

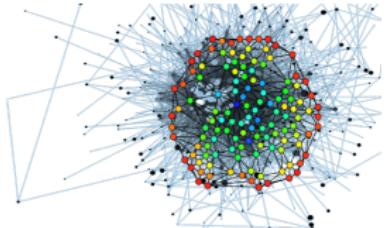
# Seminario: el proceso de ciencia de datos y problemas de clasificación (Parte 1)

Javier Sánchez-Monedero

Departamento de Métodos Cuantitativos  
Universidad Loyola Andalucía  
2 de noviembre de 2017

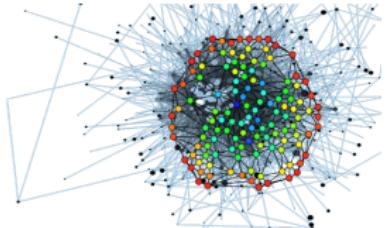


# Índice



**Introducción**  
**Visión general de la ciencia de datos**  
**y etapas**  
**Proceso de Ciencia de Datos**  
**Inteligencia Artificial y Aprendizaje**  
**automático**  
**Conclusiones**

# Índice



## **Introducción**

Visión general de la ciencia de datos  
y etapas

Proceso de Ciencia de Datos

Inteligencia Artificial y Aprendizaje  
automático

Conclusiones

# Aprendizaje automático

## Machine Learning



what society thinks I  
do

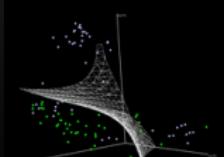


what my friends think  
I do



what my parents think  
I do

$$\begin{aligned} L_w &= \|\mathbf{w}\|^2 - \sum_i a_i j_i(\mathbf{x}_i, \mathbf{w} + b) + \sum_i a_i \\ a_i &\geq 0, \forall i \\ \mathbf{w} &\approx \sum_i a_i j_i(\mathbf{x}_i, \mathbf{w}) \sum_i a_i j_i = 0 \\ \nabla \hat{g}(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t), \\ \theta_{t+1} &= \theta_t - \eta_p \nabla \ell(x_{(t)}, y_{(t)}; \theta_t) - \eta_p \cdot \nabla r(\theta_t), \\ \mathbb{E}_{i(t)}[\ell(x_{(t)}, y_{(t)}; \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t). \end{aligned}$$



```
>>> from sklearn import svm
```

what other programmers  
think I do

what I think I do

what I really do

Aprendizaje automático, IA, aprendizaje profundo, Big Data...

# Aprendizaje automático



# Un mundo de datos I

Nuestro mundo gira cada vez más en torno a los datos:

- **Ciencia:** astronomía, genómica, medio-ambiente...
- **Industria y Energía:** redes de sensores, IoT, gestión parques eólicos, previsión de demanda, ciudades inteligentes...
- **Ciencias sociales y humanidades:** libros digitalizados, documentos históricos, datos sociales...



## Un mundo de datos II

- **Entretenimiento**: sistemas de recomendación, contenidos digitales, búsquedas multimedia...
  - **Medicina**: examen de imágenes médicas, previsión de demanda en hospitales, sistemas expertos...
  - **Finanzas y negocios**: transacciones de mercados automatizadas...

# Explosión de datos I

Aunque hace décadas que existen los analistas de datos, también hace décadas que se almacenan datos que no han podido ser procesados hasta hace pocos años:



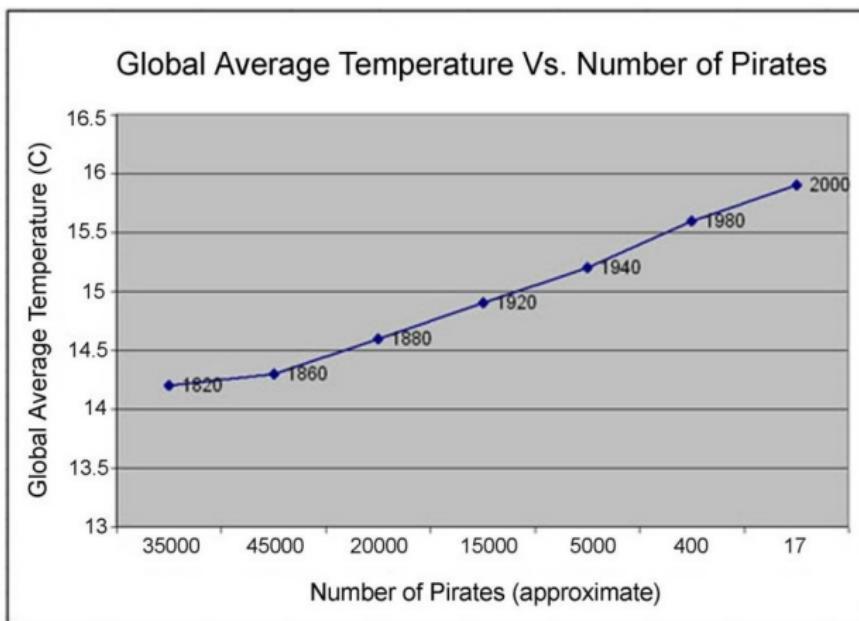
- Tecnologías de bases de datos
  - Coste del hardware de almacenamiento
  - Aumento del ancho de banda
  - Aumento capacidad de procesado
  - Software científico

## Figura: Fuente Big Band Data

## Explosión de datos II

Todo esto nos capacita para pasar de la **información** al **conocimiento**.

# Precaución



Más en <http://www.tylervigen.com/spurious-correlations>

# Precaución



**¡Para frenar el calentamiento global: hagámonos piratas!**  
**Correlación no implica causalidad**

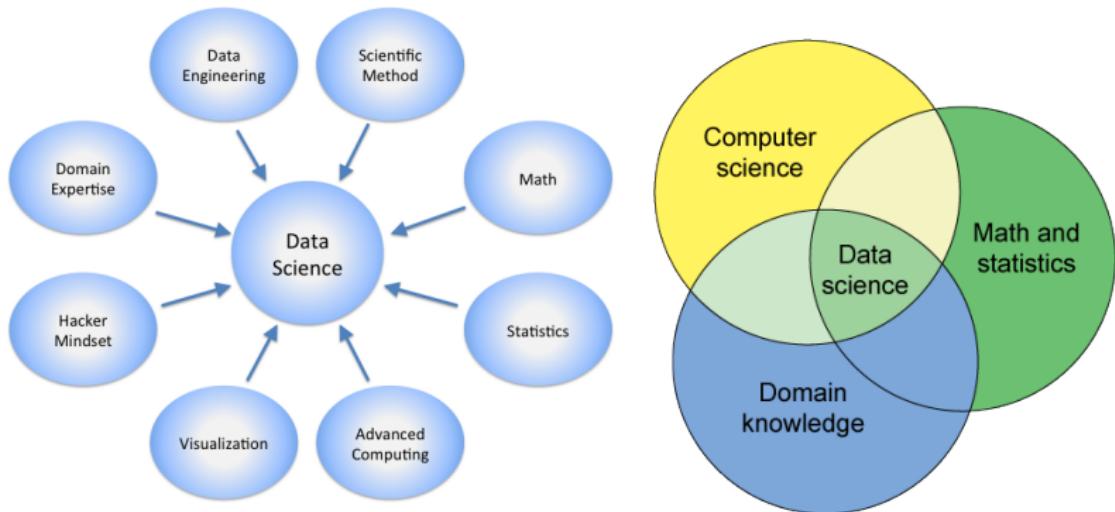
# Objetivos de la sesión

- Pequeña introducción a la ciencia de datos.
- Tipos y técnicas de minería de datos.
- Modelos y algoritmos de clasificación de referencia.

# Definición de ciencia de datos

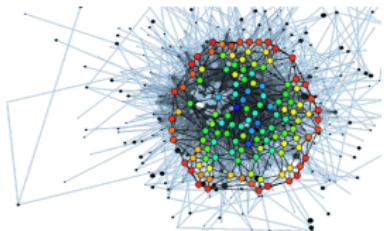
Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos

# Habilidades del científico de datos



**Figura:** Fuentes DreamHost e IBM

# Índice



## Introducción Visión general de la ciencia de datos y etapas

Proceso de Ciencia de Datos  
Inteligencia Artificial y Aprendizaje  
automático  
Conclusiones

# Minería de datos y KDD I

## Minería de datos

“La **Minería de datos (MD)** es el proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos.” [?]

Aunque *Data Science* y *Big Data* son términos más actuales, desde 1989 se denomina a actividades similares como **KDD** (*Knowledge Discovery from Databases*) o **descubrimiento de conocimiento en bases de datos**.

- El KDD es el **proceso completo de extracción de conocimiento** a partir de bases de datos

# Minería de datos y KDD II

- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La **Minería de Datos** es sólo una etapa en el proceso de KDD
- Informalmente se asocia Minería de Datos con KDD

## Aportación del término ciencia de datos

Tal vez el término “ciencia de datos” añada más actividades, como por ejemplo el énfasis en la visualización de datos, o el trabajar con datos no estructurados (algo bastante común en el área del *big data*).

# ¿Para qué?

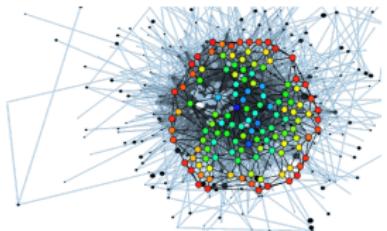
- **Resumir** una gran base de datos
- **Visualizar** datos multi-dimensionales
- **Predecir** valores ⇒ Nos centraremos en este.
- **Explicar** los datos existentes

# Orígenes de datos

Las fuentes de datos son muy variadas, a menudo incluso se mezclan, dando lugar a disciplinas como *fusión de información, extracción de características, preprocesamiento de datos*:

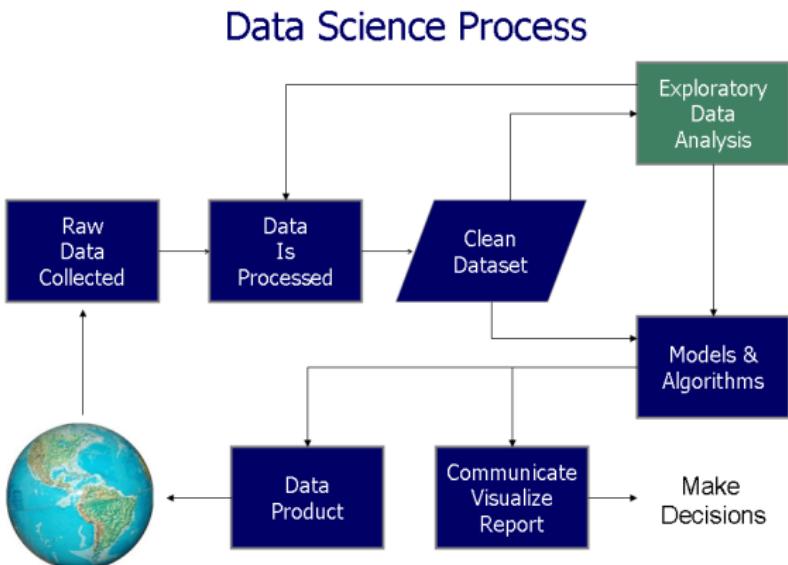
- Bases de datos relacionales
- Bases de datos espaciales y/o temporales: satélites, redes de sensores, telefonía móvil (**Cómo te espía tu centro comercial por WiFi y BlueTooth**)
- Bases de datos de documentos: archivos históricos...
- Bases de datos multimedia: imágenes, vídeos, sonidos...
- La *World Wide Web*
- Grandes volúmenes de datos no estructurados (*Big Data*)

# Índice



Introducción  
Visión general de la ciencia de datos  
y etapas  
**Proceso de Ciencia de Datos**  
Inteligencia Artificial y Aprendizaje  
automático  
Conclusiones

# Etapas en el proceso de KDD I



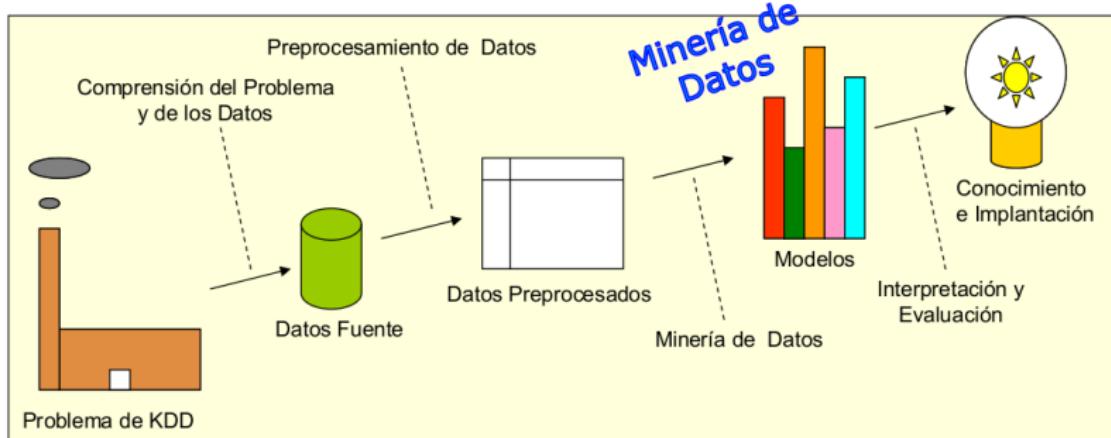
**Figura:** Fuente [https://en.wikipedia.org/wiki/File:Data\\_visualization\\_process\\_v1.png](https://en.wikipedia.org/wiki/File:Data_visualization_process_v1.png)

# Etapas en el proceso de KDD II

Según F. Herrera [?]:

1. **Integración y recopilación:** Comprensión del dominio de aplicación del problema, identificación de conocimiento a priori y creación del **almacén de datos** (*Datawarehouse*)
2. **Preprocesamiento:** Selección de datos, limpieza, reducción y transformación
3. **Selección de la técnica** de MD y aplicación de algoritmos concretos de MD
4. **Evaluación, interpretación y presentación** de los resultados obtenidos
5. **Difusión** y utilización del **nuevo conocimiento**

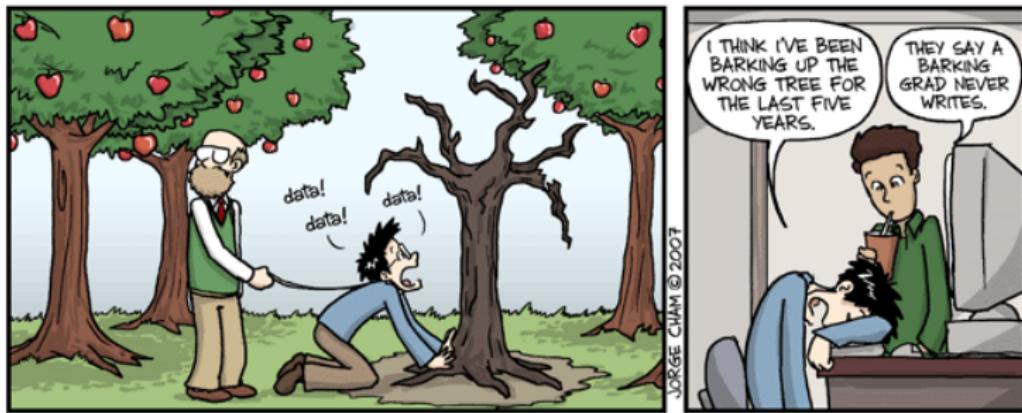
# Etapas en el proceso de KDD III



**Figura:** Etapas en el proceso de KDD, fuente [?]

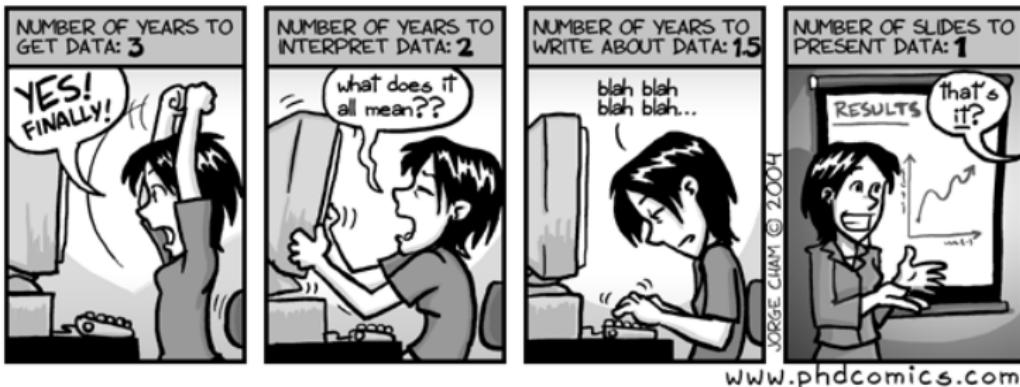
# ¿Qué etapa lleva más esfuerzo?

¿Qué etapa lleva más esfuerzo en el proceso de ciencia de datos?

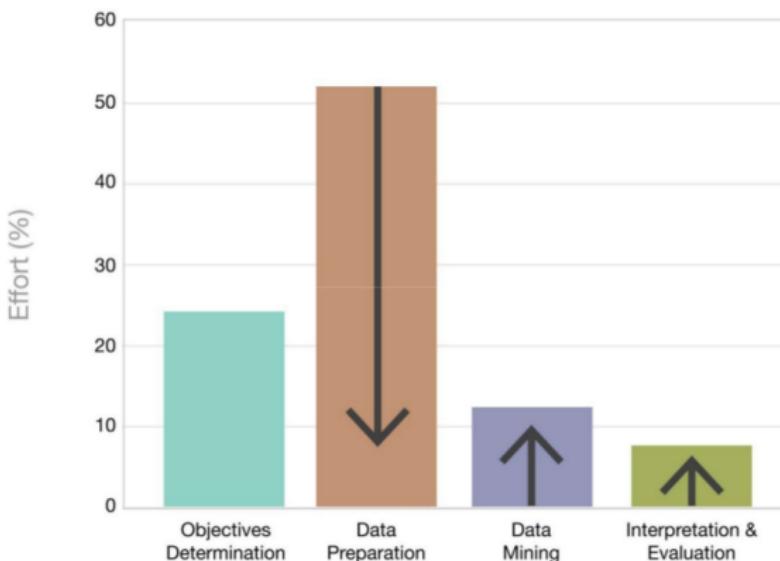


# ¿Qué etapa lleva más esfuerzo?

## DATA: BY THE NUMBERS



# ¿Qué etapa lleva más esfuerzo?

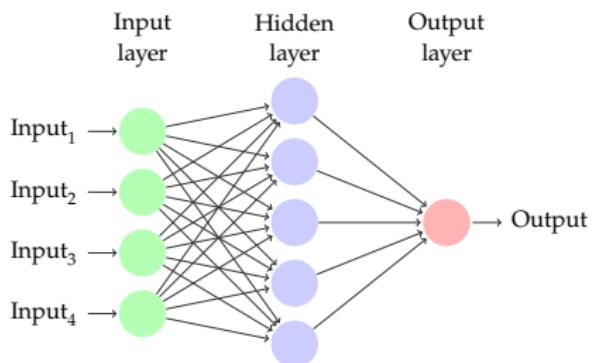


**Figura:** Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos

# Inciso



# Diferencia entre modelo y algoritmo



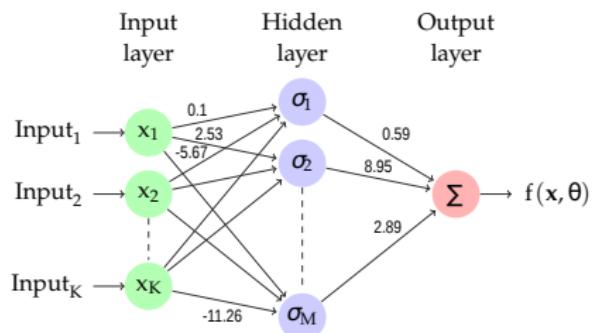
**Figura:** Modelo de red neuronal

Un **modelo** es, en general, una función o una estructura de datos, que puede representar el conocimiento subyacente en un conjunto de datos. Dependiendo del objetivo del problema, tipo de variables de entrada, tipo de salida esperada, complejidad de los datos, etc. habrá modelos más o menos apropiados.

# Diferencia entre modelo y algoritmo

- Un **algoritmo** es una secuencia de pasos con un fin. En informática todos los programas se descomponen en múltiples algoritmos que interaccionan entre si.
- En el contexto de aprendizaje automático, un **algoritmo de aprendizaje** se encarga de que **el modelo aprenda de los datos**, o expresado de otra forma, de que **el modelo de ajuste a los datos**.
- Existen algoritmos correspondientes al campo de la optimización numérica, pero también otros se basan en el aprendizaje estadístico o la inteligencia computacional.

# Modelo entrenado

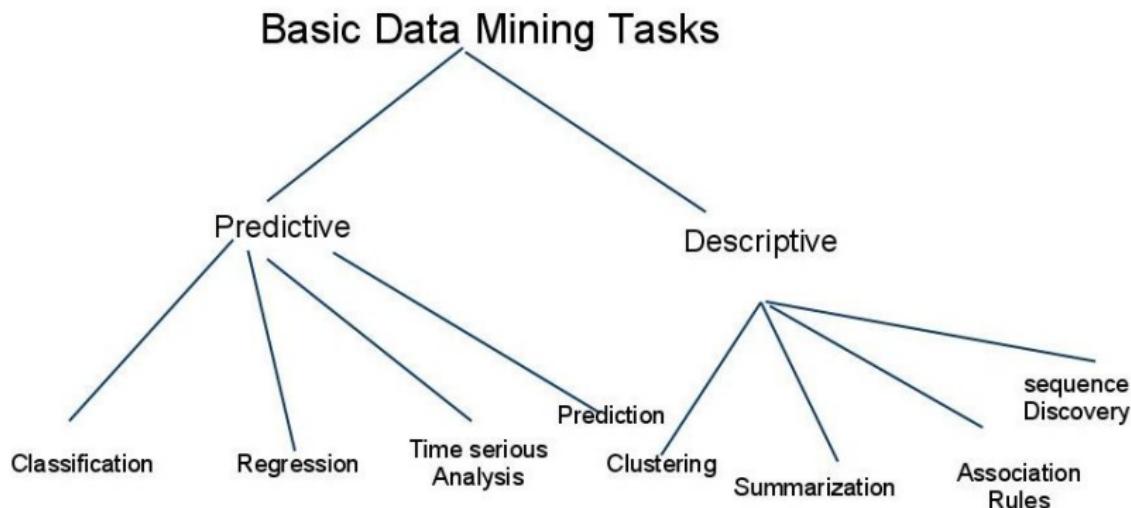


**Figura:** Modelo de red neuronal entrenado

## Ejemplo de red neuronal

En el caso de una red neuronal artificial (RNA), entrenar un modelo consiste en asignar pesos a las conexiones de la red que minimicen una función de error, por ejemplo el error cuadrático medio. El modelo resultante depende de los datos de entrada, los parámetros del modelo (función de base, número de capas y neuronas por capa...) y el algoritmo de aprendizaje.

# Técnicas de Minería de Datos

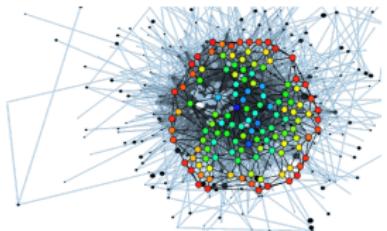


# Conocer bien las herramientas



m-IMAGENES.es

# Índice



**Introducción**  
**Visión general de la ciencia de datos**  
**y etapas**  
**Proceso de Ciencia de Datos**  
**Inteligencia Artificial y Aprendizaje**  
**automático**  
**Conclusiones**

# ¿Cómo extraer conocimiento?

La Ciencia de Datos trata de extraer conocimiento de los datos mediante:

- Técnicas estadísticas *clásicas*
- Inteligencia Artificial y Aprendizaje automático

Muchos métodos de aprendizaje automático se apoyan en **métodos de optimización** matemática y **técnicas estadísticas**, sin embargo a menudo se combinan con técnicas de inteligencia artificial para superar las limitaciones de los primeros en cuanto al entrenamiento de modelos, pero también para diseñar soluciones a problemas, crear sistemas expertos, etc.

# ¿Por qué IA?

En este nuevo milenio:

- La ciencia y la tecnología están cambiando rápido.
- Se tiene relativamente bastante conocimiento de distintos campos de la ciencia más tradicionales (p. ej. física).
- Los computadores están extendidos por todo el mundo.

Grandes retos de la ciencia y la tecnología:

- **Comprender el cerebro** (razonamiento, conocimiento, creatividad).
- **Crear máquinas inteligentes**: ¿Es esto posible? ¿Cuáles son los retos tecnológicos y filosóficos?
- IA presenta las preguntas y retos más interesantes de la informática en la actualidad.



# ¿Qué es la IA?

## Definición (controversia sobre sus límites)

**Desarrollo de métodos y algoritmos que permitan comportarse a las computadoras de modo inteligente.**

**Premisa:** Los procesos cerebrales pueden ser analizados, a un nivel de abstracción dado, como procesos computacionales de algún tipo. → **Propósito:** hacer computacional el conocimiento humano por procedimientos simbólicos o conexionistas.



# Historia de la IA

**Comienzo difuso** (no hay un sólo padre de la IA):

- Alan Turing (1950): *Propongo que se considere la siguiente pregunta, '¿Pueden pensar las máquinas?' o '¿Existirán computadoras digitales imaginables que tengan un buen desempeño en el juego de imitación?.*



- Acuñamiento del término IA (1956).
- 1955-1965: Primeros programas.
- 1966-1973: Primeras limitaciones (casi desaparece este área).
- 1969-1985: Auge de los sistemas expertos.
- 1986: El aprendizaje automático vuelve a estar de moda (ANNs).

# Hitos de la IA

- Deep Blue **venció al campeón mundial de ajedrez Garry Kasparov en 1997.**
- Un programa de IA **resolvió la conjetura matemática de Robbins** sin solución por décadas (1997).



# Hitos de la IA

- NASA usa un **sistema de planificación autónoma** para controlar las operaciones de una aeronave.
- Proverb **resuelve crucigramas** mejor que la mayoría de personas (1999).
- **Conducción autónoma:** DARPA (2003-2007).
- **Software de reconocimiento facial** en cámaras (2006).



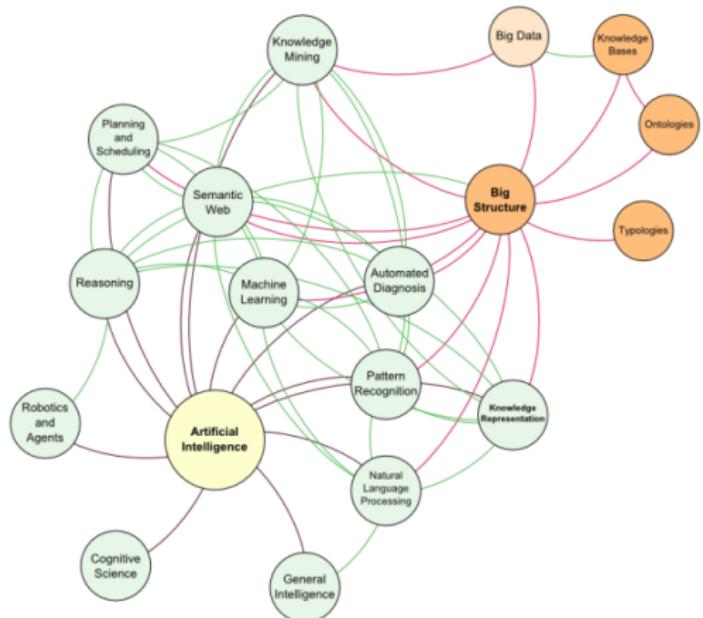
# ¿Qué se necesita?

## Conducción autónoma

Visión por computador, detección de obstáculos, análisis de señales de tráfico, mecanismo de control del vehículo, planificación de rutas, evaluación del terreno, etc.

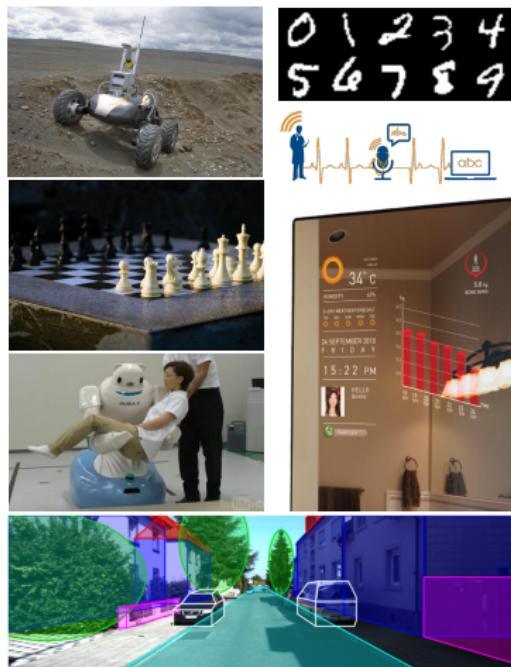
## Proyecto Proverb

Procesamiento del lenguaje natural, conocimiento extenso del lenguaje, la historia y la cultura popular, búsqueda de soluciones posibles.



# Aplicaciones de la IA

- Navegación autónoma
- Tecnologías asistivas
- Detección de objetos
- Reconocimiento de escritura/habla
- Planificación estratégica
- Inteligencia ambiental
- Sistemas de recomendación
- Medicina
- Diseño industrial



# ¿Cuándo usar IA?

Son tareas de gran impacto social, diversas y complejas.

- **No existe una solución analítica o algorítmica conocida.**
- Cuando existan demasiadas posibilidades que hagan difícil el cómputo y podamos usar **heurísticas para reducirlo**.
- Cuando es **difícil el tratamiento de la información** y posiblemente sea incompleta o imprecisa.



CommitStrip.com

# Aprendizaje automático

## Machine learning

El *aprendizaje automático* o *aprendizaje máquina* (*machine learning* en inglés) se define como “campo de estudio que proporciona a los ordenadores la capacidad de aprender sin haber sido explícitamente programados”.

El aprendizaje automático equivale a “aprender de los datos” con el fin de extraer el conocimiento necesario según diferentes propósitos

Este “**aprender de los datos**” hace que el aprendizaje automático se sitúe entre diferentes ramas que pertenecen a la inteligencia artificial, la estadística y las matemáticas

# Aprendizaje automático

Área de estudio que confiere a los ordenadores (máquinas) la habilidad de aprender sin haber sido específicamente programadas para la tarea en cuestión



**Inteligencia Artificial**

Minería de datos

**Estadística / Matemáticas**

**Aprendizaje Automático**

Visión Artificial

Robótica

# Aprendizaje automático

- El aprendizaje automático como herramienta empresarial para examinar grandes repositorios de datos de Big Data.

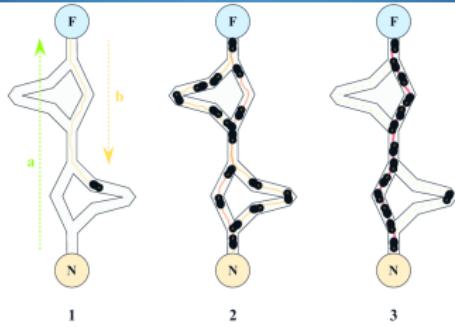
## Objetivo

Ayudar en la **toma de decisiones** descubriendo **patrones ocultos, correlaciones desconocidas, predicciones** y otra información **útil** ⇒ **ventajas competitivas** para las empresas que lo posean.

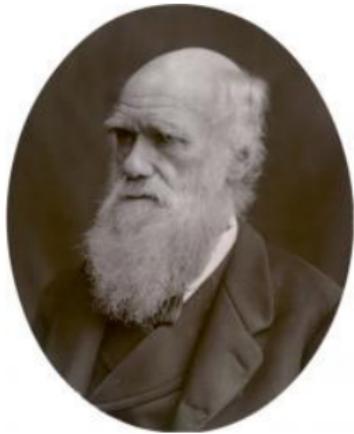
- “Algunos analistas confirman que las empresas que adopten técnicas de analítica de Big Data tendrán **una ventaja competitiva de 20 % en todas las métricas financieras** sobre sus competidores”  
Gustavo Tamaki (2012 “La hora del Big Data”).

# Aprender de la naturaleza

Algoritmos bioinspirados:



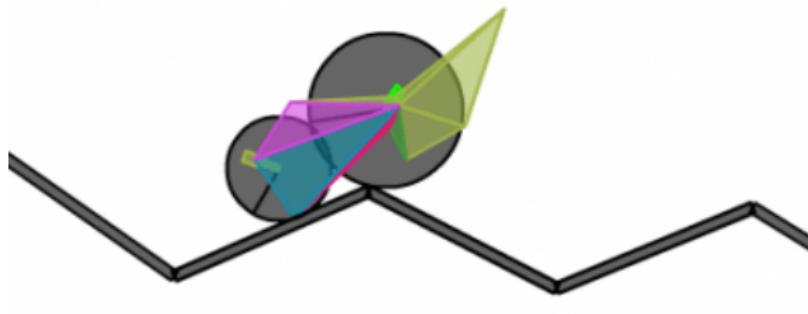
# Algoritmos evolutivos



- Son algoritmos de búsqueda **no deterministas**, que incorporan la semántica de la **evolución natural** a los procesos de optimización.
  - ▶ Darwin: Las especies **evolucionan** de acuerdo al medio y sobreviven **los mejor adaptados**.
  - ▶ Individuos de cada especie → Soluciones al problema.
  - ▶ Diseño de operadores de mutación y cruce de individuos.

# Algoritmos evolutivos

La web BoxCar 2D (<http://www.boxcar2d.com/>) permite diseñar un coche automáticamente utilizando un algoritmo genético.

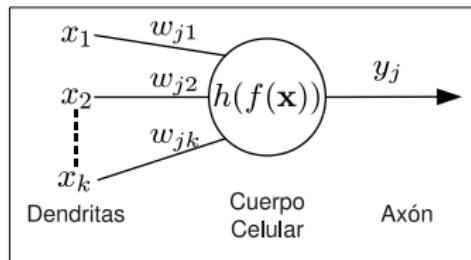
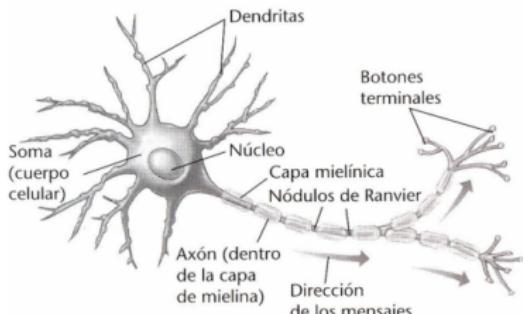


Score: 189.3

Time: 3:40

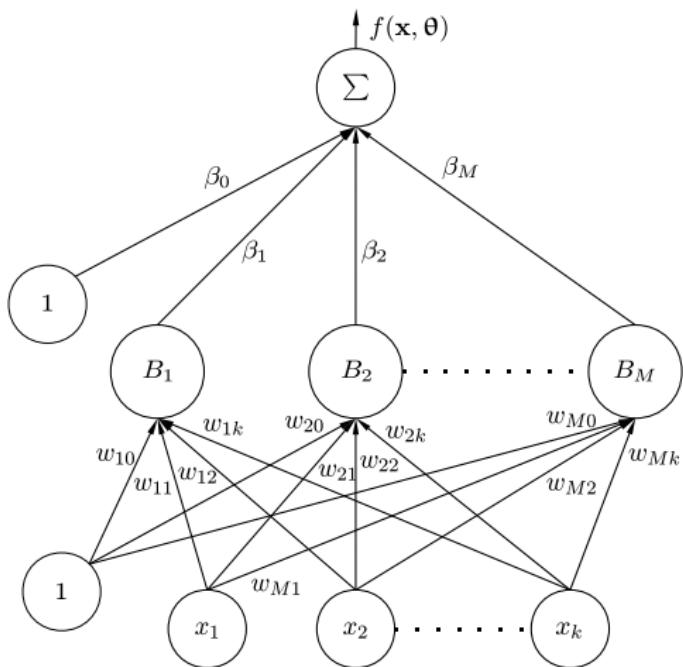
# Concepto de RNA

- Técnica de modelado fundamentada en la **emulación de los sistemas nerviosos biológicos**.



- Combina una gran cantidad de elementos simples de procesado (**neuronas**), altamente interconectados y agrupados en **capas**.
- Una red neuronal es una **relación funcional matemática** entre unas variables de entrada y unas variables de salida.

# RNA para regresión



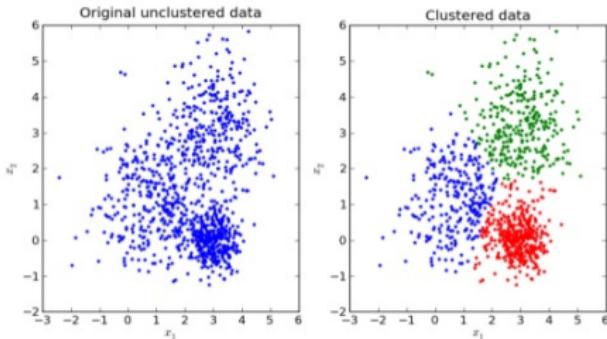
Para 1 variable de respuesta,

$$f(\mathbf{x}, \theta) = \beta_0 + \sum_{j=1}^M \beta_j B_j(\mathbf{x}, \mathbf{w}_j)$$

# Descripción: agrupamiento

## K-means Clustering

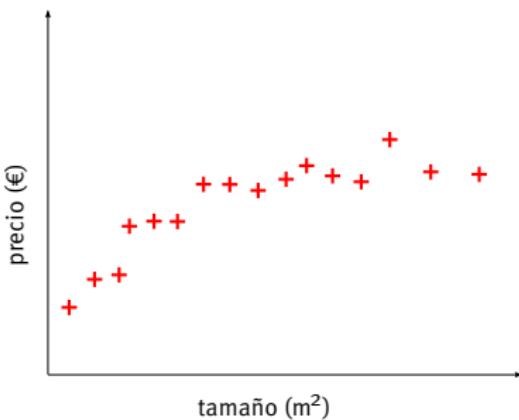
- partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- [http://en.wikipedia.org/wiki/K-means\\_clustering](http://en.wikipedia.org/wiki/K-means_clustering)



<http://pypr.sourceforge.net/kmeans.html>

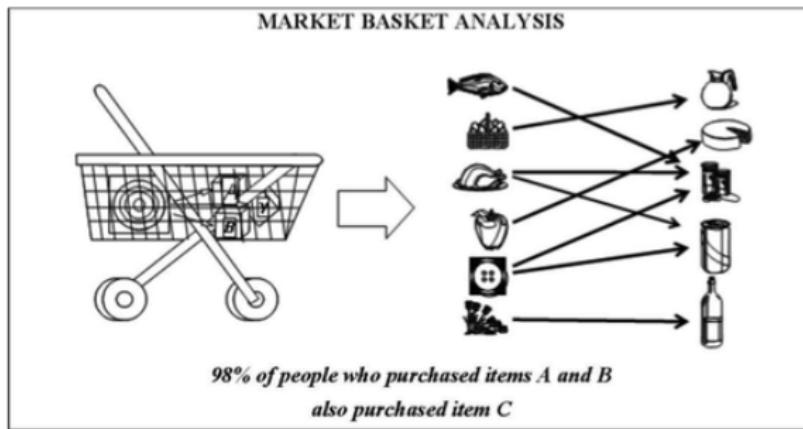
**Figura:** Ejemplo de aprendizaje no-supervisado con el algoritmo de agrupamiento k-means

# Predicción: regresión



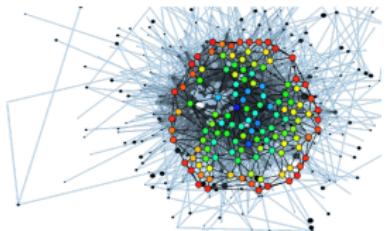
**Figura:** Ejemplo de problema de aprendizaje supervisado de regresión:  
Dados estos datos, un amigo tiene una casa de 75 metros cuadrados,  
¿por cuánto podría esperar venderla?

# Predicción: reglas de asociación



**Figura:** ¿Qué productos suelen ir juntos en las cestas de la compra? ¿Qué probabilidad hay de que una persona que compre el producto A compre el producto B?

# Índice



**Introducción**  
**Visión general de la ciencia de datos**  
**y etapas**  
**Proceso de Ciencia de Datos**  
**Inteligencia Artificial y Aprendizaje**  
**automático**  
**Conclusiones**

# Conclusiones ciencia de datos

- Introducción a la ciencia de datos, minería de datos y aprendizaje automático.
- Vista general de técnicas de minería de datos.
- Destacamos la importancia de ajustar propiamente los parámetros de los modelos y algoritmos.

# Conclusiones clasificación I

Hemos identificado una serie de objetivos contrapuestos que se cumplen por lo general:

- Simplicidad del modelo vs precisión
- Interpretabilidad del modelo vs precisión
- Escalabilidad del modelo vs precisión
- Velocidad del modelo vs precisión

La **eficiencia** puede condicionar la aplicabilidad del algoritmo:

- En 2009 el algoritmo ganador del premio de 1.000.000 \$ Netflix, no se implementó nunca debido al coste computacional.

# ¿Por qué aprender a programar?

- Preprocesamiento relacionado con la naturaleza del problema.
- Conversión de ficheros.
- Creación de nuevos modelos.
- Proporcionar métricas de validación no disponibles en el software utilizado.
- Automatizar experimentos y generación de informes.

¿Preguntas?  
¡Gracias!

