

# BIOINFORMÁTICA

## 2013 - 2014

---

### PARTE I. INTRODUCCIÓN

- Tema 1. Computación Basada en Modelos Naturales

### PARTE II. MODELOS BASADOS EN ADAPTACIÓN SOCIAL (Swarm Intelligence)

- Tema 2. Introducción a los Modelos Basados en Adaptación Social
- Tema 3. Optimización Basada en Colonias de Hormigas
- Tema 4. Optimización Basada en Nubes de Partículas (Particle Swarm)

### PARTE III. COMPUTACIÓN EVOLUTIVA

- Tema 5. Introducción a la Computación Evolutiva
- Tema 6. Algoritmos Genéticos I. Conceptos Básicos
- Tema 7. Algoritmos Genéticos II. Diversidad y Convergencia
- Tema 8. Algoritmos Genéticos III. Múltiples Soluciones en Problemas Multimodales
- Tema 9. Estrategias de Evolución y Programación Evolutiva
- Tema 10. Algoritmos Basados en Evolución Diferencial (Differential Evolution – DE)
- Tema 11. Modelos de Evolución Basados en Estimación de Distribuciones (EDA)
- Tema 12. Algoritmos Evolutivos para Problemas Multiobjetivo
- Tema 13. Programación Genética
- Tema 14. Modelos Evolutivos de Aprendizaje

### PARTE IV. OTROS MODELOS DE COMPUTACIÓN BIOINSPIRADOS

- Tema 15. Sistemas Inmunológicos Artificiales
- Tema 16. Otros Modelos de Computación Natural/Bioinspirados

# BIOINFORMÁTICA

## TEMA 14: MODELOS EVOLUTIVOS DE APRENDIZAJE

---

- 1. INTRODUCCIÓN AL APRENDIZAJE EVOLUTIVO**
- 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS**
- 3. APRENDIZAJE CON PROGRAMACIÓN GENÉTICA**
- 4. APRENDIZAJE EVOLUTIVO HÍBRIDO. SELECCIÓN EVOLUTIVA DE INSTANCIAS**
- 5. COMENTARIOS FINALES**

# **1. INTRODUCCIÓN AL APRENDIZAJE EVOLUTIVO**

---

- **INTRODUCCIÓN AL APRENDIZAJE**
- **APRENDIZAJE EVOLUTIVO**

# Introducción al Aprendizaje

---

**El objetivo de un proceso de aprendizaje es obtener una cierta regla o sistema que permita**

**clasificar objetos de un modo automático, o**

**predecir el valor de las variables de control de un sistema.**

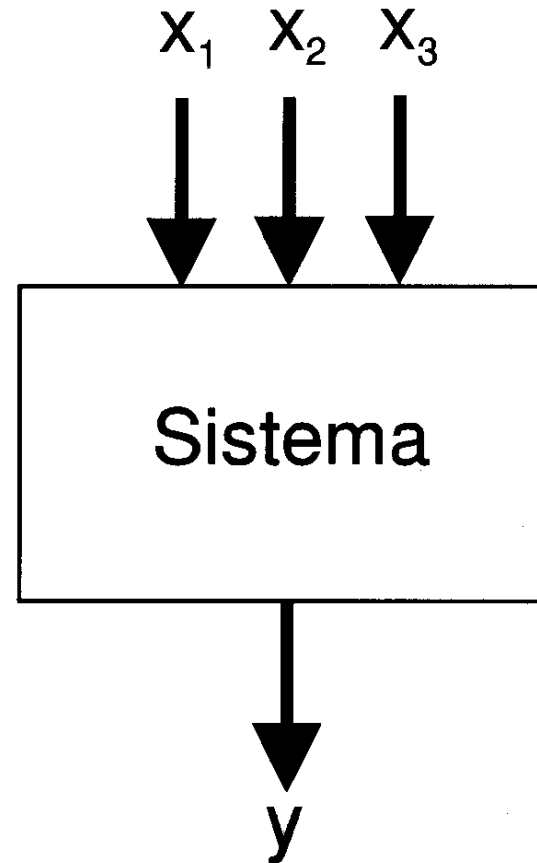
**(extraer conocimiento de una base de datos)**

# Introducción al Aprendizaje (2)

---

- **Modelado o Predicción**

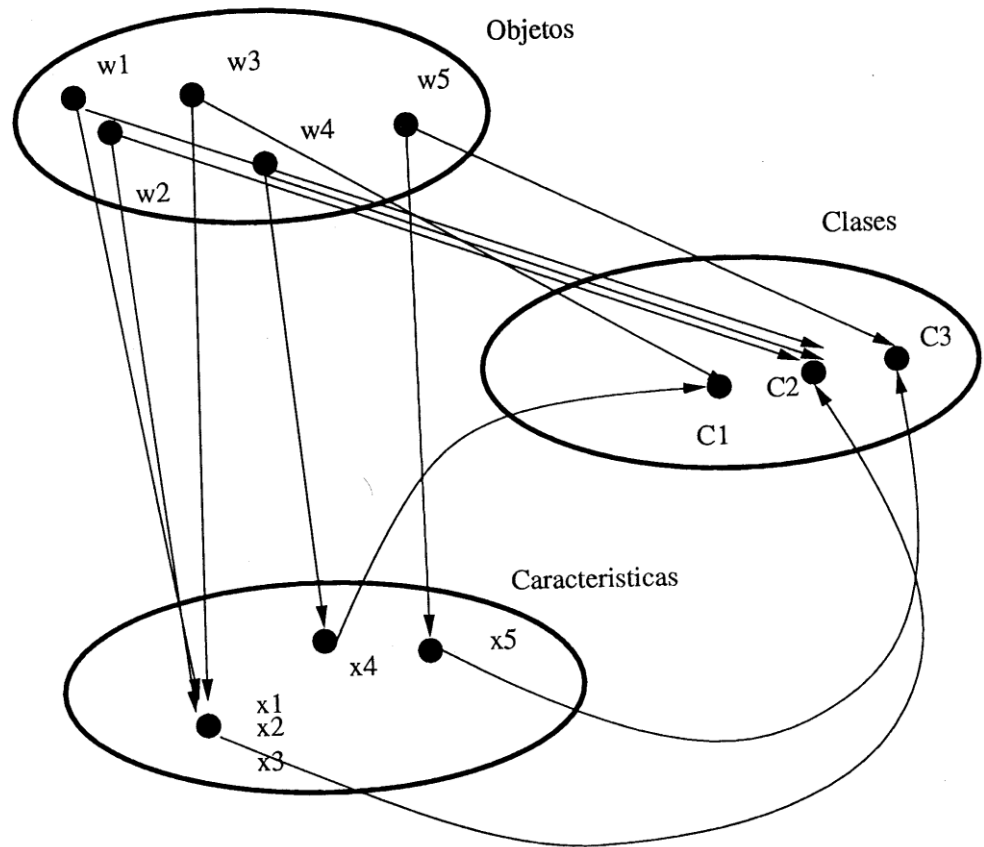
El problema fundamental de la predicción está en modelar la relación entre las variables de estado para obtener el valor de la variable de control.



# Introducción al Aprendizaje (3)

## ■ Clasificación

El problema fundamental de la clasificación está directamente relacionado con la separabilidad de las clases.



# Introducción al Aprendizaje (4)

---

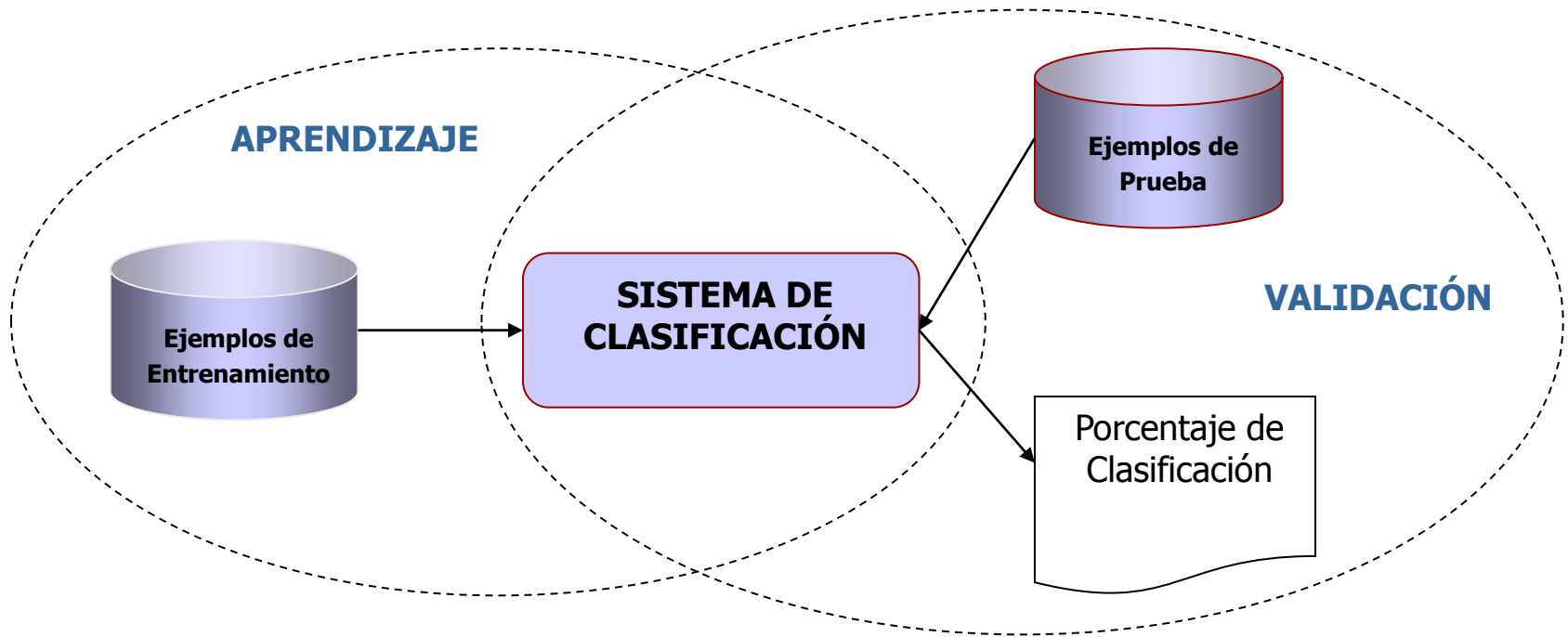
## Clasificación

- Se dispone de una muestra de objetos (patrones):  $S = \{w_1, \dots, w_k\}$
- **OBJETIVO:** Obtener una cierta regla o sistema que permita clasificar dichos objetos de un modo automático.
- Clasificación **supervisada:**
  - Se conocen las clases existentes en el problema:  $\{C_1, \dots, C_M\}$
  - Se conoce la clase concreta a la que pertenece cada objeto del conjunto de datos

$$\begin{array}{lcl} w_1 = (x_1(w_1), \dots, x_n(w_1)) & \rightarrow & C_{i_1} \\ \vdots & & \vdots \\ w_k = (x_1(w_k), \dots, x_n(w_k)) & \rightarrow & C_{i_k} \end{array} \quad i_j \in \{1, \dots, M\}, \quad j \in \{1, \dots, k\}$$

# Introducción al Aprendizaje (5)

---





# Aprendizaje Evolutivo

---

Los Algoritmos Evolutivos no fueron diseñados como modelos de aprendizaje, a diferencia de otros paradigmas como las Redes Neuronales.

Sin embargo, muchas de las propuestas de aprendizaje existentes se modelan mediante algoritmos de optimización y búsqueda.

Problema: Espacio de búsqueda grande y complejo.

Solución: Aplicar *Algoritmos Evolutivos*

# Aprendizaje Evolutivo (2)

---

**Podemos encontrar diferentes formas de uso de los Algoritmos Evolutivos en Problemas de Aprendizaje y Extracción de Conocimiento:**

**Aprendizaje Evolutivo de Sistemas Basados en Reglas (reglas intervalares, reglas fuzzy, ...)**

**Aprendizaje con Programación Genética para Regresión y Clasificación**

**Aprendizaje Evolutivo Híbrido (Redes Neuronales Evolutivas, Extracción Evolutiva de Prototipos, Clustering Evolutivo ...)**

**Aplicación en diferentes áreas de KDD: Reducción de Datos, Extracción de Reglas de Asociación en Problemas de Data Mining**

....

## **Bibliografía**

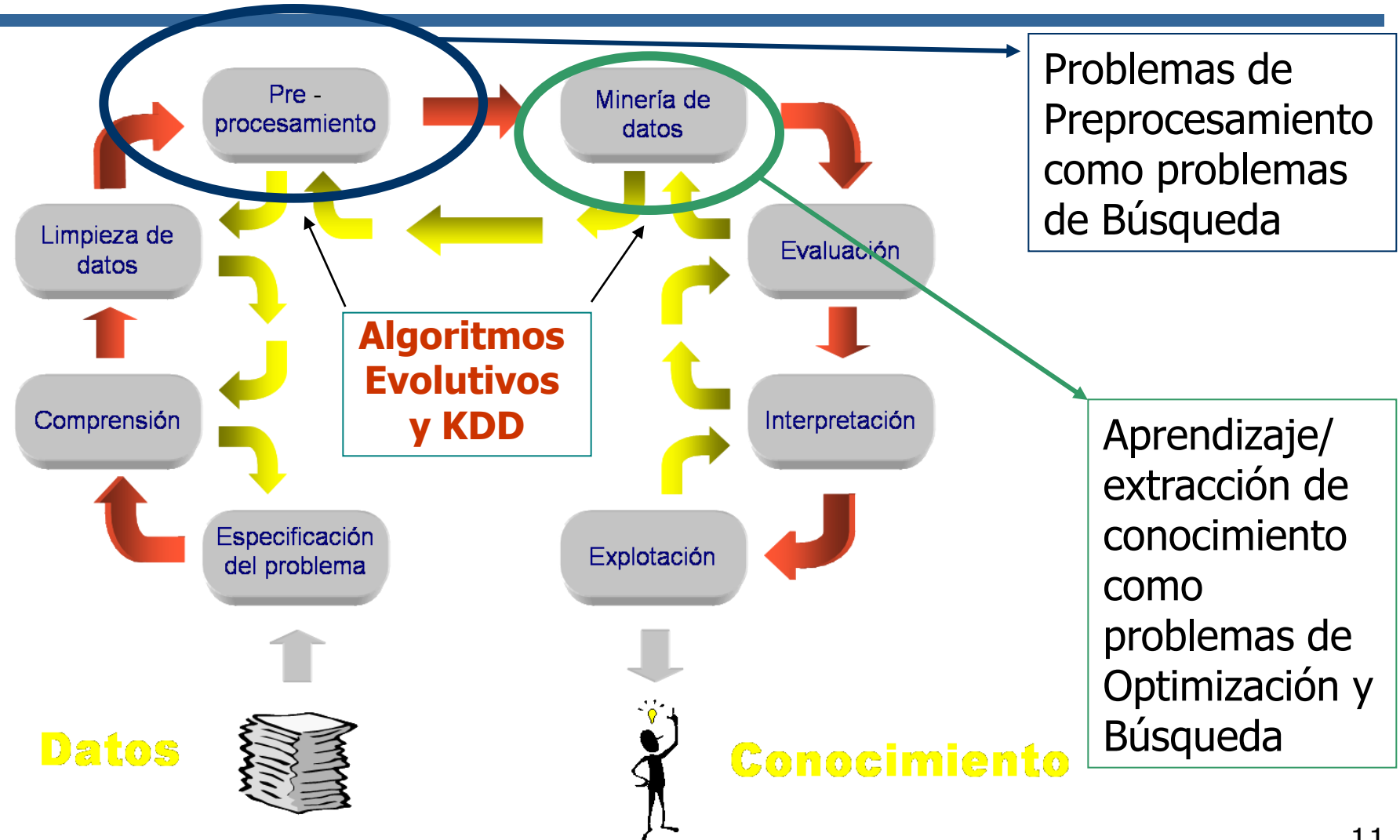
O. Cordón, F. Herrera, F. Hoffmann, L. Magdalena. GENETIC FUZZY SYSTEMS. Evolutionary Tuning and Learning of Fuzzy Knowledge Bases. World Scientific, July 2001.

A.A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag, 2002.

A. Ghosh, L.C. Jain (Eds.), *Evolutionary Computation in Data Mining*. Springer-Verlag, 2005.

# Aprendizaje Evolutivo (3)

## KDD: Proceso de Extracción de Conocimiento de Bases de Datos



## 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS

---

Los cromosomas codifican una regla o un conjunto de ellas.

Cada gen codifica una regla o parte de una regla.

**Si  $\text{cond}_1 \wedge \dots \wedge \text{cond}_n$**

**entonces Clase =  $C_i$**

## 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS (2)

### Algunos tipos de reglas para clasificación

#### Reglas intervalares

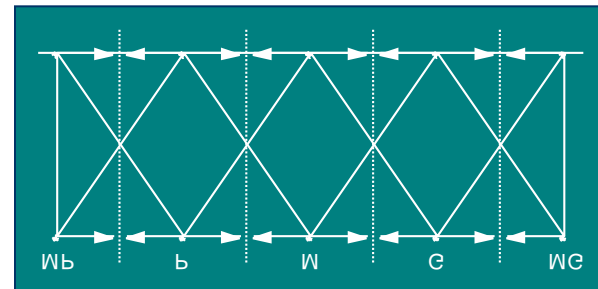
Si  $x_1 \in [a_1, b_1] \wedge \dots \wedge x_n \in [a_n, b_n]$   
Entonces Clase =  $C_i$

#### Reglas difusas

Si  $x_1$  es  $M \wedge \dots \wedge x_n$  es  $G$   
Entonces Clase =  $C_i$  con grado certeza  $r_i$   
( $M, \dots, G$  son etiquetas lingüísticas con funciones de pertenencia asociadas)

#### Referencia

**GENETIC FUZZY SYSTEMS.**  
Evolutionary Tuning and Learning of Fuzzy Knowledge Bases.  
O. Cordón, F. Herrera, F. Hoffmann, L. Magdalena  
World Scientific, Julio 2001. ISBN 981-02-4016-3



Funciones de Pertenencia

## 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS (3)

---

Existen cuatro modelos generales que están fundamentados en el uso de los AEs para aprender Sistemas Basados en Reglas (**Genetics based machine learning**)

**Cromosoma = Base de Reglas**

**Modelo Pittsburgh: GASSIST, CORCORAN, GIL**

**Cromosoma = Regla**

**Modelo Michigan (XCS, UCS)**

**LCS - Learning Classifier Systems**

**Modelo de Aprendizaje Iterativo**

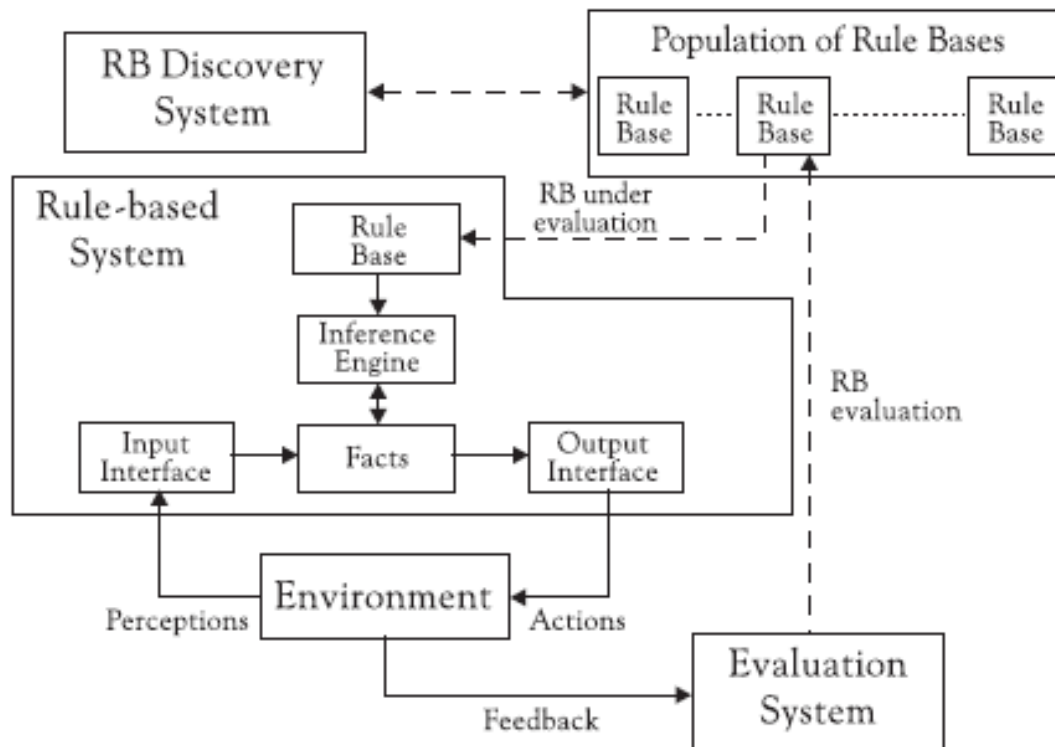
**IRL – Iterative Rule Learning (SIA, HIDER)**

**Modelo de Cooperación-Competición**

**GCCL – Genetic Cooperative-Competitive Learning (REGAL, LOGEMPRO)**

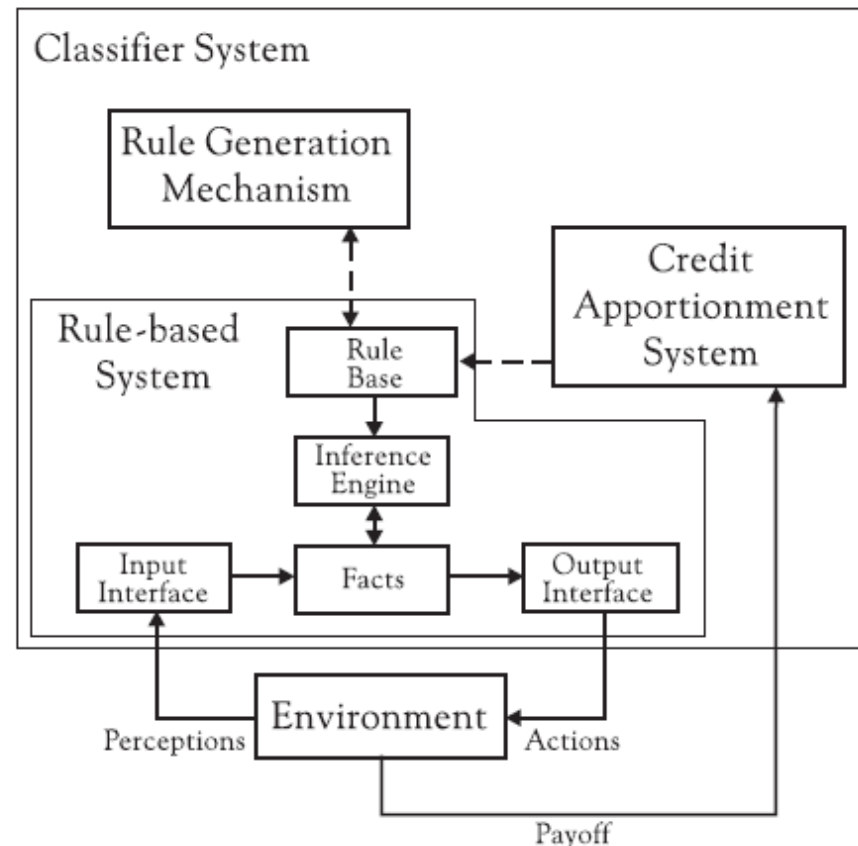
## 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS (3)

**Modelo Pittsburgh:** Cada cromosoma de la población codifica una definición distinta de la base de reglas completa. El AE aprende la mejor base de reglas asociada al problema.



## 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS (4)

**Modelo Michigan:** Cada cromosoma codifica a una sola regla. La población completa del AE compite por aprender una regla que se incorpora a una base de reglas externa. Se emplean mecanismos de refuerzo para medir la calidad de las reglas individuales (recompensa y penalización de un peso) y se sustituyen las reglas con peso bajo por nuevas reglas generadas por el proceso evolutivo.





## 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS (5)

---

**Modelo Iterativo:** Es un modelo de aprendizaje evolutivo inductivo, a partir de ejemplos.

Los cromosomas representan reglas individuales y se devuelve una única regla, la mejor. No existe cooperación entre las reglas de la población, todos los cromosomas compiten para seleccionar el mejor que formará parte de la base de reglas finalmente generada.

De la base de ejemplos se quitan los ejemplos ya cubiertos por reglas, y el algoritmo aprende nuevas reglas que cubren ejemplos diferentes. Es necesario ejecutar el AE varias veces, tantas como reglas a generar, pero se trabaja sobre un espacio de búsqueda mucho menor.

A posteriori, pueden aplicarse procesos de refinamiento para mejorar la cooperación entre las reglas mediante la selección de conjuntos de reglas cooperativas y mediante el ajuste de los parámetros de las reglas.

## 2. APRENDIZAJE EVOLUTIVO DE SISTEMAS BASADOS EN REGLAS (6)

---

**Modelo Cooperación-Competición:** Cada cromosoma codifica a una sola regla. La población completa del AE compone la base de reglas.

Existe un proceso de cooperación y competición entre las reglas.

Dispone de procedimientos de Evaluación local de las reglas y evaluación global de las reglas.

Puede trabajar con poblaciones de tamaño variable para poder modificar el tamaño de la base de reglas.

# Modelo Pittsburgh con Reglas Intervalares

---

**Referencia: A.L. Corcoran, S. Sen. Using Real-Valued Genetic Algorithms to Evolve Rule Sets for Classification. IEEE Int. Conf. on Evolutionary Computation (IEEE CEC'94), 120-124.**

**Modelo Pittsburgh:** Cada cromosoma representa una Base de Reglas completa.

Elementos a tener en cuenta:

- **REPRESENTACIÓN**
- **FUNCIÓN OBJETIVO**
- **OPERADORES**

# Modelo Pitt con Reglas Intervalares (2)

---

## REPRESENTACIÓN

Cada regla está compuesta por un conjunto de  $n$  atributos y una clase.

Cada atributo se codifica con dos variables reales que indican el mínimo y el máximo del rango de valores válidos para el atributo en esa regla.

**Si  $x_1 \in [a_1, b_1] \wedge \dots \wedge x_n \in [a_n, b_n]$  Entonces Clase =  $C_i$**

Se utilizan cromosomas de longitud fija. La longitud del cromosoma es:

$$L = m \cdot (2 \cdot n + 1) \quad (m \text{ indica el número de reglas})$$

Cuando el valor de la segunda variable es menor que la de la primera:

$$a_i \text{ (mínimo)} > b_i \text{ (máximo)}$$

se considera que ese atributo no interviene en la regla.

# Modelo Pitt con Reglas Intervalares (3)

---

## FUNCIÓN OBJETIVO

Maximización del Porcentaje de clasificación sobre los ejemplos de entrenamiento al emplear la base de reglas codificada en el cromosoma.

Cada ejemplo se clasifica con la regla que case con él. Cuando más de una regla se empareja con el mismo patrón y tienen como consecuente una clase distinta, se aplica un **sistema de resolución de conflictos**:

1. Para cada regla de la base, se determinan dos pesos, **número de ejemplos correctos-positivos** (emparejados con el antecedente y coincidiendo con la clase del consecuente) e **incorrectos-negativos** (emparejados con el antecedente y no coincidiendo con la clase).

# Modelo Pitt con Reglas Intervalares (4)

---

## FUNCIÓN OBJETIVO (2)

2. Para cada clase asociada a los consecuentes de las reglas emparejadas con el patrón, se suman el número de ejemplos correctos e incorrectos de las correspondientes reglas.
3. A la hora de clasificar, la clase seleccionada es aquella con menor número de ejemplos negativos asociados a las reglas. Si hay empate, se coge la clase con mayor número de ejemplos correctos asociados a las reglas. En caso de un nuevo empate, se coge una aleatoriamente.

# Modelo Pitt con Reglas Intervalares (5)

---

## OPERADORES GENÉTICOS

### Operador de cruce simple modificado (un punto):

Se seleccionan puntos de cruce en las fronteras de las reglas: **regla = gen**

### Operador de mutación por deslizamiento:

Se aplica a cada una de las variables de los atributos con una probabilidad llamada **ratio de deslizamiento**. Los valores de los atributos se mutan con un incremento o disminución por un valor llamado **fracción de deslizamiento**, que es una fracción del rango válido para cada atributo, ajustando los valores a los extremos si se salen de los intervalos de definición. Probabilidad de actuación alta, alrededor de 0.5.

No se aplica al consecuente de las clases.

# Modelo Pitt con Reglas Intervalares (6)

---

## OPERADORES GENÉTICOS (2)

### Operador de mutación aleatorio simple:

Se aplica sobre los atributos y clases, modificando los correspondientes valores (mínimo-máximo o valor de la clase) en sus correspondientes rangos.

Probabilidad de actuación baja, aproximadamente  $1/(2 \cdot L + 2)$ , donde L es el tamaño del cromosoma. Una actuación por recombinación.

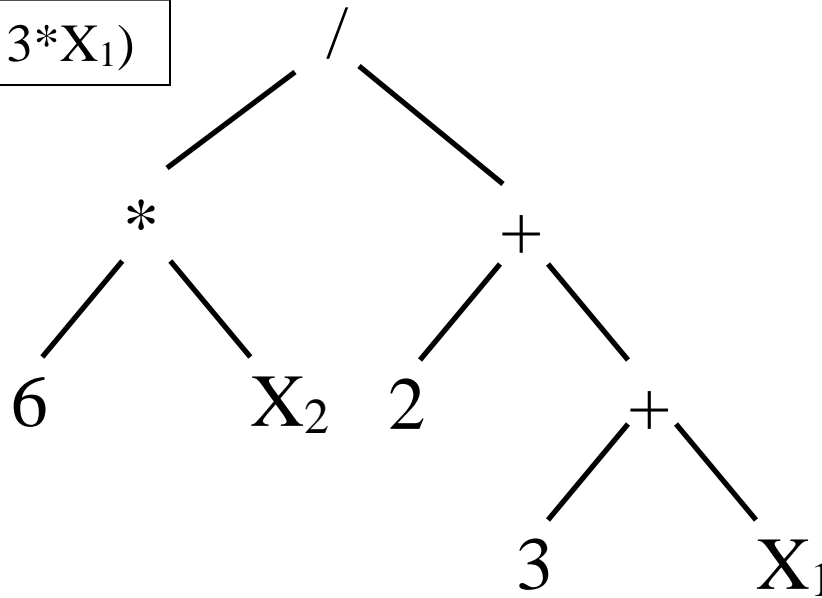
### Operador de mutación aleatorio simple:

Problema de 13 variables, gen=regla con 27 valores, cromosoma con 60 genes, tamaño de población 1500, ratio de deslizamiento 0.5, fracción de deslizamiento  $(b-a)/100$  ( $[a,b]$  int. de definición), prob. mutación  $1/28$ .



### 3. APRENDIZAJE CON PROGRAMACIÓN GENÉTICA

$$g = 6 * X_2 / (2 + 3 * X_1)$$



Se puede utilizar la Programación Genética para aprender funciones de regresión, funciones discriminantes, etc.

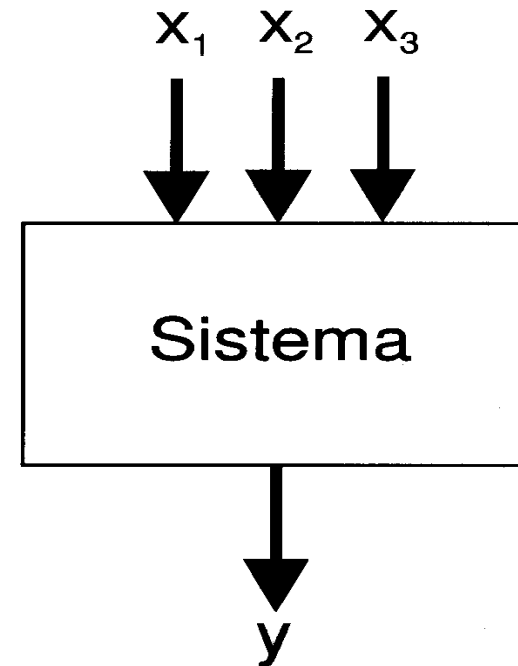
El uso de estas funciones puede dar lugar a diferentes modelos de aprendizaje.

- Símbolos terminales:  $\{X_i, \text{constantes}\}$
- Funciones:  $\{+, -, *, /, \sqrt{\quad}, \dots\}$

# 3. APRENDIZAJE CON PROGRAMACIÓN GENÉTICA

## Regresión Simbólica

Se pretende encontrar el valor de la variable  $Y$ , basándonos en  $X = (x_1, \dots, x_d)$



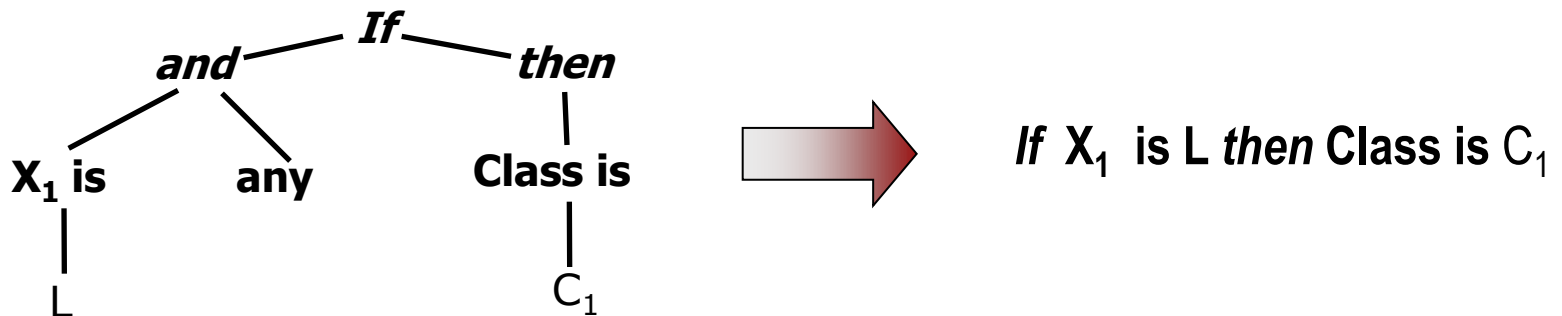
**Buscamos una función  $g$  tal que  $Y - g(X)$  sea pequeño para cualquier valor de  $X$ .**

# 3. APRENDIZAJE CON PROGRAMACIÓN GENÉTICA

## Clasificación mediante reglas

### Definición de una gramática libre de contexto

- start  $\rightarrow$  [If], antec, [then], conseq, [.]
- antec  $\rightarrow$  descriptor1, [and], descriptor2.
- descriptor1  $\rightarrow$  [any].
- descriptor1  $\rightarrow$  [ $X_1$  is] label.
- descriptor2  $\rightarrow$  [any].
- descriptor2  $\rightarrow$  [ $X_2$  is] label.
- label  $\rightarrow$  {member(?a, [L, M, H, L or M, L or H, M or H, L or M or H])}, [?a].
- conseq  $\rightarrow$  [Class is] descriptorClass
- descriptorClass  $\rightarrow$  {member(?a, [ $C_1, C_2, C_3$ ])}, [?a].

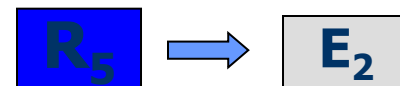
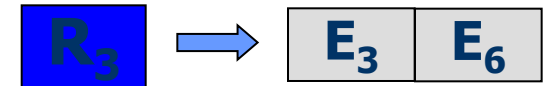
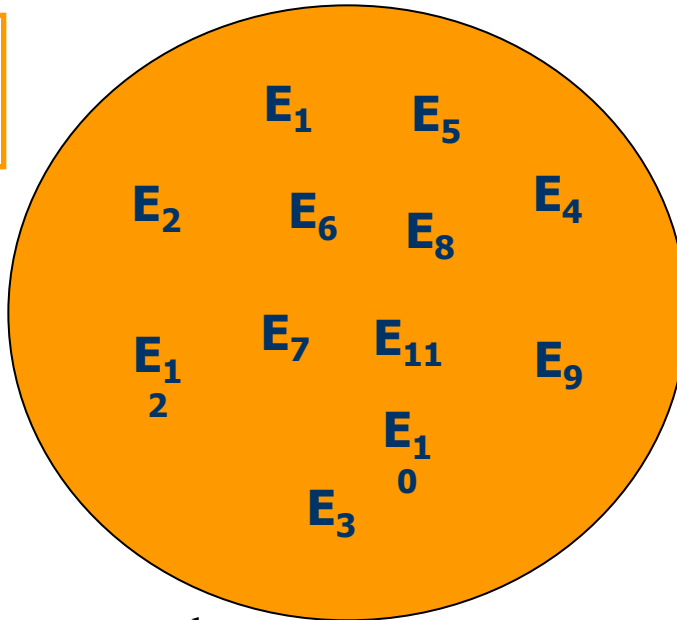


# 3. APRENDIZAJE CON PROGRAMACIÓN GENÉTICA

**Extracción Evolutiva de Reglas**  
**Modelo cooperación-competición**

## Mecanismo de Mantenimiento de la Diversidad

**Competición de Tokens**

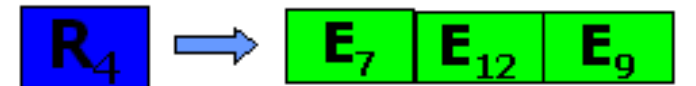
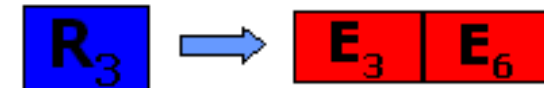
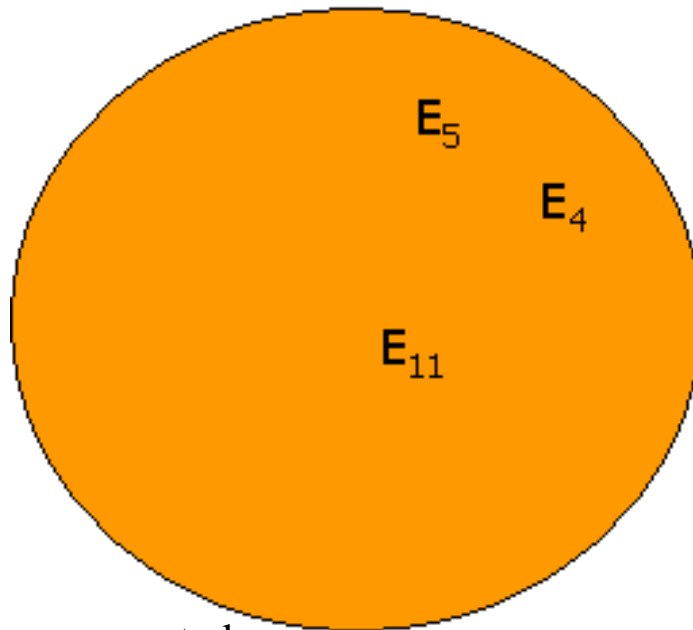


$$\text{Fitness\_modificado} = F \times \frac{\text{contador}}{\text{ideal}}$$

Burgos, 3-7 de  
 Junio de 2006

# 3. APRENDIZAJE CON PROGRAMACIÓN GENÉTICA

Extracción Evolutiva de Reglas  
Modelo cooperación-competición



$$\text{Fitness\_modificado} = F \times \frac{\text{contador}}{\text{ideal}}$$

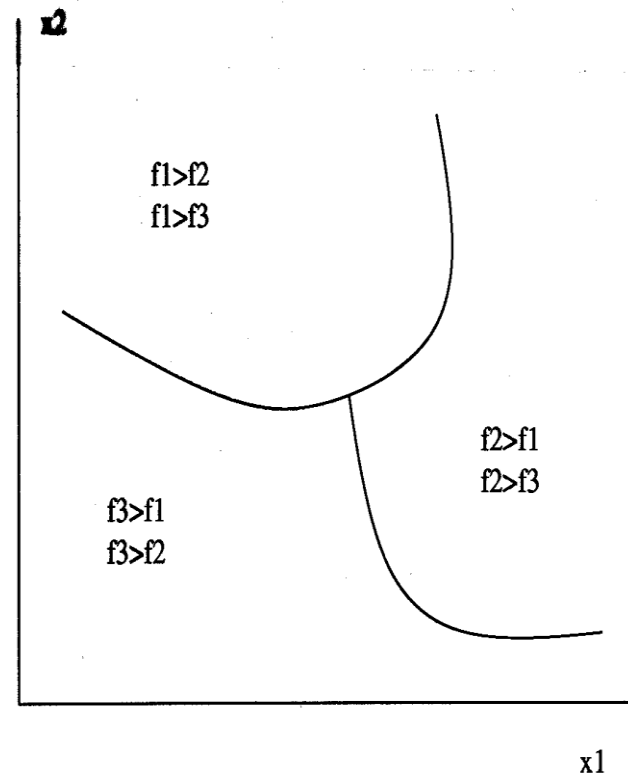
Burgos,  
Junio de

# Clasificación Mediante Funciones Discriminantes

Si se emplea el método de las funciones discriminantes, la clasificación está relacionada con la regresión simbólica.

## Modelo 1:

Se buscan  $M$  funciones  $f_1, \dots, f_M$  y se desea que  $f_k(x) > f_i(x)$ ,  $i \neq k$  cuando  $x$  es de la clase  $k$ .



# Clasificación Mediante Funciones Discriminantes (2)

---

## Modelo 2:

Si  $f_1(a) > 0$  entonces clase 1

si no

Si  $f_2(a) > 0$  entonces clase 2

si no

....

Si  $f_M(a) > 0$  entonces clase M

si no

el patrón no es clasificable

fin Si

fin Si

fin Si

## 4. APRENDIZAJE EVOLUTIVO HÍBRIDO. SELECCIÓN EVOLUTIVA DE INSTANCIAS

---

**Aprendizaje Evolutivo Híbrido.** Los AGs se utilizan conjuntamente con un modelo de aprendizaje:

sustituyendo a la técnica de búsqueda asociada:

***redes neuronales evolutivas, clustering evolutivo, selección evolutiva de prototipos, selección evolutiva de características...***

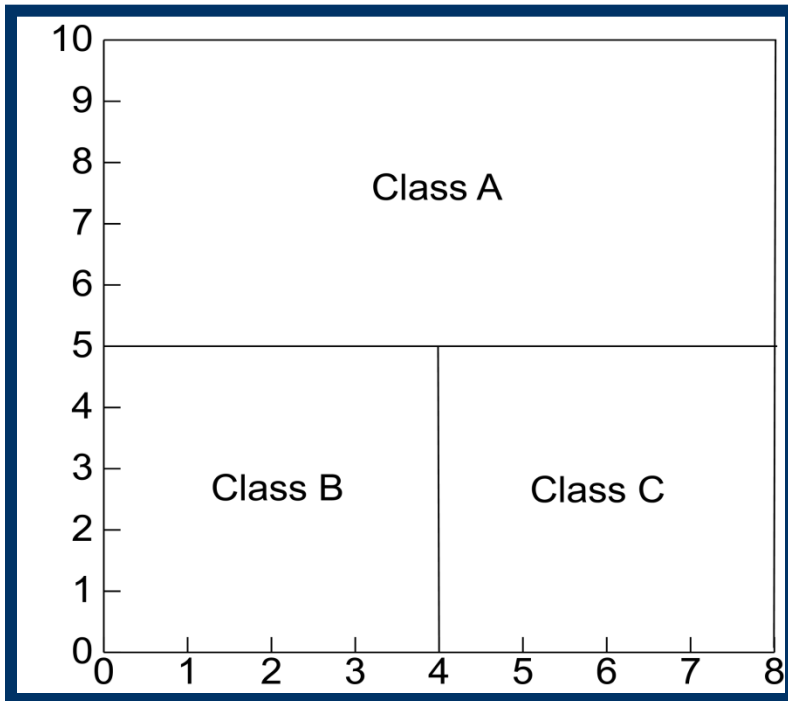
ajustando los parámetros del modelo:

***ajuste evolutivo de funciones de pertenencia para sistemas basados en reglas difusas, evolución de modelos de aprendizaje con parámetros, ...***



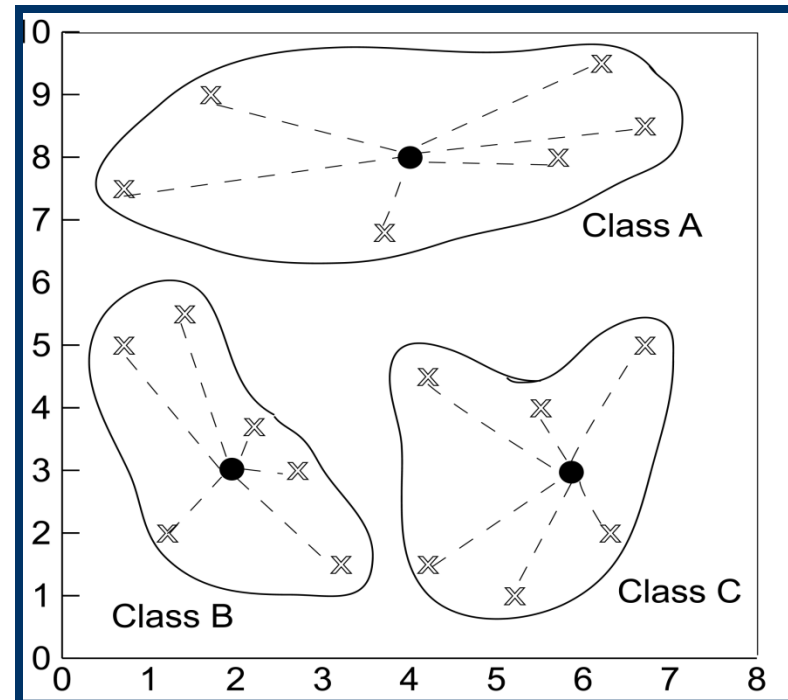
# 4. APRENDIZAJE EVOLUTIVO HÍBRIDO. SELECCIÓN EVOLUTIVA DE INSTANCIAS

## Clases Definidas



Basado en Particiones

Basado en Distancias

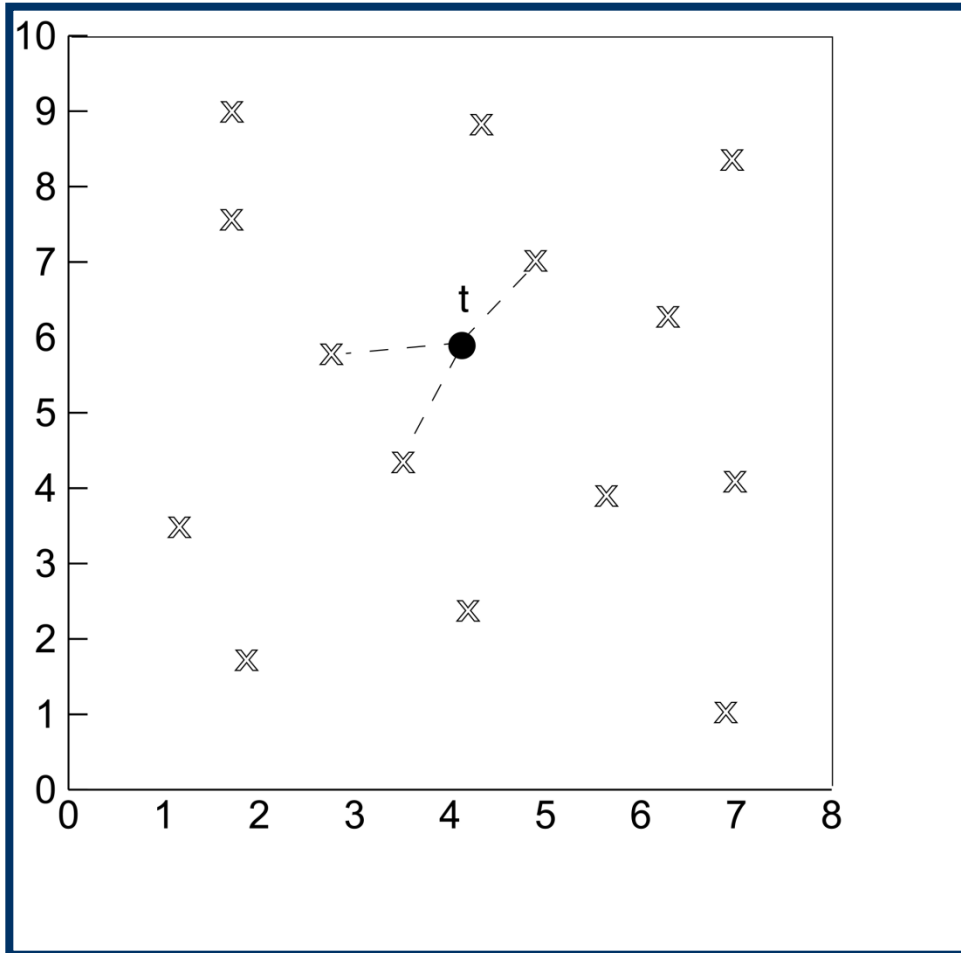


# Ejemplo de Clasificador: k-NN

---

- k-NN (*k vecinos más cercanos*) es uno de los clasificadores más utilizados por su simplicidad.
- El proceso de aprendizaje de este clasificador consiste en almacenar una tabla con los ejemplos disponibles, junto a la clase asociada a cada uno de ellos.
- Ante un nuevo ejemplo a clasificar, se calcula su distancia (usaremos la Euclídea) con respecto a los  $n$  ejemplos existentes en la tabla, y se consideran los  $k$  más cercanos.
- El nuevo ejemplo se clasifica según la clase mayoritaria de los  $k$  ejemplos más cercanos.
- El caso más sencillo es cuando  $k = 1$  (1-NN).

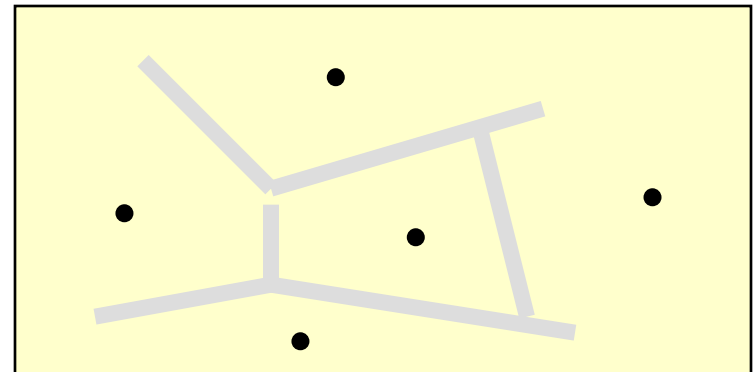
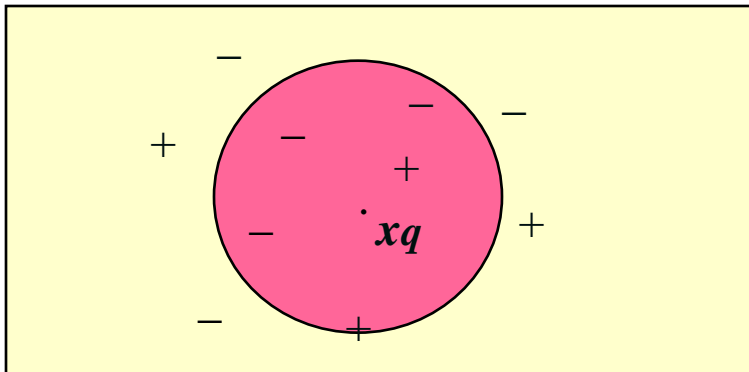
# Ejemplo de Clasificador: k-NN



$k = 3$

# Ejemplo de Clasificador: k-NN

- $k$ -NN devuelve la clase más repetida de entre todos los  $k$  ejemplos de entrenamiento cercanos a  $xq$ .
- Diagrama de Voronoi: superficie de decisión inducida por 1-NN para un conjunto dado de ejemplos de entrenamiento.



# Ejemplo de Clasificador: k-NN

---

- Para diseñar un clasificador, son necesarias dos tareas: *Aprendizaje y Validación*.
- El conjunto de ejemplos se divide en dos subconjuntos:
  - **Entrenamiento:** Utilizado para aprender el clasificador.
  - **Test:** Se usa para validarlo. Se calcula el porcentaje de clasificación sobre los ejemplos de este conjunto (desconocidos en la tarea de aprendizaje) para conocer su poder de generalización.
- Para mayor seguridad, se suele hacer varias particiones aleatorias entrenamiento-test.
- Para cada partición, se diseña un clasificador distinto usando los ejemplos de entrenamiento y se valida con los de test.
- Nosotros usaremos solo una partición.

# Ejemplo de Clasificador: k-NN

---

- **Ejemplo:** Diseño de un Clasificador para *Iris*
  - Problema simple muy conocido: *clasificación de lirios*.
  - Tres clases de lirios: *setosa*, *versicolor* y *virginica*.
  - Cuatro atributos: *longitud* y *anchura* de *pétalo* y *sépalo*, respectivamente.
  - 150 ejemplos, 50 de cada clase.
  - Disponible en <http://www.ics.uci.edu/~mlearn/MLRepository.html>



setosa



versicolor



virginica

# Ejemplo de Clasificador: k-NN

---

## Cálculo del Porcentaje de Entrenamiento en 1-NN:

- En el algoritmo 1-NN, no es posible calcular el porcentaje de acierto sobre el conjunto de entrenamiento de un modo directo.
- Si intentásemos clasificar un ejemplo del conjunto de entrenamiento directamente con el clasificador 1-NN, el ejemplo más cercano sería siempre él mismo, con lo que se obtendría un 100% de acierto.
- Para remediar esto, se debe seguir el procedimiento denominado *dejar uno fuera* (“*leave one out*”).
- Para clasificar cada ejemplo del conjunto de entrenamiento, se busca el ejemplo más cercano **sin considerar a él mismo**.
- Por lo demás, se opera igual: cada vez que la clase devuelta coincida con la clase real del ejemplo, se contabiliza un acierto.
- El porcentaje final de acierto es el número de aciertos entre el número total de ejemplos.

# Ejemplo de Clasificador: k-NN

Dado el siguiente conjunto con 4 instancias, 3 atributos y 2 clases:

$x_1$ : 0.4 0.8 0.2 positiva
$x_2$ : 0.2 0.7 0.9 positiva
$x_3$ : 0.9 0.8 0.9 negativa
$x_4$ : 0.8 0.1 0.0 negativa

Queremos clasificar con 1-NN el siguiente ejemplo:

$y_1$ : 0.7 0.2 0.1

Calculamos la distancia del ejemplo con todos los del conjunto:

$$d(x_1, y_1) = \sqrt{(0.4 - 0.7)^2 + (0.8 - 0.2)^2 + (0.2 - 0.1)^2} = 0.678$$

$$d(x_2, y_1) = \sqrt{(0.2 - 0.7)^2 + (0.7 - 0.2)^2 + (0.9 - 0.1)^2} = 1.068$$

$$d(x_3, y_1) = \sqrt{(0.9 - 0.7)^2 + (0.8 - 0.2)^2 + (0.9 - 0.1)^2} = 1.020$$

$$d(x_4, y_1) = \sqrt{(0.8 - 0.7)^2 + (0.1 - 0.2)^2 + (0.0 - 0.1)^2} = 0.173$$

**Por tanto, el ejemplo se clasificará con respecto a la clase *negativa*.**

**IMPORTANTE:** Los atributos deben estar normalizados [0,1] para no priorizarlos sobre otros.



# Definición del Problema de la Selección de Instancias

---

- En problemas reales de clasificación, los ejemplos relevantes no se conocen a priori.
- **Muchos ejemplos pueden ser redundantes o perjudiciales al concepto:**
  - Ejemplo redundante: No añade nada nuevo sobre el concepto.
  - Ejemplo perjudicial o ruidoso: Ejemplo situado en una zona equivocada.
  - Ejemplo relevante: Es aquel situado en los bordes de decisión.

# Definición del Problema de la Selección de Instancias

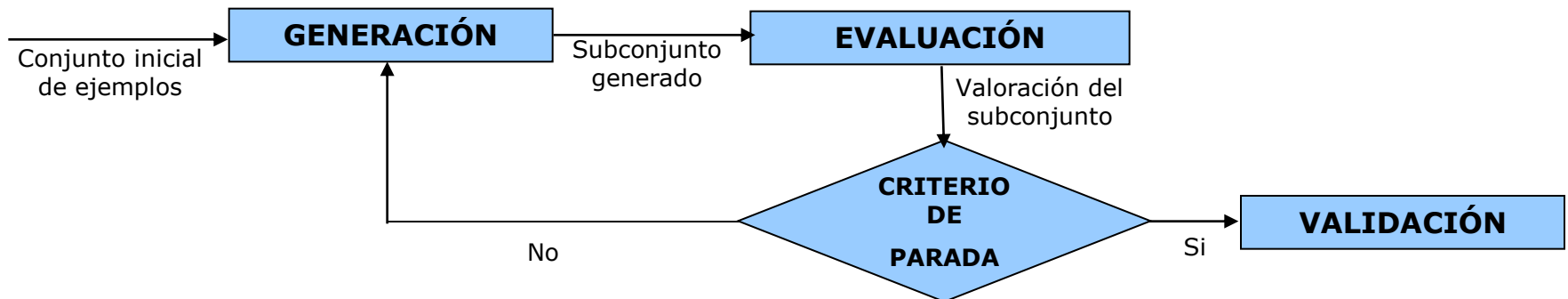
---

- **Problemas de clasificar utilizando todo el conjunto de entrenamiento:** Conjunto de ejemplos excesivamente grande:
  - El aprendizaje del clasificador es más costoso.
  - El rendimiento de dicho clasificador es peor.

*Los métodos de selección de instancias (SI) buscan obtener un subconjunto de ejemplos relevantes para la clasificación, persiguiendo como objetivo la obtención de subconjuntos muy reducidos que igualen o incluso superen el poder de clasificación.*

# Definición del Problema de la Selección de Instancias

- Un método de SI tiene cuatro componentes:
  - Un *procedimiento de generación* para construir el siguiente subconjunto de ejemplos candidato.
  - Una *función de evaluación* para evaluar el subconjunto actual.
  - Un *criterio de parada* para decidir cuando finaliza el algoritmo.
  - Un *procedimiento de validación* para comprobar la validez del subconjunto de ejemplos seleccionado finalmente.



# Definición del Problema de la Selección de Instancias

---

- El esquema de representación de soluciones es el siguiente:

$$S = \begin{array}{|c|c|c|c|c|} \hline s_1 & s_2 & \dots & s_{n-1} & s_n \\ \hline \end{array}$$

- Si asumimos un conjunto de entrenamiento  $T$  con  $n$  instancias, el espacio de búsqueda asociado con la SI de  $T$  se constituye por todos los subconjuntos de  $T$ . Para representarlos, utilizamos un vector  $S$  con  $n$  valores, en donde cada  $s_i$  puede tomar los valores 0 ó 1.
- En el caso de que  $s_i = 1$ , la instancia  $i$  se incluye en el subconjunto seleccionado. Si  $s_i = 0$ , la instancia  $i$  no está en el subconjunto seleccionado.

# Definición del Problema de la Selección de Instancias

---

- Con el esquema de representación anterior,  $S$  constituye un subconjunto de instancias de  $T$ . Para poder evaluar su calidad, definimos una función objetivo que combina dos valores:
  - Tasa de clasificación usando  $S$  ( $tasa\_clas$ ).
  - Porcentaje de reducción de  $S$  con respecto a  $T$  ( $porc\_red$ ).
- Función a maximizar:

$$F(S) = \alpha \cdot tasa\_clas + (1 - \alpha) \cdot porc\_red$$

$\alpha$  es un factor de combinación. Consideraremos un valor de 0,5.

# Definición del Problema de la Selección de Instancias

---

■ *tasa\_clas* mide el porcentaje de instancias correctamente clasificadas pertenecientes a  $T$ . Para clasificarlas usamos 1-NN con “*leave one out*”, considerando como tabla de ejemplos el subconjunto representado por  $S$ .

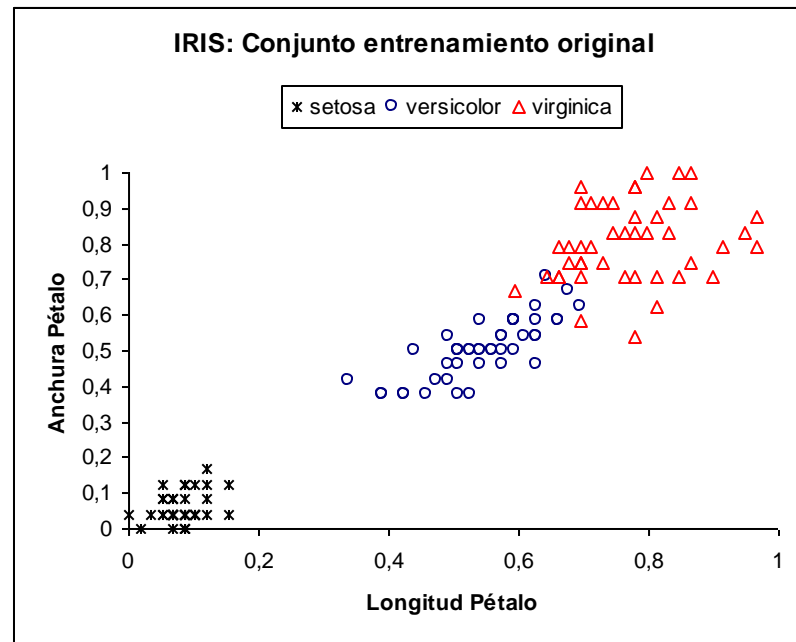
$$tasa\_clas = 100 \cdot \frac{\text{n}^\circ \text{ instancias bien clasificadas de } T}{\text{n}^\circ \text{ instancias en } T}$$

■ *porc\_red* se define como:

$$porc\_red = 100 \cdot \frac{(|T| - |S|)}{|T|}$$

# Definición del Problema de la Selección de Instancias

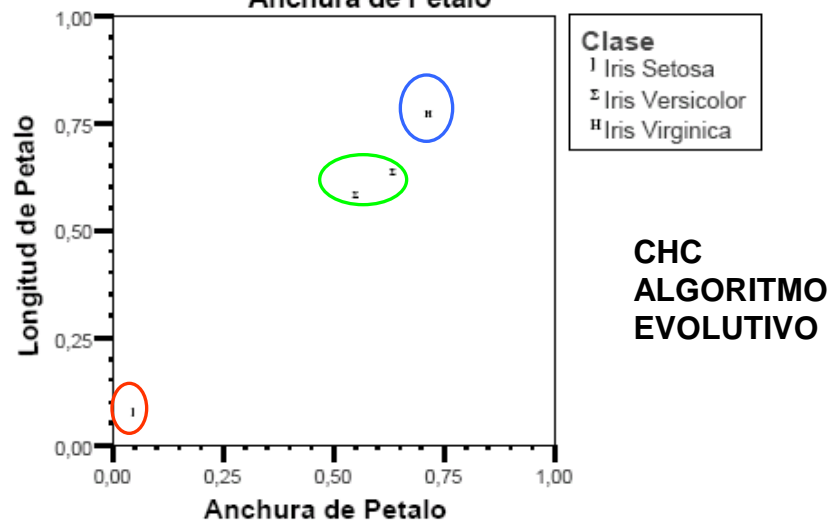
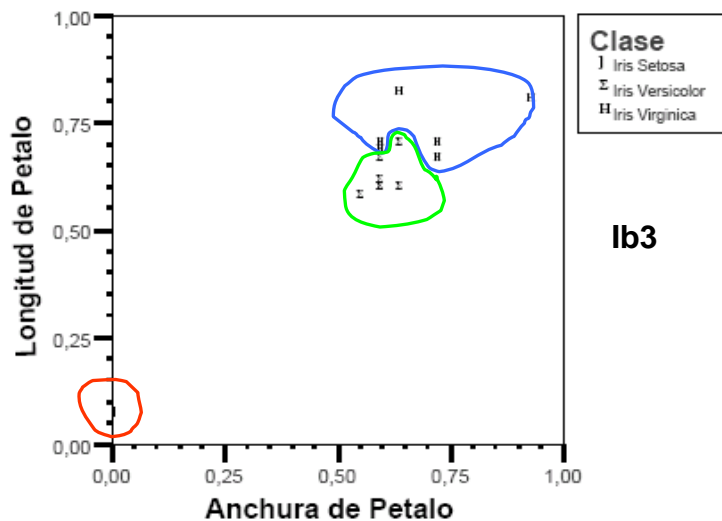
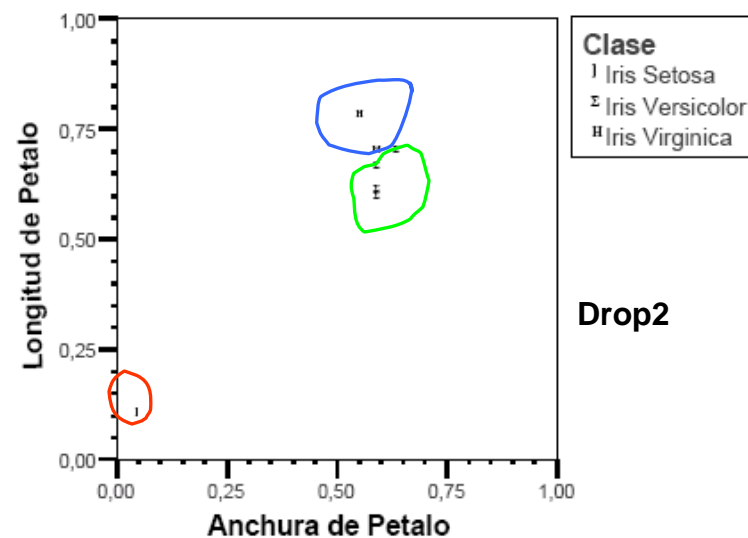
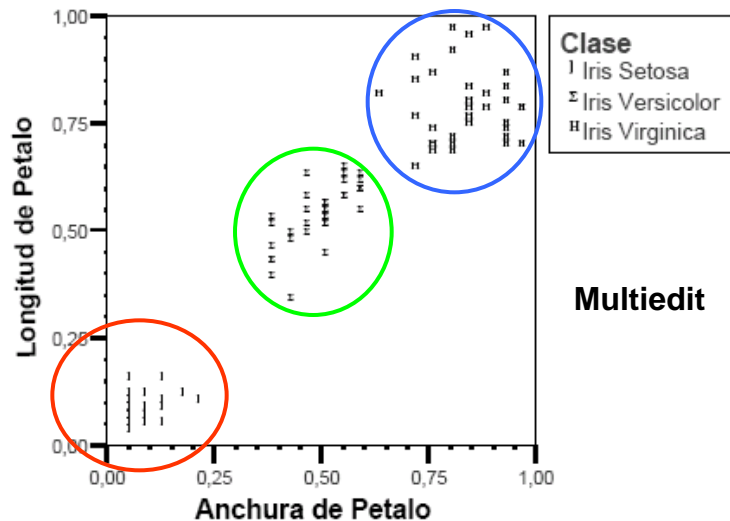
Ejemplos de conjuntos seleccionados sobre *Iris*:



Reducción: 0%. AciertoTest: 95,33%

# Definición del Problema de la Selección de Instancias

## Análisis del Mecanismo de Selección de Prototipos



**Bibliografía:** J.R. Cano, F. Herrera, M. Lozano. Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study. IEEE Trans. on Evolutionary Computation 7:6 (2003) 561-575.



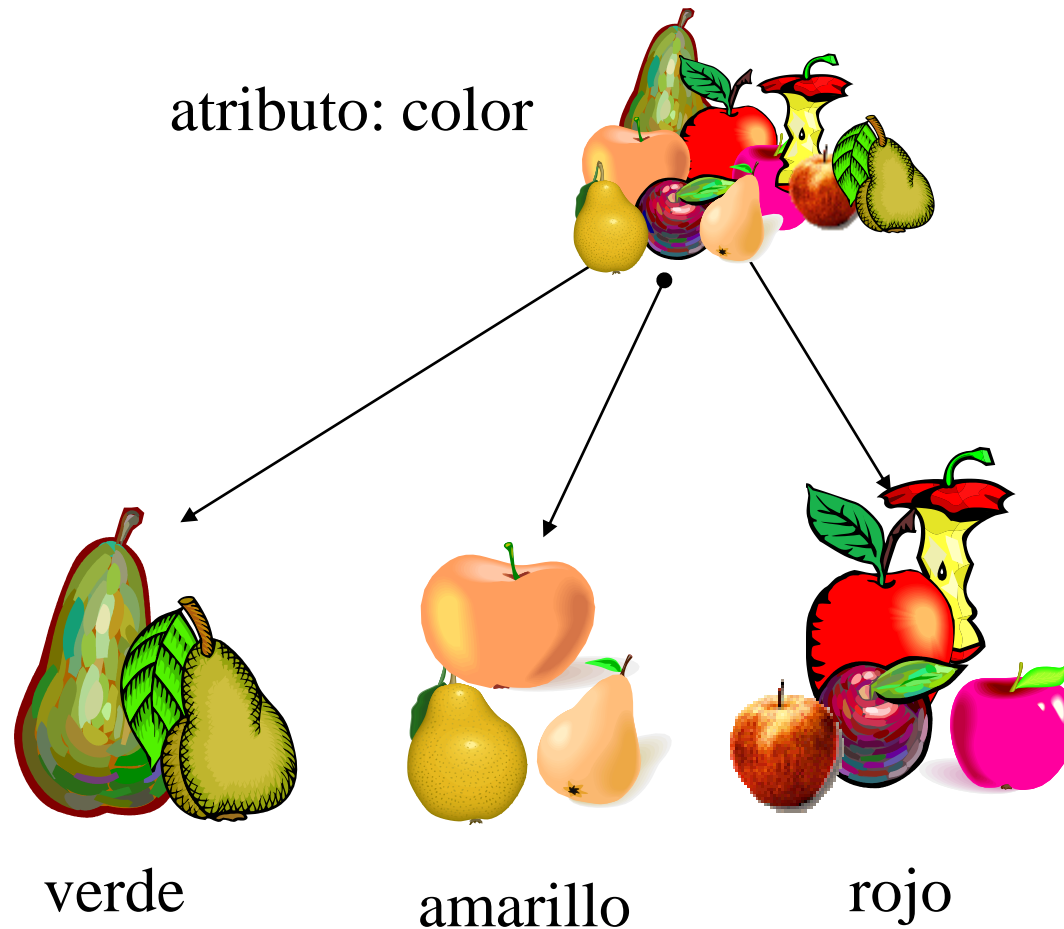
# Selección evolutiva de instancias

## Algunas aproximaciones a la selección evolutiva de prototipos (SCI<sup>2</sup>S-UGR):

---

- J.R. Cano, [F. Herrera](#), [M. Lozano](#), Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study. *IEEE Trans. on Evolutionary Computation* 7:6 (2003) 561-575.**
- J.R. Cano, [F. Herrera](#), [M. Lozano](#), Stratification for Scaling Up Evolutionary Prototype Selection. *Pattern Recognition Letters*, 26, (2005), 953-963**
- J.R. Cano, [F. Herrera](#), [M. Lozano](#), Evolutionary Stratified Training Set Selection for Extracting Classification Rules with Trade-off Precision-Interpretability. *Data and Knowledge Engineering* 60 (2007) 90-108**
- [S. García](#), J.R. Cano, [F. Herrera](#), A Memetic Algorithm for Evolutionary Prototype Selection: A Scaling Up Approach. *Pattern Recognition*, [doi:10.1016/j.patcog.2008.02.006](https://doi.org/10.1016/j.patcog.2008.02.006), 41:8 (2008) 2693-2709**
- [S. García](#), [F. Herrera](#), Evolutionary Under-Sampling for Classification with Imbalanced Data Sets: Proposals and Taxonomy. *Evolutionary Computation*, 17:3 (2009) 275-306.**
- [J. Derrac](#), [S. García](#), [F. Herrera](#), IFS-CoCo: Instance and Feature Selection based on Cooperative Coevolution with Nearest Neighbor Rule. *Pattern Recognition* 43:6 (2010) 2082-2105.**

# Ej. Selección de Instancias y Extracción de Árboles de Decisión



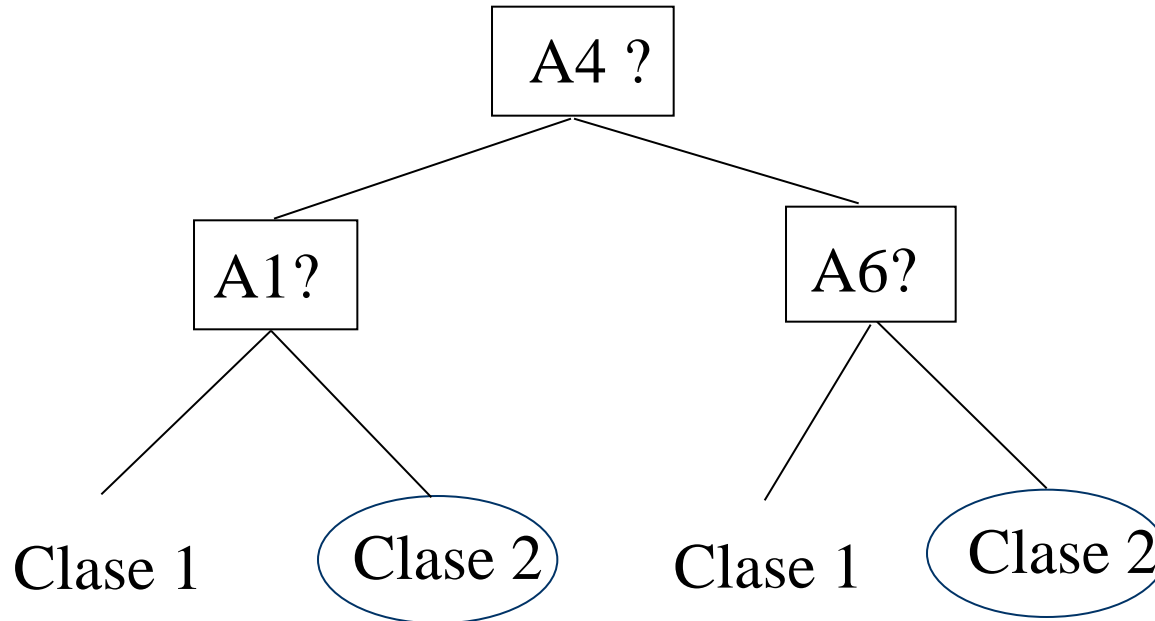
J.R. Cano, [F. Herrera](#), [M. Lozano](#), **Evolutionary Stratified Training Set Selection for Extracting Classification Rules with Trade-off Precision-Interpretability.**  
*Data and Knowledge Engineering* 60 (2007) 90-108.

# Ej. Selección de Instancias y Extracción de Árboles de Decisión

Ejemplo de árbol de decisión

Conjunto inicial de atributos:

{A1, A2, A3, A4, A5, A6}



-----> Conjunto reducido de atributos: {A1, A4, A6}

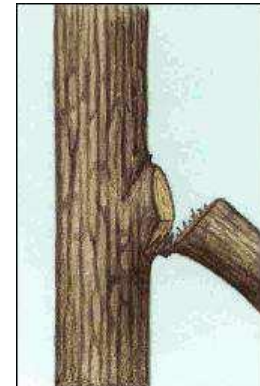
# Ej. Selección de Instancias y Extracción de Árboles de Decisión

## Ej. Selección de Instancias y Extracción de Árboles de Decisión



**Comprehensibilidad:  
Árboles de Tamaño reducido**

**Se utilizan técnicas de poda  
eliminación de nodos**



Las estrategias de selección de instancias permiten construir árboles de decisión para grandes bases de datos reduciendo el tamaño de los árboles.

Aumentan su interpretabilidad.

# Ej. Selección de Instancias y Extracción de Árboles de Decisión

Kdd Cup'99. Número de estratos: 100

	No. Reglas	% Reducción	C4.5	
			%Ac Trn	%Ac Test
<i>C4.5</i>	252		99.97%	99.94%
<i>Cnn Strat</i>	83	81.61%	98.48%	96.43%
<i>Drop1 Strat</i>	3	99.97%	38.63%	34.97%
<i>Drop2 Strat</i>	82	76.66%	81.40%	76.58%
<i>Drop3 Strat</i>	49	56.74%	77.02%	75.38%
<i>Ib2 Strat</i>	48	82.01%	95.81%	95.05%
<i>Ib3 Strat</i>	74	78.92%	99.13%	96.77%
<i>Icf Strat</i>	68	23.62%	99.98%	99.53%
<i>CHC Strat</i>	<b>9</b>	<b>99.68% (1440 instancias)</b>	98.97%	<b>97.53%</b>

# Ej. Selección de Instancias y Extracción de Árboles de Decisión

La selección de instancias nos permite obtener conjuntos de reglas más interpretables y con aporte de mayor información.

	No. Instancias N	No. Variables	No. Reglas		No. Variables/ regla		Confidencia de las Reglas $N(\text{Cond}, \text{Clas})/N$	
			C4.5	IS-CHC/ C4.5	C4.5	IS-CHC/ C4.5	C4.5	IS-CHC/ C4.5
<b>Adult</b> <b>2</b> <b>clases</b>	<b>30132</b>	<b>14</b>	<b>35</b>	<b>5</b>	<b>14</b>	<b>3</b>	<b>0.003</b>	<b>0.167</b>
			<b>9</b>					

## 5. COMENTARIOS FINALES



El uso de los Algoritmos Evolutivos como herramienta para la extracción de conocimiento es un área de investigación muy activa y en la que se está produciendo continuos desarrollos.



## 5. COMENTARIOS FINALES

### Algunos Retos

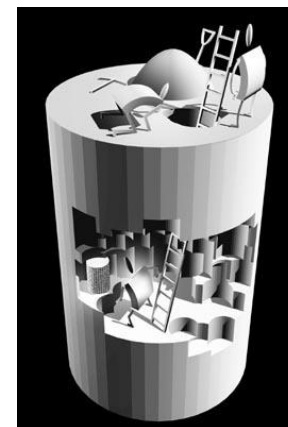
**Escalabilidad de los Algoritmos Evolutivos de Extracción de Conocimiento para su uso en grandes bases de datos (Big Data).**

**Algoritmos Evolutivos Multiobjetivo de Extracción de Conocimiento incluyendo multiples objetivos**





## 5. COMENTARIOS FINALES



PARA MÁS INFORMACIÓN



<http://www.keel.es>



## 5. COMENTARIOS FINALES

---

### **En lo que sigue:**

**Estamos en la década de los datos.**

**Surge como profesión el “Científico de Datos”.**

**Una web sobre el software libre para Inteligencia de Negocio, Ciencia de Datos ...**

**Enlaces para continuar la formación con cursos online y un buen enlace para comenzar a practicar, KAGGEL.**

**Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”.**

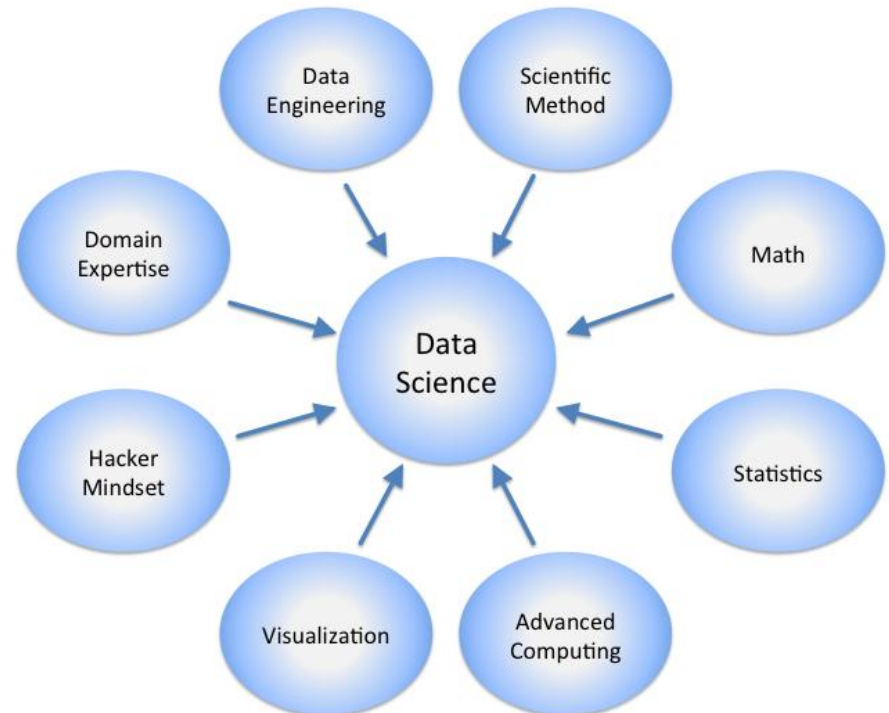
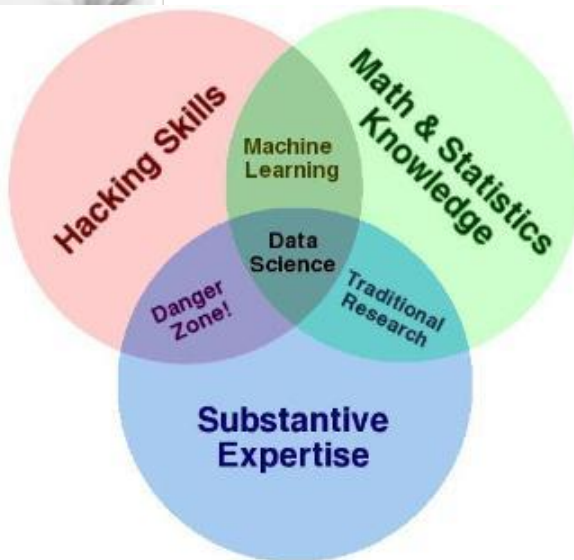


# Comentarios Finales

Surge como profesión el “Científico de Datos”

## Data Science

(Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos, big data, ...)



# Comentarios Finales

Una web sobre el software libre Ciencia de Datos ...

---

## Software (open source tools)



[BLOG](#) [BIG DATA COURSE](#) [ADVICE](#) [STARTUPS](#) [USE CASES](#) [SPEAKER](#) [OPEN SOURCE](#) [PUBLIC DATA](#) [EVENTS](#) [FORUM](#) [ABOUT](#)

---

<http://www.bigdata-startups.com/open-source-tools/>



# Comentarios Finales

Una web sobre el software libre para Ciencia de Datos ...

<http://www.bigdata-startups.com/open-source-tools/>

The image displays a comprehensive grid of logos for open-source big data tools, organized into several categories:

- Data Analysis & Platforms:** Hadoop, PARACCEL, Storm, HPCC Systems, Apache Drill, GridGain, Dremel, Hortonworks, Zettaset, calpont, ORACLE TIMESTEN, HD.
- Databases / Data warehousing:** INFOBRIGHT, Cassandra, HBASE, Hiberi, riak, Infinispan, Bigdata, orientDB, Neo4j, HYPERTABLE, Globals, redis.
- Operational:** Versant JPA, MarkLogic, mobject.
- Multivalued database:** Rocket, U2, REVELATION SOFTWARE, northgate, QM, jBASE INTERNATIONAL.
- Business Intelligence:** talend, JASPERSOFT, Jedox, SpagoBI, Palo, pentaho, BIRT Exchange, ACTUATE, KNIME.
- Data Mining:** RAPID MINER, orange, RAPID ANALYTICS, mahout, WEKA, jHepWork, KEEL, togaware, SPINF.
- Social:** Apache Kafka, ThinkUp, Corona.
- Big Data search:** Lucene, Apache Solr, elasticsearch.
- Data aggregation:** OOOOP, CHUBAZA.
- Graphs:** Gephi, InfiniteGraph, FlockDB, AllegroGraph 4.9, GraphBuilder, Gremlin, INFO GRID, HYPERGROPHDB, dex, meronomy, GraphBase, BrightstarDB.
- Multidimensional:** GT.M, SciDB, rasdaman.
- Key Value:** AEROSPIKE, leveldb, GENIE DB, Chordless, Tokyo Cabinet, Scalaris, SCALIEN, Project Voldemort, hamsterdb, RAPTOR DB, FairCom, STS DB, HyperDex, IQLECT, OpenLDAP, ioremap.net.
- Document Store:** mongoDB, Couchbase, CouchDB, Raven DB, CLUSTERPOINT, RaptorDB, EJDB, dojon DB, JasDB, SchemafreeDB, sisodb, denso db.
- Object databases:** db4objects, ZOPE, NEOPPOD, STARCOUNTER, Magma, Sterling, Picolisp, siaqodb, MORANTEX, HSS Database, EyeDB, RAMER D, NDatabase.
- Multimodel:** ArangoDB, alchemydatabase.
- XML Databases:** existdb, BASE, Qizx, sedna, xindice.
- Grid Solutions:** BIGSPACES, HAZELCAST, Galaxy.

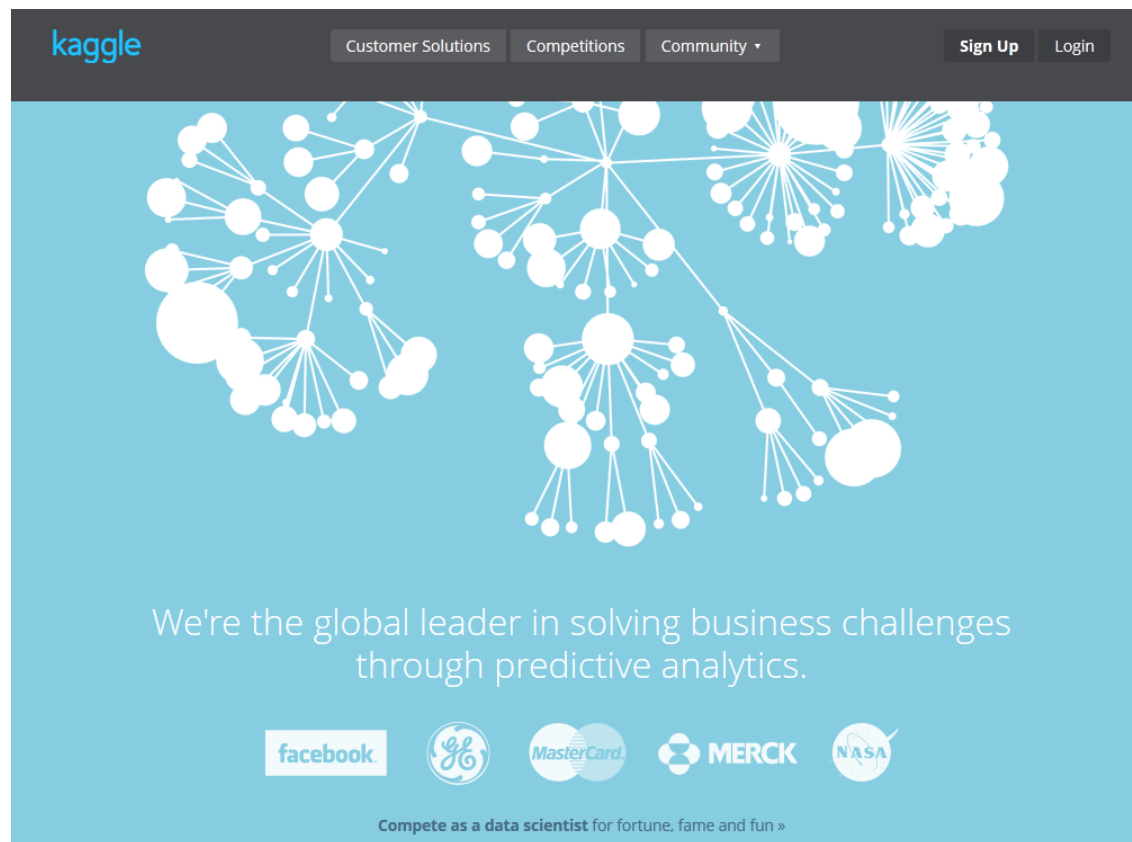
# Comentarios Finales

... y un buen enlace para comenzar a practicar, **KAGGEL**

## Kaggle: Go from Big Data to Big Analytics

<http://www.kaggle.com/>

Es una empresa con un website que ofrece competiciones, ofertas de empleo, ...



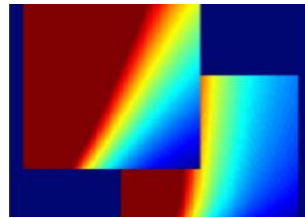
The image shows the Kaggle website homepage. At the top, there is a dark navigation bar with the Kaggle logo on the left and links for 'Customer Solutions', 'Competitions', 'Community', 'Sign Up', and 'Login' on the right. The main content area has a light blue background with a large, abstract network diagram of white nodes and lines. Below the diagram, the text reads: 'We're the global leader in solving business challenges through predictive analytics.' At the bottom, there is a row of logos for 'facebook', 'GE', 'MasterCard', 'MERCK', and 'NASA'. At the very bottom, a small tagline says: 'Compete as a data scientist for fortune, fame and fun »'.

# Comentarios Finales

Enlaces para continuar la formación con cursos online y ...

## *Cursos para continuar la formación*

Learning From Data: a free, introductory Machine Learning online course (MOOC), Caltech <http://work.caltech.edu/telecourse>



**LEARNING FROM DATA**

Machine Learning course - recorded at a live broadcast from Caltech

Coursera BETA



**Stanford University**  
**Machine Learning**

<https://www.coursera.org/course/ml>

**CS109 Data Science**

<http://cs109.org/>



Harvard  
School of Engineering  
and Applied Sciences



# Comentarios Finales

... y un buen enlace para comenzar a practicar, **KAGGEL**

## Kaggle: Go from Big Data to Big Analytics

The screenshot shows the Kaggle website interface. At the top, there is a navigation bar with the Kaggle logo, links for 'Customer Solutions', 'Competitions', and 'Community', and buttons for 'Sign Up' and 'Login'. Below the navigation bar, the breadcrumb trail reads 'All Forums » Data Science Jobs'. A search bar is present on the right. A 'Start Watching' button is located on the left. A pagination bar shows page numbers 1 through 29. The main content is a table of forum posts with columns for 'Replies', 'Views', and 'Last Post'.

431 topics, 536 posts	Replies	Views	Last Post
<b>Feedback on Jobs Board</b> <i>by joycenv, 10 months ago</i>	8	4190	<a href="#">Eyal Kazin</a> 3 days ago
<b>Welcome and How to Post</b> <i>by joycenv, 17 months ago</i>	5	6521	<a href="#">Columbia University</a> 40 days ago
<b>BioCatch is hiring</b> <i>by Lev Kadyshevitch, yesterday</i>	0	85	<a href="#">Lev Kadyshevitch</a> yesterday
<b>Data Scientist, Google [San Francisco Bay Area]</b> <i>by Gaurav Kataria, 8 days ago</i>	4	771	<a href="#">Alessandro Mariani</a> yesterday
<b>Project: Predictive Model Ultimate Fighting Championship Bet...</b> <i>by Julian Macfarlane, yesterday</i>	0	106	<a href="#">Julian Macfarlane</a> yesterday
<b>Sr Analyst, Customer Analytics, Crate and Barrel, Northbrook,...</b> <i>by Crate and Barrel, 3 days ago</i>	0	65	<a href="#">Crate and Barrel</a> 3 days ago
<b>Data Scientist @ Feedzai [Lisbon, Portugal]</b> <i>by Pedro Bizarro, 3 days ago</i>	0	169	<a href="#">Pedro Bizarro</a> 3 days ago

# Comentarios Finales

... y un buen enlace para comenzar a practicar, **KAGGEL**

## Kaggle: Go from Big Data to Big Analytics

Welcome to Kaggle, the leading platform for predictive modeling competitions. Here's how to jump into competing on Kaggle —

New to Data Science? [Visit our Wiki](#) »  
[Learn about hosting a competition](#) »  
[in-Class & Research competitions](#) »



### Enter

Find a competition & download the training data. You don't need new software/skills to submit.



### Build

Build a model using whatever methods you prefer and upload your predictions to Kaggle.



### ...Win!

Kaggle scores your solution in real time and you'll see your place on the live leaderboard.

### Active Competitions

### Active Competitions

#### All Competitions

		<b>Flight Quest 2: Flight Optimization</b> Final Phase of Flight Quest 2	34 days Coming soon \$220,000
		<b>Packing Santa's Sleigh</b> He's making a list, checking it twice; to fill up his sleigh, he needs your advice	6.1 days 338 teams \$10,000
		<b>Flu Forecasting</b>	42 days 37 teams

# Comentarios Finales

... y un buen enlace para comenzar a practicar, **KAGGEL**

## Kaggle: Go from Big Data to Big Analytics




















Playground

Active Competitions

Active Competitions

All Competitions

Getting Started

Getting Started	101		<b>Conway's Reverse Game of Life</b> Reverse the arrow of time in the Game of Life	41 days 89 teams Swag		<b>Flight Quest 2: Flight Optimization</b> Final Phase of Flight Quest 2	34 days Coming soon \$220,000		
			<b>March Machine Learning Madness</b> Predict the 2014 NCAA Basketball Tournament	54 days 39 teams Kudos			<b>Packing Santa's Sleigh</b> He's making a list, checking it twice; to fill up his sleigh, he needs your advice	6.1 days 338 teams \$10,000	
			<b>CIFAR-10 - Object Recognition in Images</b> Identify the subject of 60,000 labeled images	9 months 58 teams Knowledge			<b>Flu Forecasting</b> 	Predict when, where and how strong the flu will be	42 days 37 teams
			<b>Digit Recognizer</b> Classify handwritten digits using the famous MNIST data	11 months 421 teams Knowledge			<b>Galaxy Zoo - The Galaxy Challenge</b> Classify the morphologies of distant galaxies in our Universe	2 months 158 teams \$16,000	
			<b>Titanic: Machine Learning from Disaster</b> Predict survival on the Titanic (with tutorials in Excel, Python and an introduction to Random Forests)	11 months 1201 teams Knowledge			<b>Loan Default Prediction - Imperial College Lon...</b> Constructing an optimal portfolio of loans	53 days 71 teams \$10,000	
			<b>Data Science London + Scikit-learn</b> Scikit-learn is an open-source machine learning library for Python. Give it a try here!	11 months 181 teams Knowledge			<b>Dogs vs. Cats</b> Create an algorithm to distinguish dogs from cats	12 days 165 teams Swag	
			<b>Facial Keypoints Detection</b> Detect the location of keypoints on face images	11 months 58 teams Knowledge					











# Comentarios Finales

... y un buen enlace para comenzar a practicar, [KAGGEL](#)

## Kaggle: Go from Big Data to Big Analytics

### Kaggle Rankings

Sorted by **Rank** (Beta)

<b>1st</b> 743,673 pts  <b>Owen</b> 18 competitions NYC United States	<b>2nd</b> 669,021 pts  <b>José A. Guerrero</b> 21 competitions Spain	<b>3rd</b> 602,178 pts  <b>Leustagos</b> 30 competitions Belo Horizonte Brazil	<b>4th</b> 559,285 pts  <b>2014</b> <b>BreakfastPirate</b> 19 competitions Indianapolis United States	<b>5th</b> 427,927 pts  <b>Naokazu Mizuta</b> 29 competitions Tokyo Japan
<b>6th</b> 421,550 pts  <b>Anil Thomas</b> 20 competitions Silicon Valley	<b>7th</b> 399,836 pts  <b>Haru</b> 5 competitions Kharkov	<b>8th</b> 378,700 pts  <b>Dmitry Efimov</b> 16 competitions Moscow	<b>9th</b> 372,445 pts  <b>Xavier Conort</b> 23 competitions Singapore	<b>10th</b> 366,899 pts  <b>beluga</b> 24 competitions Budapest

Es una muy buena ventana a la resolución de problemas reales y la adquisición de habilidades en Data Science.

# Comentarios Finales

Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”

## Oportunidades en Big Data

(científico de datos es una profesión con creciente demanda)

La demanda de profesionales formados en Ciencia de Datos y *Big Data* es enorme.

Se estima que la conversión de datos en información útil generará un mercado de 132.000 millones de dólares en 2015 y que se crearán más de 4.4 millones de empleos.

España necesitará para 2015 más de 60.000 profesionales con formación en Ciencia de Datos y *Big Data*.



The image shows a screenshot of a news article from El País. The page header includes 'EL PAÍS' and navigation tabs for 'PORTADA', 'INTERNACIONAL', and 'POLÍTICA'. The main section is 'ECONOMÍA', with sub-sections for 'ECONOMÍA', 'EMPRESAS', 'MERCADOS', 'BOLSA', 'FINANZAS PERSONALES', 'VIVIENDA', and 'TECNOLOGÍA'. A secondary navigation bar lists 'ESTÁ PASANDO', 'Multa a la banca', 'Revuelo en Hacienda', 'Eléctricas y renovables', and 'Paro'. The article title is 'El maná de los datos'. A key point is highlighted: 'La conversión de datos en información útil para las empresas generará un mercado de 132.000 millones de dólares en 2015. La herramienta 'big data' sacará del mercado a quien no la use'. The author is 'SUSANA BLÁZQUEZ' from 'Madrid', dated '29 SEP 2013 - 01:00 CET'. A list of companies is provided: 'Archivado en: Citigroup, Cap Gemini Sogeti, SAP, Oracle, ING Bank, BBVA, Mapfre, Bases datos, IBM, Telefónica, Aplicaciones informáticas, Tecnología, Empresas, Programas informáticos, Economía'. The article features a graphic with the word 'SUSANA' in a stylized font, a large '@' symbol, and a globe.

[http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725\\_938376.html](http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725_938376.html)

# Comentarios Finales

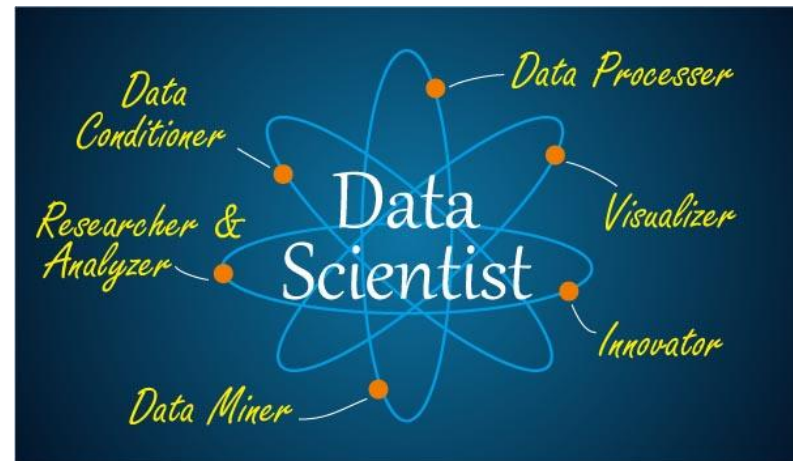
Surge como profesión el “Científico de Datos”

## Científico de Datos

**Oportunidad profesional:** En 2015, Gartner predice que 4,4 millones de empleos serán creados en torno a big data. (Gartner, 2013)

Gartner.

Fuente: <http://www.gartner.com/technology/topics/big-data.jsp>



Surge la figura del Científico de Datos



# Comentarios Finales

Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”

## Oportunidades en Big Data (en España)

[http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde\\_4377072\\_member\\_5811011886832984067#!](http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde_4377072_member_5811011886832984067#!)

España necesitará 60.000 profesionales de Big Data hasta 2015

📅 22 octubre, 2013   📍 Eventos   💬 18



España necesitará 60.000 profesionales de Big Data hasta 2015

“España va a necesitar alrededor de sesenta mil profesionales del Big Data de aquí a 2015”, así lo ha asegurado Francisco Javier Antón, Subdirector General de Tecnologías del Ministerio de Educación, Cultura y Deportes en una mesa redonda sobre beneficio y aplicación de Big Data en pymes, moderada por Daniel Tapias de [Sigma Technologies](#), celebrada durante el 4º Congreso Nacional de CENTAC de

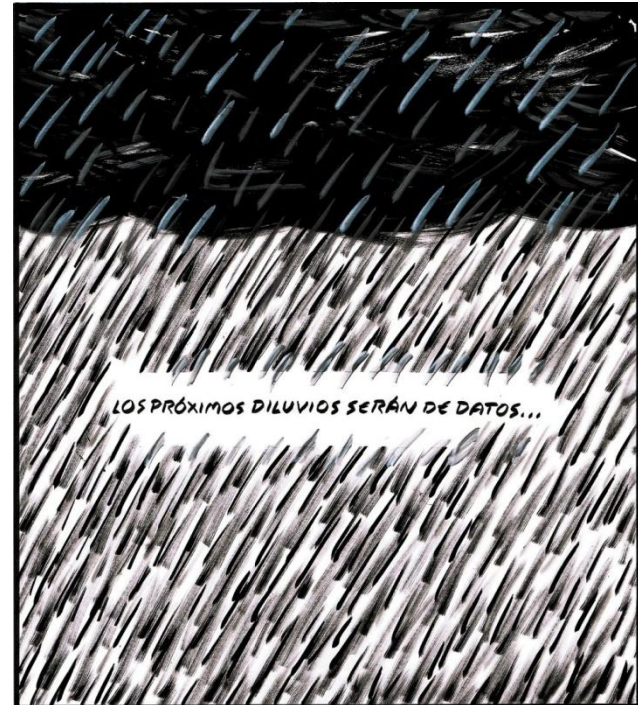
**“Existe una demanda mundial para formar a 4,4 millones de profesionales de la gestión Big Data desde ingenieros, gestores y científicos de datos”, comenta Antón. Sin embargo, “las empresas todavía no ven en el Big Data un modelo de negocio”, lamenta. “Solo se extrae un 1% de los datos disponibles en la red”, añade. “Hace falta formación y concienciación.”**

# Comentarios Finales

[http://elpais.com/elpais/2013/12/02/vinetas/1386011115\\_645213.html](http://elpais.com/elpais/2013/12/02/vinetas/1386011115_645213.html)

## El Roto

Viñeta de El Roto  
3 de diciembre de 2013



**BIG**  
DATA





# BIOINFORMÁTICA

## 2013 - 2014

---

### PARTE I. INTRODUCCIÓN

- Tema 1. Computación Basada en Modelos Naturales

### PARTE II. MODELOS BASADOS EN ADAPTACIÓN SOCIAL (Swarm Intelligence)

- Tema 2. Introducción a los Modelos Basados en Adaptación Social
- Tema 3. Optimización Basada en Colonias de Hormigas
- Tema 4. Optimización Basada en Nubes de Partículas (Particle Swarm)

### PARTE III. COMPUTACIÓN EVOLUTIVA

- Tema 5. Introducción a la Computación Evolutiva
- Tema 6. Algoritmos Genéticos I. Conceptos Básicos
- Tema 7. Algoritmos Genéticos II. Diversidad y Convergencia
- Tema 8. Algoritmos Genéticos III. Múltiples Soluciones en Problemas Multimodales
- Tema 9. Estrategias de Evolución y Programación Evolutiva
- Tema 10. Algoritmos Basados en Evolución Diferencial (Differential Evolution – DE)
- Tema 11. Modelos de Evolución Basados en Estimación de Distribuciones (EDA)
- Tema 12. Algoritmos Evolutivos para Problemas Multiobjetivo
- Tema 13. Programación Genética
- Tema 14. Modelos Evolutivos de Aprendizaje

### PARTE IV. OTROS MODELOS DE COMPUTACIÓN BIOINSPIRADOS

- **Tema 15. Sistemas Inmunológicos Artificiales**
- Tema 16. Otros Modelos de Computación Natural/Bioinspirados