



# Sistemas de Big Data

Curso de especialización en Inteligencia Artificial y Big Data.



# Programación



# Introducción - Big Data

No existe una definición precisa del término pero los términos de datos masivos o grandes volúmenes de datos hacen referencia al big data. Por este motivo, a menudo el concepto de big data es definido en función de las características que poseen los datos y los procesos que forman parte de este nuevo paradigma de computación. Esto es lo que se conoce como las Vs del Big Data.

## LAS TRES V DEL BIG DATA





# Modelos de negocio basado en datos

El modelo de negocio se refiere a cómo una organización crea y captura valor. Los modelos de negocio basados en Big Data se encuentran actualmente en una fase incipiente.

**Tabla 1.3:** Tipología de BDBM (fuente [\[WSM20\]](#))

Tipo	Fuente de valor (ejemplo)
Usuarios de datos	<ul style="list-style-type: none"><li>• Analisis BD para apoyar la toma de decisiones estratégicas</li><li>• Uso de BD para mejorar los procesos internos</li><li>• Enriquecer productos, servicios y experiencia de clientes mediante BD</li></ul>
Proveedoras de datos	<ul style="list-style-type: none"><li>• Desarrollo de nuevos productos y servicios mediante BD</li><li>• Recopilando datos primarios y vendiéndolos a terceros</li></ul>
Facilitadores de BD	<ul style="list-style-type: none"><li>• Agregando datos y empaquetando datos internos para la venta</li><li>• Ofreciendo la infraestructura a las anteriores tipos de empresa necesaria para realizar BD</li><li>• Consultoría relativa a BD</li><li>• Subcontratación de técnicas analíticas para BD (ejemplo en la nube)</li></ul>

# Complejidad computacional para el análisis de datos

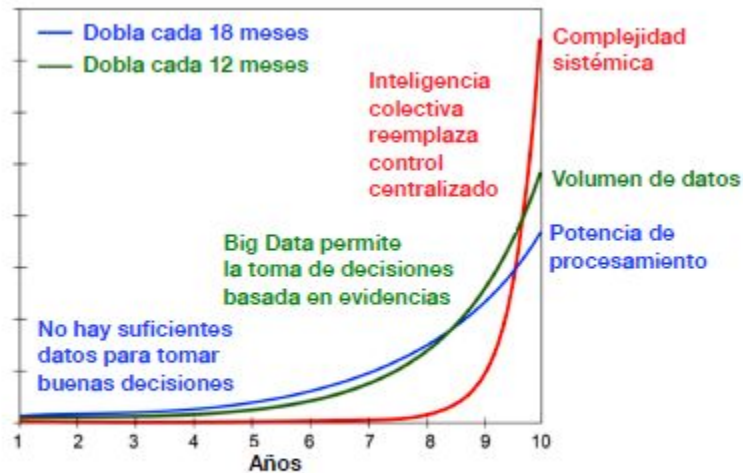


Figura 1.1: Modelo de crecimiento digital, fuente [HPG+19].



# Complejidad computacional para el análisis de datos

El crecimiento del coste de los procesos para analizar datos, aspecto que incluyen los algoritmos, tiene un crecimiento factorial cuyo consecuencia inmediata es una gran cantidad de datos que no podrán ser analizados. Esta complejidad explosiva impone un modelo de inteligencia distribuida para poder afrontar los retos del mundo y de nuestras sociedades.

El coste computacional de la ejecución de un algoritmo es un factor limitante incluso cuando no se está ante grandes volúmenes de datos.



# Complejidad computacional para el análisis de datos

Supongamos que el número de datos disponibles es  $n$ . Es evidente que si el número de datos  $n$  aumenta entonces el tiempo de ejecución de un algoritmo para procesarlos también aumente. Por tanto cualquier algoritmo tiene la propiedad que si  $n$  crece hacia infinito su coste computacional también crece hacia infinito. La cuestión esencial es que existen varias velocidades de crecer a infinito, unas significativamente más rápidas que otras. La relación matemática

$$\lim_{n \rightarrow +\infty} \frac{P(n)}{a^n} = 0$$

para cualquier  $a > 1$  y para cualquier polinomio  $P(n)$  indica que aunque tanto el numerador como el denominador tienden a infinito el denominador crece muchísimos más rápido que el numerador. El algoritmo se volverá cada vez más lento y puede ser impracticable para conjuntos de datos grandes



# Complejidad computacional para el análisis de datos

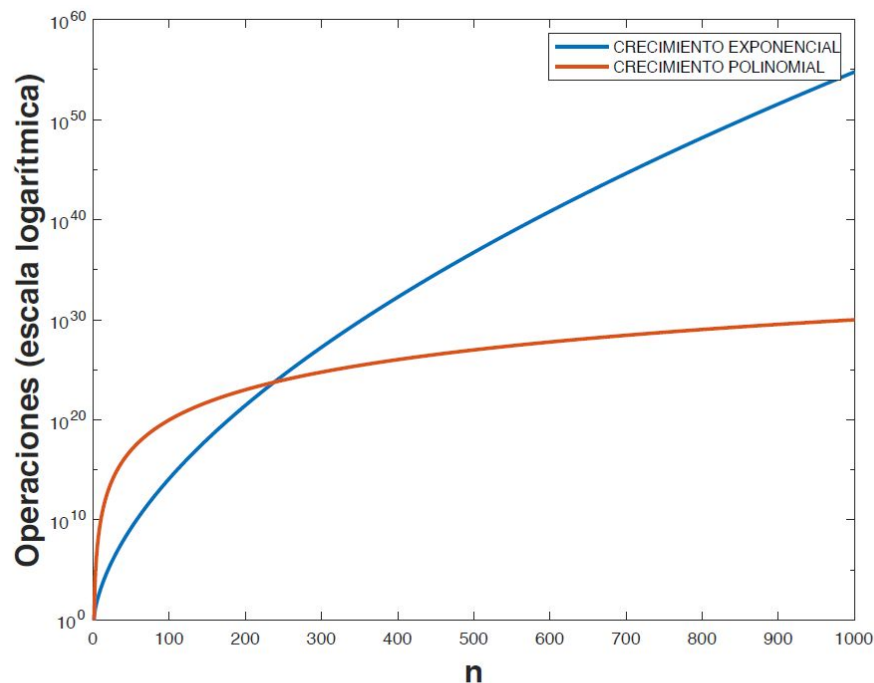
Se dice que un algoritmo tiene un tiempo de ejecución polinomial si el número de operaciones máximo que realiza para resolver un problema está acotado por un polinomio en el tamaño del problema.

Se dice que un problema está en la clase de complejidad P si existe un algoritmo para resolver el problema que tiene un tiempo de ejecución polinomial. Si no se conoce ningún algoritmo que lo resuelva en tiempo polinomial (porque no exista o porque en este momento no se haya desarrollado) se dice que el problema está en la clase de complejidad NP.



# Complejidad computacional para el análisis de datos

Crecimiento polinomial vs exponencial






## Complejidad computacional para el análisis de datos

Supongamos que disponemos del ordenador de IBM *Roadrunner* que es capaz de superar un *petaflop* de operaciones por segundo. Es capaz de realizar  $1,105 \cdot 10^{15}$  operaciones por segundo. ¿Cuánto tiempo emplearía en factorizar un número de 50 cifras?

$$\text{VelocidadCPU} = 1,105 \cdot 10^{15} \text{ operaciones / segundos}$$

$$\frac{f_A(50)}{\text{VelocidadCPU}} = 1,4907 \cdot 10^{-6} \text{ segundos}$$

El resultado muestra que tardaría *menos de una milésima de segundo*. Y si nos preguntamos cual es el máximo número de cifras  $n$  que *Roadrunner* puede factorizar durante un año. Planteamos el siguiente código (ver Listado 1.1) en Python para resolver la cuestión.



# Complejidad computacional para el análisis de datos

Listado 1.1: Código en python para calcular el máximo número de cifras de una clave RSA que se puede romper con el ordenador Roadrunner

```
1 # Carga de la libreria math para calcular log10
2 from math import *
3 # Operaciones/segundo realizadas por Roadrunner
4 VelocidadCPU = 1.105 * 10 ** 15
5 # número de dígitos
6 n=1
7 # número de operaciones
8 fA=1
9 # ¿ nº operaciones requeridas <nº operaciones en un año?
10 while fA <= (VelocidadCPU *24 * 60 * 60 *365):
11     n = n+1
12     fA = 10 ** ( pow( n * log(n,10) , 0.5 ) )
13
14 print('Numero maximo de cifras  = ' , n-1)
```

El resultado de ejecutarlo es  $n = 217$ . Hemos pasado de factorizar un número de 50 cifras en una millonésima de segundo a necesitar un año para factorizar un número con un poco mas del cuádruple de cifras. Nos podemos plantear la cuestión



# Complejidad computacional para el análisis de datos

La duplicación de los recursos computacionales no es una estrategia por sí sola que sea capaz de abordar el coste computacional de los problemas.

Este ejemplo muestra que el un tiempo de ejecución exponencial es inabordable para cualquier máquina actual. De hecho a efectos prácticos una complejidad computacional superior a  $n^3$  es dramática.

# Complejidad computacional para el análisis de datos

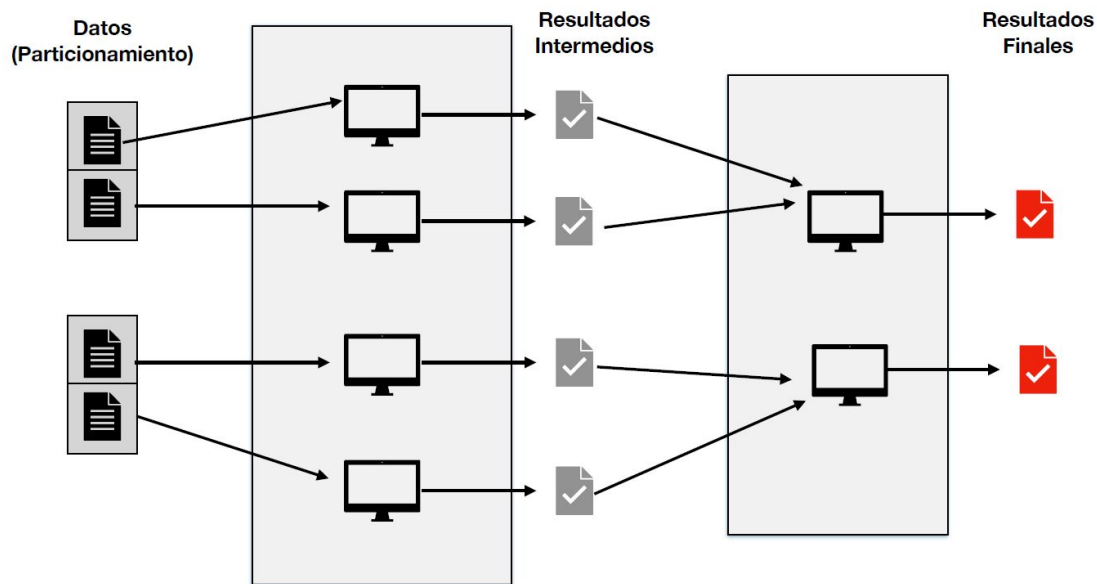


Figura 1.3: Computación distribuida.



# Complejidad computacional para el análisis de datos

Velocidad de ejecución de los lenguajes de programación. [enlace](#)

# Aplicación de técnicas de integración, procesamiento y análisis de la información



## Introducción. Técnicas integradas de procesamiento y análisis de la información

El rápido avance de las tecnologías digitales en las últimas décadas ha llevado a una generación exponencial de datos. Se estima que el 90% de todos los datos disponibles actualmente fueron generados en los últimos 2 años. Para transformar estos enormes volúmenes de datos en información útil surgen las técnicas modernas de procesamiento y análisis de información.

Estas técnicas se aplican en diversas industrias: en investigación científica permite analizar grandes conjuntos de datos experimentales para obtener nuevos insights; en salud posibilita el procesamiento de historiales clínicos para la detección temprana de enfermedades; en comercio viabiliza la personalización de contenidos y productos de acuerdo a los intereses de los usuarios; etc.



# Técnicas modernas de tratamiento de datos

Algunas de las técnicas más relevantes son:

- **Big Data:** enfocada en el almacenamiento, administración y procesamiento de enormes conjuntos de datos, tanto estructurados como no estructurados. Utiliza tecnologías como Hadoop, Spark y bases de datos NoSQL.
- **Data Mining:** busca descubrir patrones repetitivos y relaciones entre variables en grandes bases de datos, utilizando algoritmos estadísticos y de machine learning. Algunas herramientas son Orange, Weka y KNIME.
- **Business Intelligence:** conjunto de estrategias y herramientas enfocadas en el análisis de datos empresariales para facilitar una mejor toma de decisiones. Algunas soluciones populares son Tableau, Qlik Sense y Microsoft Power BI.





# Técnicas y procesos de extracción de la información de los datos

El proceso típico de extracción de información consta de:

- **Recolección:** obtención de los datos desde diferentes fuentes.
- **Limpieza:** corrección de errores, manejo de valores faltantes, etc.
- **Transformación:** conversión a formatos adecuados para el análisis.
- **Modelado y análisis:** aplicación de algoritmos para descubrir patrones y relaciones. Por ejemplo, la clasificación mediante árboles de decisión.
- **Interpretación:** traducción de los resultados del modelo analítico en insights aplicables para la toma de decisiones.

# Técnicas y procesos de extracción de la información de los datos

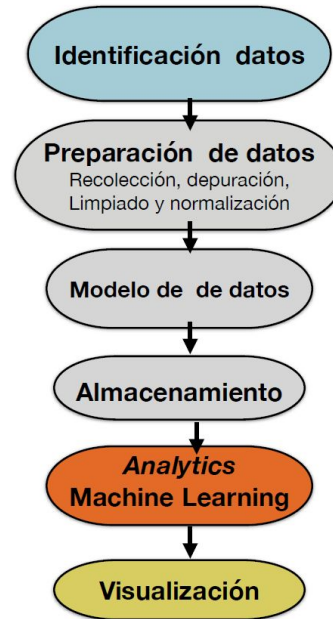


Figura 1.4: Proceso de extracción de información de los datos.



# Técnicas y procesos de extracción de la información de los datos

**Estructurados.** Los datos estructurados son aquellos que están bien definidos en cuanto su formato, longitud y significado. Ejemplos de este tipo de datos son magnitudes, fechas, cadenas de caracteres.

Su almacenamiento se registra en tablas. Si el volumen es grande se almacenan en Data Warehouse, en otros casos se emplean bases de datos relacionales o simples hojas de cálculo. En los modelos de bases de datos relacionales toda la información se almacena en tablas en las que se especifican el tipo de campos que tienen y cómo se relacionan entre ellas.



# Técnicas y procesos de extracción de la información de los datos

**No estructurados.** Los datos no estructurados no tienen un formato ni un tamaño predefinido. Ejemplos de este tipo de datos son el texto, vídeo, audio e imágenes.

Este tipo de datos se generan por múltiples fuentes como sensores, satélites, ordenadores, teléfonos inteligentes, redes sociales, por actividades en sitios webs, etc. Hoy en día este tipo de datos constituyen una fuente muy importante para las empresas para obtener información de sus clientes, de sus procesos, de sus productos, etc. El almacenamiento de este tipo de datos debe realizarse de forma organizada a través de una base de datos no relacional (NoSQL).



# Técnicas y procesos de extracción de la información de los datos

**Semi estructurados.** Este tipo de datos combinan datos estructurados y no estructurados.

Estos datos tienen un esqueleto o estructura que organiza los inputs que deben contener y define las relaciones entre los textos y objetos que contienen. Este esqueleto puede evolucionar en el tiempo y convertirse en un protocolo aceptado. Este esqueleto se define en forma de árbol, con etiquetas para facilitarte el manejo.

Ejemplos de este tipo de datos son las páginas web o servicios de correos electrónicos donde se almacenan los mensajes y ficheros adjuntos. Los metadatos permiten clasificarlos y realizar búsquedas por palabras clave.



# Técnicas y procesos de extracción de la información de los datos

El primer paso es determinar las fuentes de datos que pueden ser útiles. Estas fuentes abarcan los datos generados en el propio entorno empresarial como aquellas que son ajenas a la organización, como redes sociales, web, bases de datos compradas a otras compañías, datos geográficos, datos económicos sectoriales, etc. Tras este proceso se dispondrá de un conjunto de datos que abarcarán una o varias de las siguientes tipologías: estructurados, no estructurados y semi-estructurados.



# Herramientas de extracción de datos

Algunas herramientas útiles para extraer datos de **fuentes no estructuradas**:

- **ParseHub**: permite extraer datos de sitios web mediante un interfaz visual de arrastrar y soltar.
- **Import.io**: extrae datos de web scraping mediante un proceso guiado por el usuario.
- **MonkeyLearn**: extrae información de texto plano utilizando modelos de machine learning.

Para **datos estructurados** y bases de datos:

- **Knime**: plataforma de integración y análisis de datos con módulos para ETL, mining e informes.
- **Trifacta**: herramienta de preparación de datos que limpia, transforma y enriquece datasets.
- **Pentaho**: solución open source de business intelligence con capacidades de ETL y análisis de datos.



# Análisis en tiempo real

El análisis en **tiempo real** tiene aplicaciones como:

- **Detección de fraudes:** analiza transacciones financieras para identificar en segundos comportamientos anómalos o sospechosos.
- **Monitoreo industrial:** sensores IoT envían continuamente datos de fábricas que son analizados en tiempo real para detectar fallas o necesidades de mantenimiento.
- **Personalización de contenidos:** las interacciones de los usuarios en un sitio web se analizan al instante para ofrecer resultados de búsqueda, productos y anuncios adaptados a sus intereses.





## Costes y calidad

- **Los costes del análisis de datos pueden incluir:** licencias de software especializado, infraestructura en la nube, consultoría de expertos, capacitación de personal, que pueden variar entre algunos miles a cientos de miles de dólares dependiendo de la complejidad.
- **La calidad de los datos fuente es crucial. Los problemas más comunes son:** datos incompletos, ruidosos, duplicados, desactualizados, incorrectamente formateados. Si no se detectan y corrigen pueden sesgar los resultados del análisis.

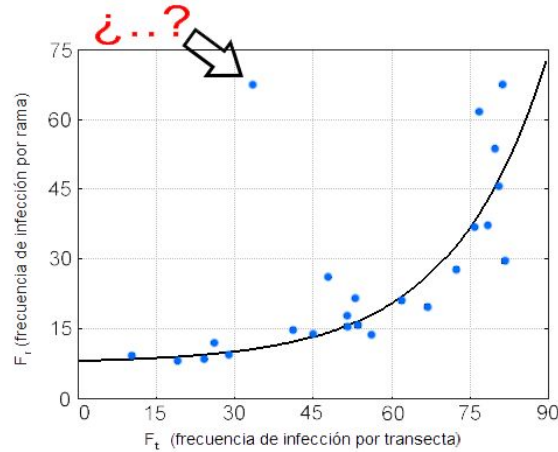
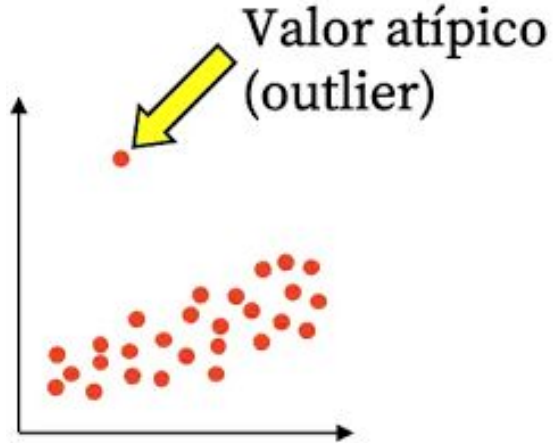


# Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

**1. Limpieza de Datos (Data Cleaning):** Durante la recopilación de datos, es común encontrar datos erróneos o inconsistentes. La limpieza de datos implica la identificación y corrección de estos problemas. Esto podría incluir la eliminación de registros duplicados, la corrección de errores tipográficos o la estandarización de formatos de datos.

**2. Detección de Outliers:** Los outliers son valores atípicos que se desvían significativamente del resto de los datos y pueden distorsionar los análisis. Se utilizan métodos estadísticos para identificar estos valores atípicos y se decide si deben ser ignorados o tratados de alguna manera (por ejemplo, reemplazándolos por valores más representativos o eliminándolos si son errores evidentes).

# Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.



Salario (en dólares):

50000

52000

55000

60000

2000000 <-- Outlier

57000

59000

# Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

**3. Manejo de Valores Faltantes (Missing Values):** Los valores faltantes en los datos son comunes y deben ser abordados adecuadamente. Esto implica tomar decisiones sobre cómo tratar esos valores. Algunas opciones incluyen:

- Relleno con la media o mediana de la columna.
- Estimación basada en otros datos relacionados.
- Eliminación de registros con valores faltantes si la cantidad de datos perdidos es pequeña y no crítica.

Datos originales:

CSS

Nombre	Edad	Puntuación
Alice	35	8
Bob	N/A	7
Charlie	42	9

Manejo de valores faltantes (usando la media):

SCSS

Nombre	Edad (Rellenado)	Puntuación
Alice	35	8
Bob	38 (media)	7
Charlie	42	9



## Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

**4. Estandarización y Normalización:** En muchos casos, es importante estandarizar o normalizar los datos para que tengan una escala similar. Esto es especialmente relevante en algoritmos basados en distancias o gradientes.

**5. Codificación de Categorías:** Si los datos contienen variables categóricas (como colores o categorías de productos), es necesario codificarlas en valores numéricos para que los algoritmos de machine learning puedan utilizarlos. Esto se hace mediante técnicas como la codificación one-hot.



## Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

**6. Selección de Características:** En ocasiones, es necesario seleccionar un subconjunto de las características (columnas) más relevantes para el análisis o el modelo. Esto puede mejorar la eficiencia computacional y reducir la complejidad del modelo.

**7. División en Conjuntos de Entrenamiento y Prueba:** Finalmente, los datos se dividen en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo. El conjunto de entrenamiento se utiliza para entrenar el modelo, mientras que el conjunto de prueba se utiliza para evaluar su capacidad de generalización.



# Modelos de datos y almacenamiento

El modelo de datos se puede entender como un lenguaje o conjunto de herramientas que utilizamos para hablar sobre cómo almacenamos y organizamos nuestros datos en una base de datos. Es como un conjunto de reglas y conceptos que nos permite describir los datos de una manera sistemática y comprensible.

Este contiene tres elementos:

- **Notación para la descripción de la estructura de los datos:** Esto significa que tenemos una forma de representar cómo se ven nuestros datos, qué tipo de información contienen y cómo están relacionados entre sí.
- **Restricciones de integridad:** reglas que indican que deben cumplir los datos para ser considerados válidos y precisos.
- **Operaciones para actualizar y recuperar los datos:** el modelo de datos también nos proporciona formas de agregar, modificar o eliminar datos, así como de buscar y obtener información específica de la base de datos.



# Modelos de datos y almacenamiento

Existe tres niveles de abstracción a la hora de analizar un modelo de datos:

- 1. Modelo de datos conceptual.** Es el nivel más alto de abstracción y en este nivel se define lo que el sistema contiene, estableciendo su organización, finalidad y reglas y conceptos del negocio.
- 2. Modelo de datos lógico.** Está situado en el nivel intermedio de abstracción y define cómo el sistema debería estar implementado, independiente del tipo de base de datos que se empleará. El objetivo es desarrollar un mapa técnico para las reglas y la estructura de los datos.
- 3. Modelo de datos físico.** Es el nivel más bajo de abstracción y describe cómo el sistema será implementado usando una base de datos específica.





# Modelos de datos y almacenamiento

Técnicas de modelización de datos empleadas. Estas técnicas de modelización son aplicables tanto a modelos relacionales como no relacionales, y se describe a continuación:

- 1. Diagrama de entidades-relaciones (ERD).** Esta técnica visual es la opción por defecto empleada en la modelización y diseño de bases de datos relacionales. Incorpora el uso de entidades, atributos, relaciones, cardinalidades, restricciones entre otros elementos, así como notación simbólica.
- 2. Diagramas de clase Unified Modeling Language (UML).** Este es una notación standard para modelizar y diseñar sistemas de información empresarial.
- 3. Diccionario de datos.** Es una representación tabular de los conjuntos de datos y sus atributos, que contienen elementos como la descripción de los elementos, relaciones entre tablas, restricciones (unicidad, valores por defecto, valores válidos).

# Modelos de datos y almacenamiento



## Diagrama de entidades-relaciones (ERD). Ejemplo

Supongamos que estamos diseñando una base de datos para una biblioteca. Utilizamos un diagrama de Entidades-Relaciones para representar la estructura de datos. Ejemplo simplificado:

Entidades:

- Libro
- Autor
- Cliente

Atributos:

- Libro: Título, ISBN, Fecha de Publicación
- Autor: Nombre, Apellido, Nacionalidad
- Cliente: Nombre, Apellido, Número de Tarjeta

Relaciones:

- Un libro puede ser escrito por uno o varios autores (relación "Escrito por").
- Un cliente puede tomar prestado uno o varios libros (relación "Prestado a").

El diagrama mostrará visualmente cómo se relacionan las entidades (Libro, Autor, Cliente) y sus atributos, así como las relaciones entre ellas (quién escribió qué libro, quién ha tomado prestado qué libro).

# Modelos de datos y almacenamiento



## 2. Diagramas de clase Unified Modeling Language (UML). Ejemplo

Estamos diseñando un sistema de gestión de pedidos en línea para una tienda. Utilizamos un diagrama de clase UML para modelar el sistema. Ejemplo simplificado:

Clases:

- Pedido
- Cliente
- Producto

Atributos:

- Pedido: ID, Fecha de Pedido, Total
- Cliente: ID, Nombre, Dirección
- Producto: ID, Nombre, Precio

Relaciones:

- Un pedido tiene un cliente asociado (relación "Pertenece a").
- Un pedido puede contener uno o varios productos (relación "Contiene").

Este diagrama UML muestra cómo se relacionan las clases (Pedido, Cliente, Producto) y sus atributos, así como las relaciones entre ellas. También puede incluir métodos y funciones que describen el comportamiento de las clases.

# Modelos de datos y almacenamiento



## 3. Diccionario de datos. Ejemplo

Supongamos que estamos documentando una base de datos para una empresa de gestión de proyectos. Utilizamos un diccionario de datos para describir los elementos de la base de datos. Ejemplo simplificado:

Tablas:

- Proyectos
- Empleados
- Tareas

Atributos (ejemplo de la tabla "Proyectos"):

- ID (Clave Primaria)
- Nombre del Proyecto
- Fecha de Inicio
- Estado del Proyecto

Relaciones (ejemplo de la tabla "Tareas"):

- Una tarea está relacionada con un proyecto (relación "Pertenece a Proyecto").
- Una tarea está asignada a un empleado (relación "Asignada a Empleado").

Este diccionario de datos proporciona una representación tabular de las tablas y sus atributos, incluyendo detalles como las claves primarias y las relaciones entre tablas. También puede incluir descripciones de atributos, restricciones (como valores únicos) y otros metadatos que son útiles para entender y gestionar la base de datos.

# Modelos de datos y almacenamiento

Taxonomía de modelos de datos: La taxonomía de modelos de datos es útil para comprender y categorizar los diversos enfoques utilizados para representar y almacenar datos en sistemas de información.

Tipo	Modelo de datos	Técnica
Basada en registros	<ul style="list-style-type: none"><li>• Jerárquico: Los datos se almacenan en registros. Los registros tienen estructura de árbol donde cada registro tiene un único <i>padre</i> (nodo superior)</li><li>• Red: Elaborado sobre un modelo jerárquico pero permitiendo que los registros tengan múltiples padres</li><li>• Registros: Se especifica la estructura de toda la base de datos, definiendo los tipos de registros. Tienen un número determinado de campos con una longitud fijada.</li><li>• Multidimensional: la estructura de datos contenida en la base de datos está mezclada con los propios datos. Es útil para fuentes de datos basadas en la web y para relacionar bases de datos de diferentes tipos.</li></ul>	ERD, Dic
Relacional	<ul style="list-style-type: none"><li>• Relacional: los datos están segmentados con la ayuda de tablas. Se emplea <i>Structured Query Language</i> (SQL) como su lenguaje canónico de base de datos.</li><li>• Modelo de entidades-relación: describe la relación entre cosas de interés en un dominio de conocimiento. Un modelo básico ER está compuesto por tipo de entidades y especifica las relaciones que pueden existir entre estas entidades.</li><li>• Modelo de datos relacional extendido (ERDM) es un híbrido del modelo relacional añadiendo funcionalidades de modelos orientados a objetos.</li></ul>	ERD, Dic

# Modelos de datos y almacenamiento

Taxonomía de modelos de datos: Ayuda a los profesionales de la informática y las bases de datos a identificar cuál es el modelo más adecuado para una tarea o aplicación específica.

	Orientados a objetos.	
Basado en objetos	<ul style="list-style-type: none"><li>• Modelos orientados a objetos: consiste de objetos que poseen sus características y métodos. Estos modelos se pueden considerar que son post relacionales debido a que no se limitan a tablas aunque las empleen.</li><li>• Modelo dato por contexto: este modelo incorpora varios modelos de datos según se necesitan, como relacional, orientado a objetos, semi estructurado entre otros. Este modelo permite varios tipos de usuarios que se diferencia en el modo de interactuar con la base de datos</li></ul>	UML ER
NoSQL	<ul style="list-style-type: none"><li>• Modelo de grafo. Se emplea una estructura de grafo con nodos, aristas y propiedades para representar los datos almacenados.</li><li>• Modelo dato multievaluado: Este modelo permite respecto al modelo relacional que cada atributo en lugar de contener un dato unitario almacene una lista de datos.</li><li>• Modelo de datos documento: este tipo permite almacenar y gestionar documento o dato semi-estructurados mas que datos atómicos.</li></ul>	Diccionario

Tabla 1.4: Taxonomía de modelos de datos



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Para el análisis de datos se aplican técnicas de aprendizaje automático o machine learning.

En el aprendizaje automático, cuando analizamos datos, tenemos dos tipos de información:

**Atributos (Características):** Estos son los detalles o características específicas de los datos que estamos estudiando. Por ejemplo, si estamos analizando información sobre coches, los atributos podrían incluir el color, la marca, el modelo y la velocidad máxima de cada coche.

**Variable Respuesta (Etiqueta):** Esta es la información que queremos predecir o entender. En nuestro ejemplo de coches, la variable respuesta podría ser si un coche es "seguro" o "peligroso" en función de sus atributos.



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

**Cómo clasificamos los problemas de aprendizaje automático depende de si tenemos o no la variable respuesta (etiqueta) en nuestros datos:**

**Aprendizaje Supervisado:** Cuando tenemos la variable respuesta en nuestros datos, estamos en un problema de aprendizaje supervisado. Significa que podemos entrenar a un modelo de aprendizaje automático utilizando los atributos y las etiquetas conocidas para predecir o entender futuros datos.

**Aprendizaje No Supervisado:** Si no tenemos la variable respuesta en nuestros datos, estamos en un problema de aprendizaje no supervisado. En este caso, estamos buscando patrones o estructuras ocultas en los datos sin utilizar etiquetas preexistentes.

La diferencia principal es si tenemos o no la información que queremos predecir o entender (la variable respuesta) en nuestros datos. Si la tenemos, es un problema de aprendizaje supervisado; si no, es un problema de aprendizaje no supervisado.



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



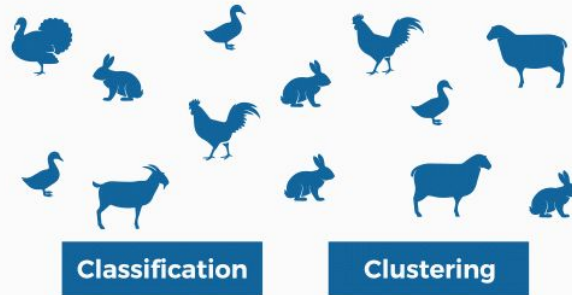
Los tres tipos de problemas que aparecen son:

**Clustering.** En este tipo de problema, tratamos de identificar grupos o conjuntos de datos que se parecen entre → patrones. Esto es, encontrar subconjuntos de observaciones que son similares entre sí. Por ejemplo, si tenemos información sobre personas y queremos agruparlas en diferentes categorías basadas en sus intereses

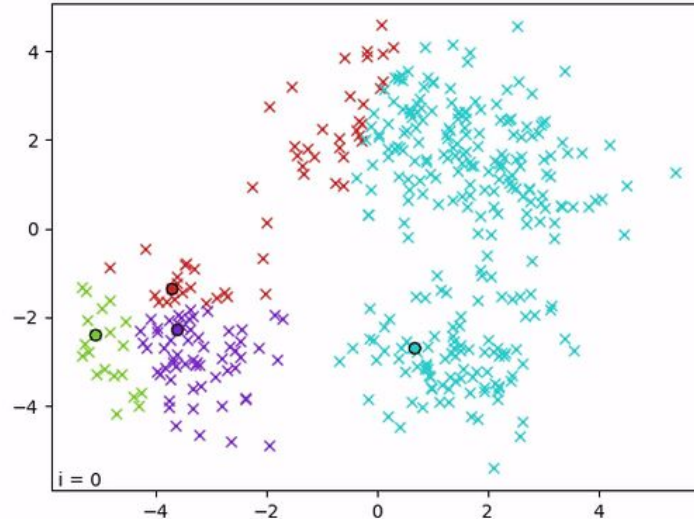
**Regresión.** Cuando tenemos un problema de regresión, lo que intentamos hacer es predecir un valor numérico o continuo basado en ciertas variables que ya conocemos. Por ejemplo, si tenemos información sobre el tamaño de casas (variable x) y sus precios (variable y), queremos encontrar una fórmula o modelo que nos permita predecir el precio de una casa en función de su tamaño.

**Clasificación.** En un problema de clasificación, nuestro objetivo es asignar una etiqueta o categoría a un conjunto de datos en función de ciertas características conocidas. Por ejemplo, si tenemos información sobre correos electrónicos y queremos etiquetarlos como "spam" o "no spam" en función de su contenido y otras características

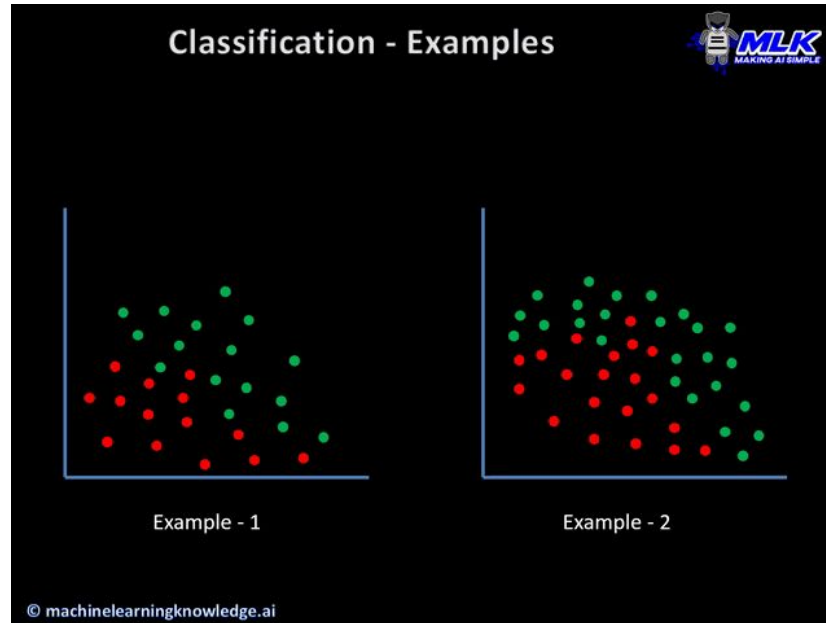
# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



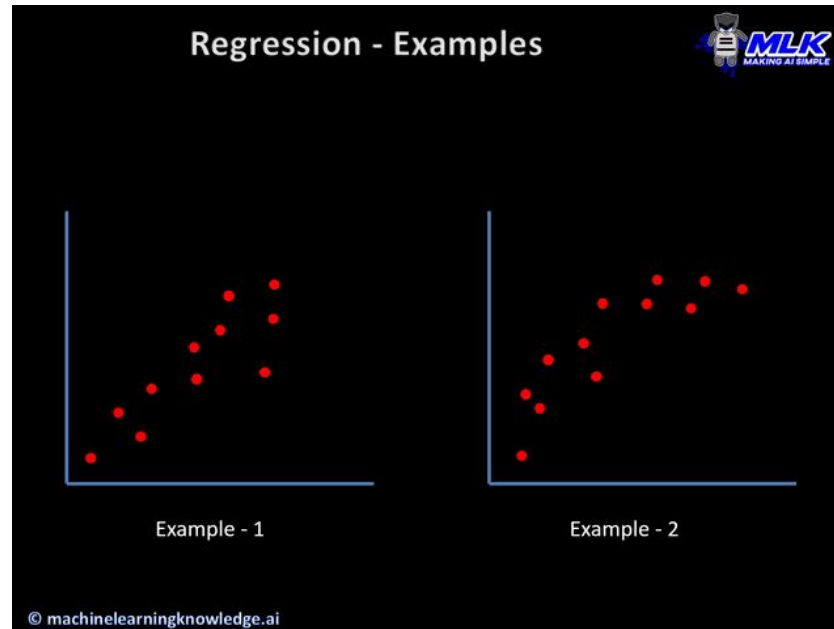
# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados





## Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Las técnicas de aprendizaje automático se aplican tanto a datos estructurados como no estructurados. La diferencia esencial es que en los datos estructurados las características  $x$  están completamente definidas a partir de los datos iniciales mientras que en los datos no estructurados hay que recurrir a procedimientos para extraer automáticamente estas características.

## Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



Figura 1.5: Clasificación de técnicas de extracción de información en datos no estructurados.



## Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

El análisis de cada tipo de dato no estructurado es en sí una disciplina distinta apareciendo el denominado procesamiento del lenguaje natural o minería de texto (texto), reconocimiento del habla (audio), reconocimiento de imágenes y procesamiento de vídeos.





# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- Aplicación en el análisis de texto incluyen:

**1. Análisis de sentimientos.** Estas técnicas analizan automáticamente el texto en busca del sentimiento de quien lo escribe (positivo, negativo, neutro, etc.) Estas técnicas permiten a las empresas analizar miles de reseñas en línea o comentarios en las redes sociales sobre ciertos productos en cuestión de minutos.

**2. El reconocimiento de entidades con nombre.** Se busca localizar e identificar las entidades con nombre dentro de un texto en categorías predefinidas como organizaciones, lugares, valores monetarios, abreviaturas, etc.

**3. Extracción de eventos.** Ejemplos de estas tareas son detectar si los eventos del mundo real han sido reportados en artículos y posts o el seguimiento de acontecimientos similares en diferentes textos.

**4. Extracción de relaciones.** Construcción de una base de datos con las interacciones entre fármacos a partir del análisis de texto bruto, o la determinación de relaciones entre personas con el objetivo de construir una base de conocimiento.



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- En el análisis de imágenes algunos de los problemas estudiados son:
  - 1. Extracción de textos y objetos**, como la matrícula de un vehículo o determinar la existencia de células cancerígenas en una imagen médica.
  - 2. Entendimiento de imágenes.** Ejemplos en esta categoría son la clasificación semántica de las imágenes.
  - 3. Análisis de imágenes geoespaciales** que van desde la distribución de cultivos a la localización de la pobreza.
  - 4. Reconocimiento facial**, permite identificar la persona de una fotografía.



## Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- Respecto al audio:

**1. Sistemas de reconocimiento de audio** que transforma el audio a formato de texto. Sistemas como Google Assistant, Siri, Alexa, Cortana, etc. implementan este tipo de sistemas.

**2. Extracción de características.** El habla contiene características prosódicas como el tono, velocidad, calidad, etc y su extracción suministra información sobre el emisor del mensaje. Hay sistemas que extraen estas características de las conversaciones en tiempo real.



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- En el análisis de videos destacan:

**1. Creación de resúmenes automáticos en formato de texto o de imágenes** en los que se identifican los momentos más destacados, por ejemplo, momentos en los el público aplaude o se pronuncian ciertas palabras clave. Estas técnicas permiten la navegación sobre los vídeos como si se trataran de colecciones de documentos escritos. También permiten la generación automática de subtítulos.

**2. Reconocimiento de lugares, objeto o acciones.** Ejemplos de esta categoría son el reconocimiento automático de una infracción de tráfico o el allanamiento de la morada en un sistema de videovigilancia.



# Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Realizar actividad: **Sentiment Analysis: Concept, Analysis and Applications**

<https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>



# Visualización

En el proceso de obtener información de los datos, la última etapa es la llamada visualización, esta debe ser entendida como una capa entre los resultados obtenidos y el decisor de modo que se facilite la interpretación y evaluación de los resultados.

**La visualización** es mucho más que una herramienta para comunicar resultados de una forma rápida y objetiva, esta permite además descubrir y comprender los patrones que se encuentran detrás de un conjunto de datos.

La visualización de datos es la parte más del proceso para corroborar hipótesis sobre fenómeno a estudiar, buscando fundamentalmente explicar los datos existentes y realizar predicciones sobre nuevos datos. Con la irrupción de BD, se ha expandido estos propósitos introduciendo aspectos exploratorios y descubrimiento de patrones, apareciendo aspectos como:

- **Resumir** bases de datos masivas para facilitar la toma de decisiones.
- **Identificación interactiva de patrones**
- **Identificación de datos** relevantes para un determinado fenómeno.



# Visualización

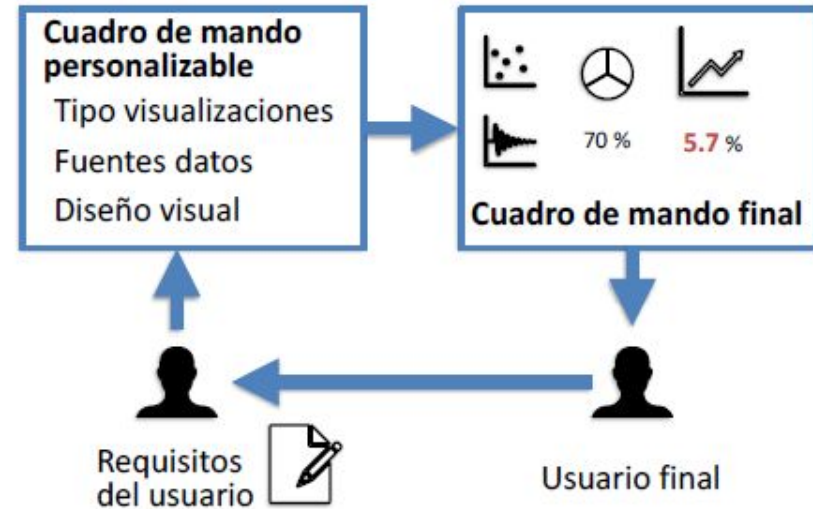
La aparición de los datos no estructurados en la era de BD ha conducido a que las técnicas de visualización deban abordar las características de las 3 Vs inherentes en su definición:

- 1. Volumen.** La gran cantidad de datos requiere a los métodos tener la posibilidad de la identificación de datos relevantes al fenómeno en cuestión. Un aspecto esencial es determinar el nivel adecuado de agregación que visualice los aspectos esenciales.
- 2. Variedad.** Debido a la existencia de datos no estructurados provenientes de múltiples fuentes se requiere la integración de los mismos en una visión de análisis.
- 3. Velocidad.** Exige la recolección y análisis en tiempo real, exigiéndoles a los métodos la inmediatez pero a la vez generando información útil a la organización.

## Visualización

Los cuadros de mando integrales son visualizaciones que permiten monitorizar los parámetros esenciales de la empresa y disponer de una imagen en tiempo real de lo que ocurre dentro y fuera de la misma. Se puede considerar una metodología de gestión estratégica utilizada para definir y hacer seguimiento a la estrategia de una organización, de tal forma que permita describir, comunicar y medir la estrategia y evaluar los resultados.

El método más sencillo para personalizar un cuadro de mando es el uso de asistentes de configuración que utilizan interfaces gráficas de usuario que facilitan la selección de los widgets y de los datos a mostrar.



**Figura 1.6:** Flujo de trabajo para la elaboración de cuadros de mando





# Visualización

Las características esenciales para el diseño de un cuadro de mando son:

- 1. KPIs:** se debe seleccionar los KPIs esenciales (entre 7 y 10) para el negocio.
- 2. Visualización:** deben ser fácilmente interpretables, hablar el mismo lenguaje del decisor y su representación gráfica la adecuada para los datos que representa y visualmente atractivos.
- 3. Análisis:** además de las KPIs el cuadro de mando debe acompañar de un análisis sobre lo ocurrido, recomendaciones y su potencial impacto sobre el negocio.