Departament de Llenguatges i Sistemes Informatics
UNIVERSITAT POLITÈCNICA DE CATALUNYA

# Torturando a los datos hasta que confiesen

**Luis Carlos Molina Félix**

**Curso de Data Mining**
Dra. Àngela Nebot
Dr . Lluís Belanche

lcmolina@lsi.upc.es

**14 junio 2001**

## Nombres al mismo problema

- Data Archeology
- Dependency Functional Analysis
- Information Recollect
- Pattern Data Analysis
- Knowledge Fishing
- KDD (1990´s → ...) (académico)
- Data Mining (1990´s → ...) (comercial)



## Data Mining - Definición

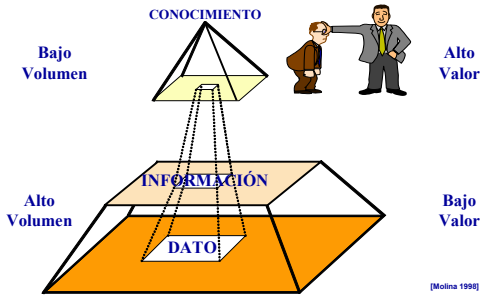
**Conjunto de áreas que tienen como propósito la identificación de conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacía la toma de decisión** [Molina 2001].

**¿Que áreas?**

- Estadística
- IA / PR
- Teoría de la probabilidad
- Teoría de la información
- Incerteza
- Teoría de grafos
- Bases de datos
- Visualización



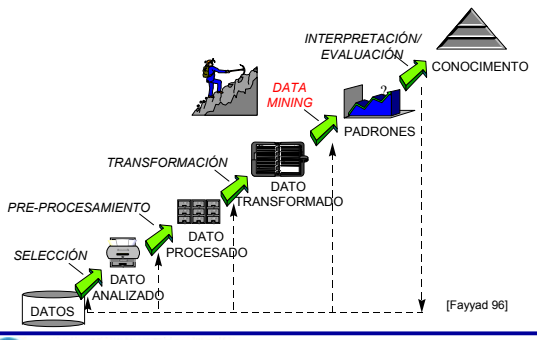
## Donde trabaja




[Molina 1998]



## PROCESO KDD




[Fayyad 96]



## Construcción de Modelos

- **Clasificación**
  - Reglas de producción[Clark:89][Holte:93]
  - Árboles de Inducción [Quinlan:86,93][Brieman et.al.:94]
- **Redes Neuronales**
  - Multi-layer perceptron
- ***k*-Nearest Neighbor**
  - Hamming
  - Euclidean
  - Mahalanobis
- **Clustering**
  - EM
  - K-means
  - Hierarchical Agglomerative Clustering
- **Híbridas**
  - NeuroRule[Lu et.al.,95] Ruleneg[Fu,99] (Reglas desde RN)

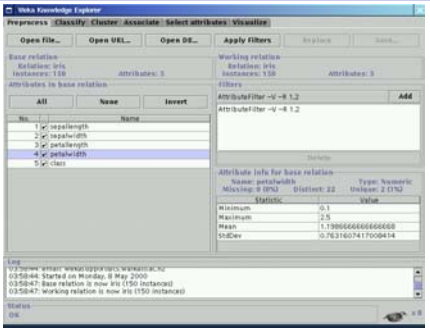
## Herramientas

- **Academicas**
  - Weka (Hall)[1]
  - MLC++ (Kohavi)[2]
  - SIPINA (Ricco)[3]
  - BKD (Ramoni)[4]
  - Mobal (Sommer)[5]
- **Comerciales**
  - MineSet (Silicon Graphics)
  - SAS (SAS Institute Inc.)
  - Clementine (SPSS)
  - Iminer (IBM)
  - Darwin (Oracle)
  - CART (Salford-Systems)
  - DataEngine (MIT GmbH)

[1] www.cs.waikato.ac.nz/~ml/weka
[2] www.sgi.com/tech/mlc/
[3] eric.univ-lyon2.fr/~ricco/sipina.html
[4] kmi.open.ac.uk/projects/bkd/
[5] ftp.gmd.de/gmd/mlt/Mobal/

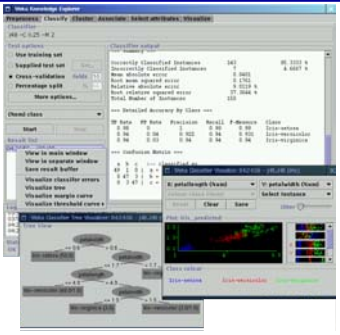


## Weka





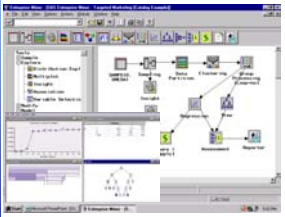
## Weka

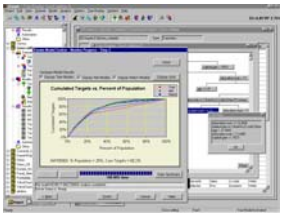




## SAS – Darwin

**SAS** **Darwin**

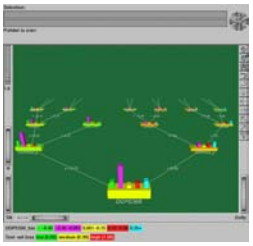


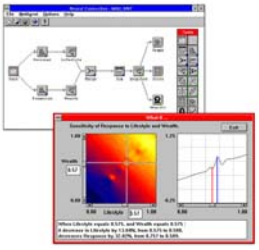




## MineSet - Clementine

**MineSet** **Clementine**







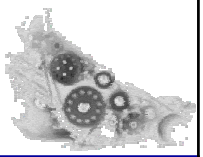
## Además de los pañales y cerveza

WAL*MART

- Averías en los coches de la Mercedes-Benz
- Detección de petróleo en el mar usando fotografías aéreas
- Proyecto Genoma
- Deep Purple

## Averías en los coches

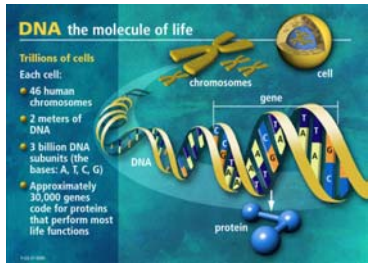
- Detecting Early Indicator Cars in an Automitive Database: A Multi-Strategy Approach [Wirth & Reinartz 96]
- Varios problemas: pintura->ciudad, taxis->uso
- Problemas de estudio: sistema de inyección
- 8 (fault profile) y 28 (life time) variables usadas
- 2628 casos en 1991
- Método EIC (ECOWEB(numeric, simbolic), CN2, C4.5)





## Proyecto Genoma

- Discovery in Human Genome Project [Schulze-Kremer 99]






## Deep Purple - Ajedrez

- Knowledge Discovery in Deep Blue [Cambell 99]
- 700.000 jugadas de Grandes Maestros
- Basado en la "Opening Theory" de los ajedrecistas para crear una "Extended Box"
- Combina 200 millones de posiciones por segundo y más de 7 factores son vistos antes de escoger una posición
- En 1996, en el 2do. juego jugaron sin la EB por error




## Detección de petróleo en el mar

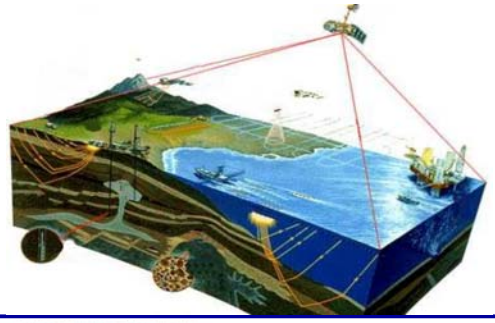
- Machine Learning for Detecting of Oil Spills is satellite Radar Images [Kubat, Holte & Matwin 98]
- 10% de los derrames en el mar se producen por causas naturales
- Fotografías desde satélites de 8.000x8.000 píxeles representando cada píxel 30x30m
- Pocos ejemplos, desbalanceados y en lotes
- SHIRINK (controla falsas alarmas) insensible al desbalance



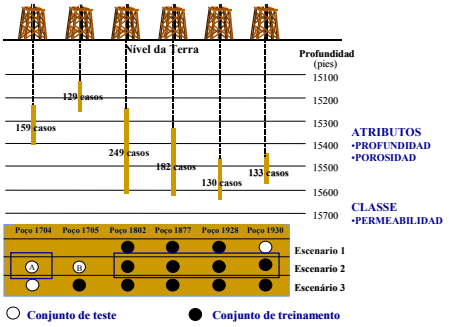
## Descanso - 5 minutos




## Base de Dados Petrolera
## Problema de Discretización

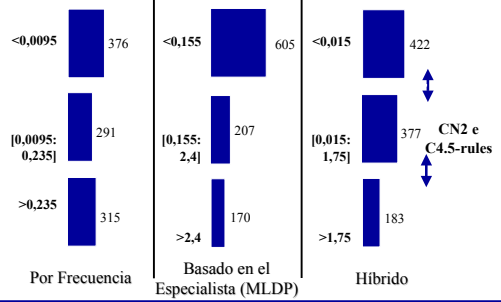


# Base de Dados Petrolera Contexto






# Métodos de Discretización

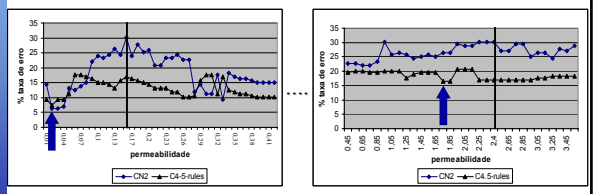





# Método de Discretización Híbrido

Fijando 2,4 y variando 0,155

Fijando 0,155 y variando 2,4






# Comparación entre las Discretizaciones

| | Discretización sugerida por el especialista del dominio 0,155 e 2,4 | | | Discretización usando el método híbrido 0,015 e 1,75 | | |
|---|---|---|---|---|---|---|
| Escenario | CN2 Tasa de Error | C4.5-rules Tasa de Error | Tasa de Error de la Clase Mayoritária | CN2 Tasa de Error | C4.5-rules Tasa de Error | Tasa de Error de la Clase Mayoritária |
| 1 | 42,1% | 28,6 % | 36,5% | 38,3% | 39,8 % | 64,4% |
| 2 conjunto de teste A | 30,2% | 17,0 % | 42,2% | 13,8% | 6,9 % | 59,9% |
| 2 conjunto de teste B | 5,4% | 7,8 % | 42,2% | 3,9% | 3,1 % | 59,9% |
| 3 | 29,6% | 18,2 % | 37,8% | 8,2% | 5,7 % | 56,7% |



# Reglas Generadas por el C4.5

**Regla 20:**
porosidad > 15,7
-> permeabilidad > 1,75 [11,1%] [112 14] [14,3%] [12 2]

**Regla 1:**
porosidad <= 2,9
-> permeabilidad < 0,015 [0,0%] [203 0] [1,6%][60 1]

**Regla 19:**
porosidad > 3,9
porosidad <= 15,7
-> permeabilidad [0,015:1,75] [17,5%] [188 40] [7,3%][76 6]

**Conjunto de Entrenamiento**
**[AA%] error de la clasificación**
**[BB CC] números de casos clasificados correcta y erróneamente**
**Conjunto de Prueba**
**[DD%] error de la clasificación**
**[EE FF] números de casos clasificados correcta y erróneamente**



# Aplicación DM. ¿Como obtener testículos grandes en los toros?

# Aplicación DM. ¿Como obtener testículos grandes en los toros?

**LUDY DE GARÇA**

*Comentários*

Mayor Destaque da Raza Nelore

Tiene Karvadi 2 veces en el Pedigrí

Toro del Año Expoinel Uberaba Diversas Veces

Gano Diversos Campeonatos de Raza Nelore

El Mejor hijo del Legendario GIM de GARÇA

El Nelore con mas hijos de mas de 1.100 Kg.

Plusmarquista - Venta de Semen

Plusmarquista - Volumen de Producción de Semen

Plusmarquista - Peso en la Expo UBERABA 1985

Plusmarquista - Progenies Premiadas na Raza

Plusmarquista - Peso de Progenies

*Premios*

Campeón Becerro 1981 Ourinhos

Campeón Becerro 1981 Marília

Campeón Becerro 1981 Baurú

Campeón Júnior 1982 Marília

Campeón Júnior 1982 Baurú

Campeón Júnior 1982 Ribeirão Preto

Campeón Toro Joven 1984 Barretos

Gran Campeón 1984 Barretos

Gran Campeón 1984 Uberlândia

Gran Campeón 1984 Pres. Prudente

Gran Campeón 1984 Ribeirão Preto

Gran Campeón 1984 Ourinhos

Gran Campeón 1984 Baurú

Gran Campeón 1984 Avaré



# Programa de Mejoramiento Genético de la Raza Nelore

- Creado a partir de la unión de criadores y de investigadores del Departamento de Genética de la Facultad de Medicina (USP-Brasil), buscando tecnologías modernas y de fácil aplicación en la pecuaria para aumentar la productividad del rebaño de corte nacional.
- Formado por 60 rebaños con ganaderos de los estados de Bahia, Goiás, Maranhão, Mato Grosso do Sul, Minas Gerais, São Paulo y Tocantins. Un total de 200.000 animales en control.



# Metodologia do Modelo Animal

- La Diferencia Esperada en la Progenie (DEP) es usada en todo el mundo para comparar el mérito genético de animales para varias características. Su objetivo es predecir la habilidad de transmisión genética de un animal evaluado como progenitor con respecto a su descendencia.

- La DEP es calculada por la característica genética de los animales a los 120, 240, 365 y 550 días y es expresada en la unidad de la característica, por ejemplo, kilogramos para peso y centímetros para perímetro testicular, con signo positivo o negativo.



# Contexto del Problema

**¿Que características deben de tener el reproductor y la matriz para tener un toro con un perímetro testicular grande?**

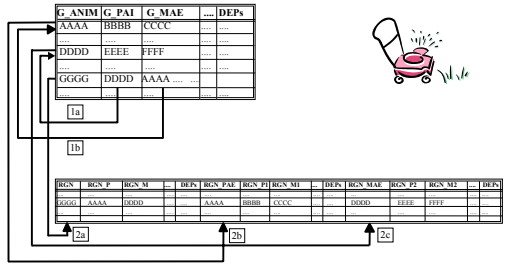


# Características da Base de Dados

| Nome do Atributo | Descrição do Atributo | Atributo |
|---|---|---|
| G_ANIM | Registro Genealógico Definitivo do Animal | Inteiro |
| G_PAI | Registro Genealógico Definitivo do Pai | Inteiro |
| G_MAE | Registro Genealógico Definitivo da Mãe | Inteiro |
| CNI | Número de Controle Interno Exclusivo | Inteiro |
| NFA | Número da Fazenda | Inteiro |
| SERIE | Série fornecida pela ABCZ, identificadora dos animais | Inteiro |
| RGN | Registro Genealógico de Nascimento do Animal | String |
| RGD | Registro Genealógico Definitivo do Animal | String |
| RC | Raça | Inteiro |
| SX | Sexo do Animal | Inteiro |
| DT_NASC | Data de Nascimento do Animal | Data |
| PAI_SER | Série fornecida pela ABCZ, relativa ao pai | Inteiro |
| PAI_RG | Registro Genealógico Definitivo do Pai fornecido pela ABCZ | String |
| MAE_SER | Série fornecida pela ABCZ, relativa a Mãe | Inteiro |
| MAE_RG | Registro Genealógico Definitivo da Mãe fornecido pela ABCZ | String |
| NOME | Nome do Animal | String |
| ANIMAL | Código interno para identificar o animal | Inteiro |
| F | Coeficiente de consangüinidade do animal | Real |
| DDPP120 | Valor da DEP direta para peso aos 120 dias | Real |
| ADPP120 | Acurácia da DEP direta para peso aos 120 dias | Real |
| DMPP120 | Valor da DEP materna para peso aos 120 dias | Real |
| AMPP120 | Acurácia da DEP materna para peso aos 120 dias | Real |
| DDPP240 | Valor da DEP direta para peso aos 240 dias | Real |
| ADPP240 | Acurácia da DEP direta para peso aos 240 dias | Real |
| DMPP240 | Valor da DEP materna para peso aos 240 dias | Real |
| AMPP240 | Acurácia da DEP materna para peso aos 240 dias | Real |
| DDPP365 | Valor da DEP direta para peso aos 365 dias | Real |
| ADPP365 | Acurácia da DEP direta para peso aos 365 dias | Real |
| DDPE550 | Valor da DEP direta para peso aos 550 dias | Real |
| ADPE550 | Acurácia da DEP direta para peso aos 550 dias | Real |
| DDPE365 | Valor da DEP direta para perímetro escrotal aos 365 dias | Real |
| ADPE365 | Acurácia da DEP direta para perímetro escrotal aos 365 dias | Real |
| DDPE550 | Valor da DEP direta para perímetro escrotal aos 550 dias | Real |
| ADPE550 | Acurácia da DEP direta para perímetro escrotal aos 550 dias | Real |
| MGT | Mérito Genético Total | Real |



# Transformación de la Base de Dados



## Reglas generadas y los trucos

**CN2** — DDPE550 mín -1.6 max 2.2

```
if:
   DDPP365-P > 5.50
   DDPE550-P > 0.95
   DMPP120-M > -1.75
   DDPP550-M > 0.35
   DDPE365 > 0.15
   -> class bueno [error 0%][casos 99]
if:
   DDPP550-P < 18.00
   DDPE550-P > 0.65
   DDPP550-M > 4.80
   DDPP365 < 12.20
   DDPE365 > 0.35
   -> class bueno [error 0%][casos 59]
if:
   DDPE550-P > 1.50
   DDPP365 > 3.00
   DDPE365 > 0.05
   -> class bueno
```

**C4.5-rules**

```
if:
   DDPE550-P > 1.4
   DDPP365-M > 0.6
   -> class bueno [error 0%][casos 50]
if:
   DDPE550-P > 0.9
   DDPP365-M > -1.4
   DMPP240 > -1.2
   DDPE365 > 0
   -> class bueno [error 4.2%][casos 119]
if:
   DDPE550-P > -0.1
   DMPP120-M > 0.7
   DDPP240-M > 5.6
   DDPP365 <= 12.9
   DDPE365 > 0.2
   -> class bueno [error 0%][casos 17]
```



## Algunos de los conocimientos obtenidos

- **Toros reproductores con perímetro testicular grande (realmente grande 40 cm.) no necesitan de vacas buenas para que puedan transmitir esa variable a los hijos.**
- **A medida que el perímetro testicular disminuye el peso de la vaca comienza a jugar una factor importante (rangos son proporcionados).**



## Para los que no se han dormido en esta charla: Ideas para trabajar

- **CBA**
  - Extrae reglas de asociación y construye clasificadores usando un subconjunto de esas reglas.
- **NeuroRules - RuleNeg**
  - Extracción de reglas a partir de Redes Neuronales.
- **Evaluación de Modelos**
  - ¿Qué modelo es mejor respecto a qué? (SAS)
- **Vector Support Machine**
  - Nueva Área