

# CURSO BIG DATA

## Contenido

MODULO 1 – Business intelligence y Advanced Analytics .....	3
DIA 1 – Ecosistema DATA .....	3
DIA 2 – Tecnología I: Transformación digital.....	3
DIA 3 – Business Intelligence .....	4
EJERCICIO: Kaggle y Looker Studio .....	6
DIA 4 – Tecnología II: Manipulación de datos. ....	6
EJERCICIOS: MySQL y Looker Studio.....	9
DIA 5 – Big Data, IA y Machine Learning .....	9
EJERCICIOS: Data Miner.....	11
MODULO 2 – Web Analytics.....	13
DIA 6 – Web Analytics .....	13
EJERCICIO 1: crea tu proyecto (definición teórica).....	13
EJERCICIO 2: análisis de competidores.....	14
EJERCICIO 3: DAFO de Samsung, DAFO de EPIs y material sanitario. ....	14
EJERCICIO 4: establecimiento de objetivos genéricos.....	15
EJERCICIO 5: establecimiento de objetivos específicos.....	15
EJERCICIO 6: establecer 5 métricas para analizar en el proyecto y 5 indicadores clave. ....	16
DIA 7 – Web Analytics .....	16
EJERCICIO 1: crear un e-commerce "SX" en el CMS Ecwid, enlazarlo con GA4 y añadir Google Ads. ....	17
EJERCICIO 2: diseñar la organización en GA4 de una cuenta para el Ecommerce SX. Decidir cuántas cuentas, propiedades y flujos de datos se necesitan. ....	18
EJERCICIO 3: analizar los informes de Usuarios con la cuenta de demostración de Google Merchandise Store y responder a diferentes preguntas relacionadas con esos datos. ....	18
DIA 8 – Web Analytics II.....	21
EJERCICIO 1: Analizar las dimensiones y las métricas que aparecen en los diferentes informes de GA4	22
EJERCICIO 2: crear un informe en blanco y preparar las variables, dimensiones (fuente de tráfico/fuente de la sesión) y métricas (sesiones). ....	26
EJERCICIO 3: crear un informe para visualizar las sesiones (fuente de tráfico/fuente de la sesión). Probar con las diferentes visualizaciones y elegir la mejor para cada caso.....	26
DIA 9 – Web Analytics III.....	26
EJERCICIO 1: realizar búsquedas de informes de Google Merchandise Store con el buscador inteligente. ....	28
DIA 10 – Entornos Analítica Web .....	35

EJERCICIO 1: Diseño un embudo de venta para tu propio proyecto SX, utilizando todas las herramientas y recursos disponibles. ....	35
EJERCICIO 2: Componer los snippets para la home de nuestra web, teniendo en cuenta las directrices para escribir unos buenos snippets y un par de palabras clave a trabajar en SEO. ....	38
EJERCICIO 3: Diseña los siguientes enlaces desde uno de los contenidos de tu ecommerce a cada uno de los diferentes destinos: .....	39
EJERCICIO 4: Averigua la autoridad de dominio y de página de dos competidores de un mismo sector. ....	39
MODULO 3 – Data Management.....	44
DIA 11 – Introducción.....	44
EJERCICIO 1: trabajar con el Excel sales_data_sample.....	45
DIA 12 y 13 – SQL Segmento 1 .....	49
EJERCICIO 2: diferentes sentencias de SQL usando diferentes datasets .....	50
DIA 14 – SQL Segmento 2 .....	52
EJERCICIO 3: sentencias SQL utilizando funciones de agregación, agregaciones condicionales y funciones fecha. ....	54
DIA 15 – Segmento SQL 3 .....	58
EJERCICIO CLASE: Crear 2 tablas en Excel con formato .CSV con delimitador de “;” .....	58
EJERCICIOS SUBQUERIES .....	58
EJERCICIO JOINS.....	60
MODULO 4 – Data Fundamentals con Python .....	62
DIA 16 – Aprendiendo a pensar como un programador .....	62
EJERCICIO: BasicosPythonOriginal.ipynb, Repasillo Estructuras.ipynb .....	62
DIA 17 – Data Fundamentals .....	64
EJERCICIO: Strings y cadenas Victor.ipynb .....	67
DIA 18 – Data Fundamentals .....	69
EJERCICIO: Practica Kaggle.ipynb, Pandas_Practicas_Alejandro.ipynb.....	71
DIA 19 – Data Manipulation .....	72
EJERCICIO: Ejercicio guiado de EDA Panda_Practicas – chipotle.ipynb .....	73
DIA 20 – Data Fundamentals con Python.....	74
EJERCICIO: Ejercicio guiado de EDA Panda_Practicas – chipotle.ipynb .....	74
MODULO 5 – Data Science y Machine Learning.....	75
DIAS 21 y 22 – Exploración y análisis de datos en Python .....	75
EJERCICIO 1: definiciones y conceptos varios.....	75
PRÁCTICA 1: Estructuras de Datos en Python .....	76
EJERCICIO 2: diferencias entre ML, AI y DL.....	77
PRÁCTICA 2: Preprocesado de un dataset con Pandas .....	80
PRÁCTICA 3: Librería Numpy .....	82

PRÁCTICA 4: Librerías para el análisis de datos.....	83
DIAS 23, 24 y 25 – Fundamentos del Aprendizaje Automático.....	84
PRÁCTICA 1: Regresión .....	87
PRÁCTICA 2: Clasificación .....	88
PRÁCTICA 3: Modelos no supervisados.....	90
PRÁCTICA 4: NLP y Redes Neuronales.....	90
MODULO 6 - Arquitecturas Cloud y Big Data .....	92
DIA 26 – Arquitecturas Cloud y Big Data .....	92
EJERCICIO 1: Python Notebook .....	92
DIA 27 – Arquitecturas Cloud & Big Data .....	93
EJERCICIO 2: Condicionales .....	93
EJERCICIO 3: Bucles .....	93
DIA 28 – Arquitecturas Cloud & Big Data .....	93
EJERCICIO 3: Bucles .....	93
EJERCICIO 4: Programación Funcional.....	93
DIA 29 – Arquitecturas Cloud & Big Data .....	93
EJERCICIO 5: SparkSession Teoría.ipynb.....	94
EJERCICIO 6: Primer RDD Teoría.ipynb.....	94
DIA 30 – Arquitecturas Cloud & Big Data .....	94
EJERCICIO 7: Transformaciones y Acciones sobre RDDs .....	94
EJERCICIO SPACE X.....	95

## MODULO 1 – Business intelligence y Advanced Analytics

### DIA 1 – Ecosistema DATA

- Introducción y contexto
- Entendiendo los datos
- Big Data
- Ejercicios (sin uso de aplicaciones)

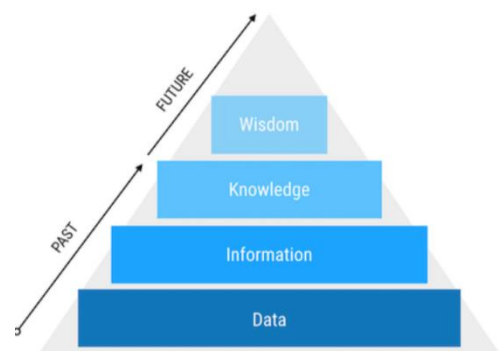
### DIA 2 – Tecnología I: Transformación digital

- [Google Trends](#) es una herramienta para explorar y comparar la tendencia de determinados términos de búsqueda y campos concretos en las consultas que han hecho los usuarios en el motor de búsqueda
- **Insight**: se utiliza en investigación de mercados, marketing, comunicación y en la empresa en general para referirse a un descubrimiento, una idea reveladora que nos da la clave para poder resolver un problema.

- **IoT:** la Internet de las cosas (IoT) describe la red de objetos físicos ("cosas") que llevan incorporados sensores, software y otras tecnologías con el fin de conectarse e intercambiar datos con otros dispositivos y sistemas a través de Internet. Estos dispositivos van desde objetos domésticos comunes hasta herramientas industriales sofisticadas.
- **Marketing Digital:**
  - SEO: todo el tráfico generado de forma orgánica proveniente de buscadores (google, yahoo, bing, duck duck go, etc.)
  - SEM/PPC: todo el tráfico generado por herramientas de publicidad online (google ads, bing ads, facebook ads, linkedin ads, etc.)
  - Social Media: tráfico generado de forma orgánica por las redes sociales en cualquier formato.
  - Marketing de afiliación: tráfico generado por una fuente externa a cambio de una compensación previamente acordada.
  - Email Marketing: tráfico generado a consecuencia del lanzamiento de campañas de email marketing.
- **Analítica digital:** Google Analytics es la herramienta más utilizada del mundo en el campo de la analítica web. Ofrece información agrupada del tráfico que llega a los sitios web según la audiencia, la adquisición, el comportamiento y las conversiones que se llevan a cabo en el sitio web.
- Búsquedas avanzadas en Google: uso de comandos INFO, SITE, OR, AND, "", INTENT, INTITLE
- <https://www.xataka.com/basics/25-codigos-funciones-trucos-para-buscar-google-exprimiendo-al-maximo-su-motor-busqueda>

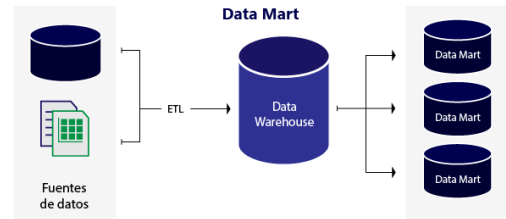
### DIA 3 – Business Intelligence

- **Estructura:**
  - **Datos:** los cuales por sí solos no aportan ninguna información y están dispersos por toda la organización y en distintos formatos.
  - **Información:** reunión de todos los datos en un formato en el que pueda leerlo. Toda la información que tengamos identificada, categorizada, etiquetada o calculada tras la recogida de datos.
  - BO: Business Operation
    - Estructura y orden básico aplicado a los datos
    - Tablas, documentos, listas, carpetas, etc.
  - BI: Business Intelligence
    - Estructura y orden aplicada a los datos
    - Data Warehouse, DataMart
  - **Conocimiento:** se deriva de las personas y es intangible y empírico.
  - **Decisiones:** implica el funcionamiento de un sistema de BI implementado que me permite tomar decisiones.
- Por el momento NO se está utilizando al 100%. No se tienen bien definidos los perfiles necesarios. Las empresas todavía lo están implementando.



- **Modelos de datos**

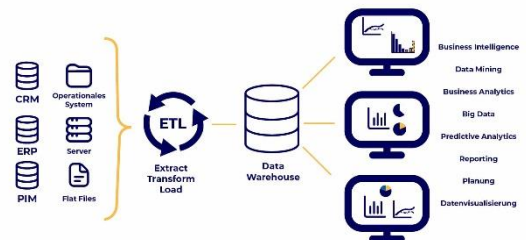
- **Data Warehouse:** almacenaje de datos a lo “bruto”
- **Data Mart:** almacenaje de datos más específicos



- Ej.: Data Warehouse = DM de Ventas + DM de Recursos Humanos + DM de Producción

- Estructura **OTLP**: Bases de datos transaccionales (las habituales); rápido de procesar, pero lento para analizar.
- Estructura **OLAP** (también denominada *Cubo OLAP*): análisis multidimensional de datos de forma veloz e interactiva. No está optimizado para transacciones; implican cargas pesadas en procesos. Formatos Estrella y Copo de Nieve

- **ETL**: extracción (datos a lo “bruto”), transformación y carga. Datos que no están optimizados. Existen procesos ELT: “más baratos” a priori.



- Las herramientas de BI nacen para dotarnos de mayor flexibilidad y homogeneidad ante la manera antigua de construir esos informes y tomar decisiones.

- **Tipos de salidas en BI:**

- **DSS**: son los llamados ‘Sistemas de soporte a la decisión’. Comprenden informes dinámicos y no requieren conocimientos técnicos. La información está dirigida y adecuada a cada perfil.
- **EIS**: son los llamados ‘Sistemas de información ejecutiva’. Ofrecen indicadores de negocio o KPI y permiten análisis de expectativas y por supuesto, apoyan la toma de decisiones empresariales.
- **CMI**: también llamados “Cuadro de mando integrales”. Orientados a la toma de decisiones por altos puestos directivos y agrupan todos los departamentos de la compañía.

- **Herramientas BI:** *Power BI - Microsoft* (líder), *Tableau* (2ª), *Qlik*.

- Google ha comprado *Looker* y lo ha integrado (*Google Cloud Platform*), con vistas a un futuro crecimiento potencial. Quejas por problemas y fallos en *Looker* (lo hemos sufrido en prácticas).

- Minería de datos: es el proceso final para interpretar y evaluar (ver módulo 5). Dos ramas:

- **Estadística clásica**: se utiliza principalmente con un fin puramente predictivo y para ello podemos hacer uso de: árboles de decisión, *clustering*, análisis de regresión, etc.
- **Moderna**: está basada en inteligencia artificial y aprendizaje automático (machine learning) y además de predecir, se usa para descubrir conocimiento. *Redes neuronales*, Agrupamiento *k-means*



EJERCICIO: Kaggle y Looker Studio

**E.2 Descarga el dataset de Kaggle ‘Netflix Movies and TV Shows’. Cárgalo en Data Studio e intenta responder a las preguntas que plantea.**

- Utilizar archivos CSV:
  - utilizar previamente una hoja de cálculo de Google para unificar formatos
  - No todos los archivos CSV están bien configurados
  - En tipo de archivo, tiene que poner Google Sheet.
- Fuente de datos: <https://www.kaggle.com/> Buscar datos con buena valoración. Utilizaremos para la práctica la base de datos de Netflix (<https://www.kaggle.com/datasets/shivamb/netflix-shows>)
- Conectar con <https://lookerstudio.google.com/datasources>
- Crear un informe genérico en Looker Studio
- Trastear añadiendo gráficos y controles

**Conexión entre tablas (Google Studio)**

- Ver video
- Se pueden conectar hasta 5
- Revisar conceptos SQL: LEFT JOIN, RIGHT JOIN, INNER JOIN (intersecciones)  
<https://programacionymas.com/blog/como-functiona-inner-left-right-full-join>

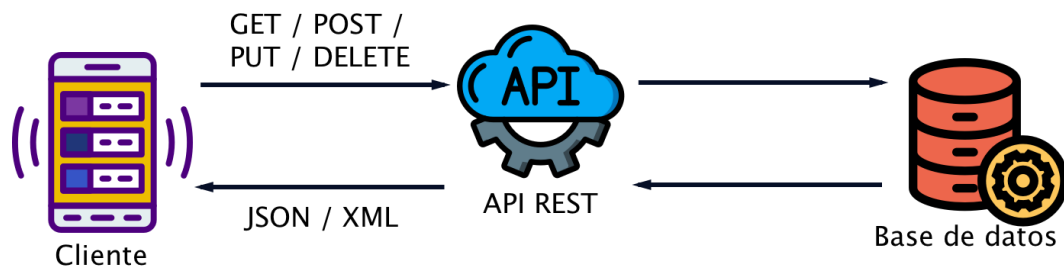
**Ejercicio: Big Data en Restauración (corrección al día siguiente).**

- Punto de partida
- Objetivos
- Recopilación de datos
- Análisis de datos
- KPIs
- Buyer Persona
- Estrategias de Marketing
- Soluciones

**DIA 4 – Tecnología II: Manipulación de datos.**

- Documentación: la documentación de un programa puede ser **interna y externa**.
  - La documentación **interna** es la contenida en líneas de comentarios.

- La documentación **externa** incluye análisis, diagramas de flujo y/o pseudocódigos, manuales de usuario con instrucciones para ejecutar el programa y para interpretar los resultados.
- Fundamentos de programación: tipos de datos, booleanos, variables, funciones, condicionales, bucles.
- Expresiones regulares: <https://regexr.com/> <https://regex101.com/>
- **JavaScript Object Notation (JSON):**
  - es un formato basado en texto estándar para representar datos estructurados en la sintaxis de objetos de JavaScript.
  - <https://www.json.org/json-es.html>
  - Utilizado para transmitir datos en aplicaciones web: enviar algunos datos desde el servidor al cliente, así estos datos pueden ser mostrados en páginas web, o viceversa.
  - Es muy común encontrarse en herramientas de analítica de datos de páginas webs y apps objetos JSON que lanzan **eventos** en cada acción que se quiera recoger, los cuales llevan incorporados muchos valores, llaves o parámetros adicionales a los que podemos acceder y extraer fácilmente para nuestro objetivo final.
- Una **API**, o interfaz de programación de aplicaciones:
  - Conjunto de reglas que definen cómo pueden las aplicaciones o los dispositivos conectarse y comunicarse -entre sí.
  - **API REST**: cumple los principios de diseño del estilo de arquitectura REST o transferencia de estado representacional. Por este motivo, las API REST a veces se conocen como API RESTful.
  - Capa intermedia entre datos (BD) y las aplicaciones web finales (cliente).



- <https://jsfiddle.net/> Test your JavaScript, CSS, HTML or CoffeeScript online with JSFiddle code editor.
- Ejercicio *json* “juego de tronos”

## Bases de datos - SQL

- La gestión de las bases de datos es fundamental para todos los trabajos de estas áreas.
- Un sistema de gestión de bases de datos (**SGBD**, por sus siglas en inglés) o **DataBase Management System (DBMS)** es una colección de software muy específico, orientado al manejo de base de datos, cuya función es servir de interfaz entre la base de datos, el usuario y las distintas aplicaciones utilizadas.
- Su uso permite realizar un mejor control a los administradores de sistemas y, por otro lado, también obtener mejores resultados a la hora de realizar consultas que ayuden a la gestión empresarial mediante la generación de la tan perseguida ventaja competitiva.



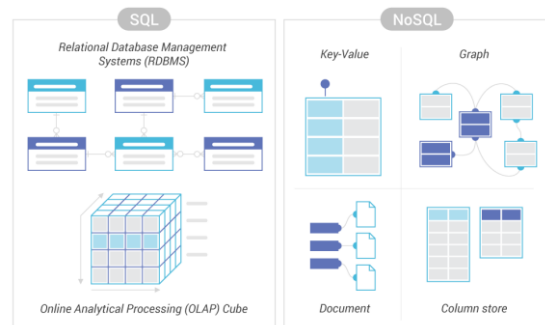
- Permite manejar los accesos diferenciados (identificación, seguridad) y permite interpretar las búsquedas para ingresar, modificar, invertir o suprimir datos.
- Se pueden diferenciar 2 grandes familias de **SGBD**: los SGBD SQL y los SGBD NoSQL.

- **SQL: Relacionales.**

- Esquema fijo y datos clasificados.
- Tipo y validez de datos muy importante
- Necesidad recurrente de escritura y modificaciones de datos sobre elementos específicos (SQL permite modificar fácilmente líneas específicas)
- Necesidad de búsquedas complejas

- **NoSQL: No relacionales. Modulares.**

- No necesitan esquema fijo.
- Necesidad de múltiples búsquedas de lectura.
- Grandes conjuntos de datos (Big Data)
- Datos distribuidos (varias fuentes)



- En programación solemos usar el término **CRUD** para referirnos a las operaciones básicas que puedes realizar sobre un conjunto de datos y por sus siglas son:
  - Crearlos: nuevos registros, insertar información.
  - Leerlos (Read): consultar esa información (un registro o una colección de estos registros).
  - Actualizarlos (Update): tomar un registro que ya existe en la base de datos y modificar alguna de las columnas.
  - Eliminarlos (Delete): tomar un registro y quitarlo del almacén.

- **Claves (Keys):**

- **Primaria o principal - PRIMARY KEY**
  - Identifica de forma única cada registro en una tabla.
  - Deben contener valores únicos y no pueden contener valores NULL.
  - Una tabla solo puede tener una clave principal, que puede consistir en campos simples o múltiples.
  - Una tabla NO tiene por qué tener una clave primaria, si bien es aconsejable.
- **Foránea - FOREIGN KEY**
  - Clave (campo de una columna) que sirve para relacionar dos tablas.
  - El campo **FOREIGN KEY** se relaciona o vincula con la **PRIMARY KEY** de otra tabla.
  - La tabla secundaria es la que contiene la **FOREIGN KEY** y la tabla principal contiene la **PRIMARY KEY**.
  - La **FOREIGN KEY** es una restricción que no permite que se agreguen o inserten datos que no válidos en la columna de foreign key, ya que los valores que se van a insertar deben ser valores que se encuentren o ya estén en la tabla con la que se quiere relacionar.

- **Sentencias SQL:**

- Create table
- Operaciones: Insert, Select, Update y Delete
- Condicionales: WHERE
  - Operadores lógicos: AND, OR, =, !=
  - Between
  - Like
- CASE – WHEN – ELSE – END



- Tablas cruzadas: LEFT JOIN, RIGHT JOIN, INNER JOIN,
- Campos calculados
- Funciones

### EJERCICIOS: MySQL y Looker Studio

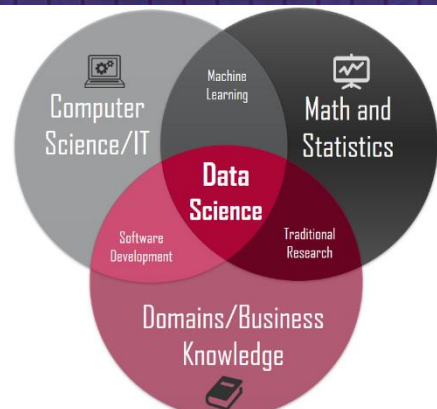
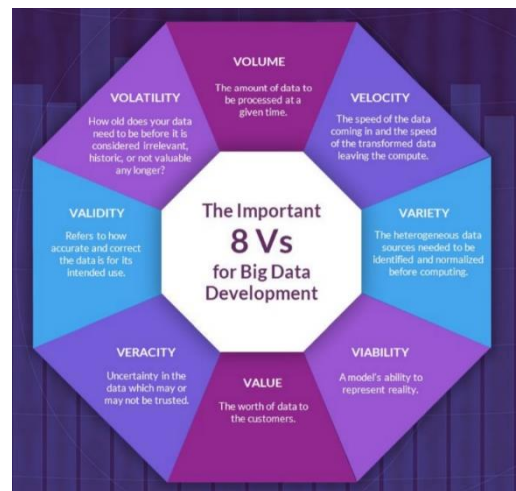
- 1- Instalación de MySQL y creación de dos tablas: Usuarios y Productos.
- 2- Ejecutar al menos dos cruzados de tablas para asignar los productos creados por cada uno de los usuarios teniendo en cuenta el ID (id de usuario como foreign key en producto)
  - a. JOIN :  
<https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58&punto=64&inicio>
- 3- Mejora el dashboard que hiciste ayer con expresiones regulares y campos calculados en Looker Studio

### Looker Studio. Ejercicios

- Funciones <https://support.google.com/looker-studio/table/6379764?hl=es>
- Expresiones regulares [https://support.google.com/looker-studio/answer/10496674?hl=es&ref\\_topic=7570421](https://support.google.com/looker-studio/answer/10496674?hl=es&ref_topic=7570421)
- Ejercicio: regiones del mundo
  - Añadir campo/dimensión en editor de fórmulas y fórmula CASE
  - Añadir métrica en editor de fórmulas y fórmula SUM

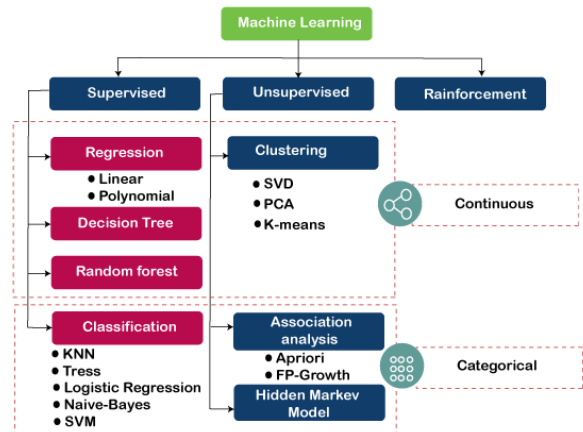
### DIA 5 – Big Data, IA y Machine Learning

- El **Big Data** no es más que un campo dentro de todo lo que conocemos como ciencias de la computación o computer science.
- **3 elementos básicos** que conforman el Big Data:
  - Velocidad a la que se consume la información.
  - Variedad de información.
  - Volumen de información.
  - Pero todo esto, no nos sirve de nada si no incorporamos un componente de: VALOR. Si los datos no sirven para aprender, descubrir o analizar, todo lo anterior no sirve de nada.
  - Viabilidad, Veracidad, Validez, Volatilidad.
- Es importante no caer en el *hype* del marketing y la innovación asociado al término de Big Data.
- **Data Science: AI y Machine Learning**
- Revisar las diferencias entre Big Data, IA, Machine Learning y Deep Learning (repaso de unidades 1 y 2)
- Cuando hablamos de **ML**, lo que aporta valor e importancia a su existencia y desarrollo es su capacidad para poder aplicar modelos predictivos → revolución de esta disciplina, enmarcada dentro de la amplia **IA**.



- Ejemplos de aplicaciones y de su crecimiento: Tik Tok, LinkedIn, etc.
- Ejemplos de publicidad programática: la compra programática nos permite el uso de data de los usuarios (histórico y en tiempo real) y nos capacita para abordar la personalización del mensaje publicitario. Esto se traduce en dos grandes ventajas: Posibilidad de segmentar de manera muy precisa el público objetivo al que queremos impactar.

- El ML basa toda su potencia en la aplicación de algoritmos para aplicar el aprendizaje automático y desarrollar así modelos predictivos. Algoritmos supervisados y no supervisados. Ejemplos.



- **Aprendizaje supervisado**: modelos más utilizados dentro del ML. Para ello, necesitamos contar con unas entradas y salidas de la información sobre la cual vamos a aplicar el aprendizaje automático.

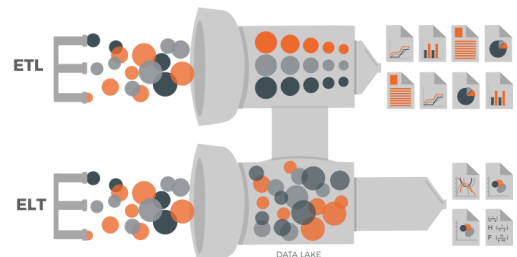
- **Clasificación**: elige entre una lista de opciones previamente definidas y limitada.
- **Regresión**: predecir números reales o números con infinitas posibilidades.

- **Aprendizaje NO supervisado**: los algoritmos se aplican para aprender de datos con elementos no etiquetados buscando patrones o relaciones entre ellos. En este caso no necesitarían delimitar el número de entradas y salidas.

- **Clustering**: clasifica en grupos los datos de salida. Concepto de *centroide*
- **Asociación**: descubre reglas dentro del conjunto de datos.

- **ELT (Extraer, cargar, transformar)**: método diferente de acercarse al flujo de datos, en el que los datos extraídos se cargan primero en el sistema de destino.

- Las transformaciones se realizan después de que carguemos los datos en el almacén de datos.
- Los datos primero se copian en el *data lake* y luego se transforman in situ.
- Funciona bien cuando el sistema objetivo es lo suficientemente potente como para manejar transformaciones a gran escala.
- ELT generalmente se usa con bases de datos NOSQL, un dispositivo de datos o una instalación en la nube.



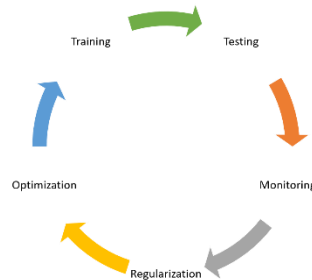
- **Data Lake**: para almacenar datos de forma no estructurada (ej: texto, imágenes, artículos, correos, etc.). Repositorio central.

- Ventajas: más flexibles, no diseñados de antemano.
- Ejemplo: Amazon S3

- ¿Ventajas e inconvenientes Data Lake VS Data Warehouse? ¿ETL VS ELT?

- Depende de coste, pero cada vez menos (cruzar con tamaño de la empresa)
- Depende del tipo y cantidad de datos.

- Depende del esfuerzo que tenga que hacer la empresa para transformar esos datos y darles valor.
- En un ELT se depende mucho de la tecnología que se utilice
- ML: Entendiendo el proceso:
  - Hacerse las preguntas adecuadas.
  - Identificar los datos y prepararlos.
  - Aplicación del algoritmo correcto.
  - Evaluación y ajuste de modelo.
  - Uso y presentación de modelo.
- Video aprendizaje supervisado VS no supervisado (VIDEO).
  - Concepto de *espacios latentes*.
  - El NO supervisado señala un camino muy prometedor
  - Cajas negras: algoritmos con matemáticas, estadísticas, etc.
- Videos de Redes Neuronales. Explicación de funcionamiento con el ejemplo de gafas VR + nachos



#### EJERCICIOS: Data Miner

Ejercicio 2: con la ayuda de la extensión 'Data Miner', haz un scraping y descarga los datos de las casas en venta en tu ciudad existentes en la web pisos.com. Ahora, haz una regresión múltiple en una hoja de Excel para predecir un precio en base a m2 y/o número de habitaciones.

- Uso de Data Miner en web [www.pisos.com](http://www.pisos.com)
- Exportar/grabar a Excel
- Realizar gráficos de dispersión

Recursos de aprendizaje

<https://www.codecademy.com/learn/paths/bi-data-analyst> Cursos

<https://8weeksqlchallenge.com/> Ejercicios SQL

(respuestas en:

[https://github.com/bcamandone/Data\\_Analysis\\_SQL/tree/main/8%20Week%20SQL%20Challenge](https://github.com/bcamandone/Data_Analysis_SQL/tree/main/8%20Week%20SQL%20Challenge))

VIDEO ejemplo Red Neuronal: <https://colab.research.google.com/drive/1QH7yhAmklHxBRi1d-dZcl3Y8uN5WNbnF?usp=sharing>

A screenshot of a web browser window displaying a confirmation message from a Zoho Public Forms evaluation. The browser's address bar shows the URL: forms.zohopublic.eu/thevalley/form/RedesExp016Lote1TestM1BIGDATAtoniGalbis/thankyou/formperma/Eku9eaxe8gtLSbferIv7-NWGVAtmCYQNkyOnZVFg?record... The page content is centered and features a green checkmark icon at the top. Below the icon, the text reads: "¡Hola Víctor Beneito López - Big Data!", "El resultado de tu evaluación de este módulo de es 10.00", and "Muchas gracias por tu participación." The browser's taskbar at the bottom shows various application icons, including WhatsApp, Calendar, and several office applications, along with the system clock indicating 17:13 on 03/02/2023.

## MODULO 2 – Web Analytics

### DIA 6 – Web Analytics

Profesor: Josema (Alicante). [josemathevalley@gmail.com](mailto:josemathevalley@gmail.com) <https://www.linkedin.com/in/josemarb/>

#### ANALÍTICA DIGITAL

##### INTRODUCCIÓN

##### Analítica en entornos digitales

- 3 tipos en base a la finalidad: descriptiva, predictiva o prescriptiva.
- Público objetivo: perfiles generacionales,

##### ¿Para qué sirve la analítica digital?

- Entender el tráfico web que recibe nuestro proyecto
- Definir mejores estrategias para comercializar nuestros productos/servicios
- Comprender el mercado y a nuestra competencia
- Optimizar nuestra estrategia de posicionamiento de marca
- Segmentar el público objetivo
- Tomar decisiones basadas en datos para modificar los procesos del negocio

##### EJERCICIO 1: crea tu proyecto (definición teórica).

- Dale un nombre a tu nuevo negocio
- Encuentra una necesidad que detectes en el mercado y piensa en un producto/servicio que la pueda satisfacer (puedes utilizar tu proyecto si cuentas ya con uno).
- Escoge los canales que vas a utilizar para llegar a tus clientes.
- Define a tu cliente ideal.

#### MEDICIÓN

##### Proceso de analítica digital

- Fase 1. Auditoría
- Fase 2. Estrategia
- Fase 3. Implementación
- Fase 4. Medición
- Fase 5. Optimización

Competidores	Página Web	Cantidad Landing Pages	*Keyword en Top3 Google	Perfiles Sociales	Tráfico web mensual	Velocidad carga web

Competidor 1						
--------------	--	--	--	--	--	--

Herramientas de medición: Keyword Sheeter; detecta qué términos podemos utilizar para realizar las búsquedas/mediciones.

## EJERCICIO 2: análisis de competidores

- Selecciona 3 de los principales competidores de tu sector
- Analiza sus recursos online clave. ¿Tienen web? ¿Tienen redes sociales? ¿Tienen un buen posicionamiento en Google?

## ENTORNOS ANALÍTICA DIGITAL

### Casos de uso:

- SEO: posicionamiento de nuestra web en motores de búsqueda como Google
- Social: relación de nuestra marca con nuestra comunidad de usuarios.
- Sales: entender y optimizar los procesos de venta en nuestro negocio.
- Web: medir y optimizar nuestra web; descubrir y comprender la intencionalidad de los usuarios que la visitan.

SEO: optimizar el posicionamiento de una página entre los resultados de los motores de búsqueda. Concepto <TopOfMind Brand> lo primero que se viene a la cabeza (ej: hamburguesería – McDonald's).

Uno de los objetivos principales de la analítica digital es permitirnos entender nuestra actividad en entornos digitales para poder tomar mejores decisiones. Una correcta decisión puede ahorrar enormes cantidades de recursos a una empresa, y permitirle obtener mayores beneficios.

**Herramientas:** análisis DAFO y benchmarking

EJERCICIO 3: DAFO de Samsung, DAFO de EPIs y material sanitario.

[https://drive.google.com/file/d/1eKyZi8GMuthl8srhkJgp4YDrwDN\\_5mud/view?usp=share\\_link](https://drive.google.com/file/d/1eKyZi8GMuthl8srhkJgp4YDrwDN_5mud/view?usp=share_link)

## BENCHMARKING

Podemos utilizar el benchmarking para comparar cualquier tipo de métrica, y resulta muy útil en analítica web para comparar algunos de los siguientes:

- Nº de visitantes únicos.
- Cantidad de enlaces que llevan a nuestra web.
- Tasa de rebote.
- Tráfico orgánico.
- Tráfico referido.
- Ventas mensuales.
- Seguidores en redes sociales.
- Nº de referencias por categoría de producto.

## OBJETIVOS

### IDENTIFICAR LOS OBJETIVOS

- Cuantitativos - Cualitativos
- SMART
- Recursos
- KPIs
- Tipos:
  - De negocio
  - De marketing & ecommerce

### EJERCICIO 4: establecimiento de objetivos genéricos

- Ejercicio 1: estableced al menos 3 objetivos en el muy corto plazo (1 mes) para cada proyecto.
- Ejercicio 2: estableced 2 objetivos en el medio plazo (hasta 1 año) para cada proyecto.
- Ejercicio 3: estableced al menos 1 objetivo a largo plazo (más de 1 año) para cada proyecto

### EJERCICIO 5: establecimiento de objetivos específicos

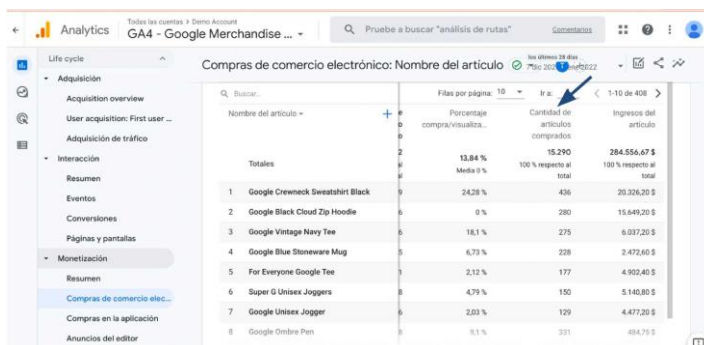
- Ejercicio 1: Estableced 5 objetivos de negocio para vuestros proyectos individuales.
- Ejercicio 2: Diseñad 5 objetivos de marketing para cada proyecto
- Ejercicio 3: Plantear 5 objetivos para cada ecommerce

## MÉTRICAS Y KPIs

- Las métricas son aquellos valores numéricos que sirven para analizar el rendimiento de una determinada acción o proceso dentro de un negocio.
- Si se puede medir, se puede mejorar.
- TODO KPI ES UNA MÉTRICA, PERO NO TODA MÉTRICA ES UN KPI

### MÉTRICAS aplicadas al marketing online - Ejemplos

- Coste de adquisición de cliente (CAC)
- Ciclo de vida del cliente o Life Time Value (LTV) – Gasto acumulado del cliente en 4 meses
- Retorno de la inversión (ROI):  $(\text{Ingresos} - \text{Inversión}) / \text{Inversión}$
- Tasa de conversión: % de acciones exitosas
- Ingresos



The screenshot shows the Google Analytics 'Compras de comercio electrónico' (E-commerce Purchases) report. The table lists various products with their respective metrics. A blue arrow points to the 'Porcentaje compra/visualiza...' column.

Nombre del artículo	Porcentaje compra/visualiza...	Cantidad de artículos comprados	Ingresos del artículo
Totales	11.84 % Media 0 %	15.290	284.556,47 \$
1 Google Crewneck Sweatshirt Black	24,28 %	436	20.326,20 \$
2 Google Black Cloud Zip Hoodie	0 %	280	15.649,20 \$
3 Google Vintage Navy Tee	18,1 %	275	6.037,20 \$
4 Google Blue Stoneware Mug	6,73 %	228	2.472,60 \$
5 For Everyone Google Tee	2,12 %	177	4.902,40 \$
6 Super G Unisex Joggers	4,79 %	150	5.140,80 \$
7 Google Unisex Jogger	2,03 %	129	4.477,20 \$
8 Google Ombre Pen	5,1 %	331	684,75 \$

- Leads o clientes potenciales: potenciales que nos dejan datos.
- Coste por lead (CPL)
- Número de visitas
- Tasa de clics o Click Trought Rate (CTR)
- Engagement: compromiso que se establece entre una marca y su audiencia en las distintas comunicaciones que producen entre sí.
- Tasa de rebote (Bounce Rate): porcentaje de visitantes que acceden a un sitio y salen sin acceder a nuevas páginas o interactuar con contenido, haciendo apenas una visualización de página. Entre mayor sea la tasa de rebote, peor es la interacción de los visitantes.
- Tasa de abandono de carritos.

#### KPI - Ejemplos

- Nº de ventas trimestrales
- Nuevos seguidores mensuales
- Tráfico web orgánico
- Tasa de rebote email marketing

EJERCICIO 6: establecer 5 métricas para analizar en el proyecto y 5 indicadores clave.

### DIA 7 – Web Analytics

#### GOOGLE ANALYTICS GA4

- Herramienta de análisis de tendencias de lo que los usuarios hacen con distintos dispositivos digitales.
- Este sistema a través de códigos en web/apps captura datos de lo que los usuarios hacen en ellas, los almacena en BBDD orientadas al Big Data y permite su consulta extrayendo sus datos o en informes y visualizaciones.
  - ¿**DÓNDE** MIDE? En cualquier sistema con conexión a internet Páginas web, Apps Mobile (iOS & Android), Juegos Mobile.
  - ¿**QUÉ** MIDE? Interacciones de los usuarios (datos independientes) que tenemos que analizar para buscar respuestas en estas mediciones y poder tomar así acciones que mejoren nuestro negocio.
- La analítica web también puede medir si los usuarios realizan las conversiones para los objetivos que te has propuesto:
  - ¿**Quién**? Quién está usando la web (ecommerce SX), app, juego, etc (definimos a quién observamos, las distintas tipologías, según su país de procedencia, o de qué campañas vienen)
  - ¿**Qué**? Qué interacciones está haciendo en nuestro ecommerce SX (Vemos que hace y lo medimos en GA4)
  - ¿**Por qué**? Analizamos el quién y el qué y evaluamos por qué lo hace .
  - **Insights**: claves/conclusiones para accionar el negocio: por qué el usuario decide realizar una acción u otra.



- **¿Desde qué área se analizan los datos?**
  - Marketing
  - SEO
  - UX/ Usabilidad
  - IT
  - Comunicación/RRSS
- **¿Para qué se analizan?**
  - DATOS: Obtener datos
  - ANÁLISIS: Entender lo que sucede
  - MEJORAR: actuar y/o tomar decisiones
  - APRENDER: alcanzar *insights*

## CREANDO UN PROYECTO WEB

- Para poder recopilar todos los datos con la información que deseamos analizar, debemos disponer de diferentes medios en los cuales implementar las diferentes herramientas que nos facilitarán la obtención.
- Estos medios a través de los cuales recopilaremos los datos se encontrarán en soportes digitales y pueden ser de diferente naturaleza (web, app, dispositivo conectado a internet, etc).
- Uno de los principales canales utilizados para recabar esta información es vía web, es decir, mediante la integración de herramientas de analítica en un sitio web para medir el tráfico y su comportamiento.
- Existen alternativas entre las que elegir para crear nuestro proyecto web, algunas más completas y flexibles y otras mucho más sencillas y rápidas de implementar.
- Podemos crear un proyecto web utilizando diferentes lenguajes de programación, siendo PHP, JAVA o PYTHON los más populares. Estos lenguajes nos permiten crear espacios web a medida, con infinitas posibilidades de personalización.
- También podemos utilizar cualquiera de los **CMS** existentes sin necesidad de tener nociones de programación. **CMS**: acrónimo de Content Management System. Sistema que nos permite crear, administrar y gestionar espacios web de manera sencilla, sin necesidad de saber cómo programar. Es una “herramienta informática” que nos facilita el diseño y creación de páginas web gracias a una interfaz amigable y a un entorno con el que es muy fácil familiarizarse.
- **Ejemplos**: Wordpress, Prestashop, Shopify, Ecwid, Magento, etc.

EJERCICIO 1: crear un e-commerce "SX" en el CMS Ecwid, enlazarlo con GA4 y añadir Google Ads.

- Crea tu propio e-commerce “SX” en la plataforma Ecwid: <http://www.ecwid.com>
- Cómo enlazarlo con GA4: [Acceso enlace con Google HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100449-Enabling-Google-Analytics-for-your-Ecwid-store" Analytics HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100449-Enabling-Google-Analytics-for-your-Ecwid-store" 4](https://support.ecwid.com/hc/en-us/articles/207100449-Enabling-Google-Analytics-for-your-Ecwid-store)
- Cómo añadir Google Ads: [Ecwid HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking" HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking" Help HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking" Center - Google HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-](https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking)

[Google-Ads-conversion-tracking"Ads, Conversion HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking" Tracking](https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking)

## RECOPIACIÓN DE DATOS DE ANALYTICS

- **Seguimiento** de un sitio web:
  - Crear una cuenta de GA4 y después agregar un código de seguimiento en el sitio web.
  - Cuando un usuario acceda a una página, el código recopila información anónima sobre cómo ha interactuado con la página y envía los datos a GA4.
- **Procesamiento e informes:**
  - Cuando este código recopila datos, los agrupa y los envía para que se elaboren informes.
  - Cuando GA4 procesa los datos, los agrega, clasifica y organiza por criterios.
- **Sesiones:** una sesión comienza cuando se accede a una página con el código de seguimiento de GA4 y termina tras 30 minutos de inactividad. Al volver a la página tras terminar una sesión, se inicia otra.
- **Cuentas, propiedades y flujos de datos:**
  - **Cuenta:** entidades a nivel legal, organizaciones, empresas.
  - **Propiedades:** modelos de negocio, unidades de trabajo, actividad comercial o marketing
  - **Flujos de datos:** distintas entradas de datos programadas por separado (cómo envío datos a la propiedad)

EJERCICIO 2: diseñar la organización en GA4 de una cuenta para el Ecommerce SX. Decidir cuántas cuentas, propiedades y flujos de datos se necesitan.

- Cuentas: sólo necesita una cuenta de empresa porque a nivel legal solo existe una entidad
- Propiedades: toda la actividad de la cuenta tiene el mismo objetivo, así que con sólo una propiedad basta
- Flujo de Datos: por cada desarrollo aplicamos un flujo, web o app

## INTERFAZ GOOGLE ANALYTICS 4 (VER PRESENTACIÓN – PANTALLAZOS)

- Cuentas y propiedades
- Informes, configuración y administración
  - **Adquisición** GA4: Cómo se han conseguido los usuarios y las visitas al ecommerce. Analizamos campañas de Marketing: Cuánta gente llega, qué hace, etc.
  - **Interacción** GA4: Podemos ver el detalle de todos los eventos que están llegando y en especial páginas /pantallas vistas. Es el sitio donde buscar acciones concretas.
  - **Monetización** GA4: Nos informa sobre el dinero que hemos ganado con nuestro ecommerce. Es el sitio donde ir a ver nuestros ingresos, info de productos, publicidad
  - **Retención** GA4: Nos habla de cohortes (grupos de usuarios analizados desde la perspectiva de su evolución) y visitas recurrentes de usuario. Es el sitio donde ir a ver cómo y cuándo los usuarios nos vuelven a visitar

EJERCICIO 3: analizar los informes de Usuarios con la cuenta de demostración de Google Merchandise Store y responder a diferentes preguntas relacionadas con esos datos.

Vamos a analizar los informes de Usuarios con la cuenta de demostración de Google Merchandise Store. Utilice el periodo comprendido entre el 1 de marzo del 2022 y el 31 de marzo del 2022 para responder a las siguientes preguntas:

1. Informe Resumen de datos demográficos, ¿qué grupo de idioma aportó el máximo número de usuarios? Inglés 56.014
2. Informe Información geográfica > Ubicación correspondiente a la India, ¿qué región aportó el máximo número de usuarios? Maharashtra 1431
3. Informe Resumen de dispositivos móviles, ¿qué porcentaje de sesiones procedió de los dispositivos móviles? Mobile 39,4%
4. En el informe de adquisición de tráfico > Fuente/Medio, al filtrar por tráfico referral, ¿qué sitio web envió el mayor número de usuarios? analytics.google.com / referral 1068
5. En el informe de adquisición de usuarios > Canales, ¿qué tres canales tuvieron el mayor número de usuarios nuevos? Direct 20.709, Organic Search 18.094, Paid Search 7.400
6. En el informe de adquisición de tráfico > en el caso de los usuarios que proceden de la fuente "google.com", ¿Cuál tuvo el máximo de sesiones? analytics.google.com / referral 1.686
7. En el informe de interacción > Páginas & pantallas, ¿cuál de estos títulos de página tiene más vistas de página: Shopping Cart, Home, Store Search results? Home 65.045 (22,51%)
8. En el informe Interacción > Páginas & pantallas, si nos centramos en el **grupo de contenido**, ¿qué páginas de destino tienen el porcentaje de vistas de página más alto? Men's Unisex (13.566) / No: not set. /Google+Redesign/Apparel/Mensv (10.694).
9. En el informe de monetización > Compras en comercio electrónico, ¿cuál es el mayor porcentaje carrito/visualizaciones por **marca** del artículo? Chrome Dino Collectible Figurines / No: Google (38.612)

**¿Qué es una Métrica?** Dato numérico exacto que dentro de una escala predefinida nos indica el volumen alcanzado de una serie de elementos en un periodo de tiempo. Recuerda: En GA4 Medimos Interacciones = Eventos.

**Práctica: busca las métricas en la cuenta de demostración de GA4:** Explorar – Iniciar una nueva exploración (en blanco) – Variables – Metricas + -

## EL ÁMBITO DE LOS DATOS

### Usuario:

- Identificación de las personas únicas a través de login, cookies, y otros sistemas
- En cada web y App eres un Usuario
- AGLUTINA SESIONES CON SUS EVENTOS ¿Cómo se creó el usuario? ¿Qué ha conseguido en su vida útil? ¿Está activo actualmente?
- **Identificación de usuarios:** USER ID, Google Signals, DEVICE ID.
  - 1. USER ID:
    - El Id de usuario de tu sistema (configuras un parámetro especial en tu código de seguimiento llamado User ID)
    - Identifica a tus usuarios cuando los tienes logueados en tu ecommerce sx. Cuando estás logueado el sistema siempre sabe que eres tú por el User Id.
    - ¿Por qué en una RRSS siempre recoges datos por User Id?
  - 2. GOOGLE SIGNALS: El id de usuario de Google.

- Google es capaz de gestionar a través de cookies en Google (publicidad, gmail, chrome, Android).
- No usa estos datos para dar datos personales de los usuarios, pero sí para mejorar lo que sabe de ellos.
- ¿Navegas en un chrome logueado?
- 3. DEVICE ID:Client Id (web)
  - El id automático que te asigna el código.
  - Cuando no tenemos User Id, el sistema crea uno por su cuenta(Es una cookie).
  - Si te cambias de dispositivo, no te puede identificar y te cuenta como otra persona.

#### **Sesión:**

- De cada persona se analizan sus navegaciones distintas en un dispositivo
- SE CONFORMA DE EVENTOS CONSECUTIVOS que se suceden en cada experiencia de uso
- ¿Cómo comenzó? ¿Qué consiguió? ¿Cuánto duró?

#### **Evento:**

- Cada interacción que hace el usuario es un evento
- TIENE LOS DATOS: clicks, páginas vistas, compras, ingresos.

#### **Métricas más importantes:**

- Número de eventos,
- Conversiones (Número de eventos + Conversión: todos nuestros KPIs van a crearse uniendo un evento a una conversión),
- Valor del evento,
- Sesiones (cuando cierras página o no interactúas en 30 min).
- Sesiones con interacción,
- Usuarios activos,
- Usuarios nuevos,
- Usuarios activos en N días
- Añadir al carrito, compras, ingresos
- Reembolsos, transacciones
- Visitas
- Salidas
- Clics de GAds, Coste GAds, Coste por clic GAd, impresiones GAds, Visualizaciones de video GAds

#### **Debate:** cómo lograr atraer tráfico web para tu Ecommerce SX.

- Ejemplo de TESCO en Corea el Sur: fotos con QR en el metro para venta online

## DIA 8 – Web Analytics II

### DIMENSIONES GA4

#### Analítica avanzada

Su enfoque es usar de manera precisa los datos de las distintas fuentes que tengamos para predecir posibles eventos y/o comportamientos que nos ayuden a afrontar cambios en el negocio

- Mejorar nuestra toma de decisiones al utilizar los datos de forma más concreta
- Automatizar procesos, ahorrándonos un tiempo valioso en tareas recurrentes además de minimizar costes
- Conseguir una mayor eficiencia al centrarnos en procesos de mayor importancia para nuestro negocio.

El uso de la analítica avanzada es esencial en la transformación digital de muchas empresas, debido a que, gracias al exhaustivo análisis de datos, podremos basar las decisiones que hagamos en información en tiempo real y no en suposiciones, instinto o puntos de vista.

Tenemos **3 fases en la analítica avanzada**:

- **Descriptiva**: a través del histórico de datos nos muestra qué ha pasado en la empresa y por qué. Así podremos visualizar una imagen general de lo que ha ocurrido hasta ahora de una manera sencilla y fácil de comprender.
- **Predictiva**: nos permite transformar esas métricas descriptivas en un conjunto de previsiones, pronósticos y tendencias muy precisas, que nos dirá lo que pasará en la empresa en un futuro.
- **Prescriptiva**: Podremos identificar las decisiones más óptimas teniendo en cuenta los grandes volúmenes de datos y las infinitas variables y restricciones que hemos generado anteriormente, pudiendo así automatizar el proceso de toma de decisiones. Este tipo de analítica utiliza inteligencia y capacidad de procesamiento para plantear propuestas, valorar todas las opciones posibles y finalmente seleccionar la más adecuada buscando el máximo rendimiento.

#### Dimensiones del evento

**Dimensiones**: cada sistema de clasificación de los datos que segmenta el total de los datos capturados en pequeños grupos a los que pone nombre. Las **dimensiones** son únicas para cada DATO. Dimensión + Métrica = Tabla de resultados.

Dimensiones más relevantes de un evento:

- Nombre del evento.
  - Los eventos marcados como conversión siguen teniendo el mismo nombre, pero pasan a poder verse en la dimensión “es conversión”, separados del resto.
- Día, Semana, Mes, Año
- Fuente, Medio, Campaña
- Campañas a nivel de usuario / Campañas a nivel de sesión
- País, Región, Ciudad
- Edad, Intereses, Sexo
- Dimensiones por geolocalización GA4
- Dimensiones por Google Signal GA4
- Dimensiones por Tecnología GA4

- Dimensiones por Página / Pantalla
- Dimensiones por Comercio Electrónico

## UNIENDO DIMENSIONES Y MÉTRICAS

### Informes con dimensiones y métricas

- Dimensión + Métrica = Tabla de resultados

EJERCICIO 1: Analizar las dimensiones y las métricas que aparecen en los diferentes informes de GA4

<https://support.google.com/analytics/answer/9143382?hl=es>

- Desarrolla 3 conclusiones para cada informe de Life Cycle.
  - Adquisición
    - Usuarios activos por semana en forma de gráfico:
      - Pico de usuarios el día 31 de enero
      - Máximo de usuarios en un día: 3.600
      - País con más usuarios. EE.UU (43.000€)
    - Usuarios nuevos por país y dispositivo
      - En España, 24 y sólo desktop
  - Interacción
  - Monetización
    - Numero de compras por semana: decreciendo
  - Retención
    - Usuarios recurrentes por ciudad: New York.
- Desarrolla 3 conclusiones para informe de Usuario.
  - Grupos demográficos
  - Tecnología
    - Usuarios recurrentes por sexo y dispositivo: más hombres y por el ordenador.
    - ¿Cuántos usuarios de Canadá con iPhone accedieron a la tienda de Google, durante noviembre y diciembre de 2022? 858 y 1261

## CENTRO DE ANÁLISIS DE GA4

Para editar o añadir comparaciones a un informe

- Las comparaciones le permiten evaluar subconjuntos de datos en paralelo (ej.: puede comparar todos los datos generados por dispositivos Android y dispositivos iOS):
  - 1) En la parte superior del informe, Haga clic en Añadir comparación.
  - 2) Seleccione Incluir o Excluir según quiera que la condición incluya o excluya los datos que determine.
  - 3) Haga clic en el campo Dimensión para seleccionar una dimensión (Plataforma).
  - 4) Haga clic en el campo Valores de dimensión para seleccionar un valor de dimensión o varios (Android o iOS).
  - 5) Haga clic en Añadir una condición nueva para seleccionar otra dimensión y otros valores para esa dimensión
  - 6) Haz clic en Aplicar.
- **Comparaciones:** Sólo se pueden comparar DIMENSIONES.
  - Para evaluar diversos valores para la misma condición, utilice el operador lógico "O"
  - Para evaluar diversas condiciones dentro de la misma comparación, utilice el operador lógico "Y"
- **Informes:** dimensiones en filas y/o columnas, métricas en valores.
- **Filas anidadas:** siempre se anida por la primera dimensión.
- **Tablas dinámicas:** cruzar con filas y columnas.

## INFORMES DE ADQUISICIÓN GA4

- Se usan dimensiones de tráfico para mostrar datos sobre cómo llegan los usuarios a su sitio web o aplicación móvil. Esta es la información que puede ver:
  - El número de usuarios nuevos y recurrentes
  - El número de sesiones y sesiones con interacción
  - La fuente y el medio desde los cuales se han iniciado las sesiones (por ejemplo, sesiones que empezaron desde un anuncio de Google)
  - El valor del tiempo de vida del cliente (TVC), que muestra el promedio de ingresos procedente de los usuarios nuevos durante sus primeros 120 días
- El informe Resumen de adquisiciones resume los datos de los informes para ayudarle a interpretar mejor cómo encuentran los usuarios su sitio o su aplicación. Esta información puede ser útil para analizar la eficacia de sus iniciativas de marketing.
- **Usuarios y usuarios nuevos**
  - Todas las personas que visitan su sitio o su aplicación, ya sean nuevos o recurrentes.
  - Los usuarios nuevos son aquellos que visitan su sitio o su aplicación por primera vez.
  - Analytics identifica a un usuario nuevo como cualquiera que no tenga una cookie de Google Analytics ni un Client-ID de su sitio, o bien un ID de instancia de su aplicación.
  - Aquellos que ya los tienen se denominan "usuarios recurrentes".
- **Valor del tiempo de vida del cliente (TVC)**
  - Muestra el promedio de ingresos procedentes de usuarios nuevos durante sus primeros 120 días.
  - Determina el valor de los usuarios en función de los ingresos adicionales que se generan.
  - Ejemplo: un usuario compra una camisa y a la semana siguiente compra unos pantalones.
  - El valor del tiempo de vida del cliente incluye ambas compras, lo que indica el éxito general de la adquisición.
- **Sesiones y sesiones con interacción**
  - Una sesión es un grupo de eventos que un usuario activa en un periodo determinado.
  - Las sesiones se inician cuando GA4 recoge un evento *session\_start* de su sitio o su aplicación.
  - Una sesión con interacción es una sesión que cumple al menos una de estas condiciones:

- Dura 10 segundos o más,
  - Ha tenido 1 evento de conversión o más,
  - Ha tenido al menos 2 vistas de una página
- Medir las interacciones de los usuarios en su sitio web o aplicación móvil.
  - La interacción incluye otras acciones aparte de las vistas de una página y las vistas de una pantalla.
  - Por ejemplo, puede medir cuándo los usuarios:
    - Se desplazan lentamente hacia abajo por una página, lo que indica que están leyendo
    - Ven los detalles de un producto o pasan tiempo en una página de producto
    - Ven un vídeo informativo
- Puede usar **eventos** para registrar cada interacción de los usuarios.
- **Adquisición de usuarios**
  - Muestra datos sobre usuarios nuevos.
  - Las dimensiones de tráfico (como el medio y la fuente) del informe incluyen las palabras "primer" y "usuario" para indicar que corresponden a usuarios nuevos.
  - "Primer medio del usuario" es el canal por el cual los usuarios nuevos llegan a su sitio o a su aplicación.
- **Adquisición de tráfico**
  - Muestra datos sobre sesiones nuevas.
  - Las dimensiones de tráfico (como el medio y la fuente) del informe incluyen la palabra "sesión" para indicar que corresponden a sesiones nuevas.
  - "Medio de la sesión" es el canal por el cual los usuarios nuevos y recurrentes llegan a su sitio o a su aplicación.
- **Campañas de Google Ads**
  - Muestra el número de usuarios que llegan a su sitio o a su aplicación desde cada uno de sus anuncios de Google.
  - Para acceder al informe, en el informe resumen de adquisiciones, haga clic en Ver campañas de Google Ads.
- **Métricas:**
  - Tiempo de interacción medio, que es el promedio de duración de las sesiones con interacción en su sitio web o aplicación
  - Sesiones con interacción por usuario, que es el número de sesiones con interacción dividido entre el número de usuarios
  - Porcentaje de interacciones, que es el número de sesiones con interacción dividido entre el número de sesiones
- **Fidelización de usuarios**
  - Los siguientes índices comparan la interacción en periodos breves con la interacción en periodos más amplios.
  - Cuanto más elevado sea el índice, mayores serán la interacción y la retención de usuarios.
  - Estos índices consisten en lo siguiente:
    - Usuarios activos al día (UAD): número de usuarios activos en las últimas 24 horas
    - Usuarios activos a la semana (UAS): número de usuarios activos en los últimos 7 días
    - Usuarios activos al mes (UAM): número de usuarios activos en los últimos 30 días
  - Por ejemplo, "UAD/UAM" indica el porcentaje de usuarios que han interactuado en las últimas 24 horas en comparación con los que han interactuado en los últimos 30 días.
- **Conversiones**
  - Las conversiones son actividades de usuario que contribuyen al éxito de su empresa.



- Estas actividades se miden usando eventos de conversión.
- Además de los eventos de conversión que se registran automáticamente, puede marcar como conversión hasta 30 eventos por propiedad.
- Una vez activados los eventos de conversión, el informe "Conversiones" muestra lo siguiente:
  - El número de veces que los usuarios activan cada evento de conversión
  - El número de usuarios que activan cada evento de conversión
  - Los ingresos procedentes de cada evento de conversión
- **Páginas y pantallas**
  - Se muestran las páginas del sitio web y las pantallas de la aplicación que los usuarios visitan, así como el número de usuarios que ven cada una de ellas.
  - Le permite medir los ingresos procedentes de compras, anuncios y suscripciones, así como el modo en que los usuarios interactúan con los artículos y los añaden al carrito.
  - También puede consultar los pasos del embudo de compra.
  - Los datos del tema Monetización pueden ayudarle a identificar el comportamiento de los consumidores y mejorar su estrategia de monetización.
  - En el informe Resumen de monetización puede ver un resumen de los datos que aparecen en los informes detallados del tema para consultar una vista unificada del comportamiento de los consumidores.
- **Compras en comercio electrónico:** muestran los datos de comercio electrónico de los eventos de venta online que implementa en su sitio web o aplicación móvil.
- **Cohortes**
  - Una cohorte es un grupo de usuarios que comparten una característica común que se identifica mediante una dimensión de Analytics.
  - Por ejemplo, todos los usuarios con la misma fecha de adquisición pertenecen a la misma cohorte.
  - En el tema Retención, los nuevos usuarios forman parte de la misma cohorte.
- **Retención de usuarios por cohorte**
  - El gráfico Retención de usuarios por cohorte muestra el porcentaje de usuarios nuevos que vuelven en su segundo y octavo día.
- **Interacción de los usuarios por cohorte**
  - El gráfico Interacción de los usuarios por cohorte muestra el tiempo medio de interacción de los usuarios nuevos que vuelven a su sitio o aplicación en su segundo y octavo día.
  - Analytics sólo incluye a los usuarios recurrentes en el tiempo de interacción medio.
- **Retención de usuarios**
  - En el gráfico Retención de usuarios se muestra el porcentaje de usuarios que regresan cada día durante sus primeros 42 días.
  - El gráfico empieza con una retención del 100 % cuando todos los usuarios visitan el sitio por primera vez. A medida que los usuarios abandonan el sitio, este porcentaje va disminuyendo desde el primer día.
- **Interacción de los usuarios**
  - En el gráfico Interacción de los usuarios se muestra el tiempo de interacción medio de los usuarios que vuelven en sus primeros 42 días.
  - Analytics sólo incluye a los usuarios recurrentes en el tiempo de interacción medio.

## CREAR NUEVOS INFORMES

- **Exploración:** consultas directas a las métricas y dimensiones de la propiedad
  - Definir lo que queremos

- Filtrarlo
- Cruzarlo (en tablas dinámicas)
- Buscarle una visualización
- Variables globales del informe (es siempre la misma para todos los informes):
  - Nombre
  - Rango de fechas
- Variables a usar en el informe:
  - Funcionan como una preselección
  - Podemos ir añadiendo las variables que vamos a necesitar para montar las distintas pestañas del informe.
  - Segmentos
  - Métricas
  - Dimensiones
- Tipología de pestaña (cambia sus opciones cuando cambiemos de técnica o de visualización):
  - Tipo de informe (explorador, embudo, flujo)
  - Opciones de visualización (tabla, gráfico de anillo, gráfico de líneas)
- Filas: definiremos con qué dimensiones formamos las filas de la tabla
- Valores: nos permite indicar las métricas que queremos ver.

EJERCICIO 2: crear un informe en blanco y preparar las variables, dimensiones (fuente de tráfico/fuente de la sesión) y métricas (sesiones).

EJERCICIO 3: crear un informe para visualizar las sesiones (fuente de tráfico/fuente de la sesión). Probar con las diferentes visualizaciones y elegir la mejor para cada caso.

- **Informes personalizados:**
  - Añadiendo varias dimensiones y métricas.
  - Visualizando dimensiones en grupo.
  - Creando tablas dinámicas.
  - Filtrando lo que se ve en los informes.
    - Las dimensiones y métricas se arrastran al cuadro de filtro y se selecciona el tipo de comparación y valor.

#### Looker Studio con Google Analytics:

- (Fallo: error de cuota)
- Importar datos para realizar consultas

## DIA 9 – Web Analytics III

### INTELIGENCIA ARTIFICIAL GA4

Google AI (Artificial intelligence) es la fusión de los equipos de investigación de Google, que combina Google Research <https://research.google/> y la división de Inteligencia Artificial <https://ai.google/>, para avanzar en machine learning o aprendizaje automático.

## ESTADÍSTICAS

El icono de estadísticas muestra tarjetas de estadísticas automatizadas (*Insights*) que proporciona Analytics Intelligence, un conjunto de funciones que utilizan el aprendizaje automático para comprender mejor los datos, detectar cambios inusuales y nuevas tendencias.

Puedes preguntar a Analytics Intelligence utilizando la barra de búsqueda superior, o hacer clic en "estadísticas" y seleccionar preguntas como las siguientes:

### Rendimiento básico:

- ¿Cuántos usuarios tuve la semana pasada?
- ¿Cuáles son mis mejores páginas y pantallas por visualizaciones?
- ¿Qué días he tenido el mayor número de usuarios?
- ¿Cuántos nuevos usuarios han llegado este año?
- ¿Cuáles son mis eventos principales por usuario?

### Grupos demográficos:

- ¿Qué idiomas son los que más emplean los usuarios?
- ¿De qué países proceden mis usuarios?
- ¿Cuáles son las ciudades principales por usuarios?
- ¿Cuáles son los mejores países por ingresos?

### Adquisición de usuarios:

- ¿Cuántos usuarios provienen de búsquedas orgánicas de los últimos 30 días?
- Comparar ingresos, usuarios provenientes de búsquedas orgánicas en comparación con búsquedas de pago

### Análisis del tráfico:

- Tendencia de usuarios mensuales el año anterior
- Aumento intermensual de usuarios

### Tecnología:

- ¿Qué dispositivos son los que más se usan?
- ¿Qué navegadores se utilizan más?
- ¿Qué versión de la aplicación utilizó la mayoría de los usuarios la semana pasada?
- ¿Qué plataformas son las que más se usan?

### Comercio electrónico:

- Tendencia de ingresos semanales durante los últimos 12 meses
- Ingresos por dispositivos este año
- Productos principales por ingresos
- ¿Cuáles son mis productos más vendidos?

## EL BUSCADOR INGELIGENTE

- Consiste en una herramienta de búsqueda que utiliza la interpretación del lenguaje de *Google* para entender lo que le estás solicitando.

- Descubrir sugerencias en el buscador inteligente de GA4

EJERCICIO 1: realizar búsquedas de informes de Google Merchandise Store con el buscador inteligente.

Para conocer:

- Ventas realizadas (ingresos): ingresos totales 16.824,69 US\$
- Nivel de tráfico: 16.412 usuarios
- Primeras visitas durante la última semana: 13.395
- Nuevos usuarios adquiridos desde España: 4
- Compras realizadas (transacciones/operaciones): 1587

## INSIGHTS AUTOMÁTICOS

- **Métricas predictivas:** detección de anomalías. GA4 detecta cambios en el tráfico aplicando modelos predictivos y ofrece estadísticas útiles.

## SEGUIMIENTO DE UNA CAMPAÑA DE MARKETING

Hay tres etiquetas de campaña que le ayudan a identificar información concreta del tráfico de la campaña: **Medio, Fuente y Campaña** son las etiquetas de campaña obligatorias.

Con las campañas podemos diferenciar el origen del tráfico que llega a nuestra web. Todo lo que sabes sobre la procedencia de tus usuarios se divide en 3 dimensiones principales:

Variables	Definición
Medio	¿Cómo? - Sistema/medio por el que entra el usuario al ecommerce. Informa del método o la forma de envío del mensaje al usuario. Puede incluir "email" si se trata de una campaña por correo electrónico, "cpc" en los anuncios de la búsqueda de pago o "social" para una red social. TIPOS: organic, cpc, referral, email, none
Fuente	¿De quién? ¿Dónde? - Fuente de la que proviene el usuario (dónde estaba antes de venir o quién nos ha traído el tráfico). Informa de la procedencia del usuario; puede ser una página web concreta o un enlace de un correo electrónico, y también puede diferenciar el tipo de medio. Si el medio es "cpc" (o tráfico de pago de "coste por clic"), la fuente puede ser "google", "bing" o "yahoo". Si el medio es "email", la fuente puede ser "boletín". TIPOS: google, facebook.com, spring_newsletter, direct
Campaña	¿Con qué acción? -Cuál es el nombre de tu acción de marketing que ha traído el tráfico. Puede informar del nombre de su campaña de marketing.

## TIPOS DE MEDIOS

- Todo el tráfico que diriges hacia tu web debe etiquetarse para poder conocer el origen/destino.
  - Directo / none: desconocido
  - Orgánico: click en un buscador
  - Referral: resto webs
  - CPC/DFP: campañas (de Google) de un producto que ha sido sincronizado con tu cuenta GA4

- Resto Medios (social media, etc.): resto de etiquetados de UTM (son fragmentos de texto que se agregan al final de las URL y permiten controlar el origen del tráfico que llega a una web) que habrá creado tu negocio con más o menos control.
- ¿Cómo decide GA4 que valores poner?
  - Referral: un dato que tu navegador puede darle a GA4 diciéndole de qué página vienes
  - El etiquetado de URLs: variables fijas que cuando se usan en las URLs ayudan a GA4 a saber qué campaña te ha traído hasta la web
  - Alterando el código de captura básico: configurando los distintos códigos de GA4 (JS, GTM, SDK) para que se recojan y traspasen esta información para que dispongas de ella.

## ATRIBUCIÓN DE CAMPAÑAS

- Agregar parámetros a las URL para identificar las campañas que refieren tráfico → puede recopilar información sobre la eficacia general de esas campañas y también comprender dónde las campañas son más efectivas.
- Cuando un usuario hace clic en un enlace de referencia, los parámetros que agrega se envían a Analytics y los datos relacionados están disponibles en los informes de Campañas.
- Los negocios de comercio electrónico pueden entender el comportamiento online de sus clientes y mejorar la comercialización de sus productos y servicios.
- GA4 compila estos datos en sus informes que podrá analizar a fondo para entender el comportamiento de los clientes y su proceso de compra.
- **Parámetros** a agregar a las URL (<https://ga-dev-tools.google/ga4/campaign-url-builder/>)
  - Obligatorios:
    - *utm\_source* - identifique el anunciante, sitio, publicación, etc. que está enviando tráfico a su propiedad
    - *utm\_medium* - el medio publicitario o de marketing
    - *utm\_campaign* - nombre de la campaña individual, eslogan, código de promoción, etc. para un producto.
  - Opcionales
    - *utm\_term* - identificar palabras clave de búsqueda de pago. Si está etiquetando manualmente campañas de palabras clave, también debe usar *utm\_term* para especificar la palabra clave.
    - *utm\_content* - Se utiliza para diferenciar contenidos similares, o enlaces dentro de un mismo anuncio.
    - *utm\_es* simplemente el prefijo requerido para estos parámetros.

## Tipos de campañas de marketing (se suelen combinar para mejorar las ventas y las conversiones):

- Anuncios de texto en resultados de búsqueda
- Anuncios de banner en sitios web de editores estratégicos
- Lanzar campañas en redes sociales
- Por correo que muestran su marca y productos a los clientes

## Realizando el seguimiento de una campaña de marketing

- Las campañas de marketing se controlan en Analytics con *etiquetas de campaña*: fragmentos de información adicionales que agrega a los enlaces de URL de sus materiales publicitarios o marketing online.
- Incluyen parámetros de seguimiento seguidos del signo igual y una palabra o varias unidas por guiones

- Cuando el usuario hace clic en un enlace con un parámetro agregado, el código de seguimiento de Analytics extrae la información del enlace y asocia el usuario y su comportamiento a la campaña de marketing.
- Así puede saber qué usuarios llegan al sitio web por sus actividades de marketing
- 3 ámbitos:
  - Ámbito Usuario - Nos indica la campaña con la se creó el usuario: evento *firts\_visit*
  - Ámbito Sesión - Nos indica la campaña con la que se creó la sesión y se mantiene incluso hasta que lleguen más campañas a la sesión: evento *session\_start*
  - Ámbito Evento de conversión - Sólo aplica a eventos marcados como *conversión + session start + firt\_visit*

## MODELOS DE ATRIBUCIÓN

- La atribución consiste en adjudicar un valor de conversión a los distintos anuncios, clics y factores que influyen en el recorrido de un usuario hasta que completa una conversión.
- Un modelo de atribución es una regla, un conjunto de reglas o un algoritmo basado en datos que determina cómo se asigna el valor de las conversiones a los distintos puntos de contacto de las rutas de conversión.
- Dos tipos de modelos de atribución en los informes de atribución de las propiedades GA4:
  - Modelos multicanal basados en reglas
    - Último clic multicanal
    - Primer clic multicanal
    - Lineal multicanal
  - Modelos de preferencia de Google Ads
    - Último clic de Google Ads
- Funcionamiento de la **atribución basada en datos**
  - La atribución usa algoritmos de aprendizaje automático para evaluar las rutas de conversión y sin conversión.
  - El modelo basado en datos resultante aprende cómo los distintos puntos de contacto influyen en los resultados de las conversiones.
  - El modelo incorpora factores como el tiempo transcurrido hasta la conversión, el tipo de dispositivo, el número de interacciones con el anuncio, el orden de exposición a los anuncios y el tipo de recursos de creatividad.
  - A través de un enfoque contrafáctico, el modelo contrasta lo que ha ocurrido con lo que podría haber ocurrido para determinar los puntos de contacto en los que es más probable que se completen conversiones.
  - El modelo atribuye el valor de contribución a la conversión a estos puntos de contacto en función de esta probabilidad.

## UTILIZANDO LA PUBLICIDAD

- Google Ads es el sistema publicitario de Google que permite generar anuncios de texto y display.
- Los anuncios de texto aparecen junto a resultados de búsqueda de Google relacionando palabras clave por las que puede pujar y búsquedas.
- Los de display constan de texto, imágenes animaciones o vídeos que aparecen en una gran colección de sitios web, la Red de Display de Google.

- <https://ads.google.com/>
- Si enlaza la cuenta de Google Analytics con la de Google Ads, podrá:
  - Ver datos de clic y coste de Google Ads junto a datos de interacción con el sitio web en Analytics,
  - Crear listas de remarketing en Analytics para usar en campañas de Google Ads,
  - Importar en Google Ads objetivos y transacciones de Analytics como conversiones,
  - Ver datos de interacción con los sitios web de Analytics en Google Ads
- Use el espacio de trabajo Publicidad de su propiedad GA4 para obtener más información sobre los recorridos más importantes de los usuarios.
- Los informes de esta sección le ayudan a:
  - Determinar más claramente cuál es el retorno de su inversión en medios de todos los canales
  - Tomar decisiones fundamentadas sobre la asignación de presupuesto
  - Evaluar los modelos de atribución.
- La atribución consiste en adjudicar un valor de conversión a los distintos anuncios, clics y factores que influyen en el recorrido de un usuario hasta que completa una conversión.
- Los informes de atribución del espacio de trabajo Publicidad le ayudan a saber cómo se complementan sus iniciativas publicitarias para conseguir conversiones. Estos informes permiten examinar diferentes modelos de atribución basados en reglas y determinar cuál podría ser el más adecuado para su empresa.
- Todos los usuarios de su propiedad GA4 pueden acceder al espacio de trabajo Publicidad.
- Los informes del espacio de trabajo tienen un aspecto y un funcionamiento ligeramente distintos a los de otros informes de GA4. El espacio de trabajo Publicidad ofrece tres informes por el momento:
  - Vista general de publicidad: consulte una vista general de las métricas de negocio y profundice en las áreas que quiera.
  - Comparación de modelos: compare la influencia de distintos modelos de atribución en la valoración de sus canales de marketing.
  - Rutas de conversión: vea qué rutas de conversión siguen sus clientes y descubra cuánto valor de conversión se atribuye a cada una según el modelo de atribución
- **Informes comparativos de modelos y rutas de atribución**
  - Informen Resumen de publicidad
  - Use el informe Comparación de modelos para comparar cómo influyen los diferentes modelos de atribución en la valoración de los canales de marketing.
  - Utilice el informe Rutas de conversión para conocer mejor las rutas de conversión de sus clientes y descubrir cómo se distribuye el valor de conversión en esas rutas según distintos modelos de atribución.

## CONFIGURANDO GA4

- Cuentas: entidades a nivel legal Organizaciones, Empresas
- Propiedades: modelos de negocio. Unidades de actividad comercial o de marketing
- Flujos de datos: distintas entradas de los datos programadas por separado
- Recogida de datos de GA4 permite:
  - Habilitar la recogida de datos de Google Signals
  - Configuración avanzada para habilitar la Personalización de Anuncios
  - Aceptar el consentimiento de recogida de datos de usuario
- Activar Google Signals

- Google Analytics recoge datos sobre el tráfico de tu web, además de los datos que ya recoge mediante la implementación estándar, para ofrecer funciones adicionales como estadísticas y audiencias multidispositivo.
- Cuando se habilita, Google Analytics recoge información de las visitas y la asocia a los datos que Google ya tiene de las cuentas de los usuarios que han iniciado sesión y han autorizado esta asociación para ver anuncios personalizados.
- <https://support.google.com/analytics/answer/7532985?hl=es#enabling&zipy=%2Csecciones-deeste-art%C3%ADculo>
- Esta información puede incluir la ubicación de los usuarios finales, el historial de búsqueda, el historial de YouTube y datos procedentes de sitios web asociados a Google, y se usa para proporcionar estadísticas agregadas y anónimas sobre cómo se comportan los usuarios en distintos dispositivos.
- Al habilitar estas funciones, aceptas la política sobre las Funciones publicitarias de Google, incluidas las reglas sobre categorías sensibles; confirmas que los usuarios finales te han otorgado los derechos y permisos de divulgación de información privada pertinentes para llevar a cabo esta asociación, y que dichos usuarios pueden acceder a estos datos y eliminarlos a través de la página
- <https://myactivity.google.com/myactivity?pli=1>

#### **Estrategia de etiquetado GA4**

- Añadir la etiqueta de Analytics a sus páginas web para que empiecen a aparecer datos en la nueva propiedad Google Analytics 4.
- Añadir la etiqueta a un creador de sitios web o a un sitio web alojado en un CMS (como WordPress, Shopify, etc.).
- Añadir la etiqueta global de sitio web directamente a las páginas web.
  - En la sección de instrucciones de etiquetado que se encuentra dentro de la pantalla de detalles del flujo de datos, encontrarás el código que debes copiar para pegarlo en la sección <head> de cada página web de la que quieras hacer un seguimiento.
  - Para verificar que la etiqueta funciona, accede a tu web y comprueba que la visita se registra en el informe "En tiempo real".
- Buscar su ID que empieza por "G-" (para cualquier plataforma que acepte un ID que empiece por "G-").
- Para desarrollar una estrategia de implementación de etiquetas, recomendamos los pasos siguientes:
  - 1. Decide qué etiquetas de tu sitio web puedes administrar mejor en el Administrador de etiquetas. Si es posible, te recomendamos migrar todas las etiquetas para que puedas administrarlo todo en un solo lugar y así evitar tener que hacer más cambios en el código de tu sitio web.
  - 2. Después, decide qué valores estáticos y dinámicos quieres transferir desde tu sitio web. Pueden ser datos de usuario, ingresos, datos de reserva o cualquier información personalizada que necesites para conocer el comportamiento de los usuarios. Te puede ser útil pensar que son "eventos", o acciones del usuario, que quieres obtener con el Administrador de etiquetas.
  - 3. Determina qué etiquetas te pueden dar los datos que necesitas basándote en las métricas y las dimensiones de tu plan de medición. Los requisitos de los datos deben estar relacionados con los objetivos empresariales generales, para que analices los datos adecuados.
- El protocolo de medición de GA4 permite hacer solicitudes HTTP para enviar eventos directamente a los servidores de Google Analytics, permitiendo medir la forma en que los usuarios interactúan con tu negocio desde cualquier entorno con HTTP habilitado.



- Puedes usar el Protocolo de medición para:
  - Asociar el comportamiento online y offline
  - Medir las interacciones del cliente y en el servidor.
  - Enviar eventos que tengan lugar fuera de la interacción estándar de los usuarios (por ejemplo, conversiones offline).
- En la guía del protocolo de medición de Google Analytics 4 para desarrolladores puedes consultar cómo enviar eventos, validarlos y verificar su implementación a través de HTTP con el Protocolo de medición.
- <https://developers.google.com/analytics/devguides/collection/protocol/ga4>

## GOOGLE TAG MANAGER

Añadir la etiqueta con **Google Tag Manager**: sistema de gestión de etiquetas que le permite actualizar en su sitio web o aplicación móvil, y de forma rápida y sencilla, códigos de seguimiento y fragmentos de código relacionados a los que se denomina de forma conjunta etiquetas.

- Una vez que haya añadido un pequeño fragmento de código de Tag Manager a su proyecto, podrá implementar opciones de configuración de analíticas y de etiquetas de medición desde una interfaz de usuario basada en la Web de forma sencilla y segura
- Google Tag Manager permite etiquetar las propiedades de Google Analytics 4 con dos etiquetas que funcionan conjuntamente:
  - Google Analytics: configuración de GA4
  - Google Analytics: evento de GA4
- Las etiquetas de evento de Google Analytics 4 permiten enviar eventos personalizados a Analytics, además de los eventos que se envían automáticamente o a través de la medición mejorada.

## MÉTRICAS PREDICTIVAS

- Google Analytics 4 enriquece los datos usando métricas predictivas, que utilizan aprendizaje automático para predecir el comportamiento futuro de los usuarios.
- Las métricas predictivas están disponibles en el generador de audiencias y en Explorar, y permiten recopilar los siguientes datos de eventos estructurados:
  - Probabilidad de compra: usuario que estuvo activo en los últimos 28 días registre un evento de conversión específico en los próximos 7 días
  - Probabilidad de abandono: un usuario que estuvo activo en la aplicación o web durante los últimos 7 días no lo esté en los próximos 7 días
  - Predicción de ingresos: Ingresos estimados de conversiones de compra en los próximos 28 días de un usuario que estuvo activo en los últimos 28 días
- Para poder entrenar modelos predictivos, la cuenta de Analytics debe cumplir los siguientes requisitos:
  - Tener al menos 1.000 ejemplos positivos y negativos de usuarios que activaron la condición predictiva durante un período de siete días.
  - La calidad del modelo debe mantenerse un período de tiempo suficiente para ser elegible.
  - Para ser elegible tanto para la probabilidad de compra como para la probabilidad de abandono, una propiedad tiene que enviar los eventos purchase y/o in\_app\_purchase (que se recopilan automáticamente).
- **Audiencia predictiva**
  - Una audiencia que cumple al menos una condición basada en una métrica predictiva, que indica una alta probabilidad de que un usuario realice una acción de conversión.

- GA4 permite crear y usar audiencias predictivas, que se comparten automáticamente con la cuenta de Google Ads vinculada
- GA4 proporciona plantillas de audiencias sugeridas como "compradores en los próximos 7 días", que incluye a los usuarios que por su comportamiento se puede prever que están a punto de realizar una compra.
- Las audiencias predictivas son especialmente útiles para segmentar las campañas de remarketing, priorizando los esfuerzos para llegar a usuarios con más probabilidad de realizar conversiones.
- Vinculación de Analytics con Google Ads, BigQuery y Search Console
  - Las nuevas propiedades de GA4 incluyen la opción gratuita de vinculación con Google BigQuery, lo que te permite exportar los eventos sin procesar de GA4 a un almacén de datos en la nube en el que puedes realizar consultas con la sintaxis de SQL.
  - BigQuery es una herramienta de Google Cloud que permite ejecutar consultas de alto rendimiento en grandes conjuntos de datos
  - Además, la vinculación de GA4 con Google Ads, permite crear audiencias para llegar a clientes con experiencias más relevantes y útiles, usando toda esta información para mejorar el retorno de la inversión (ROI).
  - La integración con Search Console permite analizar los resultados de tu sitio en la búsqueda orgánica, incluyendo su posición en los resultados de búsqueda, qué consultas conducen a clics y cómo se relacionan esos clics con el comportamiento del usuario.

## DIA 10 – Entornos Analítica Web

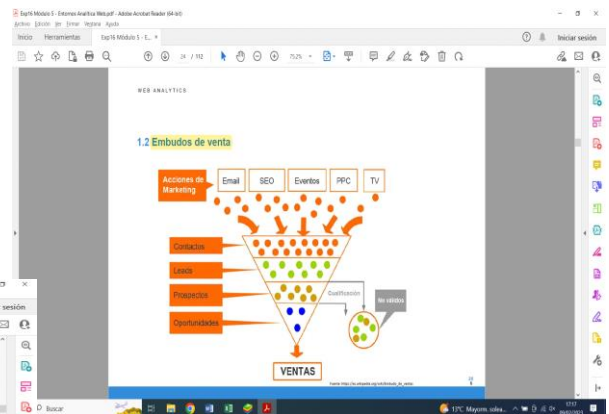
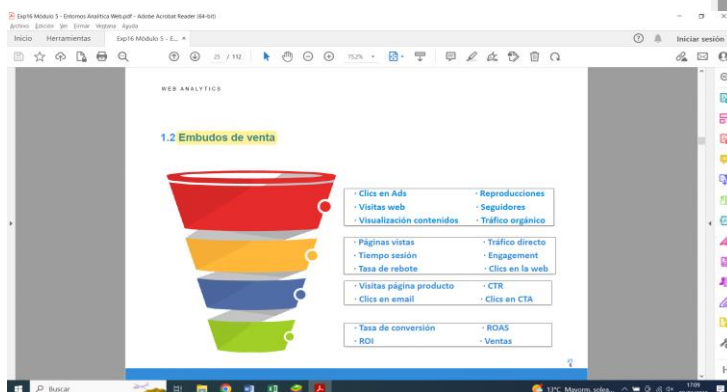
### ENTORNOS ANALÍTICOS

**Estrategias de captación de tráfico: optimizar la captación de tráfico web → aumento del ROI**

- Ya sea tráfico orgánico o tráfico de pago, a mayor volumen de usuarios visitando una página web mayor será su probabilidad de convertir ese flujo de visitantes en ventas.
- Técnicas y herramientas empleadas: SEO, SEM, Acciones PPC (pago por clic), Social Media, Inbound Marketing, Email Marketing, SMS Marketing, etc.
- Necesidad (SEO, SEM, Shopping) VS Deseo (RR.SS., Influencers)
- Proceso conversión usuario - cliente: primer contacto - tráfico web – Registro - Remarketing

**Embudos de venta:** fases por las cuales pasa un cliente durante todo el proceso de captación y venta de una compañía.

- 3 Fases: TOFU, MOFU, BOFU (top, middle, bottom).
- Diferencias en tipo de contenido, segmentación, formato, objetivo y KPIs.
- Estrategias pull



**EJERCICIO 1:** Diseño un embudo de venta para tu propio proyecto SX, utilizando todas las herramientas y recursos disponibles.

- **TOFU:** generación de tráfico. Campañas de contenido:
  - Artículos blogs: blog especializados en ocio de lujo con enlace a nuestra web
  - Publicaciones promocionadas: webs de organización de eventos (empresa, bodas, viajes, etc.)
  - SEM y SEO: palabras: eventos, lujo, caro, exclusivo, etc.
  - Redes Sociales: Facebook, Instagram: fotos y videos de eventos redirigiendo a nuestras RRSS
  - Influencers: moda y similares, contando experiencia de compra
- **MOFU:** campañas de interacción:

- Comentarios de clientes para recomendar y comentar (dándose previamente de alta en la web con datos básicos): “Cuéntanos cómo fue tu evento soñado”.
- Videos: “¿Te imaginas ser tú? Déjanos tus datos y te contaremos cómo hacerlo realidad.”
- Imágenes: coches, yates resto vehículos de lujo. Botón más información, pidiendo datos.
- RRSS: agregar perfiles que sigan/recomienden nuestras publicaciones
- **BOFU:** campañas de persuasión
  - Primera toma de contacto gratuita / primer presupuesto gratuito: formulario de datos del cliente.
  - Descuento de un % en el primer servicio al realizar X compra.
  - Campaña de fidelización: plata, oro, platino

## SEO / SEM

El objetivo esencial al hacer SEO será lograr que Google nos muestre entre los primeros resultados para determinadas búsquedas.

- Cuándo un cliente busque nuestra marca o producto/servicio
- Hay que entender la necesidad del cliente y cómo busca información
- Es imprescindible generar contenido de calidad alineado a las búsquedas de estos usuarios
- Ser creativos y persuasivos en nuestra web para generar ventas

Elección de **palabras clave:** baja competencia y gran número de búsquedas (Google Keyword Planner).

**SEO Onpage:** conjunto de acciones que podemos realizar en nuestra web con el fin de optimizar su posicionamiento entre los resultados de búsqueda.

- Lectura del código fuente de la página (etiquetas, etc.) para indexar
- Snippet:
  - Limitación en cuanto a caracteres (70 para el título 156 para la descripción).
  - Escueto, original, que incite al click y que contenga las palabras clave a posicionar.
- Imágenes en línea: para Google, mejor en formatos .png (más calidad) y .webp.
- Crawl Budget: presupuesto de rastreo (tiempo limitado, dependiendo de la “autoridad” de la web).
  - Evitar poner vídeos directos.
  - No abusar del enlazado

¿Qué debemos tener en cuenta al desarrollar nuestra web?

- Estructura web claramente jerarquizada
- Cuidar la cantidad y calidad del enlazado en cada página
- Enlaces estáticos y de texto
- Contenidos relevantes y con un enlazado coherente

- Emplear aquellas palabras clave que los usuarios utilizarían para encontrarnos
- Etiquetar las imágenes con el atributo ALT, siendo este preciso y descriptivo

¿Cómo debemos etiquetar el contenido de nuestra web?

- Meta Tittle
  - La etiqueta tittle nos permite incluir información descriptiva sobre un contenido de nuestra web, añadiendo palabras clave relevantes y haciéndola así más accesible.
  - A aquellas palabras utilizadas en esta etiqueta de título Google les dará una mayor importancia.
  - ¿Cómo estructurar los títulos de una página web?
    - 1 etiqueta H1 por cada página, destinada al título
    - Varias etiquetas H2 por página para los subtítulos de mayor relevancia
    - Varias etiquetas H3 por bloque de contenidos para subtítulos de menor relevancia
    - Etiquetas de formato párrafo <p> para el resto de contenidos no categorizados entre los anteriores.
- Meta Description
  - Esta etiqueta nos permite añadir información relevante al encabezado de cada página, incluyendo palabras clave, y es mostrada por Google como resumen en cada resultado de búsqueda.
  - Facilita al usuario un resumen del contenido que encontrará en nuestra página, y mostrar a Google la relevancia del contenido indexado para las palabras utilizadas en dicha etiqueta.
- Semántica URL
  - La utilización de palabras clave en la sintaxis de los enlaces de nuestra web nos ayudará a posicionar mejor los contenidos de las mismas y a hacer más amigables dichos enlaces.
  - [www.undominiocualquiera.com/palabra-clave/](http://www.undominiocualquiera.com/palabra-clave/)
  - ¿Cómo diseñar una URL amigable?
    - Cada URL debe contener la palabra (s) clave para posicionar el contenido de dicha página de destino.
    - Esta palabra clave debe coincidir además con alguna de las utilizadas en las etiqueta tittle y description, así como encontrarse en el contenido relevante de la página.
    - La URL debe ser corta para propiciar una mayor relevancia de la palabra clave dentro de la misma.
    - Prescindir de artículos, pronombres y preposiciones, ya que no tienen relevancia y al eliminarlos hacen la URL más amigable.
    - Dale preferencia al guión medio frente al guión bajo.
- Etiquetado de imágenes: Google no puede ver imágenes, pero puede “leerlas” a través del atributo ALT de dicha imagen. Es importante utilizar este atributo para aprovechar cada imagen en el indexado del contenido de nuestra web.
- Enlazado interno:
  - Al crear contenidos de valor para nuestra web debemos cuidar el enlazado que realizamos, tanto a sitios externos como internos de nuestra web.
  - Es importante escoger adecuadamente las palabras clave utilizadas para crear estos enlaces.
  - El anchor text (o “texto ancla”) de cada enlace debe incluir estas palabras claves, enmarcadas en un contexto coherente con el contenido al que enlazan.

EJERCICIO 2: Componer los snippets para la home de nuestra web, teniendo en cuenta las directrices para escribir unos buenos snippets y un par de palabras clave a trabajar en SEO.

Para ello usar cualquiera de estos 2 generadores:

- Merkle Mangools
- [SERP Simulator](#)

### Disfruta de la vida sin límites... sueña y lo haremos realidad.

<https://tuhedonismo.company.site/>

Diseñamos experiencias exclusivas en España para disfrutar verdaderos momentos de lujo y hedonismo. Imagina tu mejor momento y lo haremos realidad.

Disfruta de la vida sin límites. Sueña con [HYPERLINK "https://www.serpsimulator.com/](https://www.serpsimulator.com/) "lujo [HYPERLINK "https://www.serpsimulator.com/](https://www.serpsimulator.com/) " y lo haremos realidad.

<https://tuhedonismo.company.site/>

Diseñamos **experiencias** exclusivas en España para disfrutar verdaderos **momentos** de **lujo** y **hedonismo**. Imagina tu mejor **momento** y lo haremos realidad.

Disfruta de tu [HYPERLINK "https://www.serpsimulator.com/](https://www.serpsimulator.com/) "hedonismo [HYPERLINK "https://www.serpsimulator.com/](https://www.serpsimulator.com/) " sin límites. Sueña con [HYPERLINK "https://www.serpsimulator.com/](https://www.serpsimulator.com/) "lujo [HYPERLINK "https://www.serpsimulator.com/](https://www.serpsimulator.com/) " y lo haremos realidad.

<https://tuhedonismo.company.site/>

Diseñamos **experiencias** exclusivas en España para disfrutar verdaderos **momentos** de **lujo** y **hedonismo**. Imagina tu mejor **momento** y lo haremos realidad

**SEO Offline:** conjunto de acciones que podemos realizar fuera de nuestra web para optimiza el posicionamiento de la misma entre los resultados de los motores de búsqueda.

- Se debe “ser enlazados” desde otras páginas para adquirir reputación/autoridad.
- Distinguir entre enlaces “follow” y “nofollow: enlaces que interesen o no (deben de tener relación de contenido). Se puede “pedir” a Google que no te enlacen.
- **Linkbuilding:** generar una red de enlaces externos que apunten hacia una de nuestras páginas web. (ej.: mala práctica – trampear con páginas ficticias. Ej.: bloggers, periódicos, influencers).
  - **DA - Domain Authority:** refleja en una escala de 0 a 100 el nivel de autoridad de un dominio para predecir cómo se posicionará entre los resultados de búsqueda. Se compara con respecto a los competidores del sector.
  - **PA - Page Authority:** refleja un valor entre 0 y 100 para determinar el nivel de autoridad de una página de nuestro sitio web y trata de predecir cómo de bien se posicionará entre los resultados de búsqueda para una serie de palabras clave. Se compara con uno mismo para ir mejorando.
- **Linkbaiting:** generar contenidos de altísima relevancia para lograr que otras páginas externas enlacen hacia nuestro sitio web. (Ej.: Wikipedia).

- ¿Cómo de bien se ha indexado nuestra página web? SATURACIÓN =  $[\text{nº páginas indexadas} / \text{nº páginas creadas}] \times 100$ . \*Saturación = 100 ✓
- [SemRush](#) es la mejor herramienta (prueba gratuita)

EJERCICIO 3: Diseña los siguientes enlaces desde uno de los contenidos de tu ecommerce a cada uno de los diferentes destinos:

- Enlace del blog a la página de inicio:  
`<a href="tuhedonismo.company.site/">Regresar al inicio</a>`
- Enlace desde la página de inicio a un producto  
`<a href="tuhedonismo.company.site/Coches-de-lujo-y-deportivos-p529835504">Coches de lujo y deportivos</a>`
- Enlace desde el blog a un producto de una web de terceros  
`<a href="www.bershka.com" rel="nofollow"> Descubre lo que venden en Bershka</a>`
- Enlace desde un producto a tu política de privacidad.  
`<a href="tuhedonismo.company.site/pages/privacy-policy">Declaración de privacidad</a>`

EJERCICIO 4: Averigua la autoridad de dominio y de página de dos competidores de un mismo sector. Puedes utilizar herramientas como:

- MOZ Analytics
- [Website SEO Checker](https://websiteseochecker.com/#areareult) <https://websiteseochecker.com/#areareult>
- Small SEO Tools

Website Authority

Moz Website Authority	
URL	<a href="https://tuhedonismo.company.site/">https://tuhedonismo.company.site/</a>
Domain Authority (DA)	<a href="#">77</a>
Page Authority (PA)	<a href="#">36</a>
Total Backlinks	<a href="#">0</a>
Quality Backlinks	<a href="#">0</a>

Percentage of quality backlinks	<a href="#">0%</a>
MozRank	<a href="#">4/10</a>
<a href="#">1000+ HYPERLINK</a> <a href="#">"https://semrush.sjv.io/c/3986542/1208110/13053"keywords</a> <a href="#">HYPERLINK "https://semrush.sjv.io/c/3986542/1208110/13053"</a> <a href="#">lead to HYPERLINK</a> <a href="#">"https://semrush.sjv.io/c/3986542/1208110/13053"company.site</a>	<a href="#">More Info</a>

## Website Authority

Moz Website Authority	
URL	<a href="https://luxurymomentsgroup.com/">https://luxurymomentsgroup.com/</a>
Domain Authority (DA)	<a href="#">11</a>
Page Authority (PA)	<a href="#">19</a>
Total Backlinks	<a href="#">225</a>
Quality Backlinks	<a href="#">2</a>
Percentage of quality backlinks	<a href="#">1%</a>
MozRank	<a href="#">2/10</a>
<a href="#">1000+ HYPERLINK</a> <a href="#">"https://semrush.sjv.io/c/3986542/1208110/13053"keywords</a> <a href="#">HYPERLINK</a> <a href="#">"https://semrush.sjv.io/c/3986542/1208110/13053" lead to</a> <a href="#">luxurymomentsgroup.com</a>	<a href="#">More Info</a>

## Las principales diferencias entre SEO y SEM

Como hemos visto, tanto el SEO como el SEM buscan mejorar el posicionamiento de un sitio web en los motores de búsqueda, como Google. Pero hay varias diferencias entre ellos:

- La más evidente es **el tipo de inversión que requieren**. Existe el estereotipo de que "el SEO es gratis", pero esto no es exactamente así: conseguir una página web capaz de ocupar los primeros



puestos en los resultados de búsqueda requiere tiempo, esfuerzo y, en última instancia, presupuesto. Pero a excepción de algunos gastos de mantenimiento, se trata sobre todo de una inversión inicial que da resultados en el futuro. En cambio, si usas SEM, tendrás que seguir pagando por cada clic en tus anuncios.

- El **lugar** que ocupará tu sitio web en las páginas de resultados también es diferente. Los anuncios de Google Ads ocupan una banda en la parte superior y una columna a la derecha de la página, mientras que los resultados orgánicos de búsqueda copan el espacio central.
- El **tipo de contenido** con el que trabajamos es diferente entre ambas estrategias. El SEO se basa en crear contenidos de calidad, generalmente en formatos más largos, mientras que en el SEM cuentas con anuncios con un número muy reducido de caracteres y landing pages diseñadas para conseguir el máximo impacto con los mínimos elementos.
- Por último, la **temporalidad** también es distinta: con el SEM se busca conseguir resultados a corto/medio plazo, mientras que el SEO es una apuesta a medio/largo plazo.

## ANALÍTICA EN SOCIAL ADS

### Tipos de estrategia

- Redes Sociales: muchas; publicidad digital hipersegmentada.
- Segmentar el público es una variable fundamental para el éxito de una campaña.
- 3 tipos de estrategia, muy orientadas según el embudo de ventas:
  - PPC – MOFU: obtener clics para generar tráfico web, clientes potenciales, descargas
  - Visibilidad – TOFU: mostrar una propuesta de valor a la mayor cantidad posible de usuarios, con el objetivo de obtener así posteriormente la mayor cantidad de potenciales clientes.
  - Leads Ads – BOFU: obtener la mayor cantidad posible de clientes potenciales a través de diferentes tipos de anuncios. Es uno de los más utilizados actualmente.

### Medir las campañas

todas las plataformas sociales donde podemos realizar publicidad cuentan con herramientas internas y gratuitas para medir la evolución y resultados de cada acción publicitaria.

- Métricas
  - Alcance – personas
  - Impresiones – veces que
  - Clics
  - CTR (Click Trought Rate)
  - Tasa de conversión
  - Leads

## OTROS ENTORNOS

## EMAIL MARKETING

Utiliza el correo electrónico como canal para crear una comunicación directa con una base de usuarios (suscriptores) con fines comerciales y de fidelización. Esta **estrategia** cuenta con algunas ventajas:

- Mayor flexibilidad
- Menor coste
- Inmediatez de ejecución
- Resultados en un menor tiempo
- Mayor capacidad de segmentación

### Variables / elementos de la estrategia

- BBDD: alquilar, comprar, construir (más tiempo/recursos pero mejor).
- Contenido / asunto
- Envío
- Optimización

### Métricas clave

- ✉ Emails entregados
- ✉ Tasa de rebote
- ✉ Emails marcados como spam
- ✉ Tasa de apertura
- ✉ CTR / Tasa de clics
- ✉ Tasa de conversión
- ✉ Bajas de suscriptores

**Herramientas:** mailchimp, sendinblue, etc.

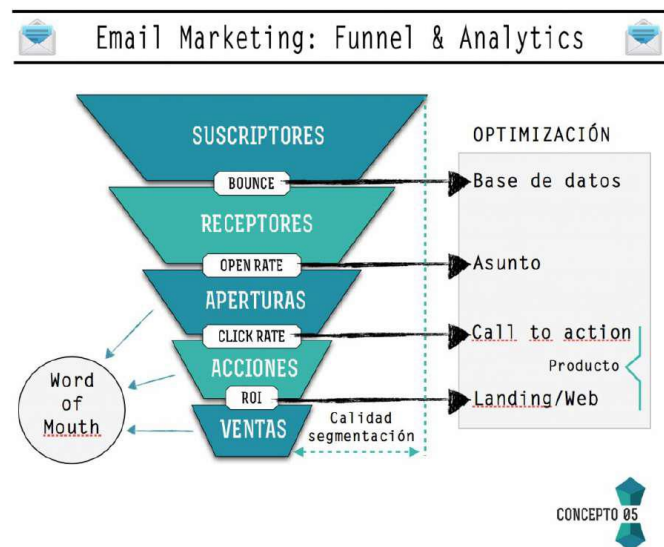
## USABILIDAD / UX

Experiencia de usuario (o User eXperience) hace referencia a:

- Proceso en el cual un usuario interactúa con una marca a través de su web, puntos de venta y/o canales de comunicación.
- Poner al usuario en el centro de toda estrategia, de forma que cada diseño, contenido y decisión tome como punto de partida la satisfacción de los usuarios.

### Aspectos:

- ✓ Usability
- ✓ Learning
- ✓ Memory



- ✓ Efficiency
- ✓ Satisfaction
- ✓ Accessibility

**Métricas** que necesitamos analizar:

- Duración de la sesión en la web
- Tasa de rebote
- Clics
- CTAs
- Páginas visitadas
- Conversiones
- Errores

**Herramientas:** hotjar (mapa de calor y pantallazos)

## MODULO 3 – Data Management

Profesor/formador: Alejandro Arranz <https://www.linkedin.com/in/alberto-arranz-ramos/>

### DIA 11 – Introducción

#### Análisis inicial de una Base de Datos BBDD

##### Datos - BBDD

**Definición:** cifra, letra o palabra que se suministra a la computadora como entrada y la máquina almacena en un determinado formato. Mínima cantidad de información. Un dato es una representación simbólica de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades.

**Tipos de datos:** texto/string, fecha/hora, número/integer, boolean. Depende del tratamiento/finalidad que se vaya a dar al dato (ej.: teléfono – string, CP – string).

**Fuentes de datos:** pueden estar en cualquier sitio → investigar.

**Bases de datos:** definición, ventajas, etc. Ejemplo: Excel NO es una BBDD.

- **Relacionales:** tablas con datos almacenados en filas y columnas relacionadas entre sí.
  - Elementos: registros (filas), variables (columna), celda/valor/dato (intersección).
  - Integridad y estructura, almacenamiento, seguridad, escalabilidad.

##### Tipos de relaciones:

- Relación 1:N
  - Es la más habitual.
  - Ejemplo: Clientes – Pedidos, Productos – Pedidos.
- Relación 1:1
  - Tablas con limitaciones.
  - Ejemplo: Cliente – NIF
- Relación N:N
  - Se da a través de tablas intermedias.
  - Ejemplo: Clientes – Productos, con tabla intermedia Pedidos.

##### Claves/Keys:

- **Primary Key (PK):** en muchos casos suele ser un identificador (ID), no nulo y con valoración incremental. Clave primaria. No se puede repetir. NO todas las tablas tienen que tener un PK. Puede estar compuesta por varios campos.
- **Foreign Key (FK):** cuando una PK se utiliza en otra tabla (ej.: clientes, pedidos). Clave foránea o secundaria. Se puede repetir en tablas destino.

**SQL:** Lenguaje para hacer consultas (técnicamente NO es un lenguaje de programación).

Utilizaremos sistemas de gestión de bases de datos (**SGBD**), algunos con GUI otros con CLI. Ventajas:

- Proporcionar herramientas para consultas, para edición de datos, es decir, agregar, o modificar o eliminar datos.
- Algunos sistemas de administración de bases de datos proporcionan una interfaz gráfica que facilita a los usuarios el manejo de las bases de datos sin conocimientos a detalle, técnicos o profundos.
- Proporcionan métodos para mantener la integridad de los datos y la concurrencia.
- Proporcionan también métodos de control de acceso a usuarios.

#### EJERCICIO 1: trabajar con el Excel sales\_data\_sample

- Hacer un análisis descriptivo de las variables incluidas
- ¿Qué nos gustaría saber con este dataset?
- Echa un vistazo a los tipos de datos y formatea los que hagan falta
- ¿Echas en falta alguna variable? ¿qué datos nos podrían venir bien?
- ¿Podríamos utilizar alguna otra fuente de datos externa para enriquecer los datos?
- Divide el dataset en tablas para crear una pseudo base de datos relacional: ¿qué posibles tablas potenciales tendríamos?

#### ¿BIG DATA?



#### INGESTA Y TRATAMIENTO

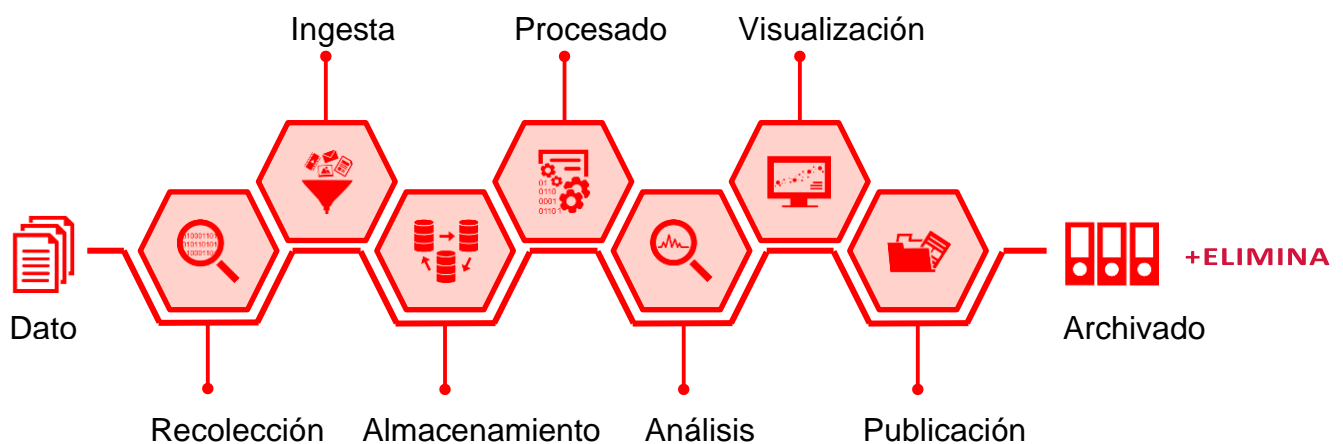
##### Tipos de datos.

- Datos estructurados:
  - Aquellos que poseen una estructura completamente definida, con un número de atributos (o columnas) fijos y con tipos de datos preestablecidos
  - Ejemplo: hojas de cálculo, bases de datos relacionales, etc.
- Datos semiestructurados:
  - Son aquellos que presentan cierta estructura, pero esta no es fija, pudiendo variar para diferentes registros.

- Ejemplos: xml, html, Word, JSON
- Datos no estructurados:
  - Son aquellos que carecen de estructura clara o interpretable, por lo que su tratamiento digital acostumbra a ser más complejo, o requiere un mayor procesamiento.
  - Ejemplos: fotos, audios, videos, texto plano

### Ingesta de datos.

- Generalmente, un dato por sí mismo no proporciona información, o esta es mínima.
- Por tanto, surge un interés en «extraer información» de un conjunto de datos, lo cual generalmente requiere de algún tipo de proceso o tratamiento sobre los mismos.
- Así, el tratamiento de datos es la serie de procesos a los que sometemos los datos para convertirlos en información relevante
- Ciclo de vida del dato: recolección, ingesta, almacenamiento, procesado, análisis, visualización, publicación, archivado



- ¿Qué es la ingesta de datos? Es el proceso por el cual se recolectan datos de varias fuentes o bases de datos y se incorporan a un entorno unificado para su posterior procesamiento.
- Big Data: 3Vs → Volumen, Velocidad, Variedad → Valor

### Limpieza y transformación.

- Durante la ingesta de datos, éstos pueden venir con formatos diversos que puede resultar conveniente convertir o dotar de estructura.
- Además, pueden contener errores o anomalías que deben ser corregidas.
- Operaciones con datos:
  - **Filtrado**: podemos querer ignorar ciertos datos que no cumplen determinadas condiciones, o que no son relevantes para nuestro sistema.
  - **Identificación**: durante la ingesta, es importante dotar a los datos ingeridos de un identificador único (ya sea propio o dependiente de la fuente de datos).
  - **Revisión**: deberían ajustarse al esquema (dominio) especificado, cumpliendo con las reglas de integridad y coherencia impuestas. Si no lo hacen, pueden omitirse, subsanarse o marcarse como inválidos.

- **De-duplicación:** cuando incorporamos datos de varias fuentes (o varias consultas), es fácil que nos encontremos con datos duplicados, incluso si su estructura no es totalmente idéntica. En este caso, es conveniente eliminar los duplicados.
- **Transformación:** cuando disponemos de datos estructurados o semiestructurados, puede ser conveniente transformar su estructura a una fija con el fin de unificar las diferentes fuentes de datos.
- **Estructuración:** cuando se dispone de datos no estructurados, puede resultar interesante tratar de dotarlos de cierta estructura para facilitar su posterior análisis (en ocasiones, el «machine learning» puede ayudar a esto).
- **El proceso ETL**
  - Extract — Transform — Load: hace referencia al concepto de extraer datos de diferentes fuentes y transformarlos para posteriormente cargarlos en algún almacén o base de datos.
  - Como proceso, está muy relacionado con la ingesta de datos, aunque históricamente se ha denominado ETL al proceso de extracción de datos estructurados disponibles «por lotes».
  - Existen herramientas que permiten facilitar el proceso de ETL mediante el diseño de «pipelines» que indican los pasos a los que se someten los datos, tales como *Talend* o *Pentaho*.
  - Datos en crudo: debido al abaratamiento de los costes de almacenamiento, puede ser interesante almacenar no solo el dato procesado, sino también el original o «crudo», por si fuera útil en el futuro.
  - **ELT** es la filosofía Extract—Load —Transform, que plantea realizar la carga de los datos tras su extracción (en «crudo»), transformándolos cuando sea necesario y del modo que resulte más apropiado en cada momento.

### Paradigmas de procesado.

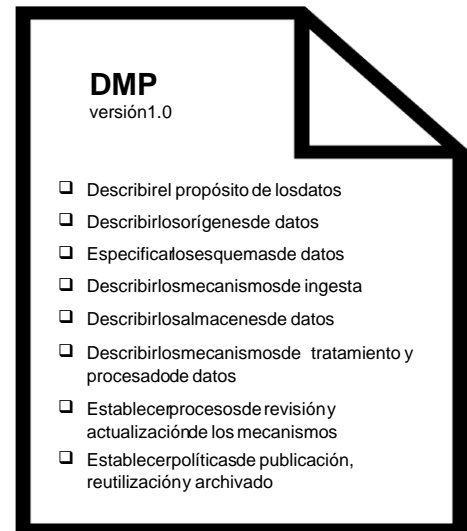
- Migración (lote único): es el proceso de realizar un procesado puntual para transformar unos datos en otros, o para tratarlos de algún modo.
- Batch (por lotes): en una aproximación batch o por lotes, grandes cantidades de datos se procesan de golpe, generalmente mediante algún enfoque distribuido. Este proceso puede repetirse periódicamente, según se dispone de nuevos datos.
- Batch (MapReduce): es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida entre diferentes servidores (o nodos).
- Streaming (tiempo real): en una aproximación en streaming, los datos llegan de forma continua y a gran velocidad, y este flujo de datos se va transformando y procesando a medida que va llegando al sistema.
- Streaming Lambda: es una arquitectura que permite el procesado de datos tanto por lotes como en tiempo real, permitiendo que ambas capas se retroalimenten, y proporcionando una salida conjunta.
- El ecosistema Hadoop: Hadoop es un ecosistema consistente en multitud de herramientas para almacenamiento, procesado, análisis y gestión integral de «Big Data», inspirado en tecnologías presentadas por Google a principios de los 2000.

### Gestión y documentación.

- En demasiados casos, las compañías no tienen documentación actualizada sobre sus fuentes de datos, mecanismos de ingesta y procesado, recursos de almacenamiento, etc.
- Es útil disponer de un documento de gestión de los datos.
- Ventajas:



- Facilita la comprensión de los procesos de datos a los nuevos empleados de la compañía, y permite la continuidad en caso de renovación del equipo de gestión de datos.
- Permite la revisión sistemática de los procesos de ingesta y tratamiento de datos, asegurándose de que las fuentes de datos permanecen correctamente conectadas y la ingesta se realiza satisfactoriamente.
- Facilita la auditoría de los datos, tanto interna como externa, ya sea con fines técnicos o regulatorios.
- Facilita la revisión de mecanismos de almacenamiento y procesado para asegurarnos de que siempre responden a las necesidades de la compañía.





## DIA 12 y 13 – SQL Segmento 1

### Métricas y dimensiones

- Una métrica es una variable donde el valor tiene un significado numérico.
- Una dimensión, por el contrario, es aquella variable (normalmente en formato texto) que expresa una propiedad cualitativa, por lo cual el rango de valores que puede adquirir es finito (limitado).
- Los términos métricas y dimensiones, comúnmente utilizados en analítica web, equivalen a los conceptos de variables numéricas y categóricas en estadística.
- Una unidad de observación es la persona o cosa a la cual corresponde la información provista en forma de métricas. Cada tabla de datos contiene una (solo una) y está compuesta por la combinación de todas las dimensiones.

### Statements & Clauses Básicos

Uso de SQLiteonline. <https://sqliteonline.com/>

Existen dos statements (comandos) básicos que son obligatorios a ser incluidos en cada query (consulta de datos): *select* y *from*.

- **Select:** qué campos (columnas) serán extraídos, separados por coma. *Nota:* las comas únicamente separan campos, por lo que la SELECT statement **no debe comenzar con coma**.
- **From:** especifica de qué tabla/s, dentro de todas las disponibles en una BD.
- **Limit:** restringe el número de filas a ser incluidas en el resultado final. *Nota:* este comando **no cambia el orden de las filas**, simplemente limita el resultado de las filas.
- **As:** visualizar con un alias las columnas que especifiquemos (renombrar los campos).
- **Distinct:** fuerza valores únicos para el total de campos de UNA columna (hay que especificarla o bien utiliza la primera). *Nota:* importante entender que el DISTINCT clause, al momento de evaluar si una fila es duplicada, tendrá en cuenta el total de las columnas en el output.

### WHERE CLAUSE

- Condicional/comparación para filtrar; condiciones separadas con operadores lógicos AND/OR (**no** están separadas por comas).
- Un operador en computación es una expresión que permite realizar ejecuciones matemáticas o lógicas. Dentro de un WHERE clause se utilizan para definir criterios de filtrado en condiciones.
  - Comparadores: =, <, >, <=, >=, <>
  - Lógicos: AND, OR, NOT, BETWEEN, IN, LIKE
- *Between* NO lleva paréntesis. *IN* SÍ lleva paréntesis.
- Precedencia de operadores: <https://learn.microsoft.com/es-es/sql/t-sql/language-elements/operator-precedence-transact-sql?view=sql-server-ver16>

### ORDER BY

- Criterio de ordenación para el resultado final.
- Ascendente por defecto. DESC – descendente. Tipo texto – alfabético.
- Se pueden utilizar varias columnas.
- Se puede poner un número como criterio, que corresponderá al orden de la columna en cuestión declarada dentro de la sentencia SELECT.

- *Nota:* las condiciones en un ORDER BY statement van separadas por coma.

## CASE WHEN ESTATEMENT

El CASE WHEN statement permite crear campos con condiciones específicas, donde el valor que adquiera cada fila dependerá del criterio establecido. El statement se invoca utilizando la siguiente sintaxis (entre el *SELECT* y el *FROM*):

*CASE*

*WHEN* condición\_1 *THEN* valor\_si\_true

*WHEN* condición\_2 *THEN* valor\_si\_true

...

*ELSE* valor\_si\_ninguno\_arriba\_es\_true

*END (AS)* nombre\_campo\_nuevo

*CASE*

*WHEN*...

...

*END (AS)* nombre\_campo\_nuevo\_2

- El nombre\_campo\_nuevo se podrá utilizar en sentencias WHERE, ORDER BY, etc.
- Se pueden generar las sentencias CASE WHEN que se quieran (generalmente seguidas). No se permite la anidación (al menos en SQLite).

## EJERCICIO 2: diferentes sentencias de SQL usando diferentes datasets

- 1) Extraer campos name, species y homeworld de la tabla de datos de Star Wars
- 2) ¿Cuáles son los planetas (homeworlds) incluidos en la tabla de Star Wars?
- 3) Extraer campos film, director, year y actor de la tabla de James Bond; filtrar por películas publicadas hasta el año 2000, cuyo director sea Lewis Gilbert o Martin Campbell. Excluir aquellas películas protagonizadas por Roger Moore.
- 4) Dataset world\_health\_org.
  - a. Extraer países que cumplan con alguna de las siguientes condiciones: (i) sean países africanos con un índice de alfabetismo entre el 25% y 75% o (ii) países europeos con un ratio de población viviendo en áreas urbanas menor al 50%.
  - b. ¿Cuáles son los 5 países africanos con mayor PIB (GIPC) per cápita?
- 5) Dataset imdb\_movies: extraer las 10 películas con mayor IMDB score, filtrar por películas publicadas a partir de la década del 80, excluir aquellas producidas en los EEUU.
- 6) Dataset loan-data
  - a. Cambia los nombres de las columnas (campos) que están en inglés por su traducción en español.
  - b. ¿Qué solicitudes de crédito tienen un plazo de devolución entre 12 y 24 meses?

- c. ¿Qué solicitudes de crédito corresponden a hombres solteros?
- d. ¿Qué solicitudes de crédito corresponden a personas que en algún momento han solicitado otros créditos y los han pagado?
- e. ¿Qué solicitudes corresponden a personas que tienen 4 créditos o más en curso?
- f. ¿Qué solicitudes de crédito corresponden a un crédito de negocio?
- g. ¿Qué solicitudes de crédito corresponden a un crédito de reparaciones?
- h. ¿Qué solicitudes de crédito corresponden a personas que viven en su vivienda de propiedad?
- i. ¿Qué solicitudes de crédito corresponden a personas con más de 60 años de edad?
- j. ¿Qué solicitudes de crédito corresponden a personas entre 35 y 50 años de edad?
- k. ¿Qué solicitudes de crédito se han aprobado?
- l. ¿Qué solicitudes de crédito se han rechazado?

## DIA 14 – SQL Segmento 2

### AGREGACIONES

- Agregar data significa realizar cálculos que permitan resumir información, abstrayendo conocimiento de data que se encuentra en estado bruto.
- Cuando realizamos agregaciones, podemos elegir agrupar observaciones en una o más variables categóricas.
- Para realizar una agregación, debemos utilizar funciones específicas (funciones de agregación)

#### Funciones de agregación

- Realizan cálculos que permiten resumir información.
- Podemos elegir agrupar observaciones en una o más variables categóricas
- NO podemos utilizar funciones de agregación sobre otras funciones de agregación.
- Debemos utilizar funciones específicas. Siempre devuelven valor de UN ÚNICO campo/columna.

#### GROUP BY Statement

- En caso de querer agrupar observaciones/filas, debemos operar utilizando el **GROUP BY** statement y devolverá diferentes registros según su contenido.
- Se pueden utilizar diversos campos.
- Cuando agrupa, ordena de forma ascendente.
- Se usa a menudo con funciones agregadas COUNT(), MAX(), MIN(), SUM(), AVG() para agrupar el conjunto de resultados por una o más columnas.

Funciones básicas (devuelven un número):

- **Count(\*)**, **Count(field)**, **Count(distinct field)**
- **Sum(field)**
- **Max(field)**
- **Min(field)**
- **AVG(field)**

#### HAVING Clause

- Para filtrar sobre el resultado, NO se puede utilizar **WHERE** sino **HAVING** y siempre DESPUÉS de **GROUP BY**.
- El **WHERE** es para filtrar sobre los datos originales – es decir, aplicará el filtrado ANTES de realizar la función de agregación - y NO sobre el resultado de una función de agregación.
- NO podemos utilizar funciones de agregación dentro de WHERE.

#### Ejemplos:

```
SELECT country, COUNT(movie_title) AS Num_Peli  
  
FROM imdb_movies  
  
GROUP BY country/1  
  
HAVING Num_Peli > 10
```

ORDER BY Num\_Peli/2 DESC;

## FUNCIONES DE FECHA

Uso para campos con formato DATE.

**SQLite** → **STRFTIME**(*date\_part*,*date\_field*): permite extraer una parte de la **fecha** y la devuelve como **texto** (string)

- STRFTIME('%m',*date\_field*): devuelve el mes
- STRFTIME('%W',*date\_field*): devuelve la semana
- STRFTIME('%d',*date\_field*): devuelve el día
- STRFTIME('%Y',*date\_field*): devuelve el año

**MySQL** → **EXTRACT**(*date\_part* FROM *date\_field*): la devuelve como **número**

- EXTRACT(year FROM *date\_field*): devuelve el año
- EXTRACT(month FROM *date\_field*): devuelve el mes

**DATE\_TRUNC**(*date\_field*, *date\_part*): esta función devuelve la fecha al inicio del 'part' y genera un resultado en formato fecha

- Ejemplo: DATE\_TRUNC(month, *date\_part*); esto devuelve la fecha al inicio del mes

## AGREGACIONES CONDICIONALES

- En una agregación condicional restringimos la variable sobre la cual buscamos operar al cumplimiento de una o más condiciones.
- En otras palabras, realizamos una operación sobre un subconjunto de observaciones (filas) que cumplan con una o más condiciones específicas.
- Para implementar una agregación condicional, simplemente necesitamos incluir un CASE WHEN statement dentro de una función de agregación.

### Ejemplo 1

SELECT

CASE

WHEN gender LIKE 'Action%' THEN 'Action Movies'

ELSE 'Rest of Movies'

END AS peliculas,

COUNT(movie\_title),

ROUND(AVG(imdb\_score),2)

FROM imdb\_movies

GROUP BY 1;

### Ejemplo 2

```
SELECT
    CASE
        WHEN artist LIKE '%god%' THEN 'GOD'
        WHEN artist LIKE '%death%' THEN 'DEATH'
        WHEN artist LIKE '%black%' THEN 'BLACK'
    END AS artist_keyword,
    COUNT(DISTINCT artist)
FROM rolling_top_albums_1
WHERE artist_keyword NOT NULL
GROUP BY 1
ORDER BY 2 DESC;
```

### Ejemplo 3

```
SELECT
    channelgrouping,
    SUM(CASE WHEN devicecategory = 'mobile' THEN sessions END) AS mobile_sessions,
    SUM(CASE WHEN devicecategory = 'desktop' THEN sessions END) AS desktop_sessions,
    SUM(CASE WHEN devicecategory = 'tablet' THEN sessions END) AS tablet_sessions,
    SUM(sessions) AS total_sessions
FROM google_analytics_formatdate
WHERE STRFTIME('%m',date) = '10' AND STRFTIME('%Y',date) = '2019'
GROUP BY channelgrouping
ORDER BY 5 DESC
```

EJERCICIO 3: sentencias SQL utilizando funciones de agregación, agregaciones condicionales y funciones fecha.

#### Ex.1: Descripción

Agregar las siguientes métricas para todos los países africanos,

- average gross income per capita
- total population

- number of countries

Tabla: world\_health\_org (`Basics.world\_health\_org`)

### **Ex.2: Descripción**

Calcular número de personajes según planeta (homeworld). Evitar personajes sin información sobre planeta de origen.

Tabla: star\_wars\_characters (`star\_wars\_characters`)

### **Ex.3: Descripción**

Calcular el total de salario percibido por cada actor en todas las películas. Omitir películas sin data sobre salario.

Tabla: james\_bond (`james\_bond`)

### **Ex.4: Descripción (dificultad media)**

¿Podemos asegurar que las películas de acción tienen de media mejor valoración que el resto de películas? Extraer total de películas y media de IMDB score para películas de acción vs. el resto (de forma conjunta).

Tabla: imdb\_movies

### **Ex.5: Descripción**

Calcular la facturación (box office) según director. Filtrar por aquellos directores que hayan generado más de 1500 en el total de facturación (todas las películas).

Tabla: james\_bond

### **Ex.6: Descripción**

Calcular número total de álbumes según sub metal genre, filtrar por aquellos subgéneros con al menos 10 álbumes.

Tabla: rolling\_top\_albums

### **Ex.7: Descripción (dificultad media)**

¿Cuántos artistas hay incluidos en el dataset cuyo nombre incluye las palabras 'god', 'death' or 'black'?

Tabla: rolling\_top\_albums

### **Ex.8: Descripción (dificultad media)**

Extraer media mensual de la cotización (open rate) y volumen de operación (volumen USD) del bitcoin desde el año 2016.

Tabla: bitcoin\_daily\_rates

#### **Ex.9: Descripción**

¿Cuál fue la semana con el valor mayor de cotización? Utilizar cotización high.

Tabla: amazon\_stocks

#### **Ex.10: Descripción (dificultad media)**

Calcular el total de sesiones según canal para octubre de 2019. Crear métricas específicas agregadas para cada dispositivo.

Tabla: google\_analytics

#### **Ex.11.1: Descripción**

¿Cuántas películas duran menos de 60 minutos?; ¿Cuántas entre 60 y 100? Y ¿Cuántas más de 100?

Tabla: imdb\_movies

#### **Ex.11.2: Descripción**

¿Cuántas películas de acción hay que duren menos de 60 minutos? Haz un listado de las mismas. Tabla: imdb\_movies

#### **Ex.11.3: Descripción**

¿Cuál sería el día de menos cotización en una tendencia alcista en el año 2018?; ¿Y la media ese mismo día?

Tabla: bitcoin\_daily\_rates\_formatdate

#### **Ex.11.4: Descripción (dificultad media-alta)**

Mostrar el conteo de las películas relacionadas con los géneros (Action, Crime, Comedy, Drama, Romance), indicando la película con mayor número de votos en cada caso (num\_voted\_users).

\*\* Utilizar el orden de (Action, Crime, Comedy, Drama, Romance) al realizar la tabla.

Tabla: imdb\_movies

#### **Ex.11.5: Descripción**

Mostrar el número de personajes que tienen el mismo color de ojos (eye\_color) y el planeta de origen (homeworld). No mostrar color de ojos desconocidos (unknown) ni planetas sin datos/nombre (NA).

Tabla: star\_wars\_characters

#### **Ex.11.6: Descripción**

Identificar y calcular el presupuesto de aquellas películas de James Bond que fueron dirigidas por John Glen y protagonizadas por Timothy Dalton.



Tabla: jamesbond

**Ex.11.7: Descripción**

¿Cuál es el monto de los créditos otorgados y no otorgados según el Status personal?

Tabla: loan-data

**Ex.11.8: Descripción**

Obtén un listado de las películas de acción con actor protagonista con más de 10000 likes en Facebook y cuyas películas hayan sido valoradas con al menos un 8 en imdb. Todo ello con fechas anteriores a 2012.

Tabla: imdb\_movies

**Ex.11.9: Descripción**

Queremos saber cuáles son las 20 películas y género al que pertenecen,

- 1) con mayor presupuesto
- 2) con mayor beneficio

Tabla: imdb\_movies

## DIA 15 – Segmento SQL 3

### SUBQUERIES

- Realizar una consulta sobre el resultado de otra consulta, ya sea por la naturaleza de la misma o por las limitaciones del lenguaje.
- La subconsulta se almacena en una tabla temporal para poder operar después con ella.
- Para implementar una *subquery*, utilizamos el **WITH statement** con la siguiente sintaxis:

WITH

Temp\_table1 AS (SELECT...FROM),

Temp\_table2 AS (SELECT...FROM)

SELECT...FROM Temp\_table1

#### Ejemplo

WITH

my\_subquery AS (SELECT Sector, SUM(Revenue) AS total\_revenue FROM fortune GROUP BY 1)

SELECT AVG(total\_revenue) AS avg\_sector\_revenue FROM my\_subquery

### UNIONES

#### UNION ALL

Combina dos tablas de forma vertical (uniendo filas). Incluye TODAS las filas, sin importar duplicados

#### UNION

Combina dos tablas de forma vertical (uniendo filas). Excluye duplicados del resultado final (aplica la cláusula DISTINCT automáticamente)

EJERCICIO CLASE: Crear 2 tablas en Excel con formato .CSV con delimitador de “;”

- Tabla 1: una con 3 registros y 2 columnas – Frutas 1
- Tabla 2: otra con 2 registros y 2 columnas – Frutas 2

#### EJERCICIOS SUBQUERIES

##### Ex.1: Descripción

Los directores de películas de James Bond han trabajado en promedio en dos películas cada uno. ¿Cómo obtendrías esta información?

Tabla: james\_bond

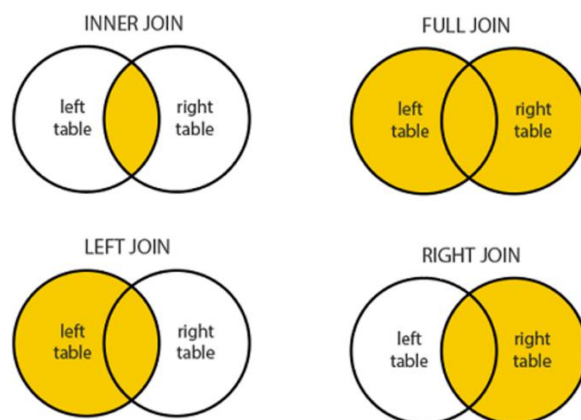
##### Ex.2: Descripción

La tabla gobierno\_paro reporta data mensual de paro para municipios de España en 2018. Calcular la media de paro de cada comunidad autónoma en 2018.

Tabla: gobierno\_paro

## JOIN

Combinar dos tablas de datos de forma horizontal (juntando columnas). Existen distintas formas de realizar joins, estas difieren en el tratamiento de las observaciones que no coinciden.



Una *join* de tipo **INNER** mantiene solo las observaciones que coinciden en ambas tablas (LEFT y RIGHT).

```
SELECT Orders.OrderID, Customers.CustomerName, Orders.OrderDate
FROM Orders
INNER JOIN Customers ON Orders.CustomerID=Customers.CustomerID;
```

Una *join* de tipo **LEFT** mantiene todas las observaciones de la tabla principal (LEFT) e integra las observaciones que coinciden de la tabla secundaria (RIGHT).

```
SELECT Customers.CustomerName, Orders.OrderID
FROM Customers
LEFT JOIN Orders ON Customers.CustomerID = Orders.CustomerID
ORDER BY Customers.CustomerName;
```

Una *join* de tipo **RIGHT** es el opuesto a una LEFT join. Mantiene todas las observaciones de la tabla secundaria (RIGHT) e integra aquellas observaciones de la tabla principal (LEFT) cuyos valores coincidan.

```
SELECT Customers.CustomerName, Orders.OrderID
FROM Customers
RIGHT JOIN Orders ON Customers.CustomerID = Orders.CustomerID
ORDER BY Customers.CustomerName;
```

Una *join* de tipo **FULL** mantiene todas las observaciones de ambas tablas (LEFT y RIGHT), sin importar si coinciden o no.

```
SELECT Orders.OrderID, Customers.CustomerName, Orders.OrderDate
FROM Orders
```

**FULL JOIN** Customers **ON** Orders.CustomerID=Customers.CustomerID;

## EJERCICIO JOINS

### Ex.3: Descripción

Extraer consolas no discontinuadas y su fecha de lanzamiento (first retail availability). Agregar ventas globales de videojuegos publicados desde el año 2000. Ordenar según ventas totales de videojuegos (descendiente).

- Table 1: videogames\_games
- Table 2: videogames\_consoles

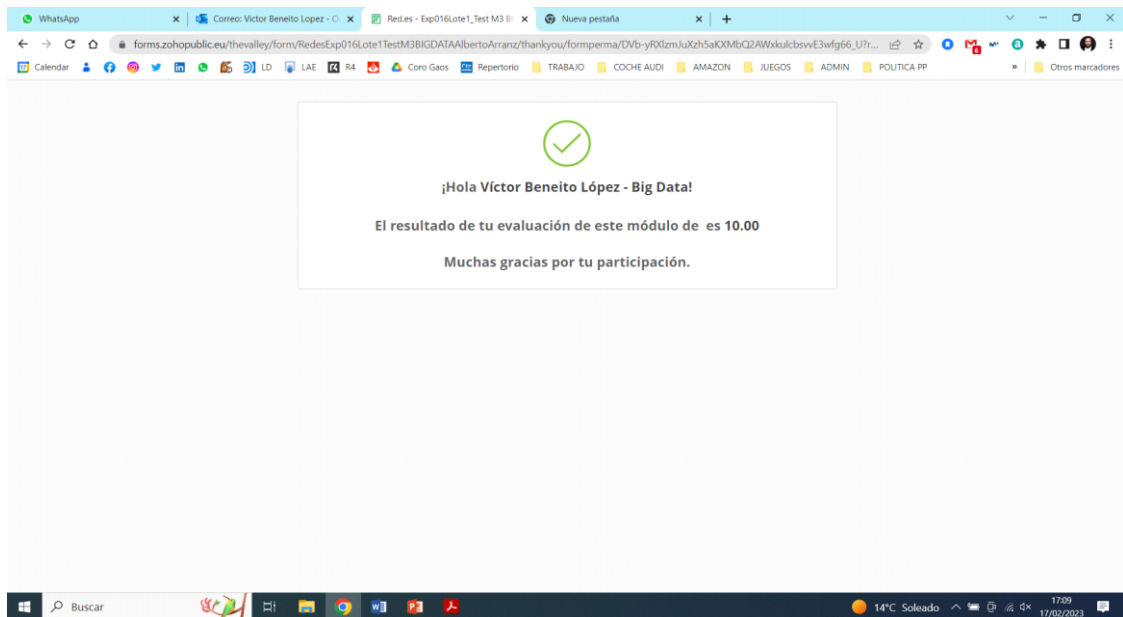
## NOCODE

<https://www.nocoders.academy/blog/que-es-nocode>

- Las plataformas de desarrollo sin código (*nocode*) permiten a los programadores y no programadores crear software de aplicación a través de interfaces gráficas de usuario y configuración en lugar de la programación informática tradicional.
- El movimiento *nocode* es una filosofía digital cuyo objetivo es permitir a cualquier persona acceder a la creación de apps y softwares sin saber programación.
- Se conoce como programación sin código o programación visual: prescindir de lenguajes de codificación para que la creación sea un proceso más visual y sencillo.
- Se utilizan diferentes herramientas de desarrollo que ofrecen un entorno visual.
- Los usuarios solo tienen que implementar métodos sencillos para crear su app o web, como por ejemplo agregar componentes con la técnica de arrastrar y soltar (también conocida como *drag&drop*).
- ¿Qué ventajas ofrece el nocode?
  - Permite que el desarrollo de software sea accesible a un público que no contaba con los conocimientos para hacerlo.
- Pero, ¿programar sin código de verdad funciona en la realidad?
  - Si, el crecimiento de las herramientas no-code así lo demuestran.
  - De acuerdo con datos de Forrester, el mercado de plataformas de desarrollo sin código crecerá a 21.2 mil millones de dólares en 2022.
  - Para entender bien su crecimiento, debes considerar que en 2017 apenas representaban 3.8 mil millones.
- Beneficios que existen el hacer uso de aplicaciones no-code
  - Agilidad:
    - Empresas, emprendedores y usuarios particulares están apostando por las no-code tools (herramientas sin código) debido a que han entendido que con ellas es más rápido crear aplicaciones web y móviles.
    - Al utilizar módulos prediseñados y elementos visuales, la creación de aplicaciones es más rápida, lo que permite una mayor agilidad y reducir el tiempo que se dedica a estas tareas.
  - Curva de aprendizaje:
    - Gracias a la facilidad de uso permitimos a más gente de nuestra organización el uso de herramientas de no code.
  - Menor coste:
    - Otra de las ventajas de poder crear aplicaciones web sin programar es que resulta más rentable.

- Contratar desarrolladores para iniciar un proyecto de programación puede ser costoso. Pero con las soluciones no-code la inversión es significativamente menor.
- No tendrás que gastar demasiado contratando un equipo de desarrolladores y además, al ser más rápido el proceso de desarrollo, también se ahorran gastos.
- Modificaciones fáciles y rápidas:
  - Cambiar una característica o funcionalidad de un software o app programando es complejo. Este procedimiento lleva tiempo y mucho esfuerzo, sobre todo cuando el lenguaje de programación es poco conocido.
  - Esto no ocurre haciendo uso de herramientas no-code, pues en el desarrollo sin código si necesitas cambiar algo, puedes hacerlo de forma sencilla y su implementación es mucho más rápida.
- **Desventajas del nocode:**
  - PERSONALIZACIÓN: al basar el desarrollo en herramientas predefinidas
  - SEGURIDAD: las herramientas nocode viven en la nube y tienen sus propios protocolos de seguridad, ajenos a los propios de nuestra compañía
  - “CAUTIVIDAD”: una vez elegimos una plataforma de nocode estamos obligados a utilizar su stack, haciendo difícil la migración a otro sistema

## Examen 17/02/2023



## MODULO 4 – Data Fundamentals con Python

### DIA 16 – Aprendiendo a pensar como un programador

#### Partes de la resolución de un problema:

- ¿Qué me están pidiendo?
- ¿Qué necesito saber?
- ¿Qué cosas conozco?
- ¿Qué resultado espero? ¿Conozco el problema?
- Resultado final

#### Cambiar cómo pensamos:

- ¿Qué pasos de la resolución del problema son similares?
- ¿Puedo ahorrarme algún paso?
- ¿Y automatizar la resolución?

Si vamos a hacer una misma operación varias veces, podemos definir una función que utilizaremos a menudo.

¿Los datos que necesito puedo generarlos u obtenerlos automáticamente?

**Actividad por grupos:** pensar y redactar distintos problemas (al menos 3) que podamos automatizar o resolver con un ordenador. Transformar la solución en pseudocódigo.

## EJEMPLOS BÁSICOS CON PYTHON

EJERCICIO: [BasicosPythonOriginal.ipynb](#), [Repasillo Estructuras.ipynb](#)

### Google Colab / Python

- No permite trabajar de forma simultánea – hay que tener precaución al trabajar en equipo. Se debe de tener un control de las versiones/cambios de los archivos.
- Los nombres de las variables deben ser significativos y tener sentido.
- El lenguaje se llama *Markdown*.
- Python es un lenguaje de tipado fuerte, si bien no hay que definir el tipo de las variables ni tampoco inicializarlas.
- Si en la última línea de código se pone una instrucción, entiende que hay que mostrar el resultado de la misma en pantalla.
- Python es sensible a minúsculas y mayúsculas.
  - Muy a tener en cuenta con valores True y False
- Python es un lenguaje tabulado (ej.: estructura If)

#### Comentarios (documentación):

- # (hashtag) → una línea o tras una instrucción

- `''' comentario'''` (tres comillas) → varias líneas. Permite que dentro vayan tanto comillas simples (') como dobles (").

Crear una **variable**: nombre de la variable = valor → no hay que declarar el tipo (pero SÍ tendrá un tipo específico) ni inicializarlas al inicio del programa.

**Funciones**: nombre de la función(a1,a2,...,aN), siempre con paréntesis y argumentos separados por coma

- La función `print` no se puede asignar a una variable

La **clase** *tuple* en Python es un tipo contenedor, compuesto, que en un principio se pensó para almacenar grupos de elementos heterogéneos, aunque también puede contener elementos homogéneos. Junto a las clases *list* y *range*, es uno de los tipos de secuencia en Python, con la particularidad de que son inmutables.

- Tupla: `variable = (valor1, valor2,...,valorN)`
- Lista: `variable = [valor1, valor2,..., valorN]`
- Rango: lista inmutable de números enteros en sucesión aritmética. `Variable = rango(0,N)`

### Librerías

- Hay librerías preinstaladas que se pueden utilizar directamente: `librería.funcion()`
- Para importarlas: `import nombre_librería`

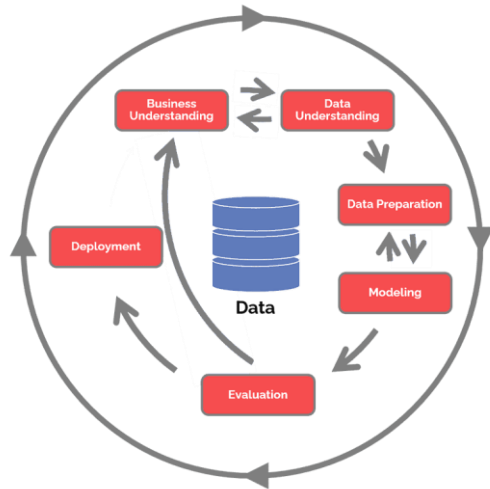
Instrucción **condicional**: `if condición1: elif condición2: else:`

### Apuntes:

- Multiplicar variables tipo string → concatena → `a = 'hola' a*3 → holaholahola`
- Los valores `False` y `True` se escriben con mayúscula; si lo hacemos con minúscula, serían variables.
- `x in "aeiou"` → para comprobar el número de vocales de una palabra.

## CRISP-DM

- [Estándar de estructuración de datos.](#)
- **Primeros pasos**
  - ¿Cuál es el objetivo?
  - ¿Qué tipo de problema es?
  - ¿Qué queremos predecir?
  - ¿Tenemos datos suficientes?
  - ¿Calidad de datos mínima?
  - ¿Qué criterio de aceptación tenemos?
  - ¿Funciona como esperábamos?
- **Fases:** business understanding → data understanding → data preparation → modeling → evaluation → deployment / \*vuelta al inicio\*
  - Primero debemos entender el modelo de negocio.
  - Después debemos entender los datos que tenemos y/o con los que trabajaremos: cantidad, estructura, exploración/captación, calidad, etc.
  - Posteriormente se pasará a la preparación y modelado de los datos.
  - Evaluación de resultados, revisión de procesos, próximos pasos.
    - Si NO se ha llegado a lo esperado → vuelta al inicio.
  - Deployment o implementación cuando se sigue todo lo anterior.
- Uso de metodologías *AGILE* o *WATERFALL* (cascada).





Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<b>Determine Business Objectives</b> <i>Background</i> <i>Business Objectives</i> <i>Business Success Criteria</i>	<b>Collect Initial Data</b> <i>Initial Data Collection Report</i>	<b>Select Data</b> <i>Rationale for Inclusion/Exclusion</i>	<b>Select Modeling Techniques</b> <i>Modeling Technique</i> <i>Modeling Assumptions</i>	<b>Evaluate Results</b> <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i> <i>Approved Models</i>	<b>Plan Deployment</b> <i>Deployment Plan</i>
<b>Assess Situation</b> <i>Inventory of Resources</i> <i>Requirements, Assumptions, and Constraints</i> <i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i>	<b>Describe Data</b> <i>Data Description Report</i>	<b>Clean Data</b> <i>Data Cleaning Report</i>	<b>Generate Test Design</b> <i>Test Design</i>	<b>Review Process</b> <i>Review of Process</i>	<b>Plan Monitoring and Maintenance</b> <i>Monitoring and Maintenance Plan</i>
<b>Determine Data Mining Goals</b> <i>Data Mining Goals</i> <i>Data Mining Success Criteria</i>	<b>Explore Data</b> <i>Data Exploration Report</i>	<b>Construct Data</b> <i>Derived Attributes</i> <i>Generated Records</i>	<b>Build Model</b> <i>Parameter Settings</i> <i>Models</i> <i>Model Descriptions</i>	<b>Determine Next Steps</b> <i>List of Possible Actions</i> <i>Decision</i>	<b>Produce Final Report</b> <i>Final Report</i> <i>Final Presentation</i>
<b>Produce Project Plan</b> <i>Project Plan</i> <i>Initial Assessment of Tools and Techniques</i>	<b>Verify Data Quality</b> <i>Data Quality Report</i>	<b>Integrate Data</b> <i>Merged Data</i>	<b>Assess Model</b> <i>Model Assessment</i> <i>Revised Parameter Settings</i>		<b>Review Project</b> <i>Experience</i> <i>Documentation</i>
		<b>Format Data</b> <i>Reformatted Data</i>			
		<i>Dataset</i> <i>Dataset Description</i>			

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

## RIRO-GIGO

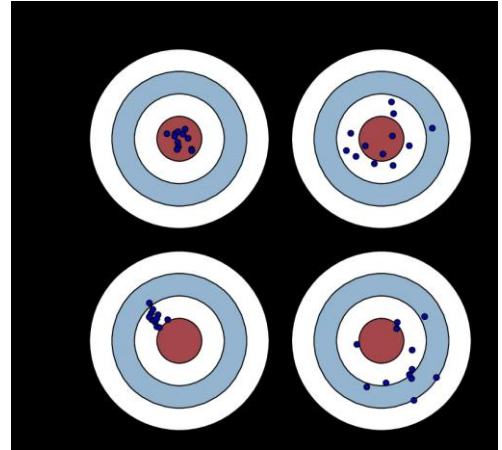
- Rubbish In Rubbish Out
- Los científicos de datos pasan el **80% del tiempo limpiando datos**.
- La **limpieza** de datos es fundamental. Da igual el modelo si los datos no son correctos.
  - ¿Tenemos datos suficientes?
  - ¿Calidad de datos mínima?
  - ¿Qué tipo de problema es?
  - ¿Qué criterio de aceptación tenemos?
  - ¿Funciona como esperábamos?

## PROBLEMAS

### El sesgo

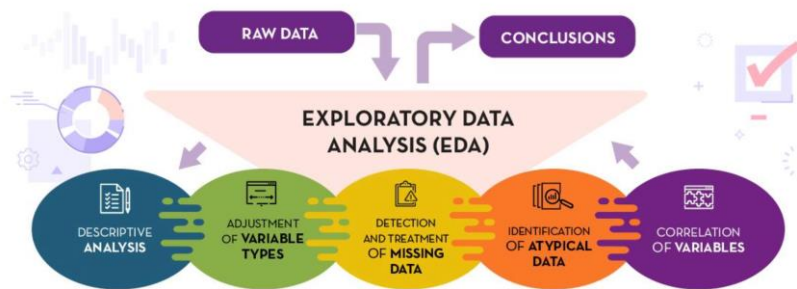
- Aunque las respuestas son técnicamente correctas, por alguna razón tenemos un sesgo para preferir una de ellas.

- No todas las personas compartirán ese sesgo; lo que percibimos y cómo respondemos está influenciado por nuestras normas.
- **Equidad:** evaluar un modelo de aprendizaje automático requiere hacer más que solo calcular las métricas de pérdida. Antes de implementar un modelo en producción, es fundamental auditar los datos de entrenamiento y evaluar las predicciones para determinar si existen sesgos.
- Diferentes tipos de sesgos cognitivos humanos se pueden manifestar en los datos de entrenamiento, por lo que debemos proporcionar estrategias para identificarlos y evaluar sus efectos.
- Modelado Data Science: varianza (*variance*) y sesgo (*bias*).



### Datos incoherentes

- Exploratory Data Analysis (EDA)
- Conceptos: histograma, distribuciones, campana de gauss, datos atípicos, etc.



### Datos incompletos y/o insuficientes

- Los datos reales tienen restricciones: coste, protección/privacidad, leyes, etc.
- Datos sintéticos: generados artificialmente.
  - Más fácil de generar a partir de datos reales.
  - Más baratos.
  - Mantiene la estructura de los originales, pero NO es la misma.
  - <https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/>

### LIBRERÍAS Y PIP

- Import y funciones, librería PIP → para saber lo que hay detrás de las librerías y cómo exportar funciones...

### Tipos de datos complejos

CHULETA	Tupla	Lista	Diccionarios
<b>Definición</b>	<code>mi_tupla = ('texto', 20, 1275.48)</code>	<code>mi_lista = ['texto', 20, 1275.48]</code>	<code>mi_dict = {'clave uno':'texto', 'clave dos':20, 'clave tres':1275.48}</code>
<b>Obtener uno de los valores</b>	<code>print mi_tupla[0] # imprime texto print mi_tupla[1] # imprime 20 print mi_tupla[2] # imprime 1275.48</code>	<code>print mi_lista[0] # imprime texto print mi_lista[1] # imprime 20 print mi_lista[2] # imprime 1275.48</code>	<code>print mi_dict['clave uno'] # imprime texto print mi_dict['clave dos'] # imprime 20 print mi_dict['clave tres'] # imprime 1275.48</code>
<b>Modificar uno de sus valores</b>	<b>NO SE PUEDE</b>	<code>mi_lista[0] = 'cambió'</code>	<code>mi_dict['clave dos'] = 34</code>

- Las claves de los diccionarios son de tipo texto.
- Las posiciones en diccionarios NO importan, ya que el orden lo da la clave.

## Funciones

- Te permite definir un bloque de código reutilizable que se puede ejecutar muchas veces dentro del programa.
- Una de las grandes ventajas de usar funciones en el código es que nos permitirá reducir el número total de líneas de código en el proyecto.
- **Nativas:** el intérprete de Python tiene una serie de funciones y tipos incluidos en él que siempre están disponibles. <https://docs.python.org/es/3.9/library/functions.html>
- Los paréntesis van pegados a la función para su ejecución/invocación.
- NO es imprescindible que tengan argumentos o que devuelvan algún valor.

## Librerías PIP

- <https://pypi.org/> Instalar librerías Repositorio Python Package Index
- Para crear datos:
  - Jupyter <https://jupyter.org/>
  - Pandas <https://pandas.pydata.org/docs/>
  - NumPy <https://numpy.org/doc/stable/>
  - Plotly <https://plotly.com/>
- Las librerías pueden tener dependencias de otras librerías y/o versiones..
- <https://www.anaconda.com/products/distribution>

## EJERCICIO: [Strings y cadenas Victor.ipynb](#)

- Strings y cadenas
- %... para no traducir. Pasarlo directamente al servidor/ordenador
- %pip list → listado de librerías

- Código para montar la carpeta del Google Drive
  - ```
from google.colab import drive
```
  - ```
drive.mount('/content/drive')
```
- Instrucción bucle **FOR** *variable* **IN** *cadena* → recorre cada valor/posición de la cadena
- Función strip() → uso en strings para quitar espacios al principio y al final del texto
- Slices (rodajas) → uso para recorrer la cadena de diferentes formas:
  - print(variable[:]) → toda la cadena
  - print(variable[:5]) → hasta el 5 carácter (incluido)
  - print(variable[4:]) → del 5 carácter en adelante
  - print(variable[::-2]) → saltando de 2 en 2 de forma inversa

## DIA 18 – Data Fundamentals

### NUMPY Y PANDAS

#### NumPY

- Nos ofrece objetos array n-dimensionales (vectores/matrices).
- Herramientas para la integración de código C/C++ y Fortran.
- Funciones de álgebra lineal, transformadas de Fourier y funciones de aleatoriedad.

También puede utilizarse como un contenedor multidimensional de datos generales, lo cual permite que se conecte de forma sencilla y rápida a distintas bases de datos.

Muy útil. Aconsejable importarlo habitualmente.

#### Pandas

Proporciona estructuras de datos rápidas, flexibles y expresivas, diseñadas para que trabajar con datos relacionales sea fácil e intuitivo.

- Es como NumPY pero con más elementos, más potente.
- Fallo o aspectos mejorables: son poco interactivos, poco intuitivos.

#### Jupyter

Nos permite la creación de documentos y programas con mayor versatilidad.

Ventaja: crear archivos jupyter notebooks (ej: Google Colab).

### ESTRUCTURAS DE DATOS

Ver apuntes de días anteriores.

CHULETA	Tupla	Lista	Diccionarios
Definición	<code>mi_tupla = ('texto', 20, 1275.48)</code>	<code>mi_lista = ['texto', 20, 1275.48]</code>	<code>mi_dict = {'clave uno':'texto', 'clave dos':20, 'clave tres':1275.48}</code>
Obtener uno de los valores	<code>print mi_tupla[0]</code> # imprime texto <code>print mi_tupla[1]</code> # imprime 20 <code>print mi_tupla[2]</code> # imprime 1275.48	<code>print mi_lista[0]</code> # imprime texto <code>print mi_lista[1]</code> # imprime 20 <code>print mi_lista[2]</code> # imprime 1275.48	<code>print mi_dict['clave uno']</code> # imprime texto <code>print mi_dict['clave dos']</code> # imprime 20 <code>print mi_dict['clave tres']</code> # imprime 1275.48
Modificar uno de sus valores	<b>NO SE PUEDE</b>	<code>mi_lista[0] = 'cambió'</code>	<code>mi_dict['clave dos'] = 34</code>

#### Operaciones con listas

- `lista.append(elemento)` añade un elemento al final de la lista
- `lista.insert(posición, elemento)` inserta un elemento en la posición y desplaza el resto de elementos.
- `list.extend(elemento1, elemento2, ..., elementoN)` añade lista/tupla de elementos al final de la lista.
- `list.index(elemento)` busca y muestra la primera posición donde coincida el elemento.

- `list.remove(elemento)` busca y elimina la primera posición donde coincida el elemento. NO devuelve valor.
- `list.pop(posición)` saca el elemento – lo devuelve - y lo elimina de la lista.

### Operaciones con diccionarios

Recordatorio: el diccionario se ordena mediante sus claves. Muy útiles para crear JSON.

- Añadir: `diccionario[nueva clave] = valor`
- Actualizar: `diccionario[clave] = nuevo valor`
- Eliminar: `del diccionario[clave]`

### Operaciones con arrays (recordemos que NumPy se basa en este objeto)

- Shape: (valor\_dimension1, valor\_dimension2, valor\_dimension3,..., valor\_dimensionN)
- Ejemplo: `shape(4 filas, 2 columnas, 3 profundidades)`

## EDA (EXPLORATORY DATA ANALYSIS)

Empezamos preguntándonos... “¿de qué se trata?”, patrones y distribuciones. Nos permitirá:

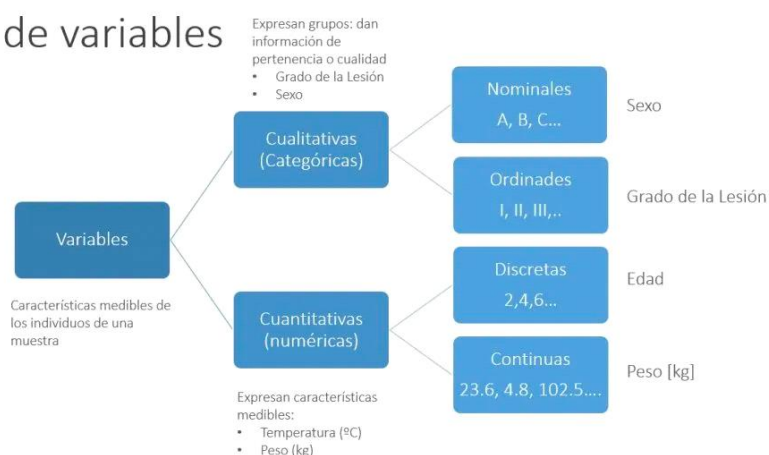
- Conocer los Datos y entenderlos
- Identificar patrones e incoherencias
- Fijarnos en los *Outliers* o valores atípicos e identificarlos
- NO se manipulan datos.

### Preguntas básicas:

- ¿Cuántos registros hay?
- ¿Son demasiado pocos?
- ¿Son muchos y no tenemos capacidad (CPU+RAM) suficiente para procesarlo?
- ¿Están todas las filas completas ó tenemos campos con valores nulos?
- En caso que haya demasiados nulos: ¿queda el resto de información inútil?
- ¿Cuáles parecen ser *features* importantes? ¿Cuáles podemos descartar?
- ¿Hay correlación entre *features* (características/campos/columnas)?

## Pandas

### Tipos de variables



- ¿Qué datos son discretos (se pueden medir por unidades) y cuáles son continuos?
- ¿Siguen alguna distribución?
- Obtenemos datos de casi cualquier fichero y podemos volcarlo a otro fichero.
- Creación de columnas a partir de otras.
- “Graficación” de datos sencilla implementada.
- Selección de datos específicos.

EJERCICIO: [Practica Kaggle.ipynb](#), [Pandas Practicas Alejandro.ipynb](#)

Análisis de datos de un fichero CSV.

Tipos en Pandas: series (un conjunto índice/elemento) y dataframes (tablas y similares).

- <https://www.analyticslane.com/2021/07/15/pandas-cambiar-los-tipos-de-datos-en-los-dataframes/>

Práctica Kaggle → fichero de Morningstar. Problema: demasiado complejo para poder trabajar.

Pandas Practicas Alejandro → fichero de Chipotle. Más práctico para trabajar.

## DIA 19 – Data Manipulation

Continuar con la práctica de ayer.

<https://datacarpentry.org/python-ecology-lesson-es/04-data-types-and-format/>

### DATA MANIPULATION

Ver ejemplos y prácticas anteriores.

#### Data Wrangling

Hay más contenido teórico-práctico para manejar mejor los datos:

- Data Cleansing / Cleaning
- Data Preparation

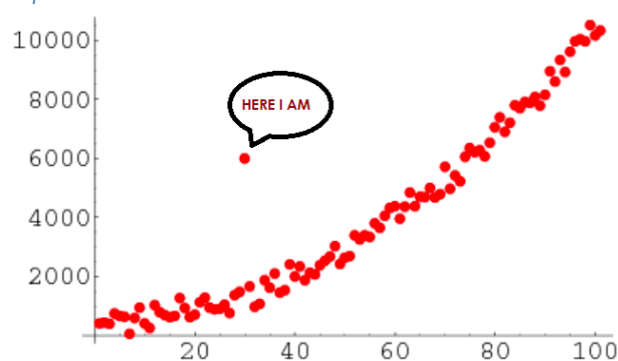
Una vez tenemos bien explorados los datos, debemos identificar los valores:

- Incompletos
- Nulos
- Incoherentes
- Inexistentes

#### Incompletos y/o nulos

- Tenemos que identificarlos y “marcarlos”.
- -1, Null, 0, Símbolos... Tenemos que saber dónde están y colocar una señal.
- Habrá veces que no tendremos buena distribución de los datos; cuando esto ocurra diremos que están no balanceados / desbalanceados o que no tienen dispersión suficiente.

#### Valores Atípicos



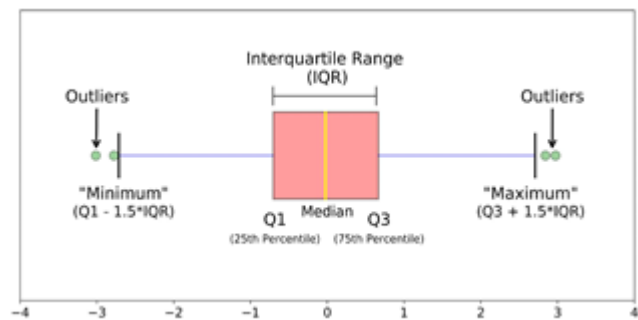
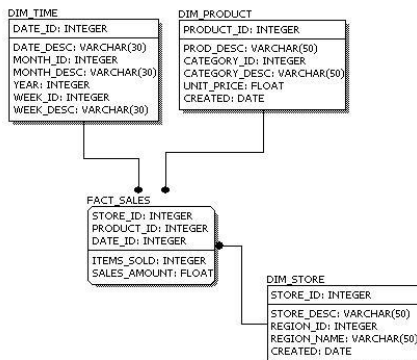
- Valores fuera de la normal general → *OUTLIERS*
- Utilizamos gráficas de dispersión
- Medidas de posición: cuartiles, deciles y percentiles
- Depende de la distribución de los datos. Generalmente suponemos que es normal.



### Incoherentes

- ¿Cómo son los datos?
- ¿Cuál es el rango de valores posible?
- ¿Existe un orden?

¡Un Diccionario de Datos es la respuesta! análisis exploratorio de datos.



EJERCICIO: Ejercicio guiado de EDA [Panda Practicas – chipotle.ipynb](#)

- Buscamos Outliers
- Identificamos datos incompletos
- Creamos un diccionario de Dato
- Generamos datos sintéticos para los imbalances

### Ejercicios:

- calcular percentiles / cuartiles con la función `.describe()`
- recuperar datos de tabla con la función `.iloc(f,c)`
- suma de datos con la función `.sum()`
- comprobación de nulos con la función `.isnull()`
- recuperar datos no nulos con la función `.notnull()`

### Filtros

- `serie = tabla['índice'] <, >, ==, etc valor de comparación tabla[serie]`

- `tabla[tabla['campo'] <, >, ==, etc. valor de comparación]`

### Plotear

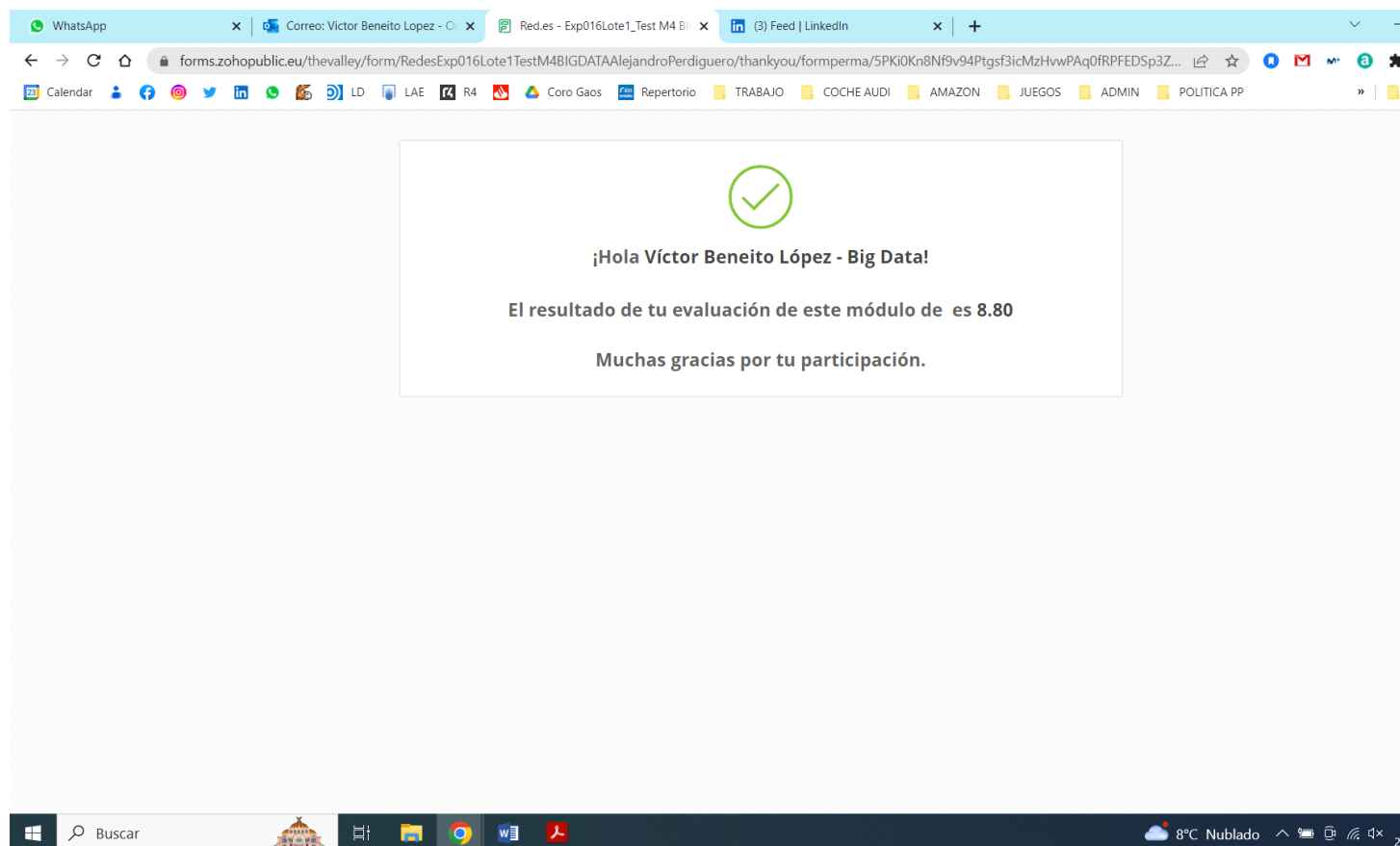
- `import plotly.express as px`
- `px.box(tablaSurveys, y='weight')`

## DIA 20 – Data Fundamentals con Python

- Revisión de enlaces de interés del documento [Links de interés](#)
- DavidMcCandless - La belleza de la visualización de datos sub: español  
<https://www.youtube.com/watch?v=YUAUwUt9ggg>
- Data, information, knowledge: we distil it into beautiful, useful graphics & diagrams.  
<https://informationisbeautiful.net/>

EJERCICIO: Ejercicio guiado de EDA [Panda Practicas – chipotle.ipynb](#)

Continuación.



The screenshot shows a web browser window with multiple tabs. The active tab is titled "forms.zohopublic.eu/thevalley/form/RedesExp016Lote1TestM4BIGDATAAlejandroPerdiguero/thankyou/formperma/5PKi0Kn8Nf9v94Ptgsf3icMzHvwPAq0FRPFEDSp3Z...". The browser's address bar shows the URL. Below the address bar, there is a navigation bar with various icons and labels: "Calendar", "LD", "LAE", "R4", "Coro Gaos", "Repertorio", "TRABAJO", "COCHE AUDI", "AMAZON", "JUEGOS", "ADMIN", and "POLITICA PP". The main content area of the browser displays a confirmation message in a white box with a green checkmark icon. The message reads: "¡Hola Víctor Beneito López - Big Data!", "El resultado de tu evaluación de este módulo de es 8.80", and "Muchas gracias por tu participación.".

WhatsApp x | Correo: Víctor Beneito Lopez - x | Red.es - Exp016Lote1\_Test M4 B | (3) Feed | LinkedIn x | +

forms.zohopublic.eu/thevalley/form/RedesExp016Lote1TestM4BIGDATAAlejandroPerdiguero/thankyou/formperma/5PKi0Kn8Nf9v94Ptgsf3icMzHvwPAq0FRPFEDSp3Z...

Calendar | LD | LAE | R4 | Coro Gaos | Repertorio | TRABAJO | COCHE AUDI | AMAZON | JUEGOS | ADMIN | POLITICA PP

¡Hola Víctor Beneito López - Big Data!

El resultado de tu evaluación de este módulo de es 8.80

Muchas gracias por tu participación.

Windows taskbar: Buscar | 8°C Nublado

## MODULO 5 – Data Science y Machine Learning

DIAS 21 y 22 – Exploración y análisis de datos en Python

Profesora: Aurora Cobo Aguilera

<https://www.linkedin.com/in/aurora-cobo-aguilera-9a8823117/>

[https://drive.google.com/drive/folders/1VaF-JxNvb5uBK8M5medu2ZnWuOWSS7YA?usp=share\\_link](https://drive.google.com/drive/folders/1VaF-JxNvb5uBK8M5medu2ZnWuOWSS7YA?usp=share_link)

### Contextualización

Diferentes **roles**:

- Big Data Architect / Developer
- Data Engineer
- Data Analyst
- **Data Scientist**: Data Science & Machine Learning

**Datos**: pequeñas pinceladas

- Identificarlos
- Obtenerlos
- Transformarlos
- Almacenarlos
- Consumirlos
- Analizarlos

**Business Intelligence**: ¿BI vs Big Data / Data Science? Ambas se alimentan de datos masivos, pero ...

- Big Data: captura, almacenamiento y procesamiento de datos masivo
- Data Science: análisis predictivo y prescriptivo de esos datos
- Business Intelligence: aprovechamiento de los datos para optimizar las decisiones y el reporting de una compañía

**Python**, ¿Por qué?

- Más fácil
- Multipropósito
- Sencillo de interpretar
- Muchas librerías
- Tipado fuerte y dinámico

Además, muy interesante para:

- IA, para hacer modelos de ML y algoritmos
- Big Data: Transformaciones de datos (ETL) y pequeños scripts

**EJERCICIO 1**: definiciones y conceptos varios.

- **¿Qué es una regresión lineal?**

En estadística, la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente  $Y$ ,  $m$  variables independientes  $X_i$  con  $\epsilon$  + término aleatorio. Este método es aplicable en muchas situaciones en las que se estudia la relación entre dos o más variables o predecir un comportamiento, algunas incluso sin relación con la tecnología. En caso de que no se pueda aplicar un modelo de regresión a un estudio, se dice que no hay correlación entre las variables estudiadas.

- **¿Y un dato perdido?**

Son aquellos que no constan debido a cualquier acontecimiento, como por ejemplo errores en la transcripción de los datos o la ausencia de disposición a responder a ciertas cuestiones de una encuesta. Los datos pueden faltar de manera aleatoria o no aleatoria.

- **¿Qué significa entrenar un modelo?**

El proceso de entrenamiento de un modelo de ML consiste en proporcionar datos de entrenamiento de los cuales aprender a un algoritmo de ML (es decir, el algoritmo de aprendizaje). El término modelo de ML se refiere al artefacto de modelo que se crea en el proceso de entrenamiento.

Utilizaremos el set de datos de entrenamiento para ejecutar nuestra máquina y deberemos de ver una mejora incremental (para la predicción). Recordar inicializar los “pesos” de nuestro modelo aleatoriamente, los pesos son los valores que multiplican o afectan a las relaciones entre las entradas y las salidas, se irán ajustando automáticamente por el algoritmo seleccionado cuanto más se entrena. Revisar los resultados obtenidos y corregir (por ej. inclinación de la pendiente) y volver a iterar...

- **¿Qué es un algoritmo?**

Conjunto ordenado de operaciones sistemáticas que permite hacer un cálculo y hallar la solución de un tipo de problemas. Un algoritmo informático es un conjunto de instrucciones definidas, ordenadas y acotadas para resolver un problema, realizar un cálculo o desarrollar una tarea. Es decir, un algoritmo es un procedimiento paso a paso para conseguir un fin.

- **¿Y minibatch?**

El Mini-batch Stochastic Gradient Descent es uno de los métodos más utilizados en el uso de las redes de neuronas profundas durante el Deep Learning, puesto que su proceso práctico resulta efectivo y sencillo en comparación con otros métodos de la teoría de optimización.

PRÁCTICA 1: [Estructuras de Datos en Python](#)

## Introducción al Aprendizaje Automático

### Motivación.: la Era del ‘big data’

Estamos presenciando una explosión de aplicaciones donde el análisis de datos juega un papel fundamental. Tan importantes como las fuentes de datos son los algoritmos que extraen la información relevante.

Ejemplos de aplicaciones:

- Máquinas de búsqueda en Internet.
- Biomedicina (clasificación automática de pacientes y enfermedades).
- Predicción de mercados financieros / Inversión.
- Sistemas de recomendación (ej.: Amazon).
- Contenidos en Redes Sociales.
- Asistentes de voz.

- Actuales smartphones.
- Sistemas de atención al cliente.
- Filtros de spam.
- Automatizaciones en el hogar.

**Aprendizaje Automático o Machine Learning (ML):** es un conjunto de métodos que pueden automáticamente detectar patrones en datos, y utilizar los patrones descubiertos para predecir datos futuros, o realizar otros tipos de toma de decisiones bajo incertidumbre.

Utiliza herramientas de teoría de probabilidad. La teoría de la probabilidad se puede aplicar a cualquier problema que involucre incertidumbre. Muy relacionado con estadística.

- ¿Cuál es la mejor predicción sobre el futuro dados algunos datos del pasado?
- ¿Cuál es el mejor modelo para explicar ciertos datos?
- ¿Cuál medida debería ser la siguiente?

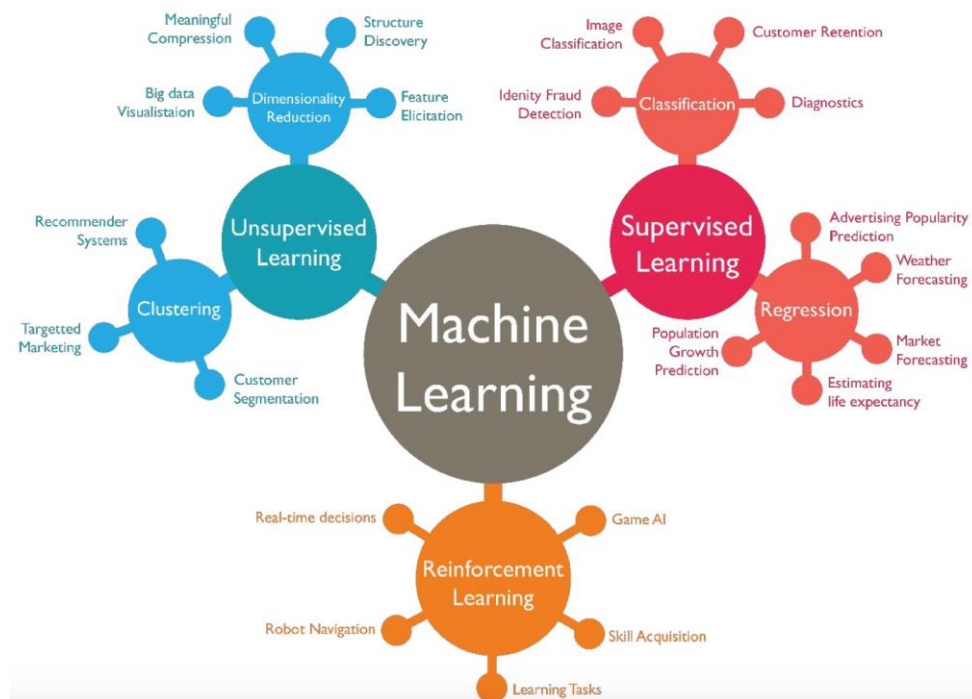
## EJERCICIO 2: diferencias entre ML, AI y DL

Diferencias entre... ¿Qué son?

- **Machine Learning:** algoritmos que permiten conseguir el aprendizaje automático a través de los datos. Es el resultado de datos + variables + algoritmos.
- **Artificial Intelligence:** ciencia que tiene como objetivo imitar la mente humana.
- **Deep Learning:** máquina aprende por sí sola a través de un gran conjunto de algoritmos que imitan la red de neuronas del cerebro humano.

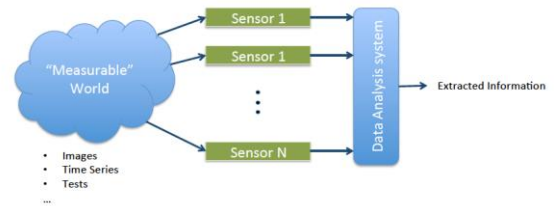
## Conceptos fundamentales del ML

Clasificación de los modelos de ML



## Diagrama de bloques de un sistema de aprendizaje

- La información deseada no se puede acceder directamente, así que hay que usar una serie de variable estadísticas relacionadas.
- El análisis de datos explota esta información estadística para proporcionar resultados precisos.
- Algunos errores son generalmente inevitables.



## Elementos básicos

Construir modelos que se ajustan a una colección de datos

- **Modelo:** objeto de una clase (programa informático) o función matemática (modelo estadístico) con parámetros libres.
- **Conjunto de entrenamiento:** Conjunto de ejemplos sacados de la distribución de datos que se pretende modelar.
- **Optimizador:** Método que ajusta los valores de los parámetros libres del modelo (entrena) para que capturen patrones informativos contenidos en los datos.

## Objetivo: Generalización

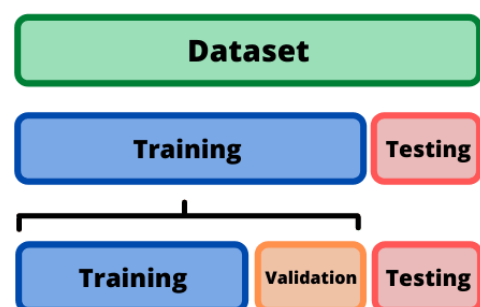
- Conseguir que el modelo pueda hacer predicciones más o menos correctas cuando se le presenten datos que no se hayan usado durante el entrenamiento.
- Evitar el *sobreentrenamiento (overfitting)*, que provoca un error pequeño en el entrenamiento, pero elevado en test. <https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/>

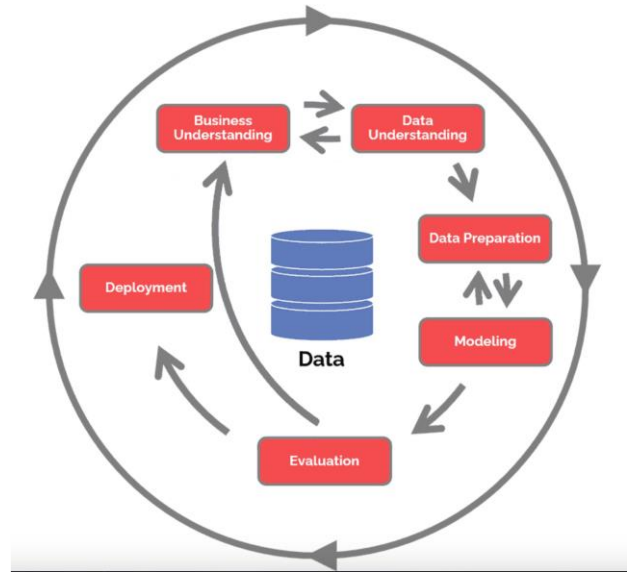
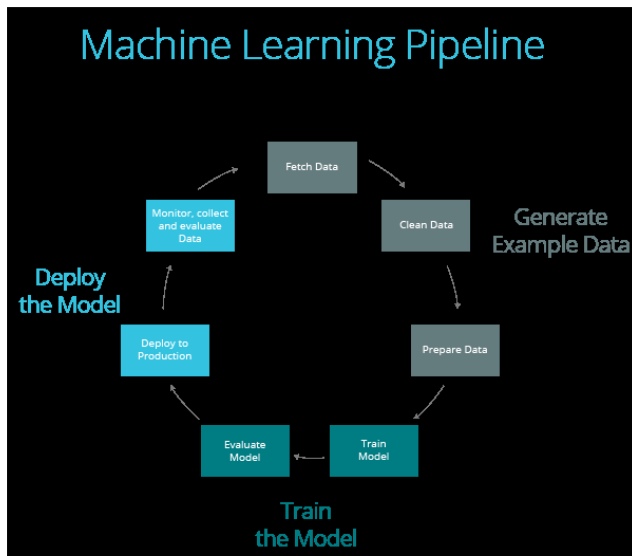
## Conjuntos de...

**Entrenamiento:** Conjunto de datos usado durante el entrenamiento del modelo.

**Test:** Conjunto de datos que nunca se usa durante el entrenamiento ni durante la optimización de los parámetros, sólo para calcular la puntuación de test tras la finalización de dicho proceso.

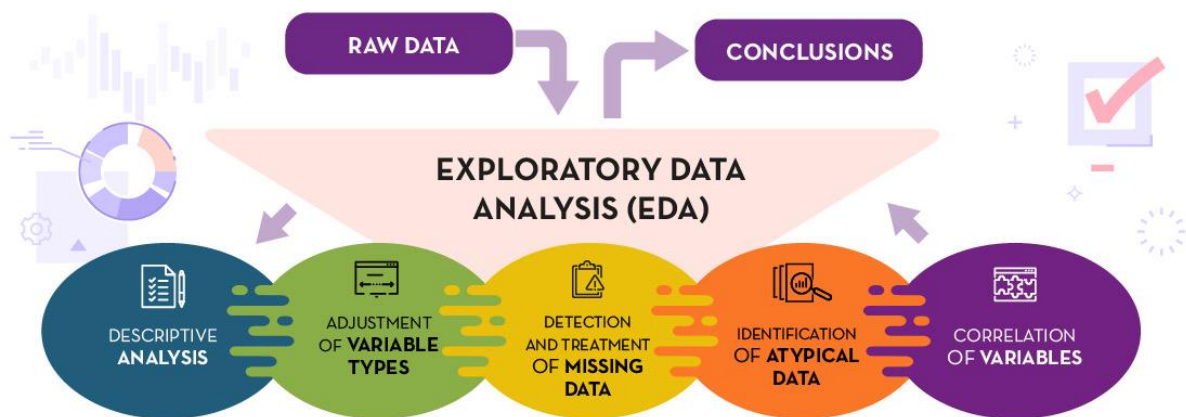
**Validación:** Conjunto de datos usado para evaluar el modelo durante el entrenamiento y elegir el mejor valor para los parámetros, los que obtengan mejor puntuación en dicho conjunto.





## Repasemos EDA y Data Preparation

**EDA:** Exploratory Data Analysis. Conocer los datos, identificar patrones y detectar *outliers*.

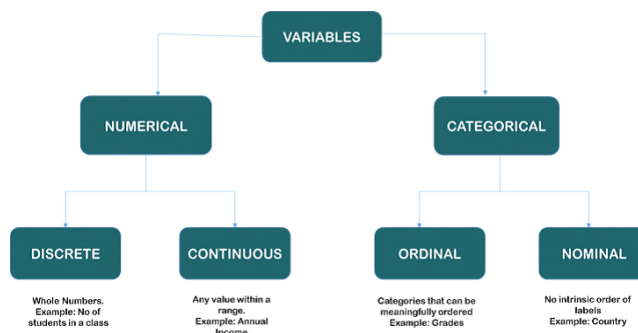


### Definiciones importantes

- **Observaciones:** también se les denomina registros. Hace referencia a cada una de las filas de la BD.
- **Características/variables:** también se les denomina *features*. Hace referencia a cada una de las columnas en una BD.
- **Missing data:** Son los valores perdidos de la base de datos. Se les puede llamar nulos. A veces son campos vacíos, NaN, None, 0, -1, ...
- **Outliers:** son observaciones de la base de datos que se alejan de la distribución del resto. Es decir, son muestras muy diferentes a las demás.

## Variables. Tipos de datos:

- Int : numérica discreta.
- Float y double : numérica continua.
- String : categórica nominal u ordinal.
- Bool : categórica ordinal.



## Data preparation: proceso

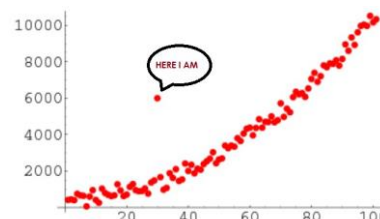
Una vez que tenemos los datos explorados, empezamos a detectar:

- **Nulos (missing):** debemos detectarlos y colocar una señal que identifiquemos en nuestro código como nulos. Unificar todos los nulos de la misma forma (N/A, NULL, NaN). Opciones que podemos hacer, dependiendo del tipo y significado de los datos:

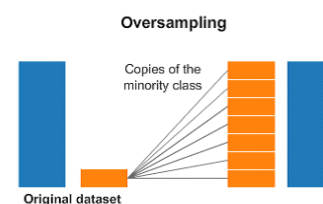
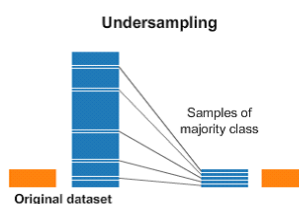
- Completar con una constante
- Completar con la media o la mediana
- Completar con el valor más frecuente
- Poner un *flag* (por ejemplo -999)
- Otro método de imputación
- Borrar las filas o columnas correspondientes (no recomendable!)

- **Valores atípicos (Outliers):** valores fuera de la norma general / de la distribución de los datos. Opciones que podemos hacer:

- Detectarlos: stats.zscore()
- Cambiarlos o
- Eliminarlos



- **Datos incompletos:** hay veces que, aun teniendo todo el histórico de los datos, no es suficientemente buena la calidad de los mismos y no es representativa la distribución que tenemos. Ejemplo: datos no balanceados.



- **Datos erróneos:** una mala calidad en los datos también se ve reflejada en errores en las bases de datos.

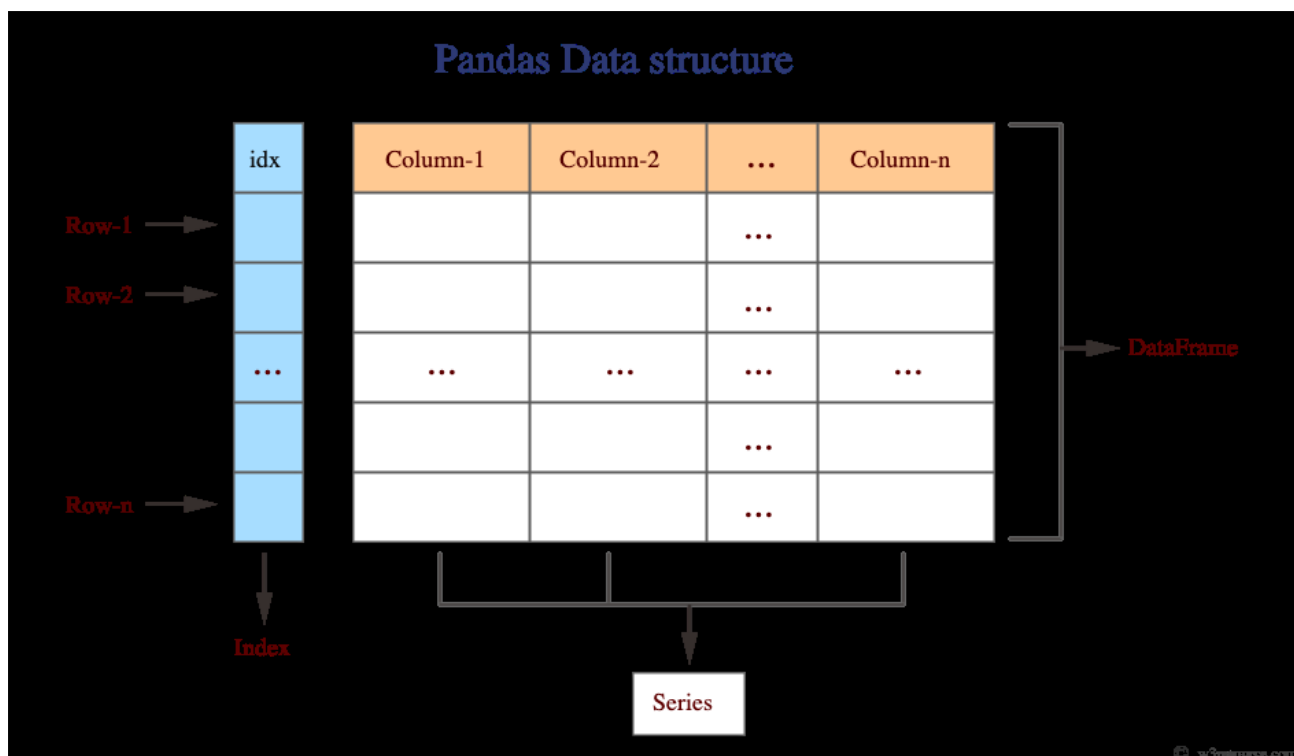
- ¿Cómo son los datos?
- ¿Cuál es el rango de valores posible?
- ¿Existe un orden?

**BAD DATA IS NO  
BETTER THAN NO  
DATA.**

## PRÁCTICA 2: [Preprocesado de un dataset con Pandas](#)

Repasemos Pandas con Python e incorporemos algún concepto nuevo de ML.





## Herramientas para el ML

- Aplicación fundamental en el área de las matemáticas.
  - Álgebra lineal, cálculo y probabilidad son los 'lenguaje' en los que el ML está escrito.
  - El aprendizaje de estas materias ayudará a entender los algoritmos y a modificarlos o crear unos nuevos.
  - Forma la columna vertebral de muchos algoritmos de ML.
- **Álgebra:** para el tratamiento de los datos vamos a trabajar con vectores (1D), matrices (2D) y tensores (mayores dimensiones) y operaciones entre ellos.
  - **Cálculo:** es imprescindible en el paso de la optimización de una función de pérdidas para ajustar los parámetros del modelo. Optimización de una función de coste/error/pérdidas para ajustar los parámetros del modelo. Derivadas parciales: cómo se altera la función de pérdidas con individuales cambios en cada parámetro. Las derivadas se agrupan en matrices para un cálculo más directo.
  - **Programación:** es la herramienta para analizar los datos, entrenar los modelos y visualizar los resultados. Usaremos *notebooks* de *Python* para estudiar todo el módulo, ver la teoría, procesado de datos y aplicación del modelo de ML con ejemplos prácticos. Conocidos del módulo anterior. Google Colab.
  - **Teoría de la probabilidad:** es la base teórica de un modelo de ML para cuantificar la incertidumbre.

## Álgebra lineal en ML

- Vectorización como una manera de paralelizar operaciones.
- Notación de bucle reformulada con ecuaciones matriciales ofreciendo ganancias en eficiencia computacional.
- Uso en librerías de Python como Numpy, SciPy, Scikit-Learn, Pandas, tensorflow o Pytorch.
- Las GPUs se han diseñado para realizar operaciones de álgebra lineal optimizadas.

- El crecimiento explosivo de Deep Learning se debe en parte de la naturaleza de paralelización en los algoritmos sobre hardware GPU.

### Factorización de matrices

- Problemas: *overflow* (desbordamiento) y *underflow* son los límites de representar computacionalmente números extremadamente grandes o pequeños.
- Solución: Por ejemplo usando técnicas de factorización de matrices.
- Permiten representar matrices en otras más estructuradas y simples que tiene propiedades computacionales muy útiles.
- Las descomposiciones LU o SVD son componentes intrínsecos de algoritmos como Linear Least Squares (LLS) o Principal Components Analysis (PCA), que veremos al final del módulo.

### Vectores y matrices

- Son las entidades primarias, y son ejemplos de una entidad más general conocida como un tensor.
  - **Escalar:** Tensor de orden cero. Definimos el conjunto al que pertenece. Ejemplo:  $x \in \mathbb{R}$ ,  $x \in \mathbb{N}$ ,  $x \in \mathbb{Z}$
  - **Vector:** Tensor de una dimensión. Miembros de espacios vectoriales. Ejemplo:  $x \in \mathbb{R}^1$
- En ML, los vectores representan las características de los datos.
- Ejemplo: la importancia de cada palabra que forma un documento, la intensidad de los pixels en una imagen o los valores de precios históricos para una muestra de instrumentos financieros.
- **Matriz:** Tensor de 2 dimensiones,  $m \times n$ , con m filas y n columnas. Ejemplo:  $A \in \mathbb{R}^{m \times n}$ 
  - Por defecto un vector es una matriz  $1 \times n$ . Definimos como vector columna a una matriz  $m \times 1$
- **Tensor:** es el término general para más de 2 dimensiones. Muy utilizado en Deep Learning, por ejemplo: para los parámetros de las redes neuronales o para describir datos como una imagen con datos de intensidad en múltiples canales (colores RGB).

### PRÁCTICA 3: Librería Numpy

¿Conoces la librería Numpy de Python? <https://numpy.org/doc/stable/>

### Cálculo en ML

#### Función de pérdidas/coste.

- Es una función que mapea el resultado de un modelo (predicción) en un número real que representa el coste asociado al modelo.
- El problema de optimización busca minimizar dicha función: calcular las derivadas e igualar a cero.
- Si en vez de función de pérdidas tenemos función de ganancias o de probabilidad, se buscará maximizar dicha función.
- Ejemplo: MSE (Mean Square Error o Error cuadrático medio).

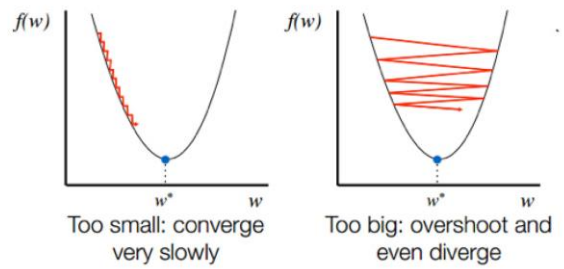
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

**Optimización:** buscar los parámetros que minimizan una función de pérdidas.

**Función de pérdidas:** en función de 2 parámetros, por ejemplo

**Descenso por gradiente.**

- Cuando la optimización de una función no tiene forma cerrada: no podemos despejar los parámetros en la ecuación, hay que hacer uso de algoritmos iterativos que buscan esa solución del mínimo en la función.
- Descenso por gradiente es uno de ellos, muy conocido y utilizado en ML.
- Tiene un parámetro muy importante: el 'learning rate'.



Ejemplo práctico: reconocimiento de dígitos manuscritos.

#### PRÁCTICA 4: [Librerías para el análisis de datos](#)

Practiquemos con *datasets* reales el uso de las librerías más importantes.

Leer datasets de ficheros, separar variables de entrada y salida, hacer gráficas e incluso entrenar el primer modelo.

## DIAS 23, 24 y 25 – Fundamentos del Aprendizaje Automático

### Tipos de modelos de ML

**D-dimensional vector:** características / features.

#### Aprendizaje supervisado

- Aprender un mapeo de entradas  $x$  a salidas  $y$ , dado un conjunto etiquetado de pares de entrada-salida.
- Aprende una relación 1 a 1 entre observaciones y targets.
- Ejemplos:
  - Scores para recomendaciones de netflix, spotify, etc.
  - Sistemas de concesiones de créditos, seguros, etc.
  - Reconocimiento de caras, huellas, etc.
  - Filtrado de correos electrónicos, noticias, etc.
  - Sistemas de ayuda a la diagnosis en aplicaciones clínicas

$$D = \{\mathbf{x}_i, y_i\}_{i=1}^N$$

**Regresión:** aprender una función que para cada observación prediga un número real o score.  $Y_i$  es un número real.

**Clasificación:** aprender una función que para cada observación prediga a qué clase, dentro de un conjunto finito y establecido a priori, pertenece esta observación.  $Y_i$  es una categoría

#### Aprendizaje no supervisado

- Sólo se dan datos de entrada.
- Objetivo: encontrar 'patrones interesantes' en los datos.
- No se dispone de un target/salida.
- Ejemplos:
  - Agrupar colecciones de datos
  - Limpieza de *outliers*
  - Segmentar vídeo o audio
  - Segmentación de clientes
  - Organización de colecciones de documentos
  - Aprendizaje de funciones de densidad de probabilidad

$$D = \{\mathbf{x}_i\}_{i=1}^N$$

**Agrupamiento o clustering:** dividir un conjunto de datos (observaciones) en  $K$  partes que contengan datos parecidos entre sí y diferentes de los datos contenidos en los otros grupos.

**Detección de novedad:** aprender el soporte de la distribución de datos para ayudar a decidir si un dato de test pertenece a esa distribución o no.

**Densidad de probabilidad:** aprendizaje de densidades de probabilidad a partir de una muestra finita de datos

**Reducción de dimensionalidad:** aprender transformaciones de las variables de entrada para conseguir un conjunto con menos variables

#### Otros aprendizajes

Existen **otros paradigmas** como el aprendizaje semi-supervisado o el **aprendizaje por refuerzo**:

- Los datos están formados por secuencias de observaciones/decisiones que desembocan en una recompensa
- Se aprende una estrategia para encadenar una secuencia de decisiones que maximicen la recompensa global
- Aplicaciones:
  - Conducción autónoma
  - Robótica
  - Videojuegos y juegos de mesa
  - Diseño de estrategias de trading

### Aprendizaje incremental vs batch

Condicionado por el problema de optimización que haya que resolver:

- **Batch:** se dispone de una vez del conjunto completo de datos de entrenamiento. Generalmente es el caso cuando buscamos una optimización que nos dé un modelo globalmente óptimo
- **Incremental:** no se dispone del conjunto de datos de entrenamiento completo. Estrategia a emplear cuando el procesador en el que se ejecuta el algoritmo de entrenamiento no puede con todos los datos a la vez

### Modelos paramétricos

- Número fijo de parámetros
  - Más rápidos
  - Suposiciones más fuertes sobre la naturaleza de las distribuciones de los datos
- Se suele dar a la función estimación una forma paramétrica a priori, y el propósito de diseño es encontrar los valores de los parámetros de acuerdo a cierto objetivo.
- Ejemplos: regresores lineales, regresión logística, agrupamiento k medias, ...

### Modelos no paramétricos

- El número de parámetros crece con la cantidad de muestras de entrenamiento.
  - Más flexibles
  - Intratables computacionalmente para datasets más grandes.
- La forma analítica del modelo no se asume a priori.
- Ejemplos: KNN, SVM, ...

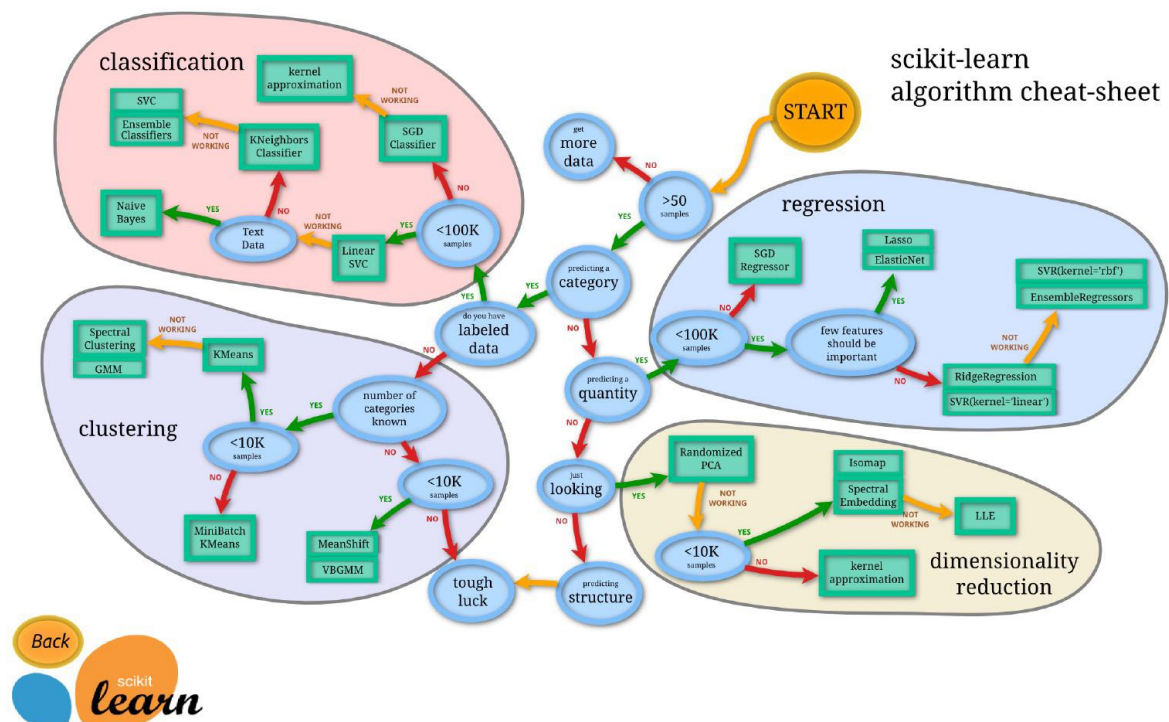
### K nearest neighbor KNN

Mirar los K vecinos del conjunto de entrenamiento más cercanos a la muestra de test, contar los miembros de cada clase y devolver la fracción empírica como la estimación. Distancia: ejemplo la euclídea.

La maldición de la dimensionalidad

- KNN funciona bien con datos de baja dimensionalidad, pero no alta.
- El método deja de ser local a altas dimensiones.
- El problema de mirar vecinos que están tan lejos es que puede que no sean buenos predictores.

Solución: modelos paramétricos, que hacen suposiciones sobre la naturaleza de las distribuciones de los datos.



## Regresión (aprendizaje supervisado)

- La variable de respuesta es continua  $Y_i \in \mathbb{R}$ .
- Objetivo: predecir valores (número real).
- Aplicaciones del mundo real:
  - Predecir el precio del mercado de valores de mañana dadas las condiciones del mercado actual y otra posible información lateral.
  - Predecir la edad de un espectador viendo un video de YouTube dado.
  - Predecir la localización en 3D de un brazo robot dadas señales de control enviadas a sus motores.
  - Predecir la cantidad de antígeno específico de la próstata en el cuerpo como una función de un número de diferentes medidas clínicas.
  - Predecir la temperatura en cualquier localización dentro de un edificio utilizando datos meteorológicos, tiempo, sensores, ...
- Ejemplos: lineal vs polinómica
  - Lineal:  $y = W_0 + W_1 * X_1 + \dots + W_n * X_n$
  - Polinómica:  $y = W_0 + W_1 * X + W_2 * X^2 + \dots + W_n * X^n$

## Ajustando una curva polinómica

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j$$

- Función lineal en los coeficientes,  $\mathbf{w}$ .
- Ajustar el polinomio a los datos de entrenamiento.
- Minimizar una función error, que mide la diferencia entre la función  $y(x, \mathbf{w})$ , para unos valores de  $\mathbf{w}$  dados y los datos del conjunto de entrenamiento.
- Función error ejemplo: suma de los errores cuadráticos.

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

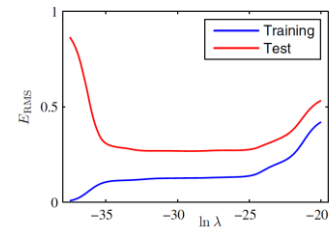
- Función no-negativa
- Minimizar
- Calcular derivadas con respecto a los parámetros e igualar a cero
- Solución: lineal, solución única
- Grado M? Selección del modelo / model selection

- **Sobre-entrenamiento: Overfitting**

- ¿Cómo evitarlo? Ajustar número de parámetros al número de muestras

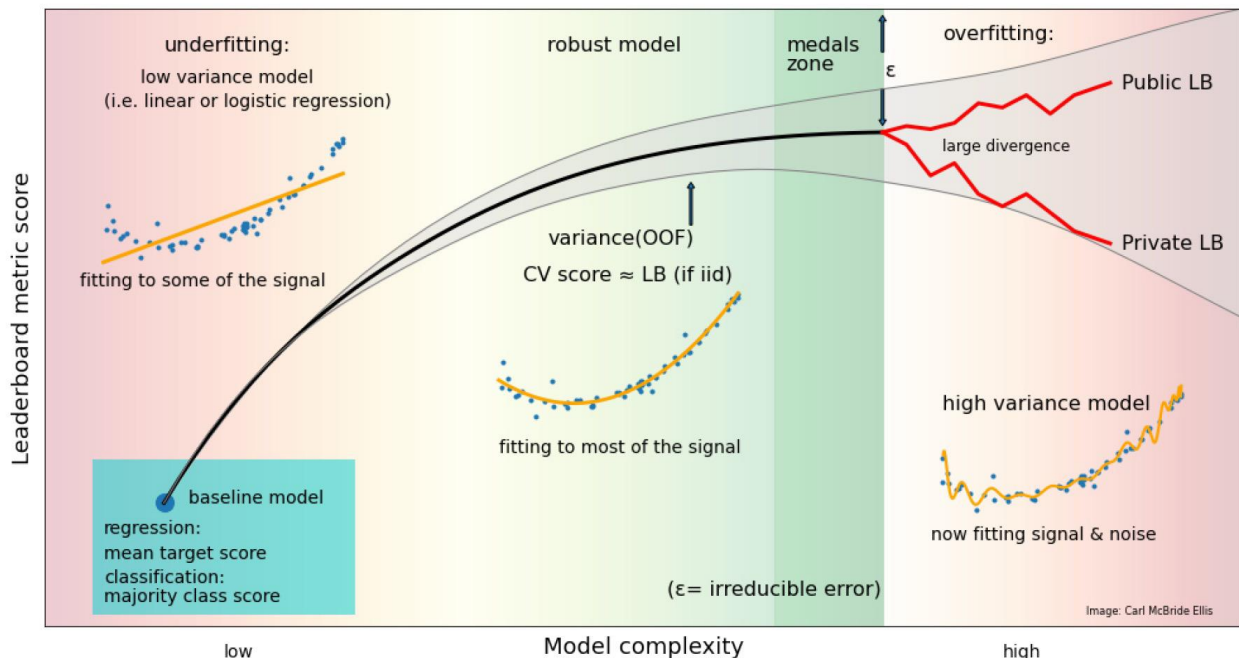
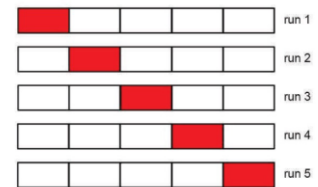
- **Regularización:**

- Regresión Ridge: Añadir un término de penalización para que los parámetros no alcancen valores grandes.
- Conjunto de entrenamiento: usado para calcular los coeficientes,  $w$ .
- Conjunto de validación: usado para seleccionar el modelo,  $M$  o  $\lambda$



- **Validación cruzada / cross validation (CV)**

- Dividir el conjunto de entrenamiento en K particiones.
- Para cada partición, entrenar en todas menos en una, donde testamos.
- Error de test: Calculamos el error promedio en todas las particiones.
- Útil cuando hay pocos datos.
- Leave-one out cross validation.



## PRÁCTICA 1: Regresión

- 1- [Regresión lineal](#)
- 2- [Regresión KNN](#)
- 3- [Regresión lineal y polinómica](#)

Partes imprescindibles de un programa para visualizar datos (ejemplo completo de predicción en un modelo de regresión):

- Cargar datos ( `read_csv()` ) - *pandas*
- Separar (si necesario) X (entrada) e Y (salida)
- Seleccionar las columnas
- Dividir entrenamiento y test (`train_test_split` – módulo *sklearn*)
  - Xtrain
  - Xtest
  - Ytrain
  - Ytest
- Normalizar datos ( `standarScaler()` ) - *sklearn.preprocessing.StandardScaler*
- Aplicar el modelo: entrenar, validar, testear.
  - `.train`
  - `.score`
  - `.predict`  $\hat{Y}$

## Clasificación (aprendizaje supervisado)

Varían los modelos, ya que la Y tiene que ser una clase/categoría y NO un número real.

Aprender un mapeo de entradas a salidas, donde  $Y \in \{1, \dots, C\}$

- Si  $C=2$ , clasificación binaria
- Si  $C>2$ , clasificación multiclase
- Si las etiquetas no son mutuamente exclusivas, clasificación multi-etiqueta

Necesidad de predicciones probabilísticas:

- Para manejar casos ambiguos, es deseable retornar una probabilidad.
- Corresponde a la moda de la distribución:  
MAP (Maximum a posteriori)
- Evaluación del riesgo: importante especialmente en medicina y finanzas.

$$\hat{y} = \hat{f}(\mathbf{x}) = \underset{c=1}{\operatorname{argmax}}^C p(y = c | \mathbf{x}, \mathcal{D})$$

Aplicaciones del mundo real

- Clasificación de documentos y filtrado de spam
- Bolsas de palabras / bag of words
- Clasificación de flores
- Detección y reconocimiento facial

Machine learning \ Manual counting	True	False
	True	False
True	True Positive (TP)	False Positive (FP)
False	False Negative (FN)	True Negative (TN)

Equations:

$$\text{False positive rate (FPR)} = \frac{FP}{FP+TN}$$

$$\text{False negative rate (FNR)} = \frac{FN}{FN+TP}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Youden index} = \text{Sensitivity} + \text{Specificity} - 1$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

## PRÁCTICA 2: Clasificación

- 1- [Regresión Logística](#)
- 2- [Clasificación KNN](#)
- 3- [Más clasificadores](#)
- 4- [Clasificadores multiclase](#)
- 5- [Otros clasificadores](#)

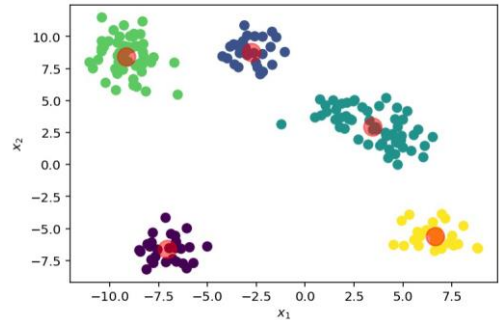


## Modelos no supervisados

### Agrupamiento o clustering

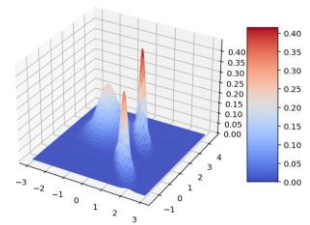
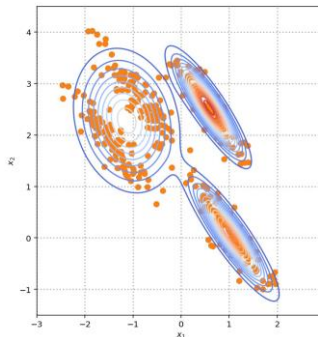
**Recordemos:** en los datos no supervisados NO tenemos los datos de salida Y.

- El agrupamiento es quizá el problema más inmediato que se puede resolver con aprendizaje automático.
- El algoritmo de agrupamiento más referenciado posiblemente sea el Kmedias: divide los datos en K grupos (K es un parámetro que se debe fijar a priori) encontrando los K centroides o prototipos que mejor representan estos grupos.



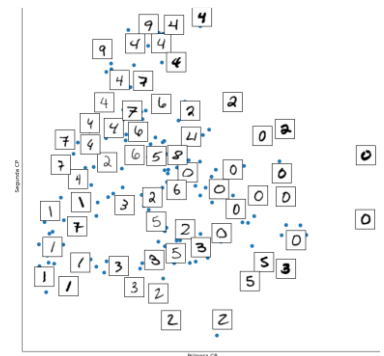
### Aprendizaje de densidad de probabilidad

- Caracterización de los datos. La caracterización más completa desde un punto de vista matemático es mediante la función de densidad de probabilidad que modela el proceso generativo de los datos.
- Estimar densidades de probabilidad es complejo, especialmente si no se dispone de información a priori sobre el proceso real responsable de la generación de los datos.
- Frecuente: mejora considerablemente las prestaciones de los usos de densidades aproximadas en la construcción de los modelos mismos.
- Quizá el método más habitual para aprender densidades de probabilidad que se ajusten a una colección de datos definidos en términos de variables numéricas continuas sea la mezcla de Gaussianas.



### Reducción de dimensionalidad

- Simplificar la carga computacional de los optimizadores al emplear datos con menos dimensiones
- Eliminar componentes ruidosas o que no estén alineadas con el patrón que queremos capturar
- El algoritmo más comúnmente empleado para este tipo de problemas es el Análisis en Componentes Principales o PCA.

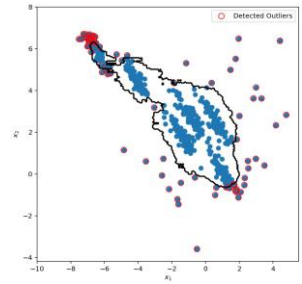


### Detección de novedad

- Muy relacionada con el aprendizaje de densidades de probabilidad.
- Evaluar en qué medida cada una de las observaciones puede ser un *outlier*, es decir, es lo suficientemente diferente del resto de observaciones del conjunto de entrenamiento como para

considerar que es una observación ruidosa o que no se ha generado con el modelo que suponemos ha generado la mayoría de los datos.

- No hace falta aprender una densidad de probabilidad que represente los datos muy fielmente, sólo nos basta con encontrar una métrica que nos permita decidir si un dato está suficientemente lejos de otras zonas más densamente pobladas de observaciones.

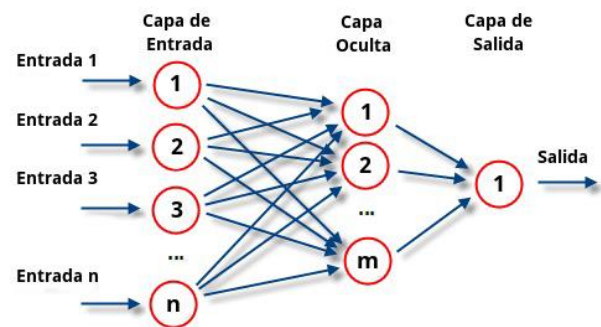
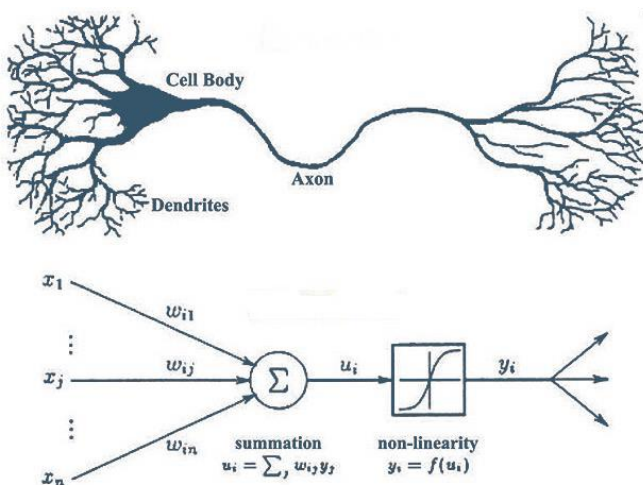


### PRÁCTICA 3: Modelos no supervisados

- 1- [Principales modelos no supervisados](#)
- 2- [Otros modelos no supervisados](#)

### Redes neuronales

- Posiblemente el método más representativo del aprendizaje automático. Siempre vuelven al foco.
- Son un método general que puede adaptarse a problemas de clasificación, regresión, detección de patrones, etc.
- Inspiración en las redes de neuronas naturales.
- Conexión con neurociencia.
- Una neurona es una unidad elemental de cálculo que recibe unas entradas, las combina linealmente y dispara con una función sigmoide.



### PRÁCTICA 4: NLP y Redes Neuronales

- 1- [Redes neuronales recurrentes \(RNNs\) para análisis de opinión.](#)

A screenshot of a web browser window. The address bar shows a URL from 'forms.zohopublic.eu'. The page content is a white box with a green checkmark icon at the top. Below the icon, the text reads: '¡Hola Víctor Beneito López - Big Data!', 'El resultado de tu evaluación de este módulo de es 7.60', and 'Muchas gracias por tu participación.' The browser's taskbar at the bottom shows various icons including Windows, search, and several open applications like Chrome and Word. The system clock in the bottom right corner shows '17:17' and '03/03/2023'.

## MODULO 6 - Arquitecturas Cloud y Big Data

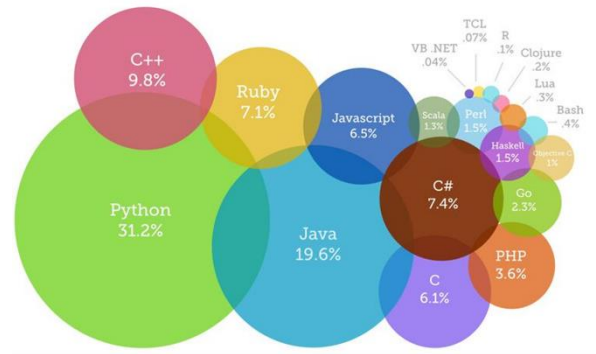
### DIA 26 – Arquitecturas Cloud y Big Data

#### Presentación 1: Arquitecturas Cloud & Big Data

- El lenguaje de **Base de Datos** (relacionales) es **SQL**. Lenguaje de consultas.

- **Python** se utiliza en el 31,2% de los casos como lenguaje de programación.

- Lenguaje de alto nivel, fácil de aprender.
- Expresivo y legible.
- Sintaxis elegante y tipado dinámico y fuerte.
- Multiparadigma.
- Interpretado: se lee línea por línea y se va ejecutando.
- Multiplataforma.



- **IDE**: ¿Qué es un IDE? Software que nos permite desarrollar de una forma más fácil (ej.: Anaconda, Google Colab, Visual Code, etc.).
  - Autocompletado de código
  - Coloración sintáctica
  - Navegación de clases, objetos, funciones
- **CRISP-DM**: esquema. 80% del tiempo en Data Preparation.
- **Variables**:
  - Donde se almacenan y se recuperan los datos de un programa.
  - No utilizar palabras reservadas.
  - Los nombres tienen que ser significativos.
- En programación, siempre debemos empezar a contar “en 0” (problemática de **índices**).
- **Funciones**:
  - Permite definir un bloque de código reutilizable, el cual se puede ejecutar muchas veces dentro del programa.
  - Ventaja: reduce el número total de líneas del proyecto.
  - Las hay con parámetros o sin parámetros (o argumentos).
  - Funciones nativas y propias
- **Librerías**: NumPY, Pandas, Plotly, Jupyter, Matplotlib, etc. Buscador: [www.pypi.org](http://www.pypi.org)
- Búsqueda de errores para solucionarlos: [www.stackoverflow.com](http://www.stackoverflow.com),
- Inteligencia Artificial. Machine Learning. Redes Neuronales. ImageNet.
- Google Colab: repaso al contenido de Python
  - Operadores aritméticos. Novedad: // para división entera (int), % para el resto de una división.

#### EJERCICIO 1: [Python Notebook](#)

## DIA 27 – Arquitecturas Cloud & Big Data

### Presentación 2: Arquitecturas Cloud & Big Data

- [GitHub](#): repositorio multifunción. Actualmente se valora como un CV técnico.
- **Práctica**: crear repositorio en GitHub llamado Curso Big Data y subir todos los ejercicios resueltos, separados por carpetas con un nombre descriptivo del ejercicio. Además, creamos un archivo README.md donde deberemos escribir el lenguaje utilizado y la lista de ejercicios con una descripción acotada

### EJERCICIO 2: Condicionales

### EJERCICIO 3: Bucles

## DIA 28 – Arquitecturas Cloud & Big Data

- Colocar archivos en repositorio GITHUB

### EJERCICIO 3: Bucles

### EJERCICIO 4: Programación Funcional

- Funciones **LAMBDA** (función anónima). Útil para llamar directamente a las funciones como parámetro.
- Función **MAP**: aplica una función (1º arg) a todos los elementos de una lista (2º arg). Crea un elemento de tipo "generador", por lo que para su manejo habría que transformarlo a listas.
- Función **FILTER**: aplica una función de comparación (1º arg) a todos los elementos de una lista (2º arg) filtrando el resultado.

## DIA 29 – Arquitecturas Cloud & Big Data

- Corrección de la práctica 4
- Concepto FIFO

## Introducción a PySpark: RDDs

### Presentación

[http://ferestrepoca.github.io/paradigmas-de-programacion/paralela/paralela\\_teoría/index.html](http://ferestrepoca.github.io/paradigmas-de-programacion/paralela/paralela_teoría/index.html)

<https://geekytheory.com/apache-spark-que-es-y-como-funciona/>

- **Apache Spark** es un motor de código abierto desarrollado específicamente para el procesamiento y análisis de datos a gran escala.
- Objetivo: procesar de manera distribuida grandes cantidades de información.
- *Cluster*: conjunto de ordenadores trabajando entre sí, cooperando para lograr un objetivo.
- Estructura maestro-esclavo (Master Daemon-Worker Daemon). El Master divide la estructura de datos y designa los Workers a cada una de las divisiones mediante variables compartidas (tipo Broadcast)
- **SPARK**: aportaciones
  - Plataforma de computación para clústers

- Es de propósito general
- Desarrollo simplificado
- Trabaja en memoria
- Rápido
- Permite trabajo interactivo, streaming
- **DRIVER:** programa que corre el MASTER (main).
- En Apache Spark:
  - Una **sesión** es una conexión de un cliente con el cluster Spark (ej.: una empresa). Sólo se puede tener una sesión abierta en Google Colab.
  - Un **contexto** es un objeto que representa una conexión a un cluster Spark y proporciona un punto de acceso a todas las...
- **RDDs (Resilient Distributed Dataset):**
  - Colección inmutable y distribuida de elementos.
  - Spark automáticamente distribuye los datos y paraleliza las operaciones.
  - Los RDD realmente cargan colecciones de datos.
  - **Closure:** lo que hay que hacer con los RDDs (acciones a realizar).

EJERCICIO 5: [SparkSession Teoría.ipynb](#)

EJERCICIO 6: [Primer RDD Teoría.ipynb](#)

## DIA 30 – Arquitecturas Cloud & Big Data

### Presentación

- **Transformaciones:** cambios/tareas que hay que realizar dentro de cada *closure*. Tras aplicar una transformación, obtenemos un nuevo y modificado RDD basado en el original.
- `nuevo_rdd = rdd_ejercicio.map(lambda x : x*3)`
- **Acciones:** cuando el Master les proporciona las transformaciones a realizar a los Workers. Es aplicar una operación sobre un RDD y obtener un valor como resultado, que dependerá del tipo de operación.
- `nuevo_rdd.collect()` y `sc.stop()`

#### RDDs: Transformaciones más comunes

Transformación	Descripción
<code>map(func)</code>	Crea un nuevo RDD a partir de otro aplicando una transformación a cada elemento original
<code>filter(func)</code>	Crea un nuevo RDD a partir de otro manteniendo solo los elementos de la lista original que cumplan una condición
<code>flatMap(func)</code>	Como map pero cada elemento original se puede mapear a 0 o varios elementos de salida
<code>distinct()</code>	Crea un nuevo RDD a partir de otro eliminando duplicados
<code>union(otroRDD)</code>	Une dos RDD en uno
<code>sample()</code>	Obtiene un RDD con una muestra obtenida con reemplazamiento (o sin) a partir de otro RDD.

#### RDDs: Acciones más comunes

Acción	Descripción
<code>count()</code>	Devuelve el número de elementos del RDD
<code>reduce(func)</code>	Agrega los elementos del RDD usando <i>func</i>
<code>take(n)</code>	Devuelve una lista con los primeros n elementos del RDD
<code>collect()</code>	Devuelve una lista con todos los elementos del RDD
<code>takeOrdered(n[,key=func])</code>	Devuelve n elementos en orden ascendente. Opcionalmente se puede especificar la clave de ordenación

EJERCICIO 7: [Transformaciones y Acciones sobre RDDs](#)

## EJERCICIO SPACE X

### Carpeta

- 1- [Data Collection API](#)
- 2- [EDA](#)
- 3- [EDA using SQL](#)
- 4- [EDA with Data Visualization](#)
- 5- [Interactive Visual Analytics con Folium](#)
- 6- [Machine learning predictions](#)