

# Detección de Intrusos mediante técnicas de MInería de DAtos

## DIMIDA

Antonio Bella Sanjuán

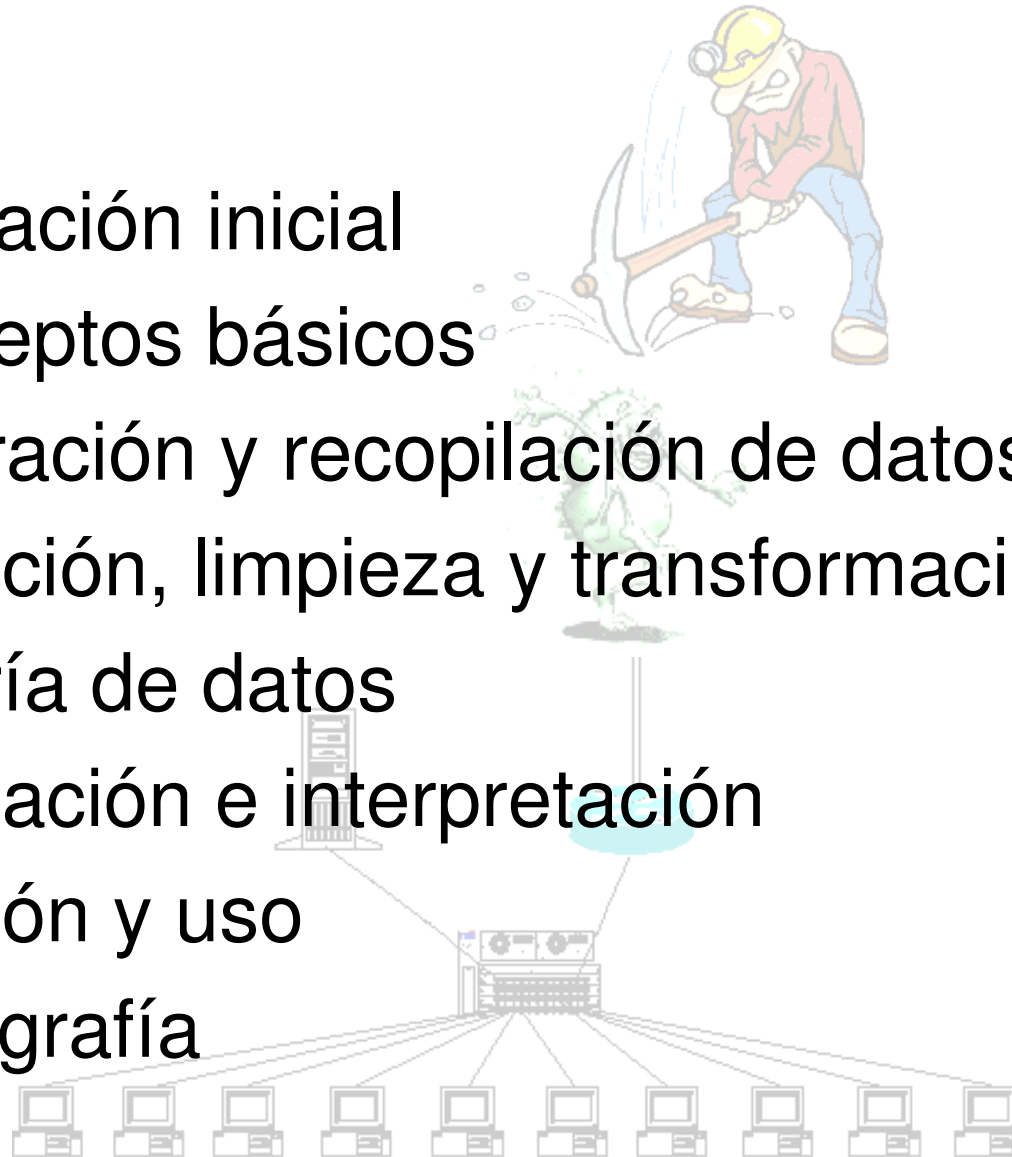
[abella@dsic.upv.es](mailto:abella@dsic.upv.es)



**Máster Universitario en Redes Corporativas  
e Integración de Sistemas**

# Índice

1. Motivación inicial
2. Conceptos básicos
3. Integración y recopilación de datos
4. Selección, limpieza y transformación
5. Minería de datos
6. Evaluación e interpretación
7. Difusión y uso
8. Bibliografía



# 1. Motivación inicial (I)

- Tesina. ¿Qué mejor que aplicar minería de datos a algún problema de redes?
- A la detección de intrusos.
- La búsqueda “*data mining*” AND “*intrusion detection*” da 230.000 entradas en el google.
- Todo inventado,... pero se podrían realizar experimentos si se dispusieran de datos reales y etiquetados (¿intrusión o no?)



# 1. Motivación inicial (II)

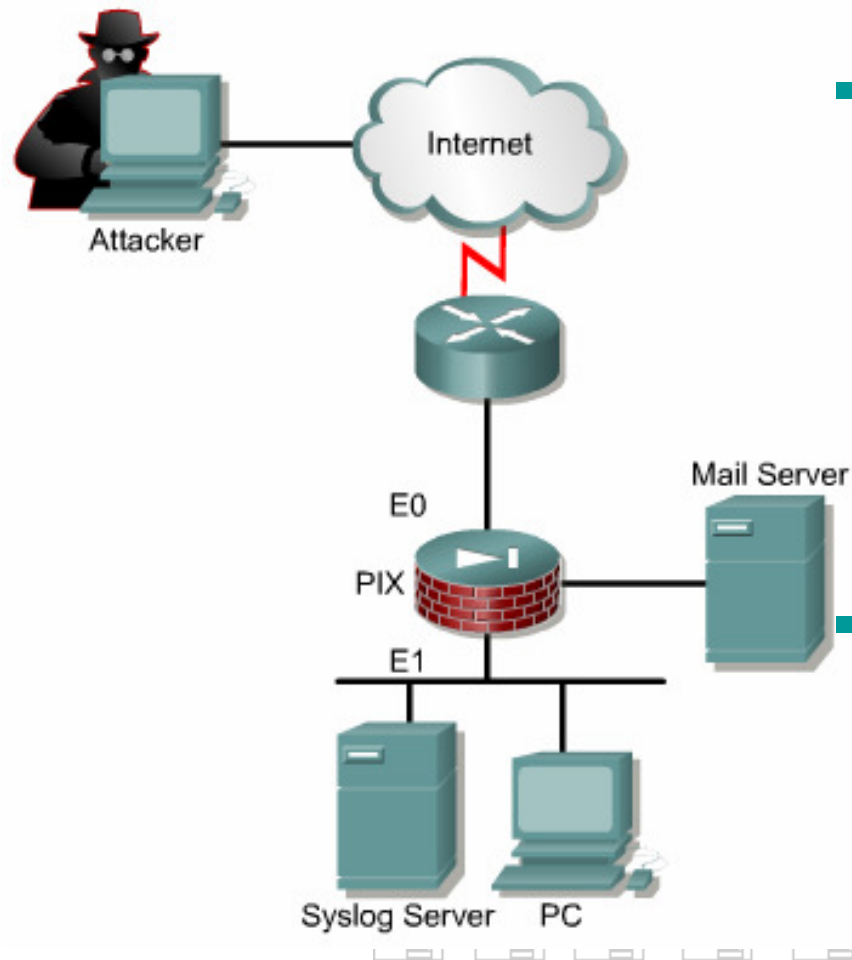
- Contrariamente a lo que normalmente sucede,... existen datos reales, correctamente etiquetados y GRATIS.



- KDD Cup 1999 Data [1]



## 2. Conceptos básicos

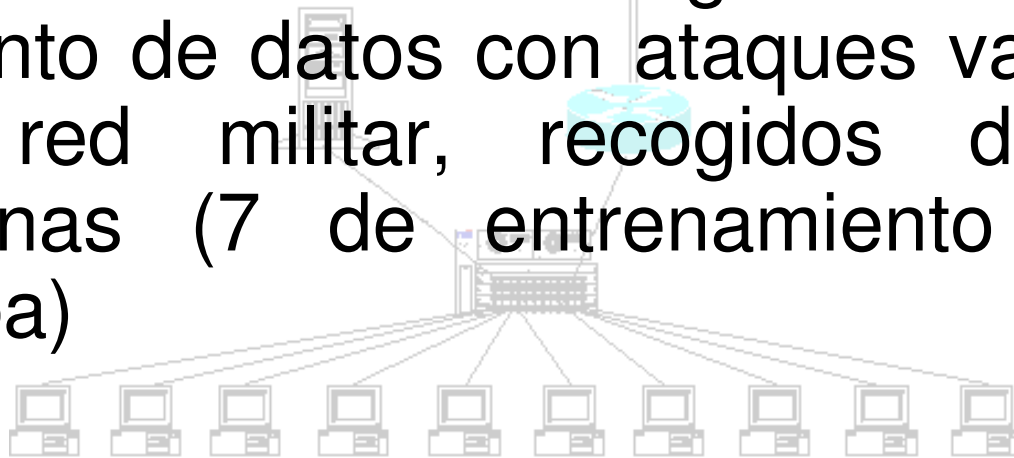


- Se puede definir ***intrusión*** como cualquier conjunto de acciones que tratan de comprometer la integridad, confidencialidad o disponibilidad de un recurso. [2]

- La ***detección de intrusos*** es la capacidad de detectar ataques en una red, incluyendo dispositivos y computadores. [3]

### 3. Integración y recopilación de datos (I)

- La tarea de la KDD Cup era conseguir un detector de intrusos, un modelo predictivo capaz de distinguir entre las conexiones “malas”, llamadas intrusiones o ataques, y las “buenas” o normales.
- La base de datos original contiene un conjunto de datos con ataques variados en una red militar, recogidos durante 9 semanas (7 de entrenamiento y 2 de prueba)



### 3. Integración y recopilación de datos (II)

- Cada conexión (registro) está etiquetada como “normal” o como ataque, con un tipo específico de ataque, que se podrán agrupar en:
  - DOS: denegación de servicio
  - R2L: acceso no autorizado a una máquina remota
  - U2R: acceso no autorizado a los privilegios de superusuario.
  - Probe: monitorización.
- Los datos de prueba no tienen la misma distribución de probabilidad que los de entrada, para hacer la tarea más realista.



### 3. Integración y recopilación de datos (III)

- 42 atributos (100 bytes / registro)
- Duration: tiempo de conexión (continuo)
- Protocol\_type: tipo de protocolo (discreto)
- Service: servicio de red en el destino (discreto)
- Src\_bytes: nº bytes de datos de fuente a destino (continuo)
- Dst\_bytes: nº bytes de datos de destino a fuente (continuo)
- ...
- Attack\_type: Tipo de ataque (clase)





## 4. Selección, limpieza y transformación

- Fichero entrenamiento: 743 MB
- Fichero 10 % entrenamiento: 75 MB
- Transformar el ***attack\_type*** a los 5 valores:
  - back → DOS, buffer\_overflow → U2R, ftp\_write → R2L, guess\_passwd → R2L, imap → R2L, ipsweep → probe,...

### 1. Resample 10 % unsupervised

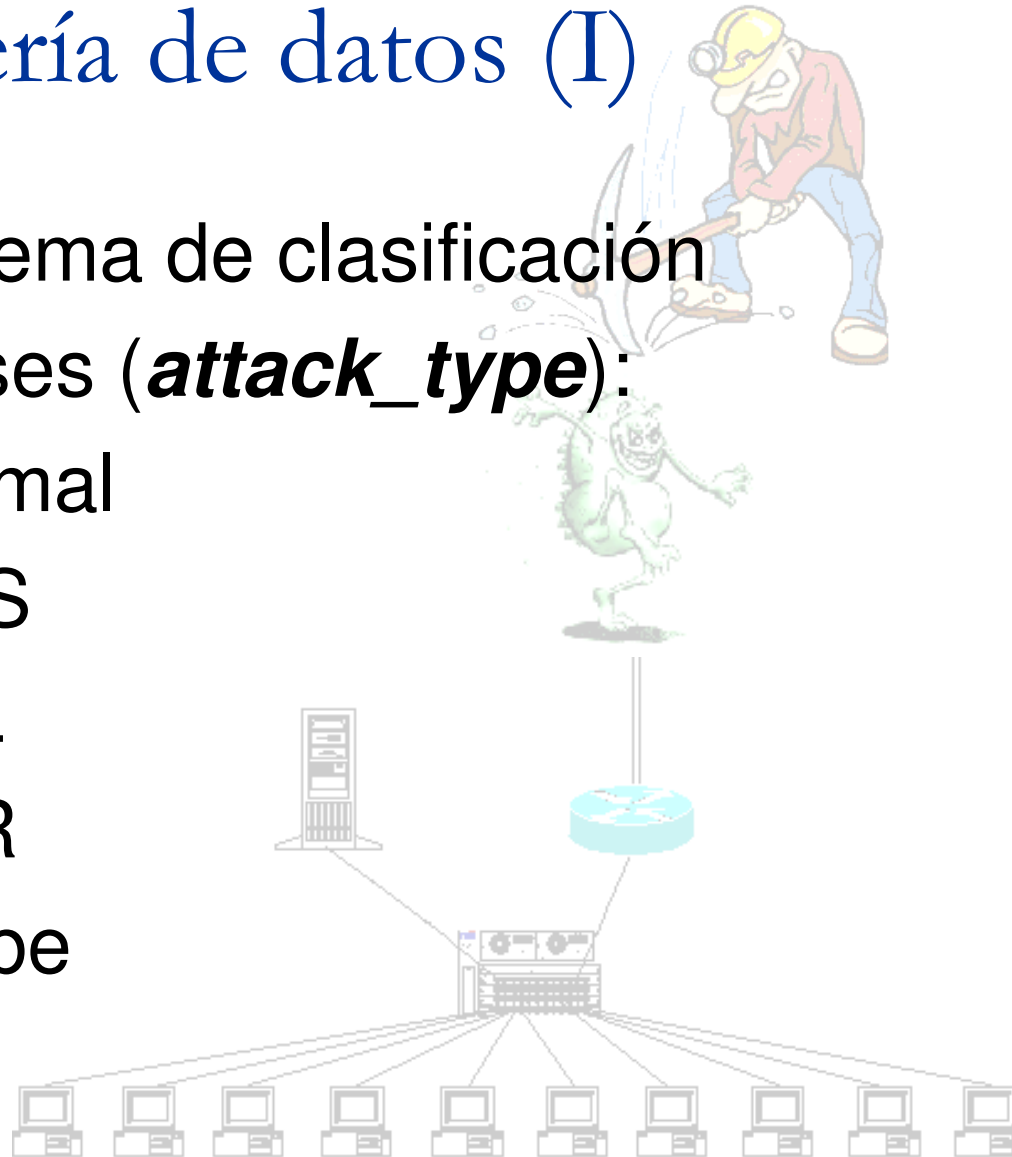
1. Selección de atributos: de 42 a 8 (src\_bytes, dst\_bytes, logged\_in, count, srv\_diff\_host\_rate, dst\_host\_count, dst\_host\_srv\_diff\_host\_rate, ***attack\_type***)

### 2. Resample 10 % supervised

1. Selección de atributos: de 42 a 5 (src\_bytes, dst\_bytes, logged\_in, dst\_host\_count, ***attack\_type***)

## 5. Minería de datos (I)

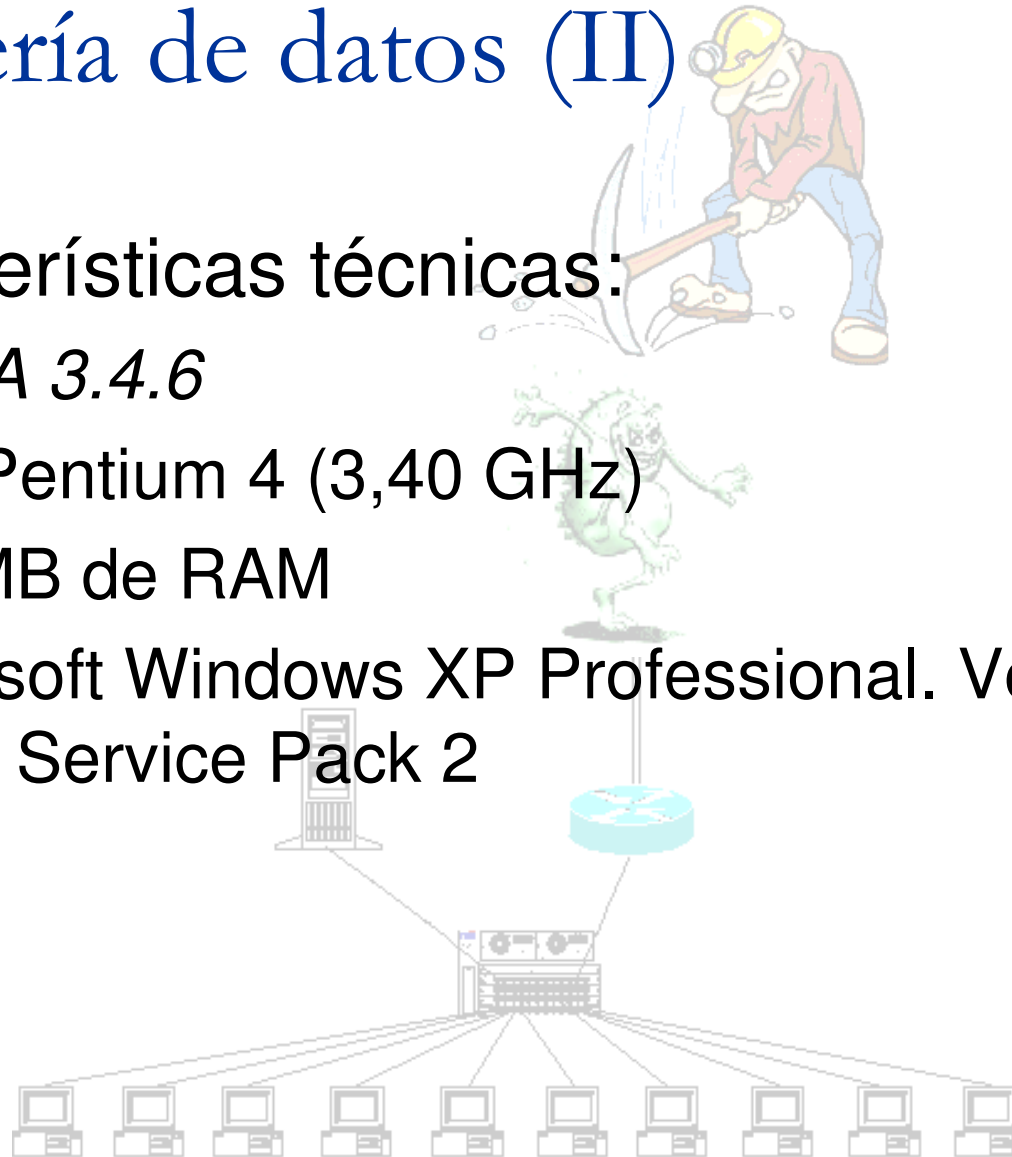
- Problema de clasificación
- 5 clases (***attack\_type***):
  - Normal
  - DOS
  - R2L
  - U2R
  - Probe



## 5. Minería de datos (II)

### ■ Características técnicas:

- ❑ *WEKA 3.4.6*
- ❑ Intel Pentium 4 (3,40 GHz)
- ❑ 992 MB de RAM
- ❑ Microsoft Windows XP Professional. Versión 2002. Service Pack 2



## 5. Minería de datos (III)

### ■ Métodos utilizados:

#### □ Rules

- ZeroR
- Ridor
- PART
- Decision Table
- Conjunctive Rule

#### □ Bayes

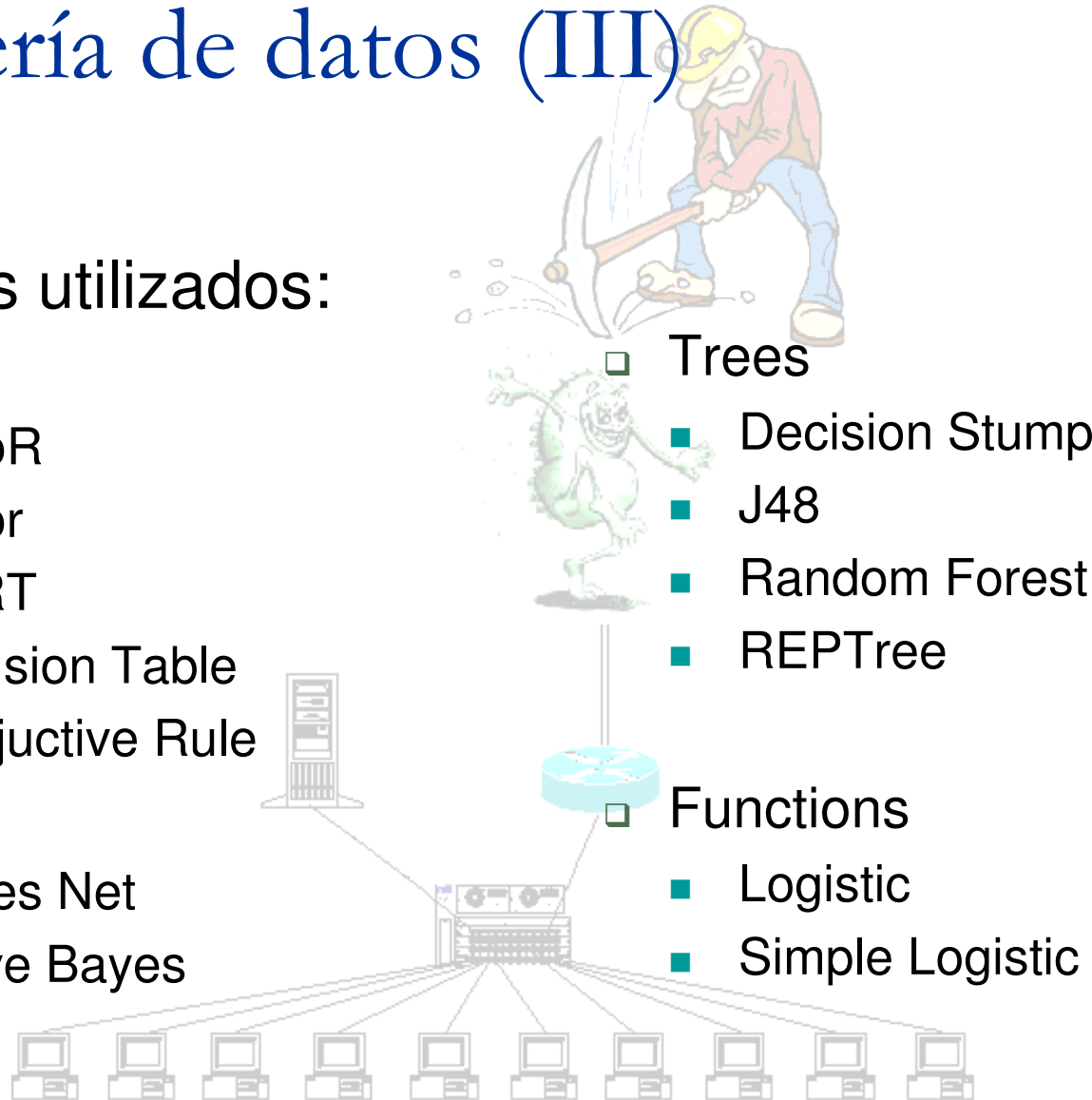
- Bayes Net
- Naive Bayes

#### □ Trees

- Decision Stump
- J48
- Random Forest
- REPTree

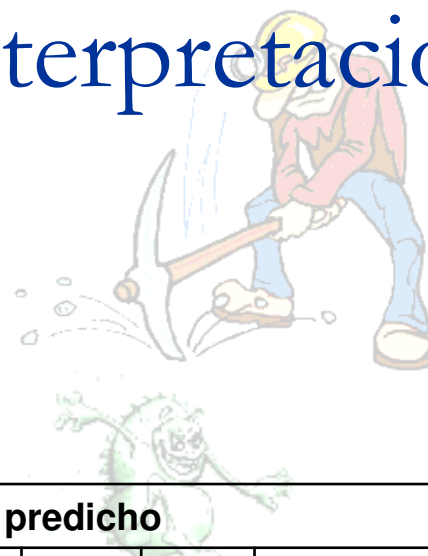
#### □ Functions

- Logistic
- Simple Logistic

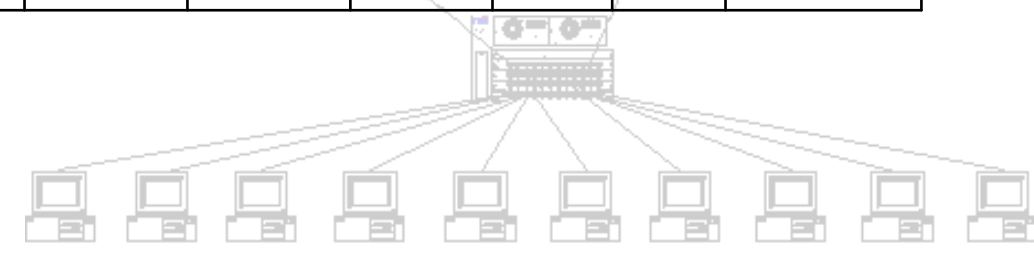


## 6. Evaluación e interpretación (I)

- Matriz de coste:



		predicho				
		normal	probe	DOS	U2R	R2L
real	normal	0	1	2	2	2
	probe	1	0	2	2	2
	DOS	2	1	0	2	2
	U2R	3	2	2	0	2
	R2L	4	2	2	2	0

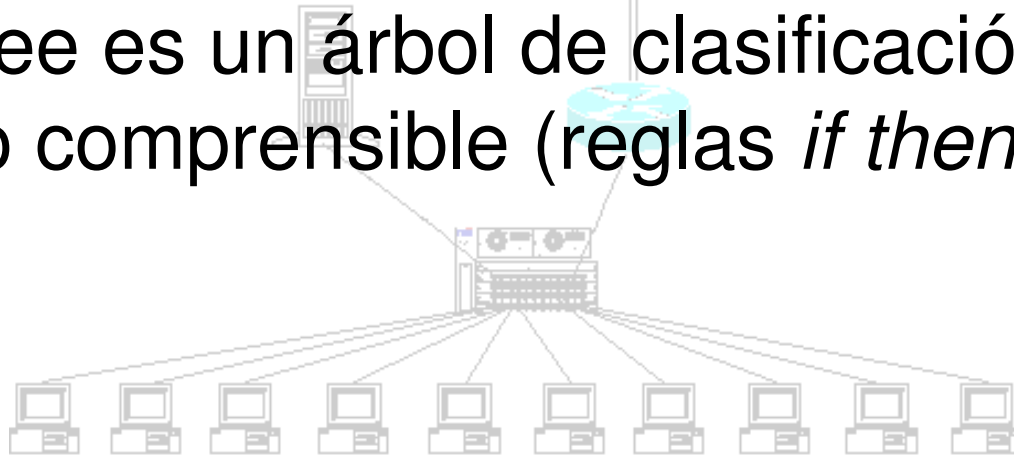


## 6. Evaluación e interpretación (II)

		8 atributos		5 atributos	
		Acertos	Coste Unitario	Acertos	Coste Unitario
Rules	ZeroR	74,42%	0,3647	74,42%	0,3647
	Ridor	92,09%	0,1339	89,50%	0,1648
	PART	92,04%	0,1336	91,67%	0,1529
	DecisionTable	92,02%	0,1369	89,78%	0,1626
	ConjunctiveRule	90,53%	0,1491	89,91%	0,1684
Trees	DecisionStump	90,10%	0,1533	89,57%	0,1718
	J48	92,07%	<b>0,1286</b>	91,38%	0,1471
	RandomForest	92,19%	<b>0,1280</b>	91,62%	0,1450
	REPTree	92,32%	<b>0,1245</b>	91,51%	0,1424
Functions	Logistic	90,68%	0,1480	86,23%	0,2303
	SimpleLogistic	90,64%	0,1485	86,15%	0,2311
Bayes	BayesNet	91,30%	0,1477	89,29%	0,1661
	NaiveBayes	85,00%	0,2663	84,26%	0,2708

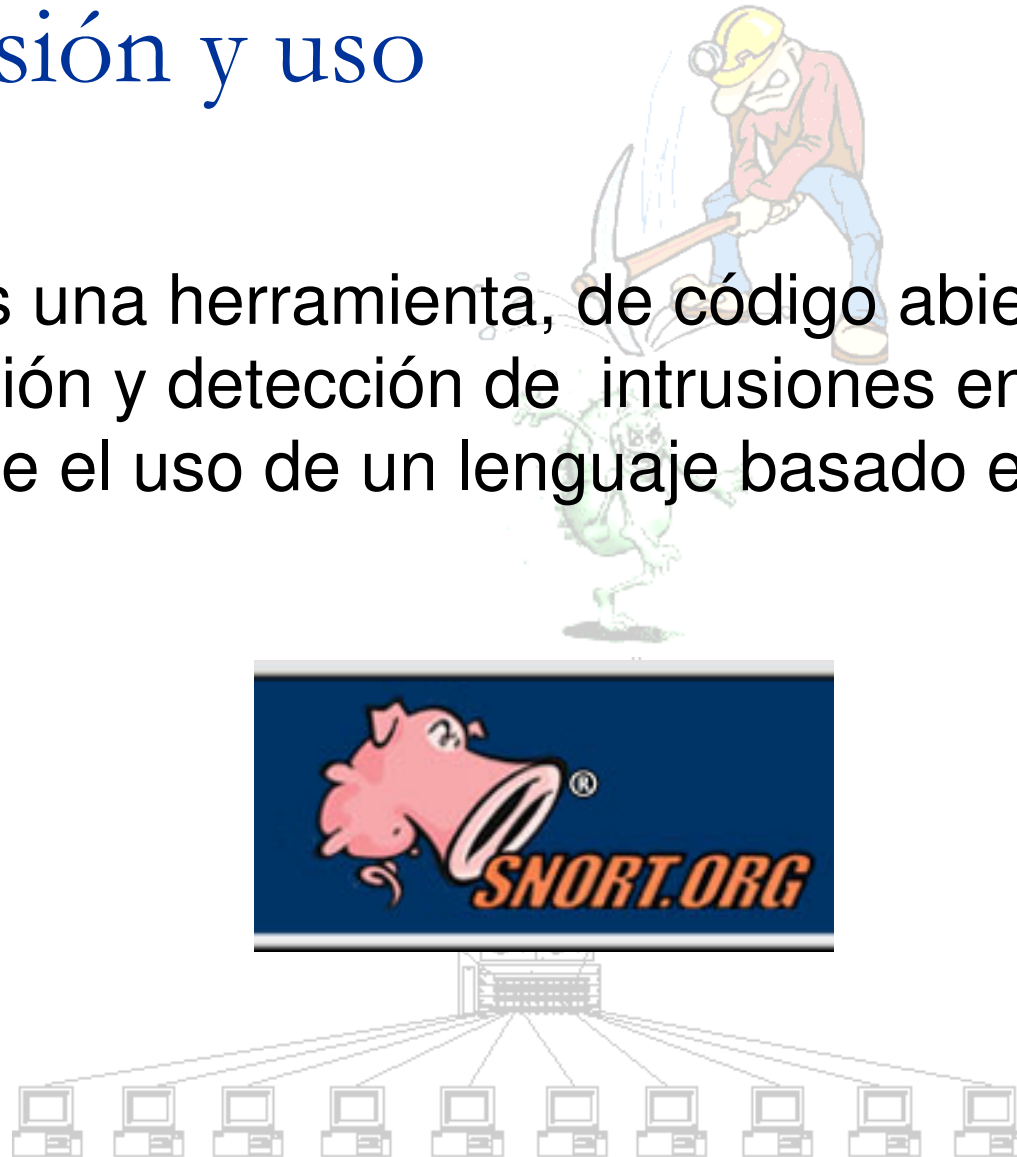
## 6. Evaluación e interpretación (III)

- Mejor coste unitario obtenido en la KDD'99 0,2331.
- Mejoramos en 0,1086.
- Sólo 8 atributos de los 42.
- REPTree es un árbol de clasificación con modelo comprensible (reglas *if then else*)



## 7. Difusión y uso

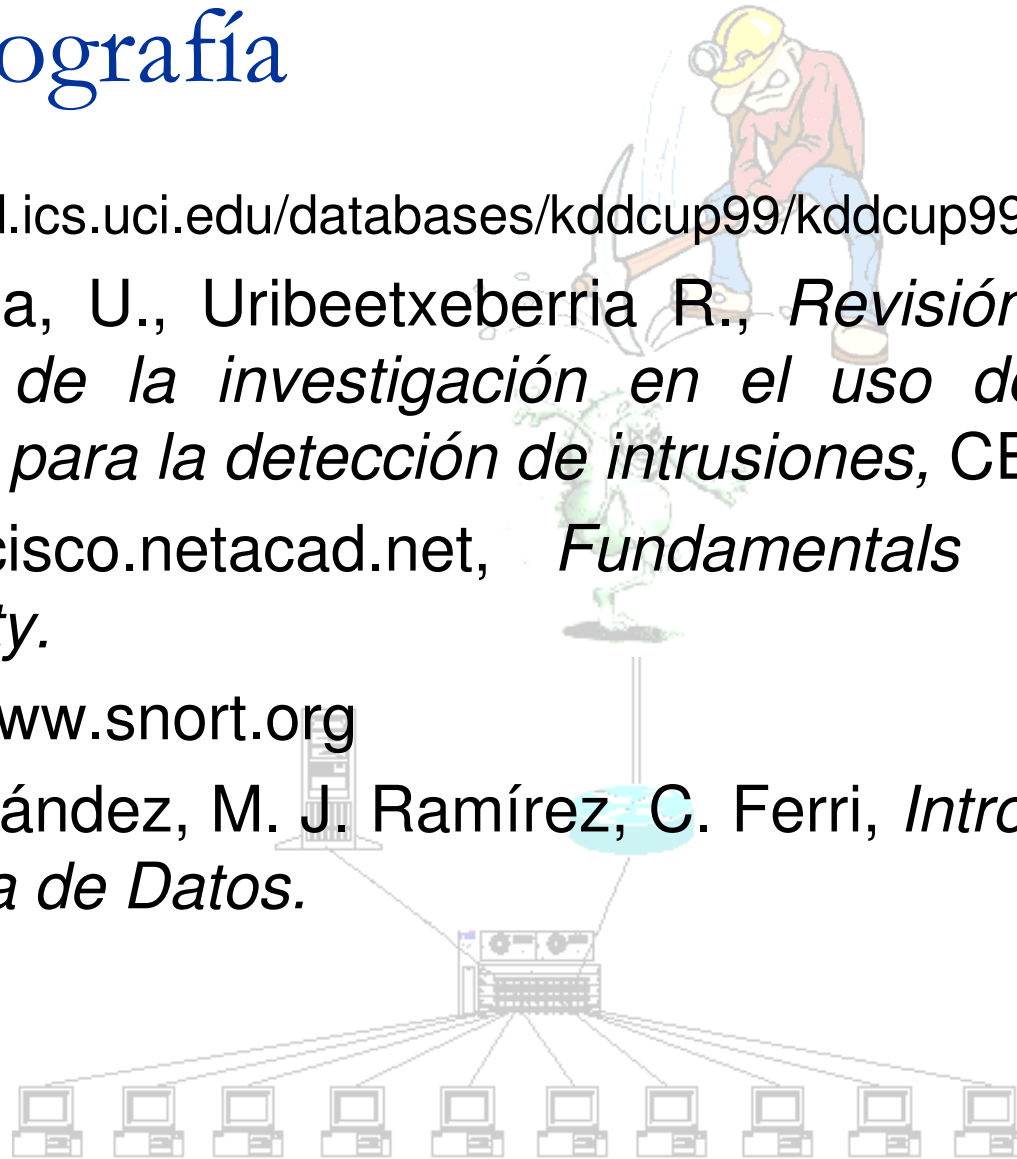
- Snort es una herramienta, de código abierto, para la prevención y detección de intrusiones en una red, mediante el uso de un lenguaje basado en reglas [4]





## 8. Bibliografía

- [1] <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
- [2] Zurutuza, U., Uribeetxeberria R., *Revisión del estado actual de la investigación en el uso de data data mining para la detección de intrusiones*, CEDI'05.
- [3] <http://cisco.netacad.net>, *Fundamentals of Network Security*.
- [4] <http://www.snort.org>
- [5] J. Hernández, M. J. Ramírez, C. Ferri, *Introducción a la Minería de Datos*.



¿Preguntas?

¿Sugerencias?

