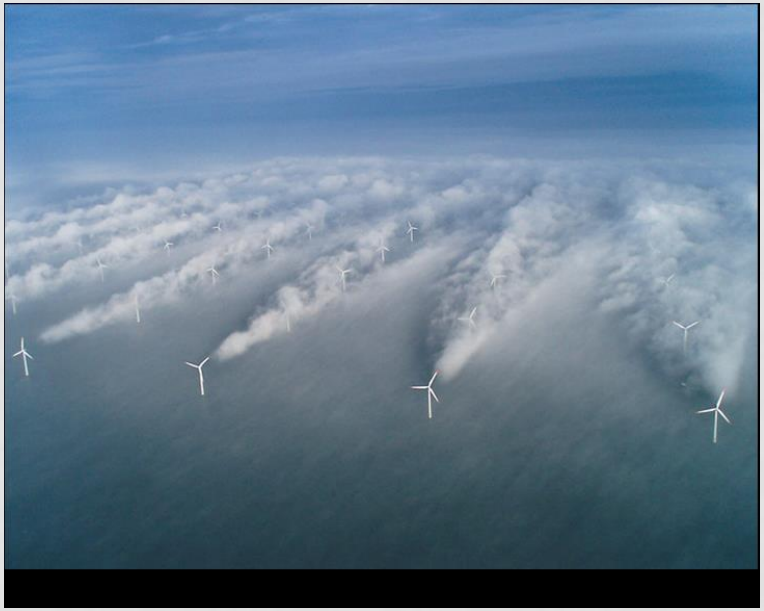
- If the class is continuous, it is a **regression** problem
- Models are typically mathematical functions *y=g(x)*
  - Linear: y = ax+b
  - Non linear: y = a*x^2+bx+c / y = log(sin(x))

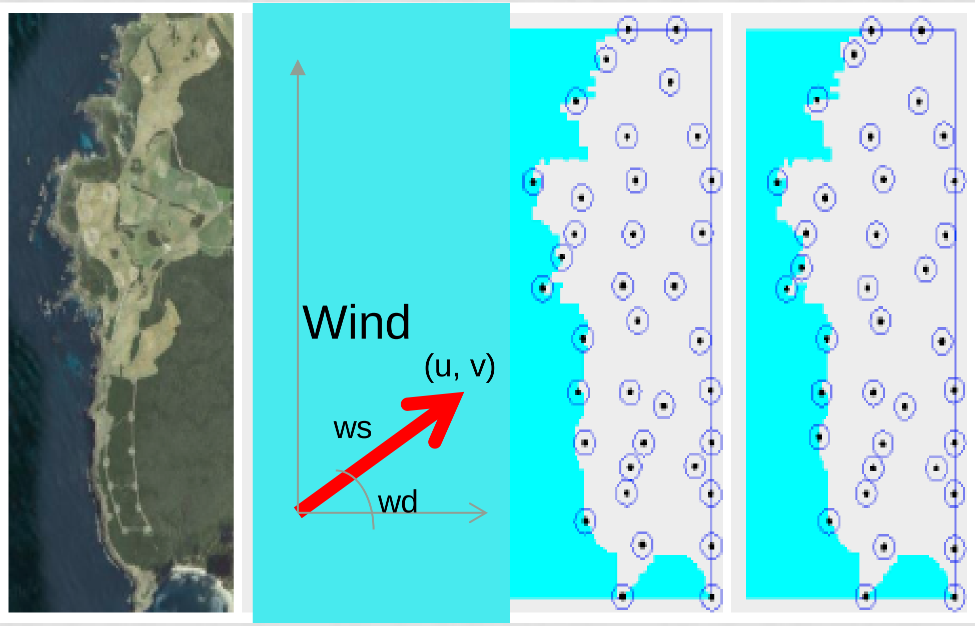




REGRESSION EXAMPLE

- **A wind power forecasting problem: predicting hourly power generation at 7 wind farms**





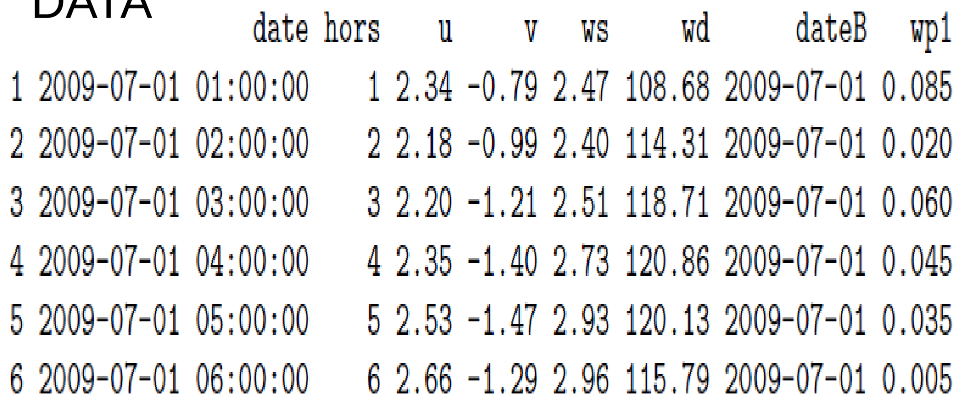
Some input variables:

- ws: wind speed
- wd: wind direction
- (u,v): wind direction vector

Model to estimate electricity production from ws, wd, u, v?
wp = f(ws, wd, u, v, …)

REGRESSION EXAMPLE

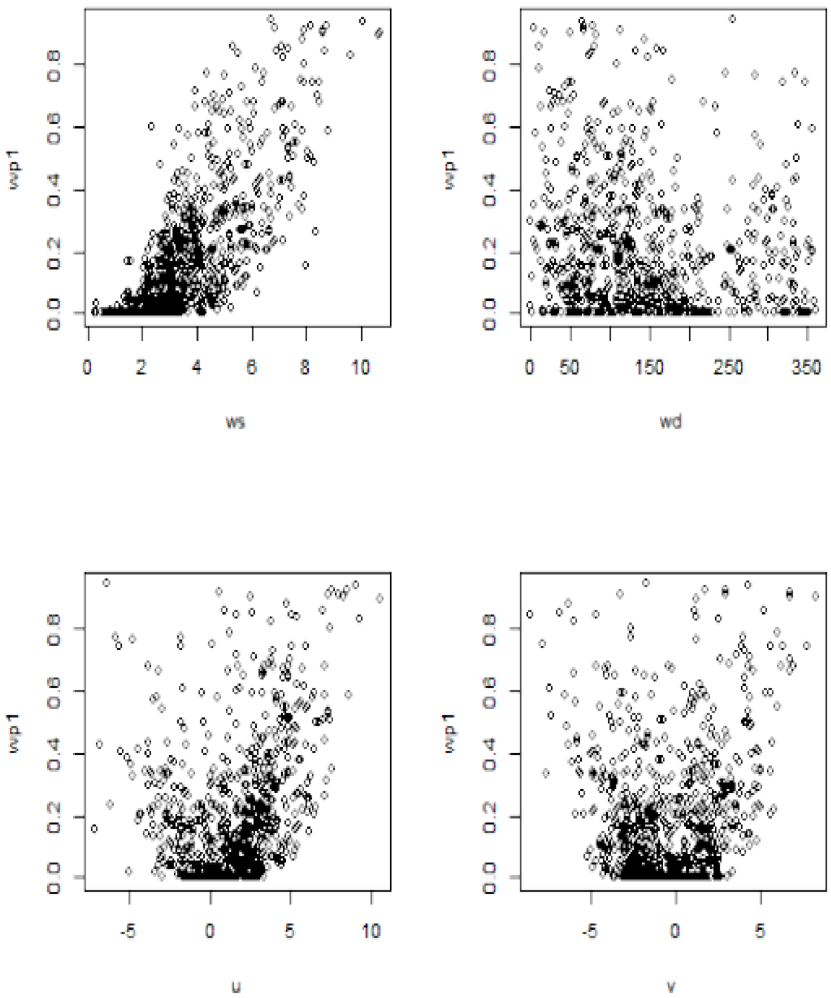
DATA



```
                 date hors    u     v   ws     wd      dateB   wp1
1 2009-07-01 01:00:00    1 2.34 -0.79 2.47 108.68 2009-07-01 0.085
2 2009-07-01 02:00:00    2 2.18 -0.99 2.40 114.31 2009-07-01 0.020
3 2009-07-01 03:00:00    3 2.20 -1.21 2.51 118.71 2009-07-01 0.060
4 2009-07-01 04:00:00    4 2.35 -1.40 2.73 120.86 2009-07-01 0.045
5 2009-07-01 05:00:00    5 2.53 -1.47 2.93 120.13 2009-07-01 0.035
6 2009-07-01 06:00:00    6 2.66 -1.29 2.96 115.79 2009-07-01 0.005
```

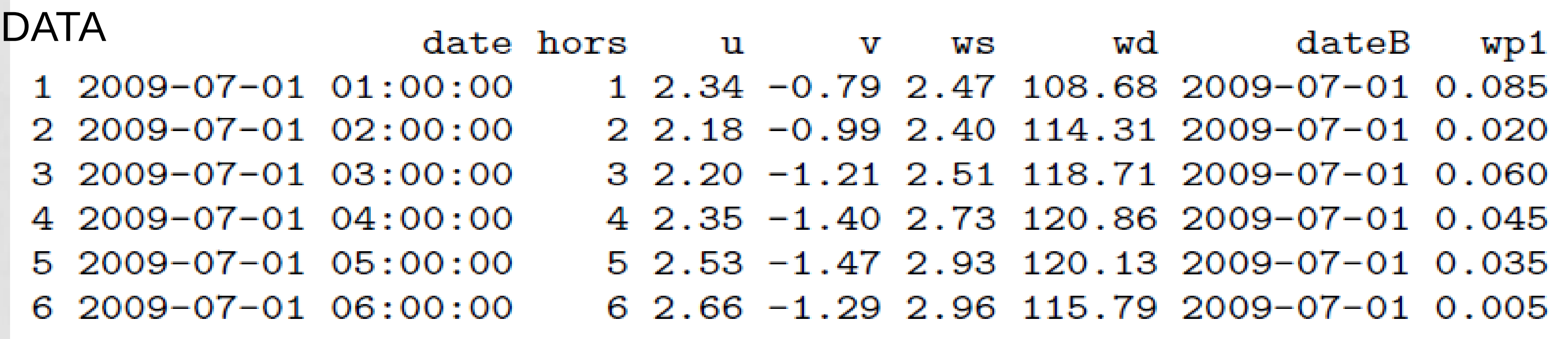
Some input variables:

- ws: wind speed
- wd: wind direction
- (u,v): wind direction vector



REGRESSION EXAMPLE

DATA



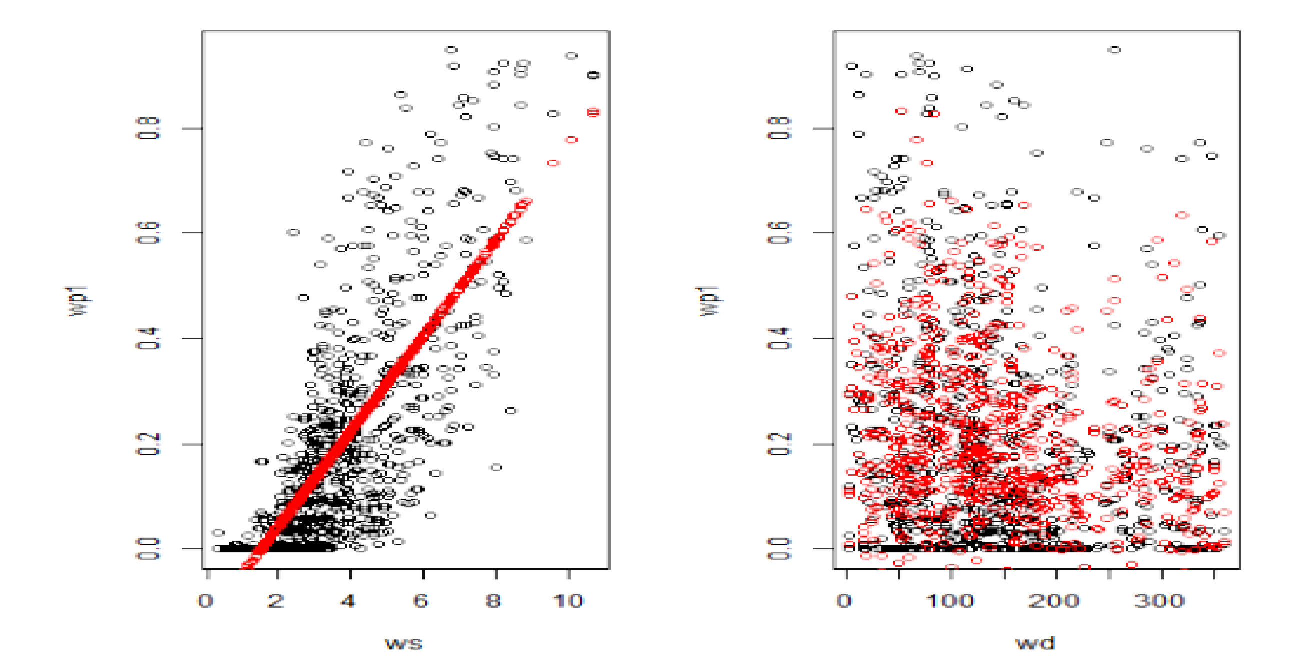
```
                 date hors    u     v   ws     wd      dateB   wp1
1 2009-07-01 01:00:00    1 2.34 -0.79 2.47 108.68 2009-07-01 0.085
2 2009-07-01 02:00:00    2 2.18 -0.99 2.40 114.31 2009-07-01 0.020
3 2009-07-01 03:00:00    3 2.20 -1.21 2.51 118.71 2009-07-01 0.060
4 2009-07-01 04:00:00    4 2.35 -1.40 2.73 120.86 2009-07-01 0.045
5 2009-07-01 05:00:00    5 2.53 -1.47 2.93 120.13 2009-07-01 0.035
6 2009-07-01 06:00:00    6 2.66 -1.29 2.96 115.79 2009-07-01 0.005
```

Linear model:

wp = f(ws, wd, u, v)

$$wp = a_1*ws + a_2*wd + + a_3*u + a_4*v + b$$

Obviously, a non-linear model could do better

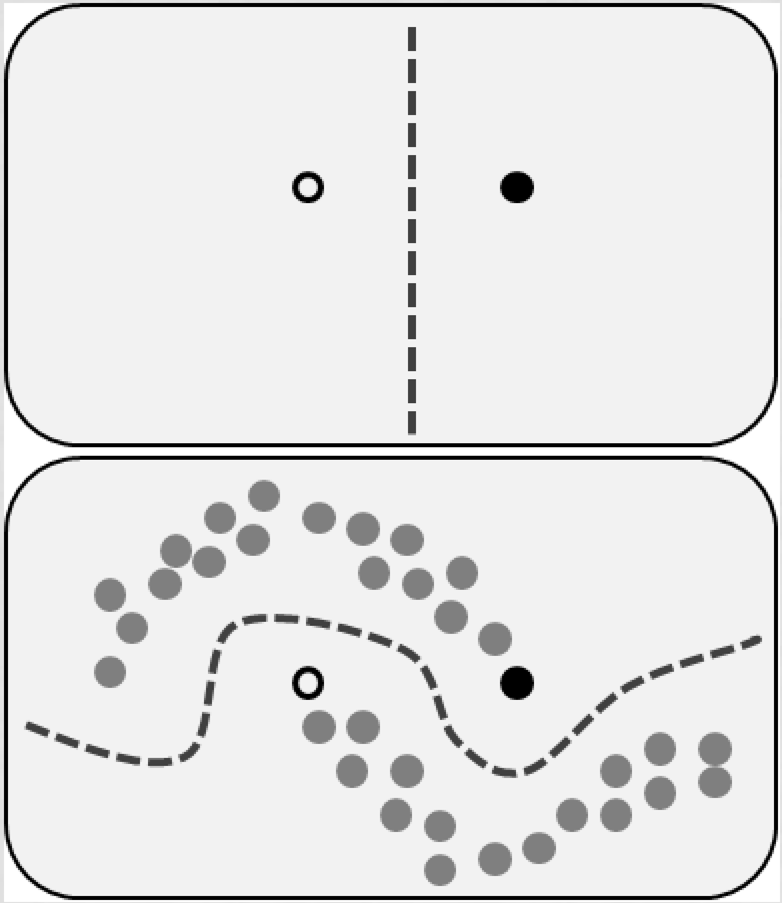


TASKS

- **Inductive learning**(from instances)
  - Supervised learning:
    - Classification
    - Regression
  - **Semi-supervised learning**
  - Unsupervised learning:
    - Clustering
    - Association
  - Reinforcement learning

SEMISUPERVISED LEARNING

- When both labelled and unlabelled instances are available
- Why: labelling instances may be costly (ex: to perform a biopsy to determine if a person has cancer)

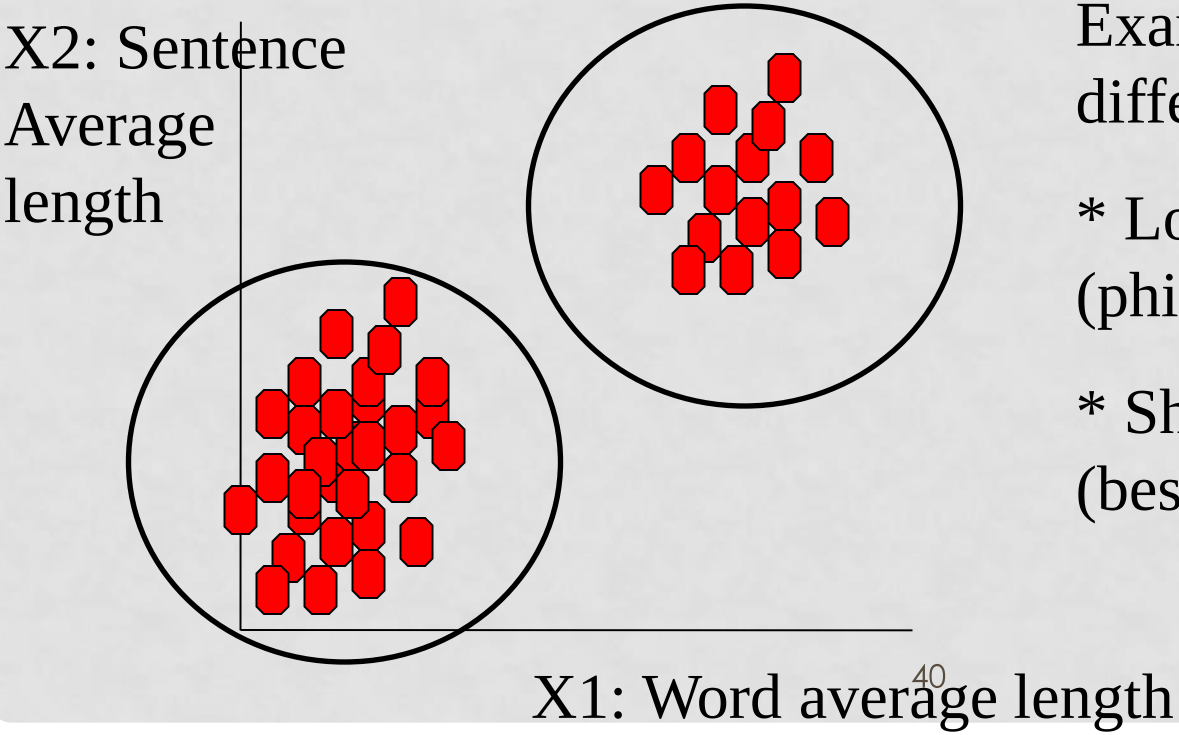


TASKS

- **Inductive learning**(from instances)
  - Supervised learning:
    - Classification
    - Regression
  - Semi-supervised learning
  - **Unsupervised learning:**
    - **Clustering**
    - Association
  - Reinforcement learning

UNSUPERVISED LEARNING (NO LABELS): CLUSTERING

- To determine natural clusterings in instance space, based on the input attributes (no labels)



Example:each point is a different book. 2 groups:

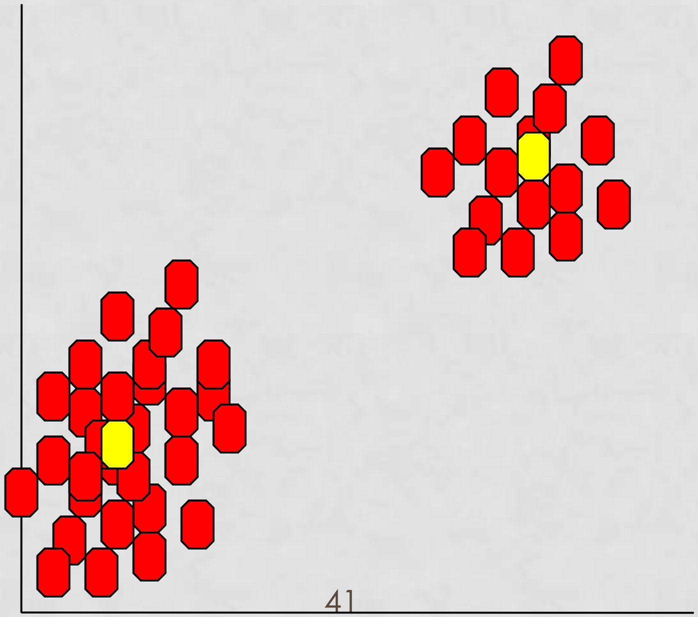
* Long words and sentences (philosophy?)

* Short words and sentences (best-sellers?)

CLUSTER REPRESENTATION

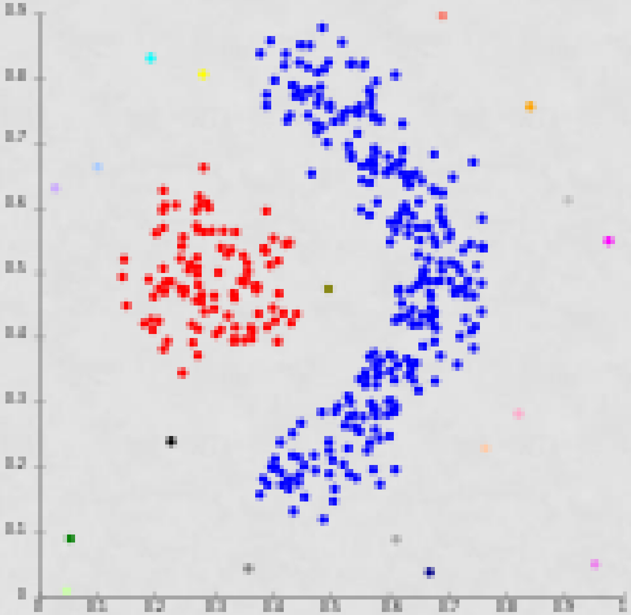
- Most commonly: centroids (ex: k-means algorithm)

K-MEANS: http://www.youtube.com/watch?v=74rv4snLI70



CLUSTERING

- Clustering is not so well defined as classification: clustering based on neighbourhood or connectivity?



CLUSTERING EXAMPLE

- Human resources department would like to cluster employees in order to understand the different types of employee and treat them accordingly (fire problematic workers? ☺).

CLUSTERING EXAMPLE. TRAINING DATA

Id	Salary	Married	Car	Offspring	Own-house	Syndicate	Sick leave	Years working	Sex
1	1000	Yes	No	0	No	No	7	15	M
2	2000	No	Yes	1	No	Yes	3	3	F
3	1500	Yes	Yes	2	Yes	Yes	5	10	M
4	3000	Yes	Yes	1	No	No	15	7	F
5	1000	Yes	Yes	0	Yes	Yes	1	6	F
...	...	...	...	...		...	...	...	...

MODEL (CLUSTERS)

	GROUP 1	GROUP 2	GROUP 3
Salary	1535	1428	1233
Married (No/Yes)	77%/22%	98%/2%	0%/100%
Car	82%/18%	1%/99%	5%/95%
Offspring	0.05	0.3	2.3
Own-house	99%/1%	75%/25%	17%/83%
Syndicated	80%/20%	0%/100%	67%/33%
Sick leave	8.3	2.3	5.1
Years working	8.7	8	8.1
Sex (M/W)	61%/39%	25%/75%	83%/17%

MODEL (CLUSTERS)

- Cluster 1: No offspring and rented house. Low level of syndication. Lots of sick leaves
- Cluster 2: No offspring and own-car. High syndication level. Few sick leaves. Tipically women living in rented houses
- Cluster 3: Married men with children and own-car and own-houses. Low syndication level

TASKS

- **Inductive learning**(from instances)
  - Supervised learning:
    - Classification
    - Regression
  - Semi-supervised learning
  - **Unsupervised learning:**
    - Clustering
    - **Association**
  - Reinforcement learning

MARKET BASKET ANALYSIS (ASSOCIATION)

- A supermarket needs to know customer behavior.
  - Ex: if customer buys X then s/he also buys Y
- Service might be improved (putting together products bought together, etc.)

TRAINING DATA (CUSTOMER BASKETS)

Id	**Eggs**	**Oil**	**Napies**	**Wine**	**Milk**	**Butter**	**Salmon**	**Lettuce**	**...**
1	Yes	No	No	Yes	No	Yes	Yes	Yes	...
2	No	Yes	No	No	Yes	No	No	Yes	...
3	No	No	Yes	No	Yes	No	No	No	...
4	No	Yes	Yes	No	Yes	No	No	No	...
5	Yes	Yes	No	No	No	Yes	No	Yes	...
6	Yes	No	No	Yes	Yes	Yes	Yes	No	...
7	No	No	No	No	No	No	No	No	...
8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	...
…	…	…	…	…	…	…	…	…	…

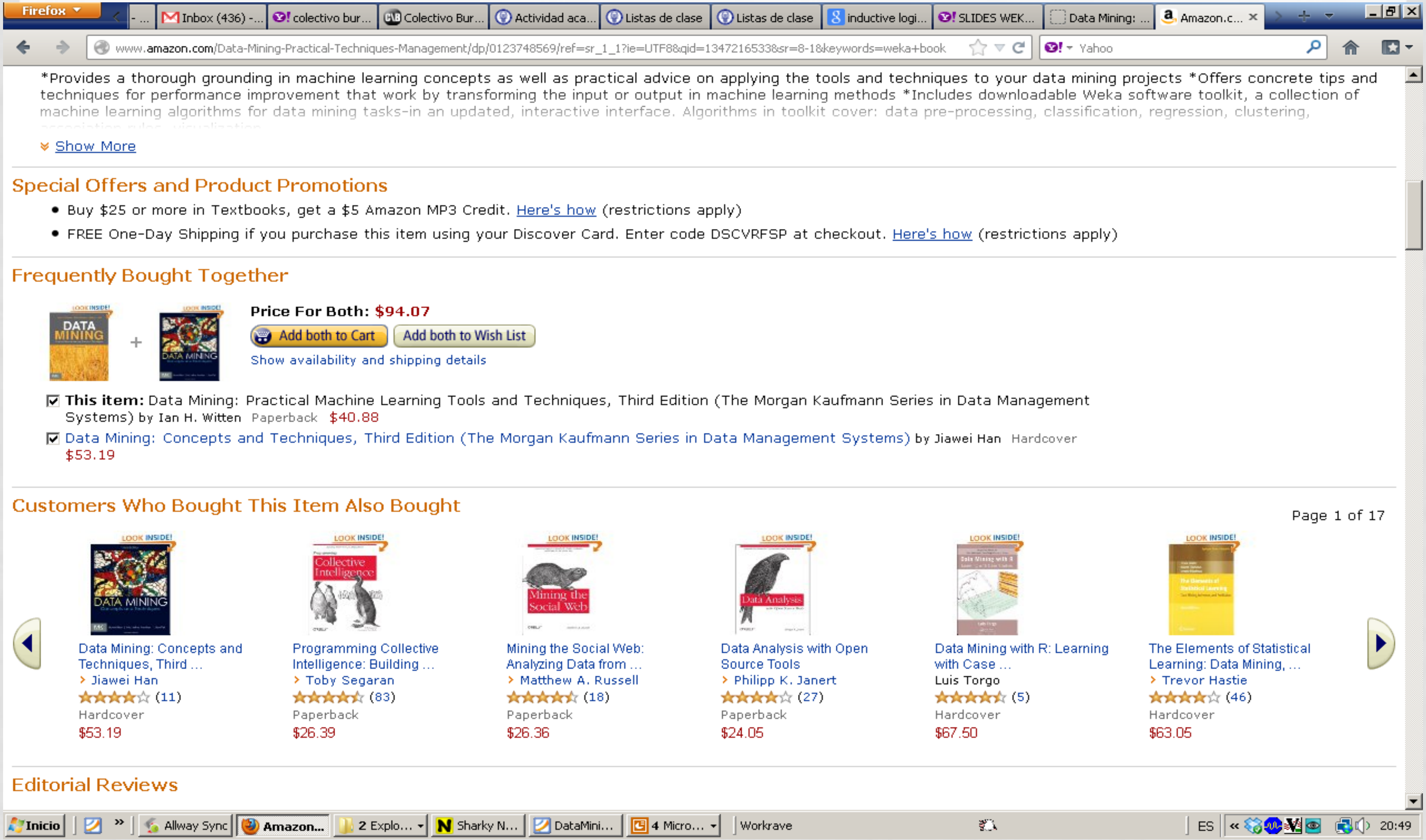
MODEL

- Rules **IF** At_1=a **AND** At_2=b y ... **THEN** At_n=c
  - **IF** nappies=Yes **THEN** milk=Yes
  - **IF** butter = Yes **AND** salmon = Yes **THEN** wine = Yes

- Also: **IF** At_1=a AND At_2=b **THEN** At_n=c, **At_4=D**

Service might be improved (putting together nappies and milk, etc.)

ASSOCIATION



*Provides a thorough grounding in machine learning concepts as well as practical advice on applying the tools and techniques to your data mining projects *Offers concrete tips and techniques for performance improvement that work by transforming the input or output in machine learning methods *Includes downloadable Weka software toolkit, a collection of machine learning algorithms for data mining tasks-in an updated, interactive interface. Algorithms in toolkit cover: data pre-processing, classification, regression, clustering, association rules, visualization

Show More

Special Offers and Product Promotions

- Buy $25 or more in Textbooks, get a $5 Amazon MP3 Credit. Here's how (restrictions apply)
- FREE One-Day Shipping if you purchase this item using your Discover Card. Enter code DSCVRFSP at checkout. Here's how (restrictions apply)

Frequently Bought Together

Price For Both: $94.07

Add both to Cart | Add both to Wish List

Show availability and shipping details

☑ **This item:** Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) by Ian H. Witten Paperback $40.88

☑ Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han Hardcover $53.19

Customers Who Bought This Item Also Bought

Page 1 of 17

Data Mining: Concepts and Techniques, Third ...	Programming Collective Intelligence: Building ...	Mining the Social Web: Analyzing Data from ...	Data Analysis with Open Source Tools	Data Mining with R: Learning with Case ...	The Elements of Statistical Learning: Data Mining, ...
Jiawei Han	Toby Segaran	Matthew A. Russell	Philipp K. Janert	Luis Torgo	Trevor Hastie
(11)	(83)	(18)	(27)	(5)	(46)
Hardcover	Paperback	Paperback	Paperback	Hardcover	Hardcover
$53.19	$26.39	$26.36	$24.05	$67.50	$63.05

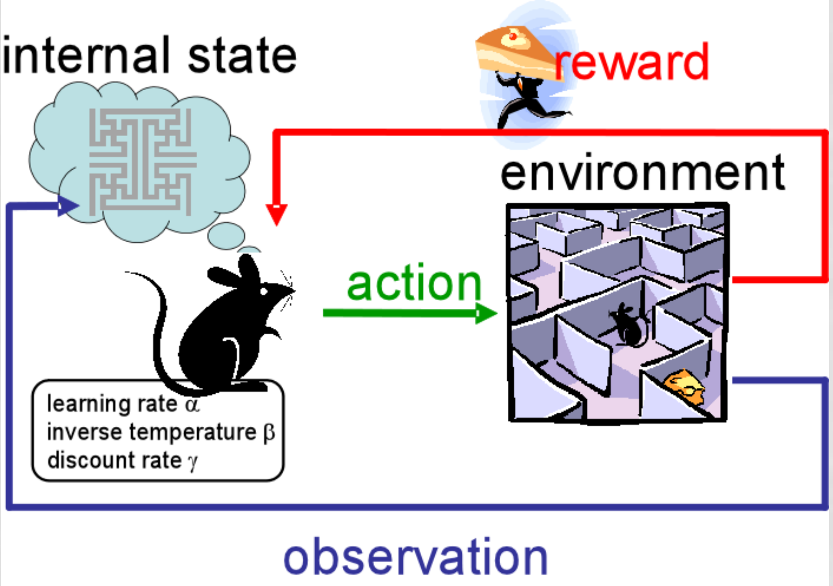
Editorial Reviews

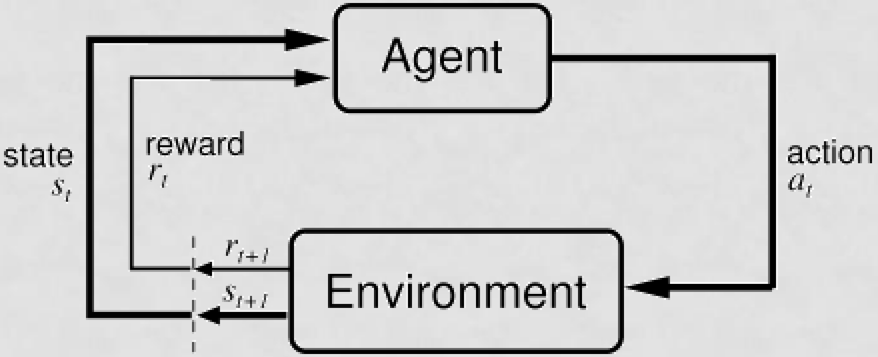
TASKS

- **Inductive learning**(from instances)
  - Supervised learning:
    - Classification
    - Regression
  - Semi-supervised learning
  - Unsupervised learning:
    - Clustering
    - Association
  - **Reinforcement learning**

TASK: REINFORCEMENT LEARNING

- The goal of learning is a “policy” π so that the agent (mouse) knows what to do at each situation (in the case of the mouse, a situation is a particular location within the maze). Robotics.
- Actions:
  - forward
  - turn left
  - turn right





TASKS

- **Inductive learning**(from instances)
  - Attribute-value models
    - Supervised learning:
    - Semi-supervised learning
    - Unsupervised learning:
    - Reinforcement learning
  - **Relational learning**

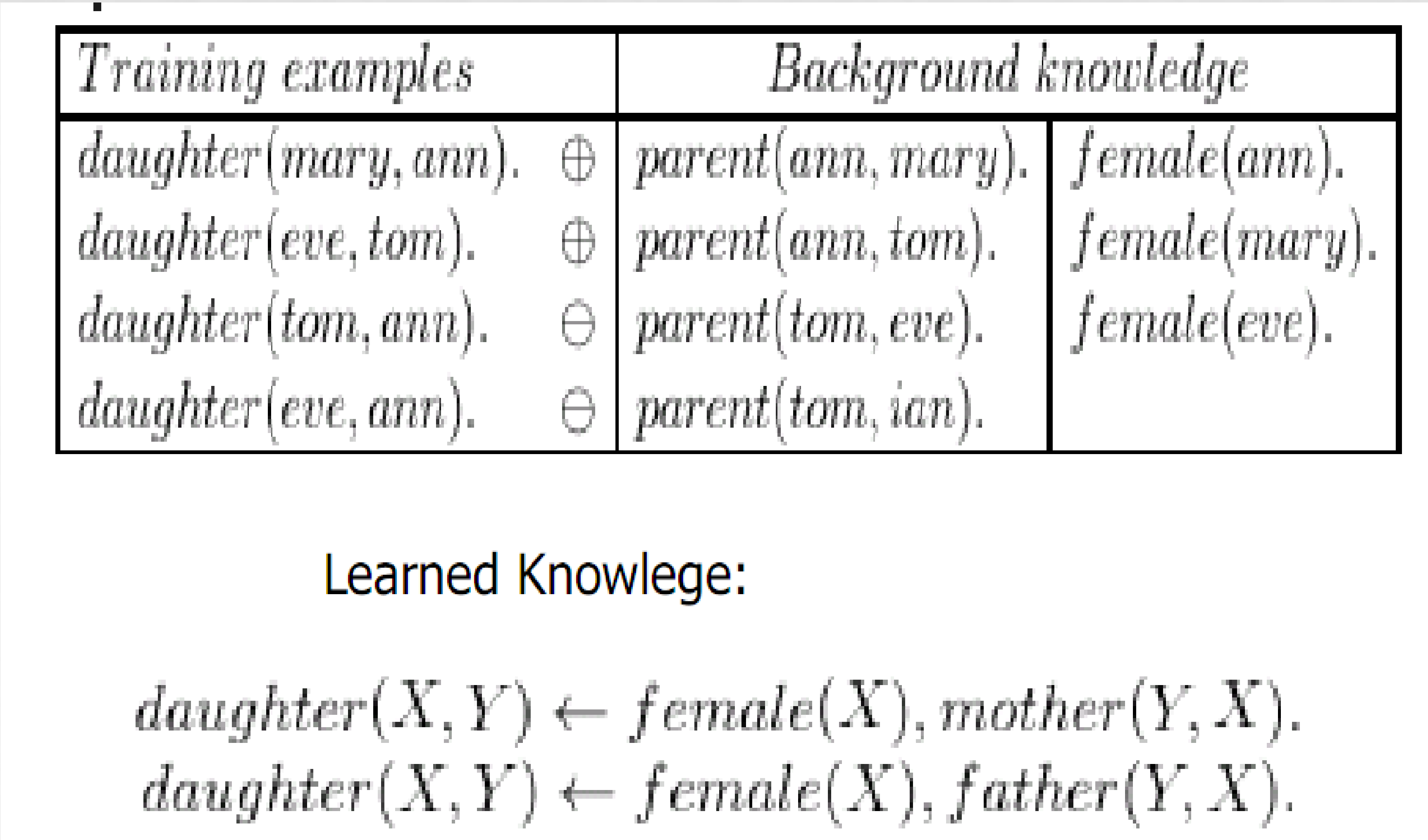
Relational Learning

- For instance, learn the concept of “being a daughter”
- **IF** X is female **AND** Y is the mother of X **THEN** X is a daugther of Y
- Compare this rule with:

IF Overdue Accounts ==0 **AND** Salary >2500 **THEN** Repay loan = Yes

- Relational rules use variables (X, Y) and relations

Relational Learning: ILP (inductive logic programming



Training examples		*Background knowledge*	
$daughter(mary, ann).$	$\oplus$	$parent(ann, mary).$	$female(ann).$
$daughter(eve, tom).$	$\oplus$	$parent(ann, tom).$	$female(mary).$
$daughter(tom, ann).$	$\ominus$	$parent(tom, eve).$	$female(eve).$
$daughter(eve, ann).$	$\ominus$	$parent(tom, ian).$	

Learned Knowlege:

$$daughter(X, Y) \leftarrow female(X), mother(Y, X).$$
$$daughter(X, Y) \leftarrow female(X), father(Y, X).$$

Bibliography

SCIKIT-LEARN

- Learning scikit-learn: Machine Learning in Python:
  - http://m.proquest.safaribooksonline.com/hd/catalog?isbn=9781783281930
- Mastering Machine Learning with scikit-learn:
  - http://m.proquest.safaribooksonline.com/hd/catalog?isbn=9781783988365
- scikit-learn Cookbook
  - http://m.proquest.safaribooksonline.com/hd/catalog?isbn=9781783989485

SPARK

- Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, Learning Spark, O'Reilly Media, 2015. ISBN: 978-1-449-35862-4
- Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, Advanced Analytics with Spark, O'Reilly, 2015. ISBN: 978-1-491-91276-8