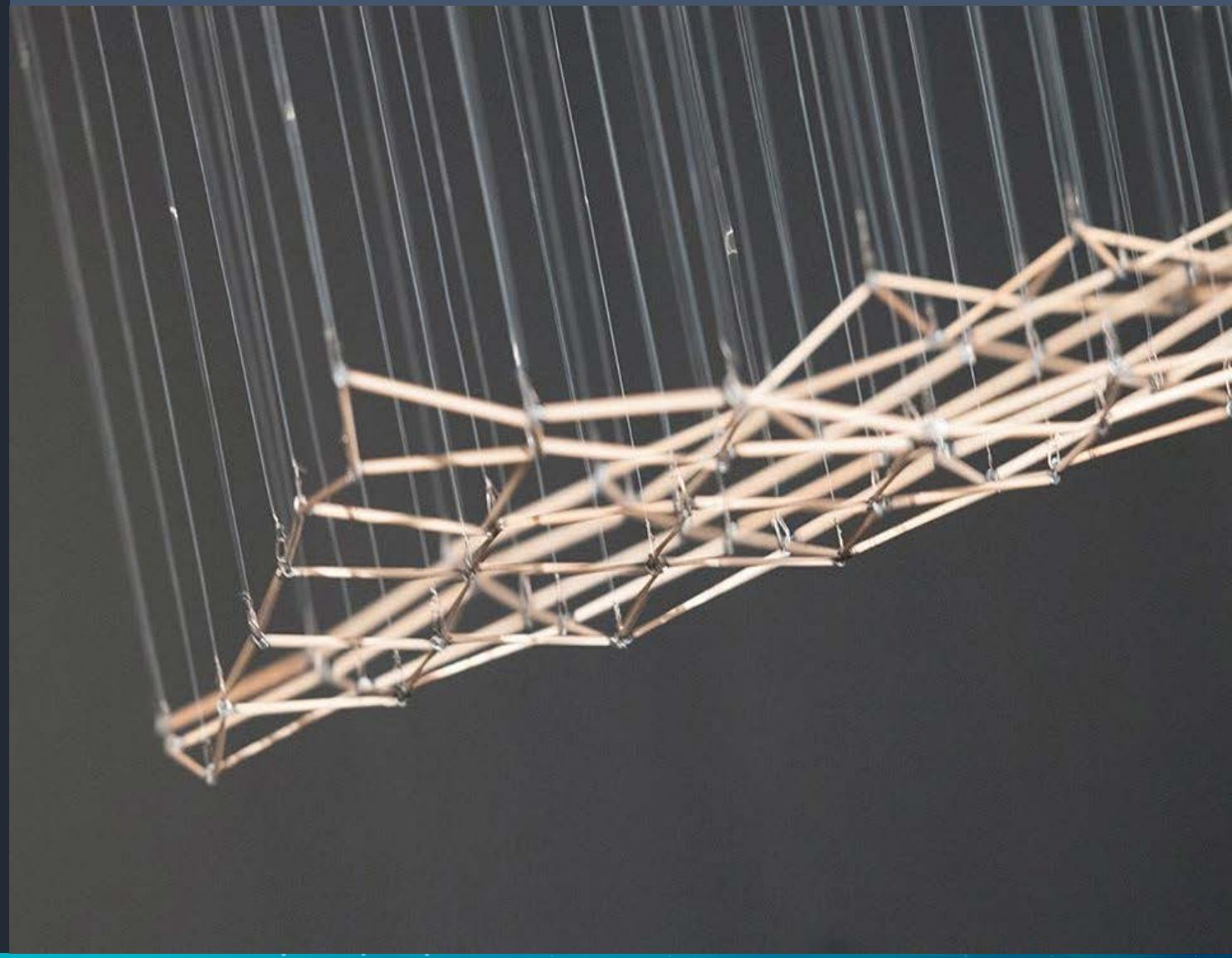
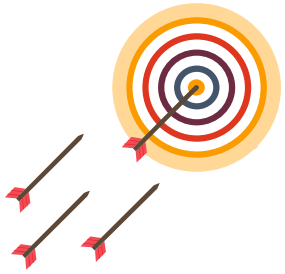
Sistema de Gestión de Datos y de la Información

Enrique Martín Martín
Rafael Caballero Roldán

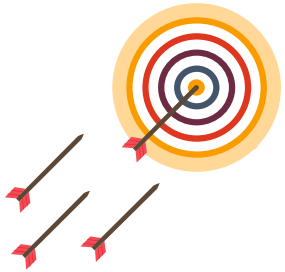
Introducción





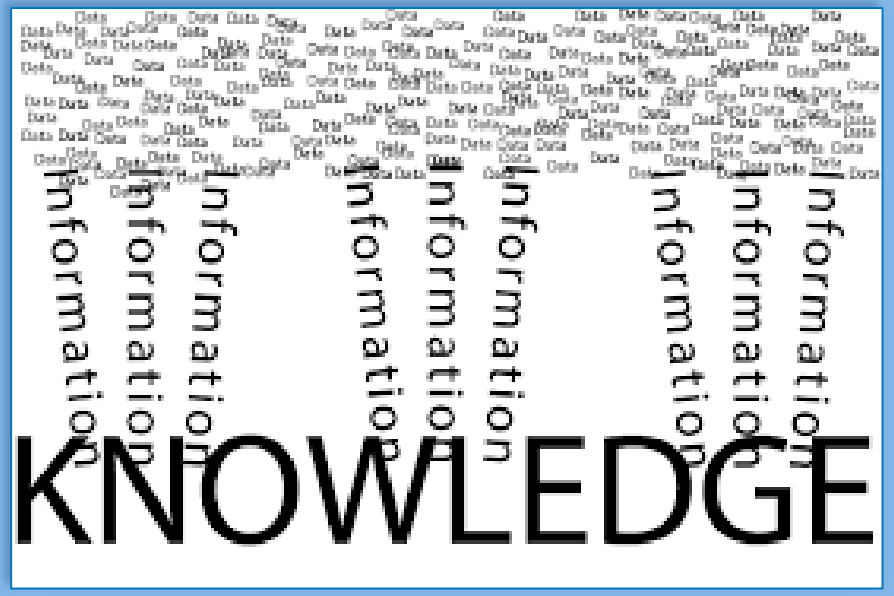
Objetivo del curso

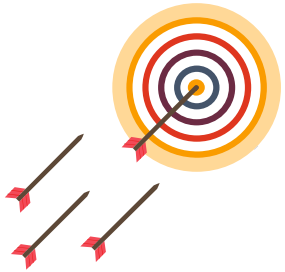
Lo que pretendemos lograr



Objetivo del curso (II)

Un punto de vista más ambicioso





Objetivo del curso (III)

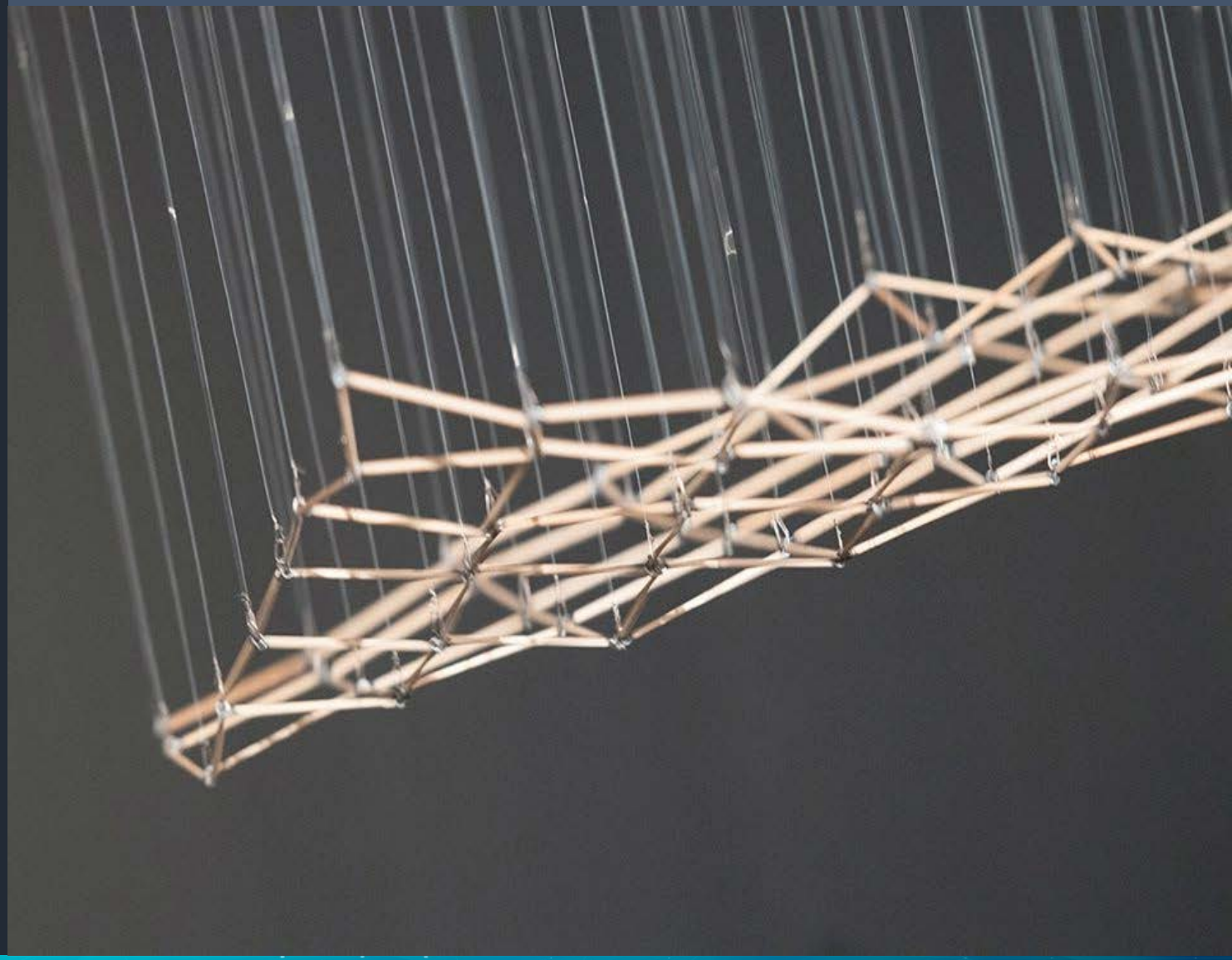
Pirámide DIKW: ¡Tampoco hay que pasarse!

1 Introducción

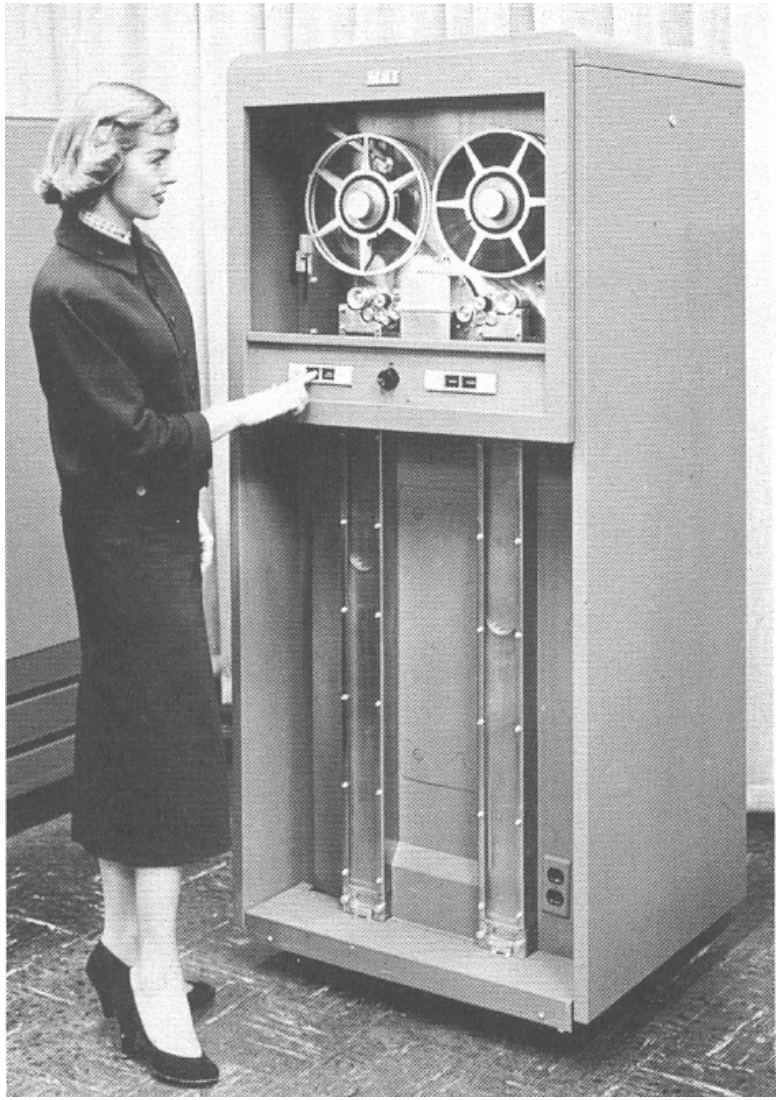
2 Definición de Big Data

3 Análisis científico de datos

Introducción

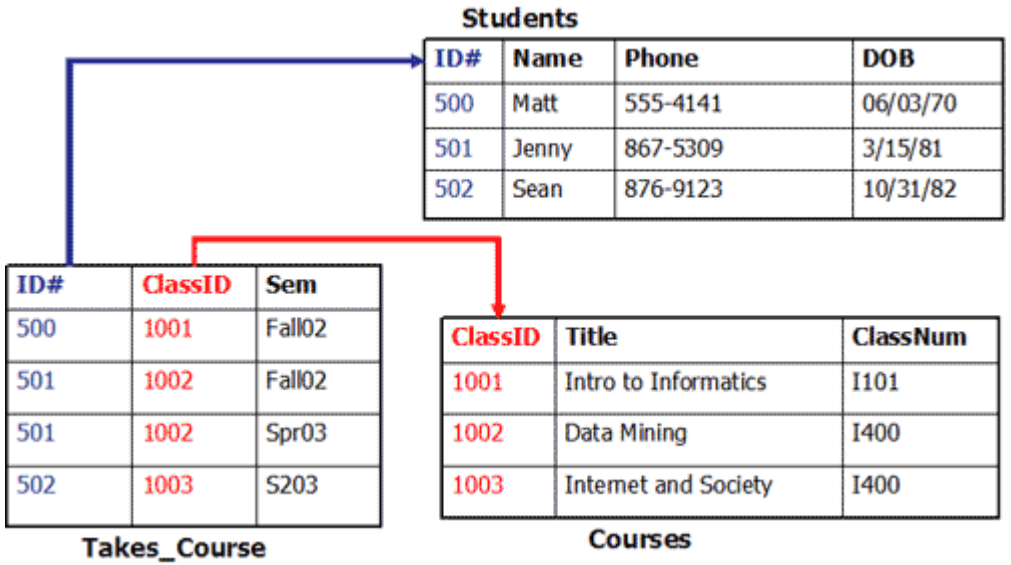


Prehistoria: 1940-1980



- ✓ Origen → censo, aseguradoras
- ✓ Propósito → almacenar y consultar
- ✓ Formato → ninguno
- ✓ Problema → inconsistencias

Bases de datos relacionales: 1980-20…



Students

ID#	Name	Phone	DOB
500	Matt	555-4141	06/03/70
501	Jenny	867-5309	3/15/81
502	Sean	876-9123	10/31/82

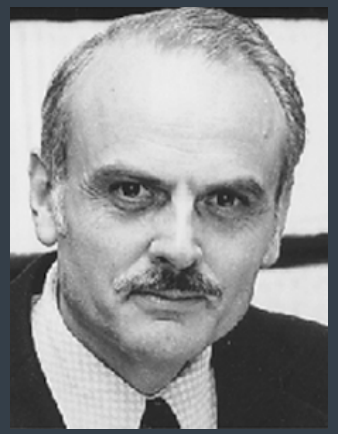
Takes_Course

ID#	ClassID	Sem
500	1001	Fall02
501	1002	Fall02
501	1002	Spr03
502	1003	S203

Courses

ClassID	Title	ClassNum
1001	Intro to Informatics	I101
1002	Data Mining	I400
1003	Internet and Society	I400

- ✓ Origen → empresas en general
- ✓ Propósito → almacenar y consultar, se empieza con “inteligencia de negocio”
- ✓ Formato → Tablas, lenguaje: SQL
- ✓ Problema → casi desconocidos, hasta hace poco



Dr. Edgar F. Codd

1923-2003

Edgard F. Codd estudió Matemáticas y Químicas en Oxford. Piloto aéreo con la RAF durante la segunda guerra mundial, en 1948 marchó a Estados Unidos para ingresar en IBM, aunque marchó a Canadá en 1953 en desacuerdo con la “caza de brujas” del senador McCarthy. Regresó 10 años más tardes, doctorándose por la Universidad de Michigan con una tesis sobre autómatas celulares. Tras su regreso a IBM propuso las bases de lo que hoy es el modelo relacional. Sin embargo sus ideas no tuvieron mucho éxito. Las ideas expuestas en su artículo fueron rechazadas inicialmente tanto como publicación científica (hoy tiene casi 10000 citas) como por su empresa. Al final, un poco a regañadientes, IBM terminó creando un lenguaje inspirado en las ideas de Codd al que llamaron SEQUEL, y que fue pronto mejorado por Larry Ellison el creador de Oracle dando lugar a SQL

¿Puntos débiles del modelo relacional?

Dificultad para manejar enormes cantidades de datos (Big Data)

Dificultad para manejar datos heterogéneos originados en Internet, IoT

Problemas del modelo relacional

Dificultad para tratar muchos datos

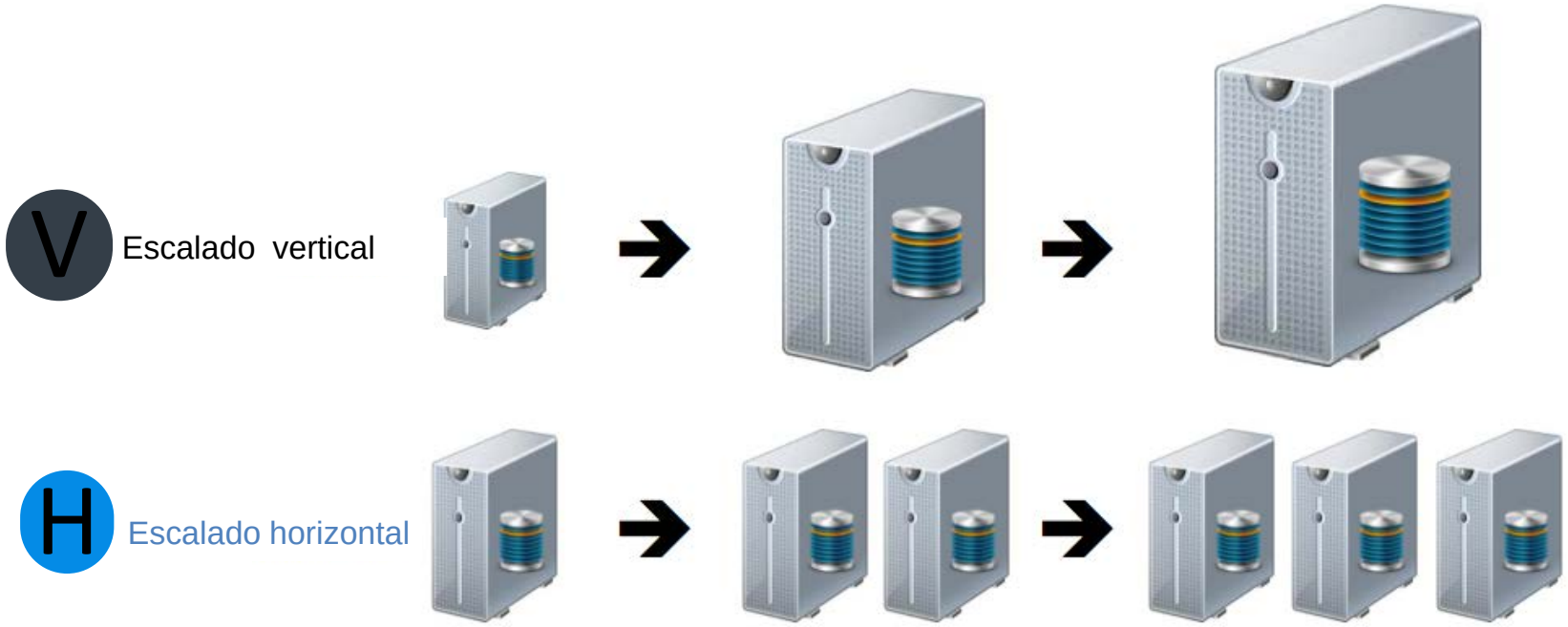
Cuando aparecen demasiados datos tenemos dos posibilidades

V Escalado vertical

Se sustituye el equipo por otro mayor, más potente (disco, memoria, etc.)

H Escalado horizontal

Se crea un clúster de equipos de gama baja, añadiendo más según demanda



Problemas del modelo relacional

Dificultad para tratar muchos datos

El escalado horizontal es más barato, pero poco adecuado para las bases de datos relacionales habituales

OjO: no es tan barato debido a las réplicas; cada dato se copia al menos 3 veces

Problemas del modelo relacional

Dificultad para tratar muchos datos

Pero,

¿dónde almaceno tanto equipo?

Problemas del modelo relacional

Dificultad para tratar muchos datos

Pero,

¿dónde almaceno tanto equipo?

La solución más ´común y más económica:

Problemas del modelo relacional

Dificultad para tratar muchos datos

Pero,

¿dónde almaceno tanto equipo?

Las grandes compañías: Centros de Proceso de Datos

Centro de proceso de datos de Zara, Arteixo (Coruña)

Problemas del modelo relacional

Dificultad para tratar muchos datos

Pero,

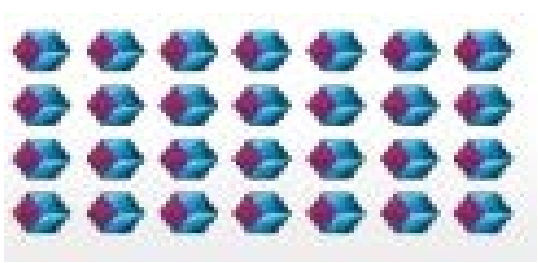
¿dónde almaceno tanto equipo?

Las grandes compañías: Centros de Proceso de Datos

Centro de proceso de datos por dentro

Definición de Big Data: Las 3Vs, 4Vs, 5Vs…

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. Doug Laney (Gartner)



Gran Volumen

Tanto como para necesitar de varios ordenadores (clúster).
La idea de Big Data se asocia a la de información distribuida

Gran Velocidad

En teoría la velocidad se refiere tanto a la escritura como a la lectura (consultas)

Gran variedad

Se piensa en datos heterogéneos, semi-estructurados, que no pueden ser tabulados con facilidad

Veracidad y/o Valor

A menudo se añade la “veracidad” indicando la “falta de veracidad”, es decir que algunos de los datos pueden ser poco fiables. El Valor (la 5ª V) se refiere a que debe tratarse de datos valiosos

Si estas definiciones no nos gustan podemos ver estas 12 definiciones en Forbes

Problemas del modelo relacional

Datos heterogéneos, bases de datos heterogéneas

La diversidad en las fuentes de datos y los nuevos tipos de análisis generan nuevas posibilidades y demandan nuevas formas de tratar los datos

Dispositivos

Móviles, páginas web, etc. son interminables fuentes de datos.

Y están a punto de unirse el mundo de los sensores; datos constantes que desde su inicio deben ser pensados en relación con Big Data

Redes sociales

El acceso a información personal es de máxima importancia para detectar el gusto de los consumidores.

Más aún, el poder agrupar gente con gustos comunes permite análisis hasta ahora impensables

Predicción de stock

La gestión de almacenes de una gran cadena se reduce al mínimo si se es capaz de predecir en detalle qué necesidades tendrá cada tienda en cada momento

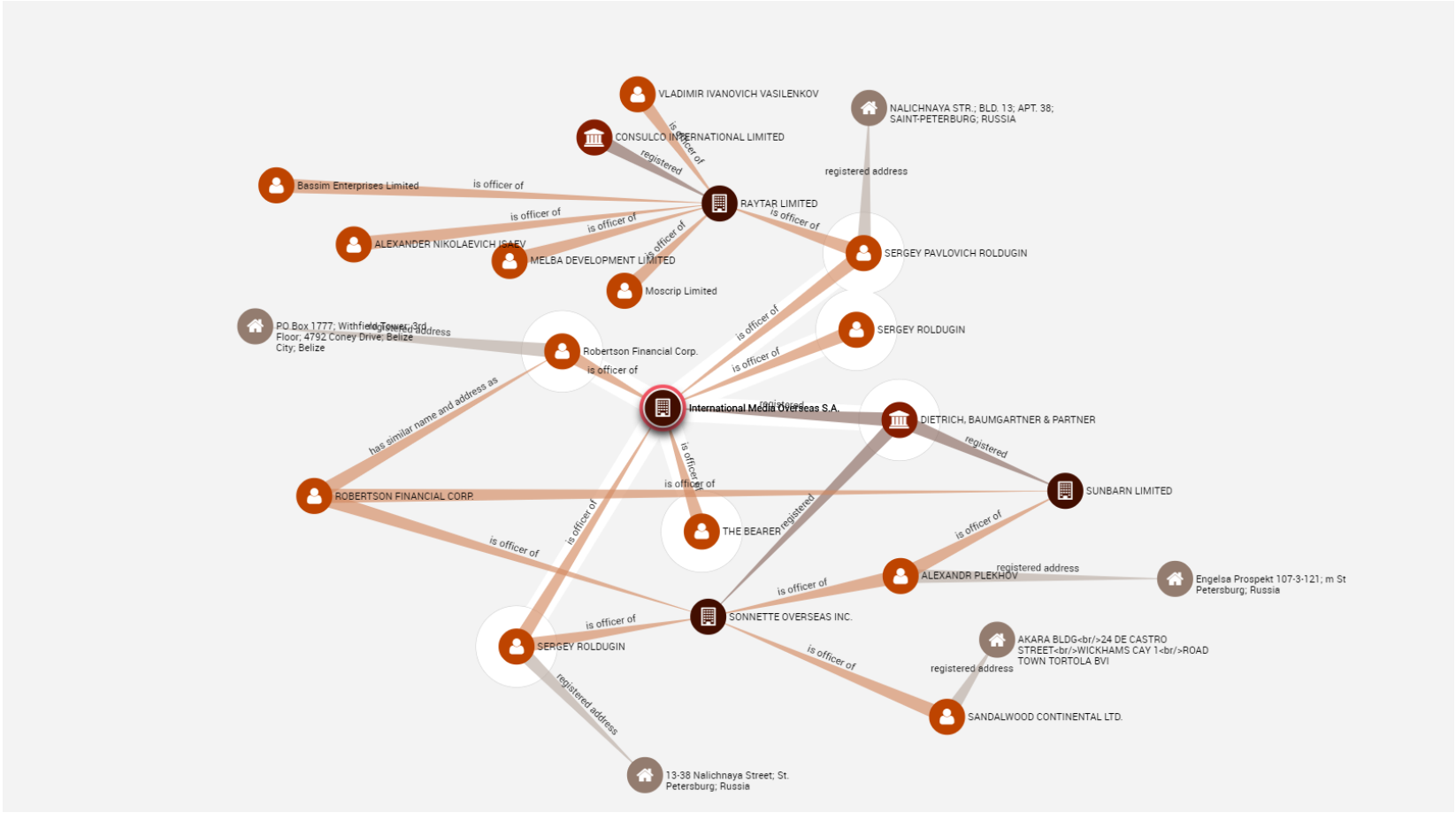
Detección de fraude

Los defraudadores, tanto bancarios, como de seguros o simples timadores, tienden a seguir patrones, que pueden ser descubiertos mediante bases de datos orientadas a grafos

Problemas del modelo relacional

Datos heterogéneos, bases de datos heterogéneas

Ejemplo: los papeles de Panamá



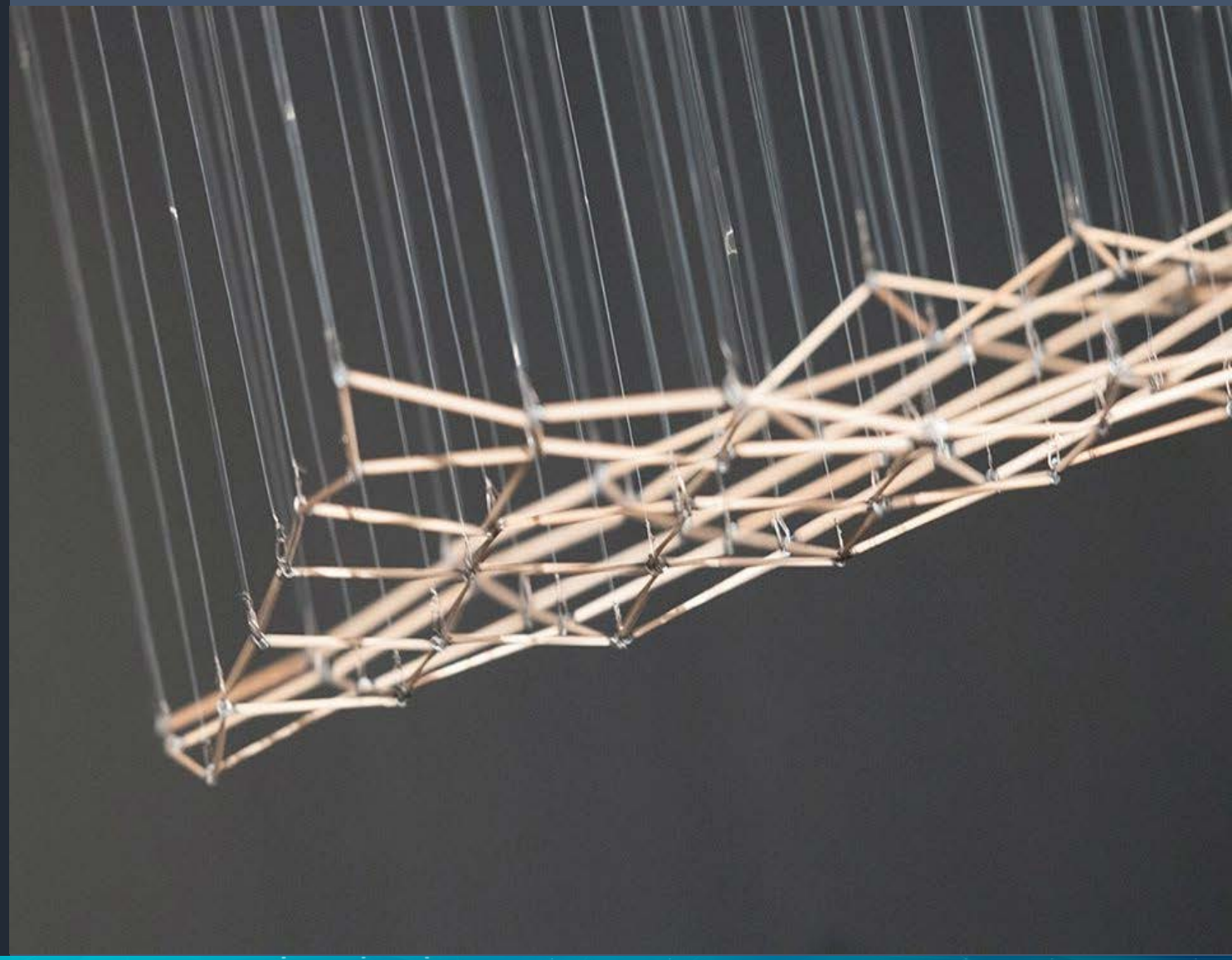
1 Introducción

2 Definición de Big Data

3 Análisis científico de datos

Introducción

Al análisis de datos en Big Data



¿Qué es eso del análisis científico de datos?

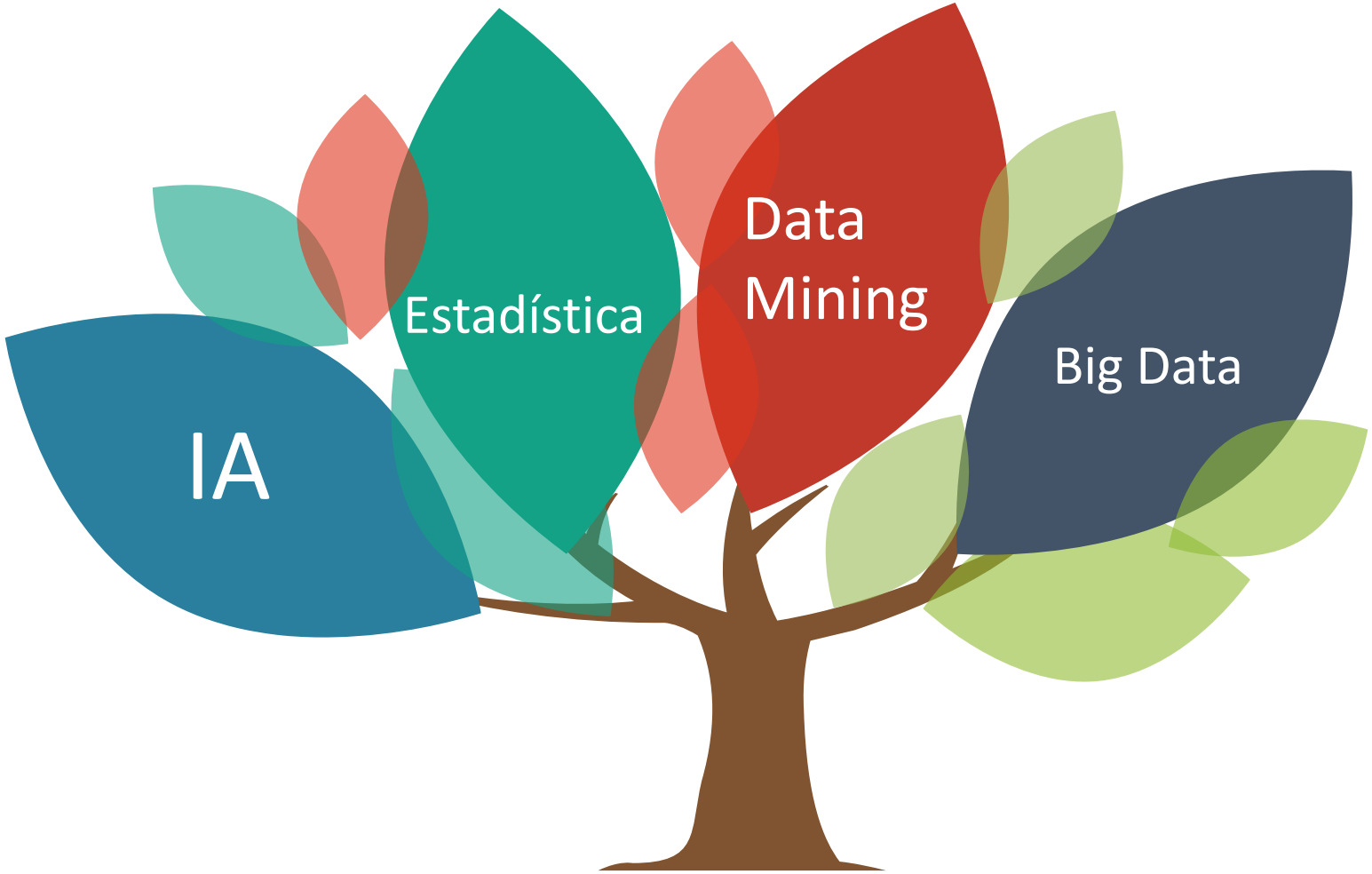
El análisis de datos consiste en la creación de modelos que representen un conjunto de datos con el objetivo de
- descubrir información útil
- realizar predicciones

y en general servir de apoyo para la toma de decisiones

El análisis de datos se puede ver desde muchos enfoques, aquí vamos a utilizar el conjunto de técnicas conocidas como **Machine Learning (Aprendizaje Automático)**

Machine Learning

Amigos y parientes



Machine Learning

Un poco de historia

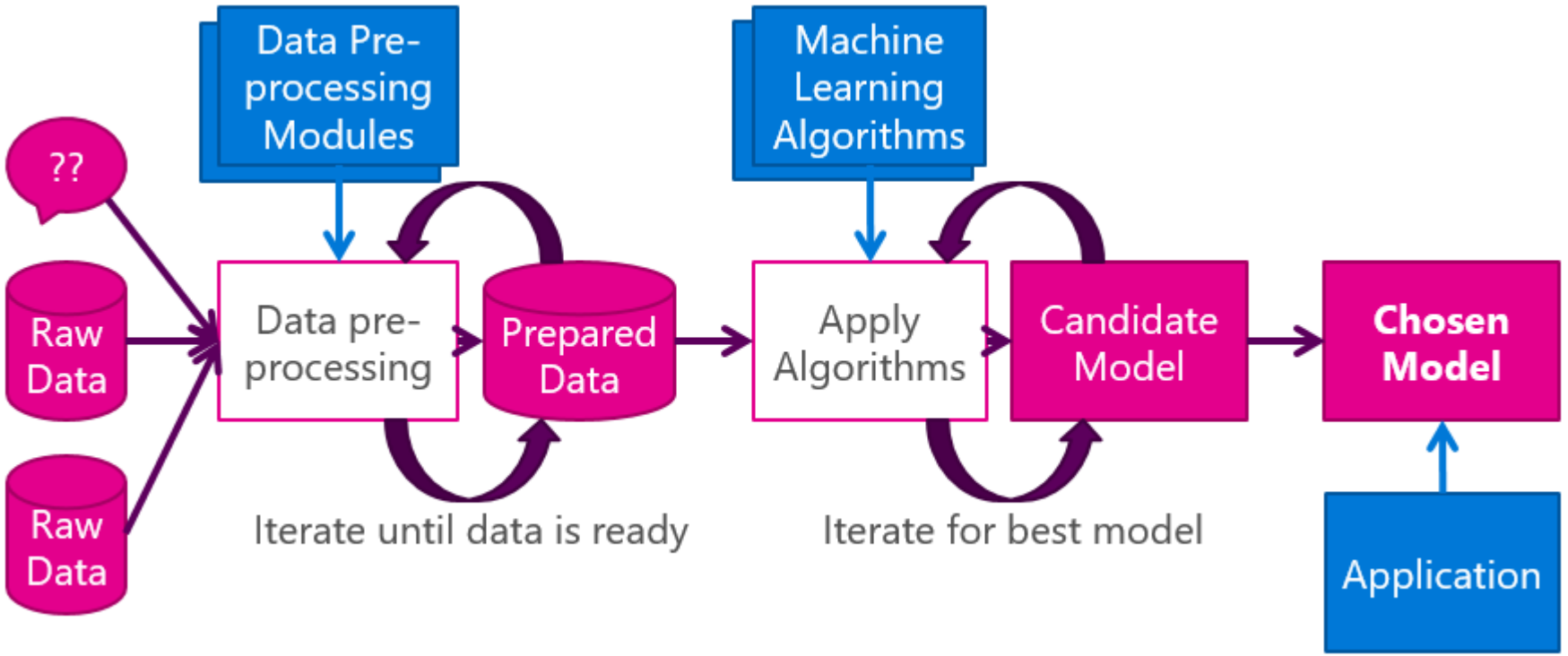
El origen

80's: AI se dedicaba a los sistemas expertos:

- Sin Estadística
- Poca atención a las redes neuronales
- Grupo de investigadores aislados continuaban en ANN: BackPropagation (86)

90's: se empieza a hablar de ML → resolver problemas prácticos. Uso de métodos estadísticos

Machine Learning: etapas



https://blogs.msdn.microsoft.com/martinkearn/2016/03/01/machine-learning-is-for-muggles-too/

La definición de Mitchell

La más formal y la más usada

Decimos que:

Un programa aprende de la experiencia E con respecto a:

a) Una clase de tareas T , y

b) Una medida de eficiencia P

Si la eficiencia en las tareas en T, medida según P, aumenta con la experiencia E

Ejemplo: recomendador

Como instancia de la definición general

Def. general

Un programa aprende de la experiencia E con respecto a:

a) Una clase de tareas T , y

b) Una medida de eficiencia P

Si la eficiencia en las tareas en T, medida según P, aumenta con la experiencia E

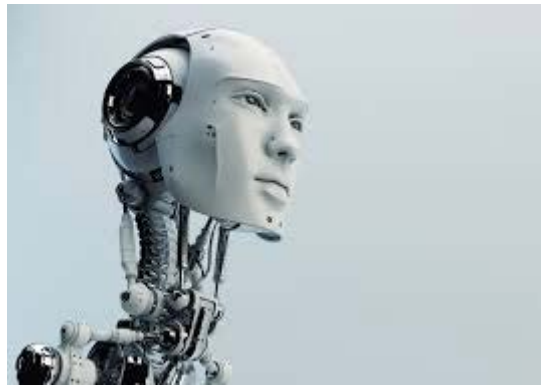
Instancia

Experiencia E: Conjunto de valoraciones de productos de los clientes

Tareas T : Recomendación de un producto a un usuario

Eficiencia P: Minimizar error de los mínimos cuadrados (diferencia entre la valoración estimada y la que realmente se ha dado) para un conjunto test subconjunto E

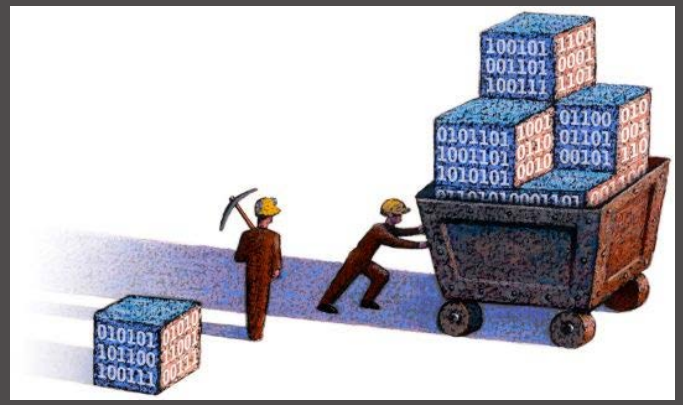
Machine Learning



Predecir

Predicción a partir de propiedades conocidas, aprendidas a partir de conjuntos de entrenamiento

Data Mining



Descubrir

Descubrir nuevas propiedades del conjunto de datos

Aprendizaje supervisado

El conjunto de entrenamiento incluye resultados esperados para cada dato

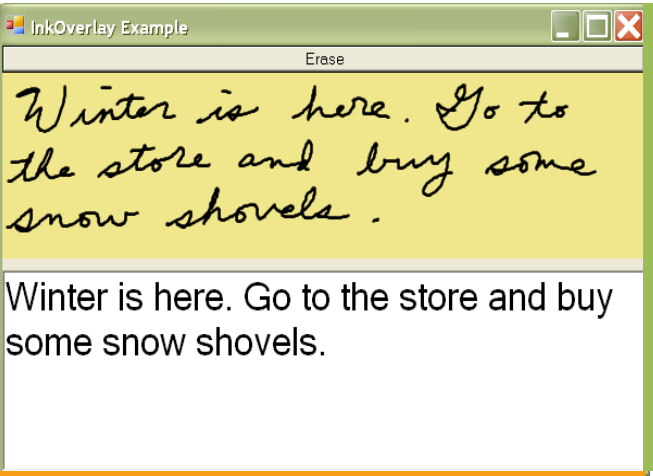
Tarea: Dado un dato, encontrar otro valor asociado
Ejemplo: Dado un dato de colesterol en sangre indicar la probabilidad de infarto

Experiencia E: conjunto de datos de entrenamiento, que incluyen para cada dato un valor asociado útil para calcular el que deseamos
Ejemplo: Personas con edad > 80 años, su nivel medio de colesterol entre los 50 y los 80 y el valor true si han sufrido algún infarto y false en otro caso.

Se dice supervisado porque tenemos una idea de los valores que se deben obtener para el conjunto de entrenamiento → útil para la eficiencia P
Ejemplo: Tras el "aprendizaje" a las personas que han sufrido un infarto les debe corresponder una probabilidad más alta que a las que no en el modelo

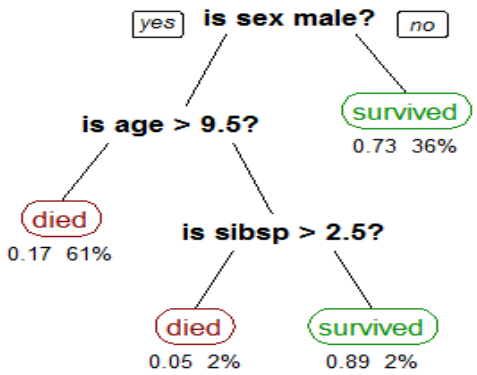
Tipos de aprendizaje supervisado

Solo unos pocos tipos



Regresión

Buscar una función que aproxime la relación de dependencia entre varias variables

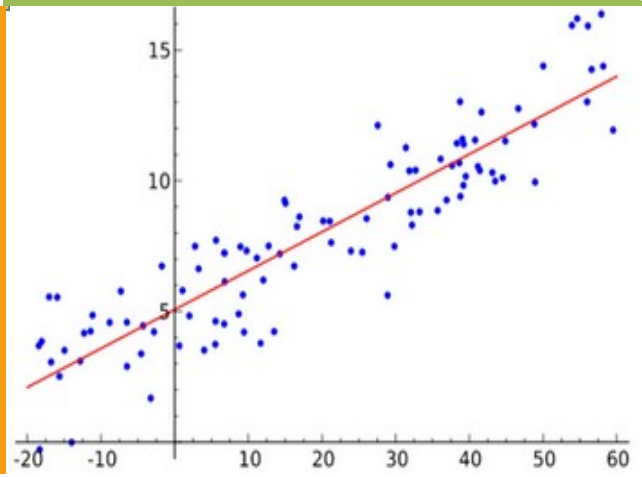


Modelos Bayesianos

Determinan los parámetros de un modelo o determinan qué modelo se ajusta mejor a unos datos

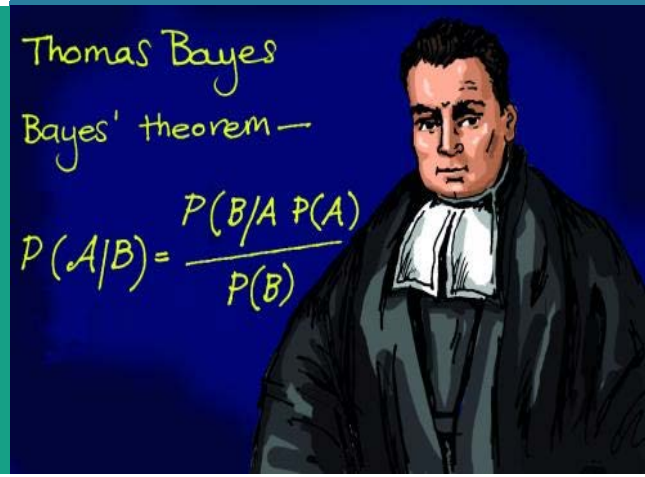
Redes neuronales

Determinar funciones que dependen de una gran cantidad de entradas con retropropagación



Árboles de decisión

Clasificación de elementos dada un serie de test consecutivos



Aprendizaje supervisado: ejemplo

Aprendiendo a conocer las frutas: https://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/

E: conjunto de frutas; para cada una observamos su tamaño y color, y además nos dicen su nombre
T: Determinar el nombre de nuevas frutas
P: Número de aciertos con frutas nuevas

No.	SIZE	COLOR	SHAPE	FRUIT NAME
1	Big	Red	Rounded shape with a depression at the top	Apple
2	Small	Red	Heart-shaped to nearly globular	Cherry
3	Big	Green	Long curving cylinder	Banana
4	Small	Green	Round to oval,Bunch shape Cylindrical	Grape

Llega una fruta que es determinada como “Big”, “Green” y “Rounded shape with a depression at the top”
¿Qué hacemos?

Aprendizaje no supervisado

El conjunto de entrenamiento incluye resultados esperados para cada dato

Se trata de elaborar un modelo a partir de unas observaciones (Experiencia E)

No hay conocimiento a priori y se trata de descubrir una estructura interna oculta en los datos

La técnica más común es clustering: agrupación de elementos con características similares

Otras técnicas: compleción de matrices, búsqueda de estructuras de grafos

Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek

¿Cómo los agruparías?

Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek






Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek

Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek

Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek

¿Para Qué?

Debemos tener en cuenta el uso que se hará de los resultados

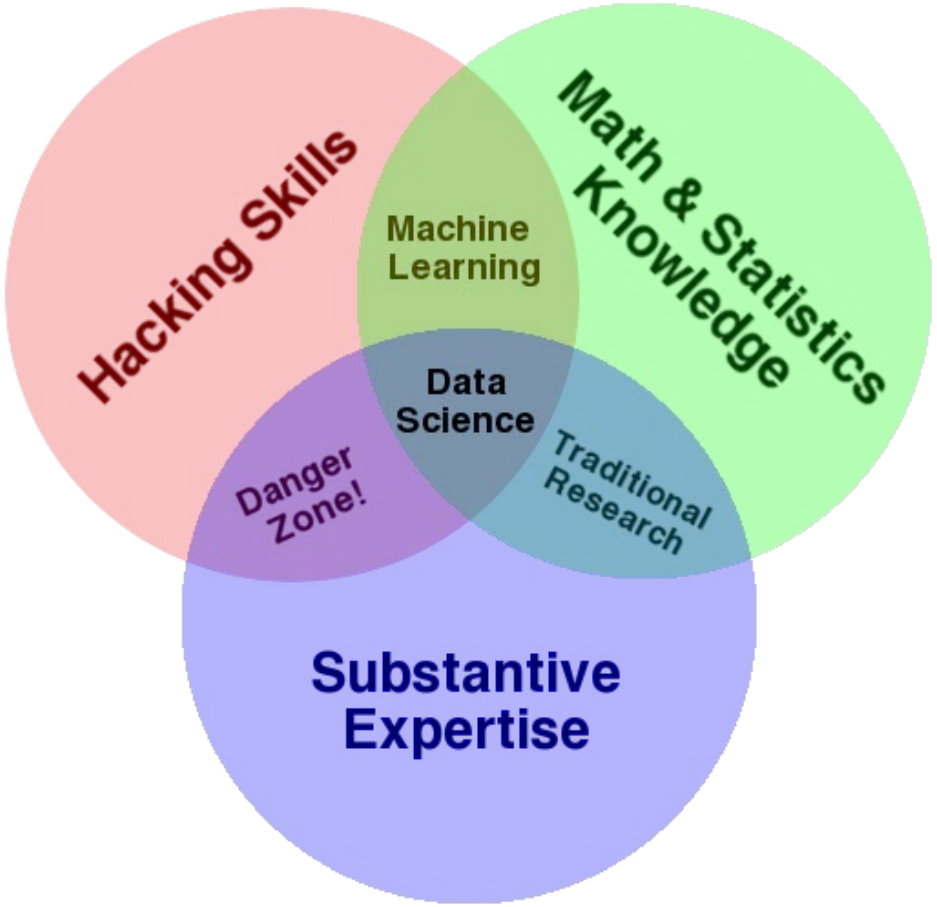
El destino que vayamos a dar a los resultados puede determinar el método y la tecnología

- ✓ ¿Respuesta rápida? ¿Llegada de datos constante? ¿Recálculo del modelo?
- ✓ ¿Nivel de error permitido?
- ✓ ¿Se prefieren falsos positivos o falsos negativos?
- ✓ ¿Accuracy, precission, recall?

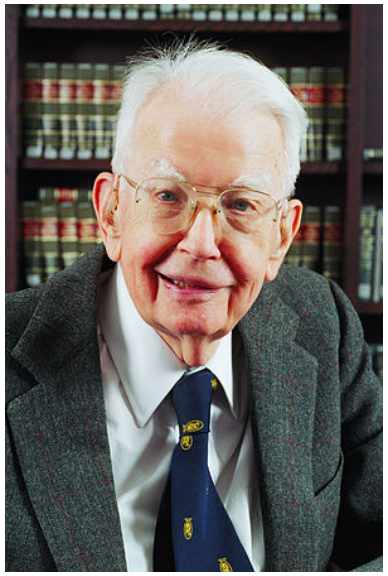
¡El análisis de datos puede ser peligroso!

El diagrama de Drew Conway

El peligro de quedarse solo con una parte



If you torture the data long enough, it will confess



Ronald H. Coase (1910-2013)

¡Gracias!