



<https://forms.office.com/r/yB2ry3aV19>

Quiz 29-ene-2026



TAREA

- Revisión artículos Boletín EconomIA



PRIMER AVANCE DEL PROYECTO

1. Presentación en clase sobre:

- Título del proyecto
- Integrantes
- Objetivo
- Descripción
- Desafíos
- Entrega de valor
- Stakeholders*
- Técnicas que se utilizarán
- Fuentes de datos
- Variables

Duración: Max. 10 minutos de presentación + 5 minutos de comentarios / preguntas del resto de compañeros y docente.



* En español, “partes interesadas” o “grupos de interés”. Son todas las personas, grupos u organizaciones que tienen interés o se ven afectados por las actividades, decisiones o resultados de un proyecto, empresa o política pública.



PRIMER AVANCE DEL PROYECTO



2. Documento escrito con:

- Título del proyecto
- Integrantes
- Objetivo
- Descripción
- Desafíos
- Entrega de valor
- Stakeholders*
- Técnicas que se utilizarán
- Fuentes de datos
- Variables

Extensión: Max. 1.000 palabras.



* En español, “partes interesadas” o “grupos de interés”. Son todas las personas, grupos u organizaciones que tienen interés o se ven afectados por las actividades, decisiones o resultados de un proyecto, empresa o política pública.



PRIMER AVANCE DEL PROYECTO

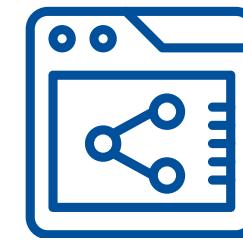


3. Creación de repositorio en Github para el proyecto.

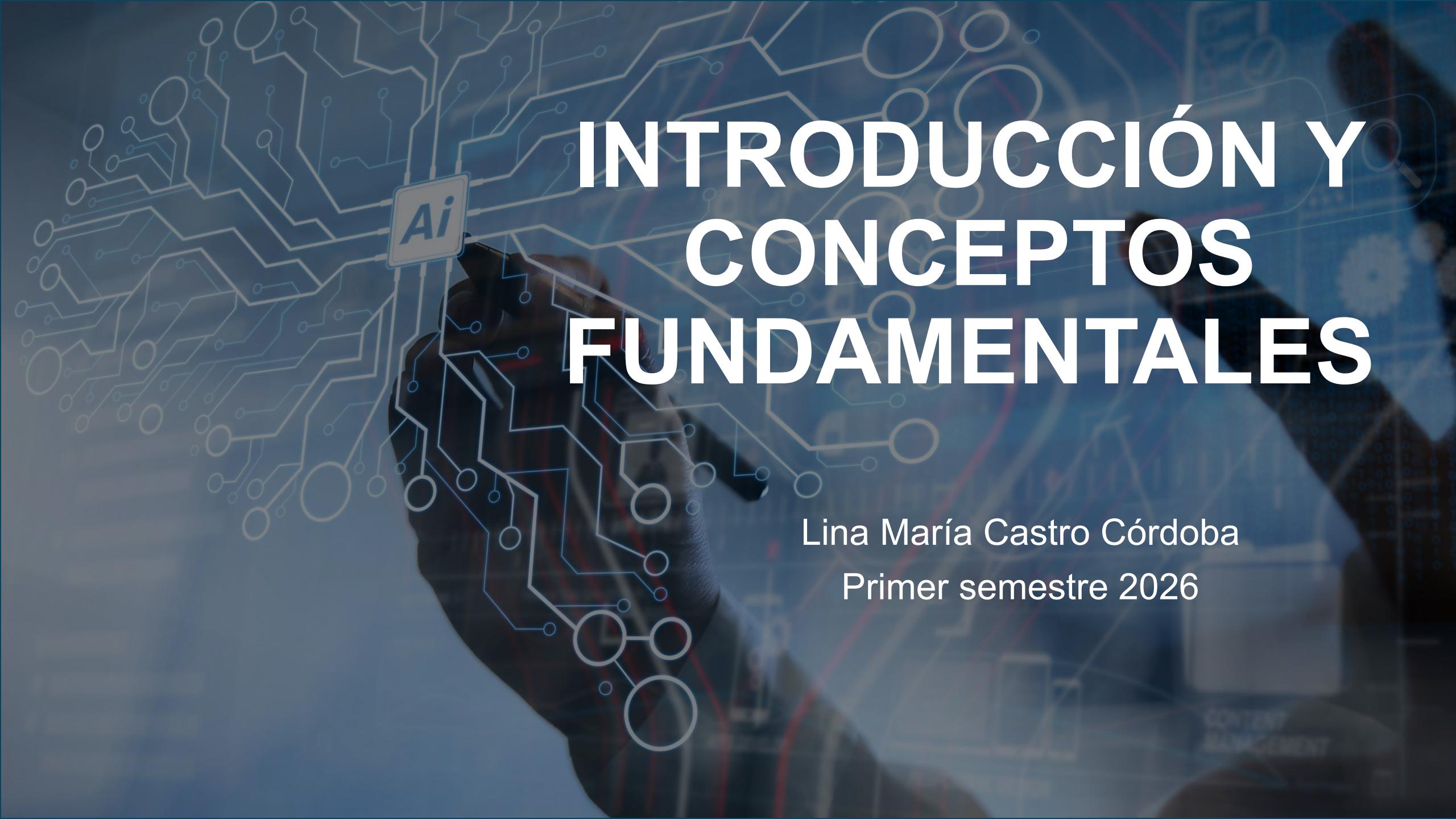
Debe contener:

- Archivo requirements.txt
- Archivo README con la redacción del planteamiento del proyecto
- El repositorio debe ser público



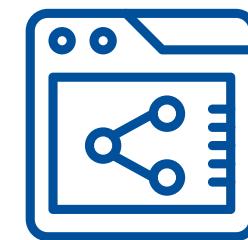


EJEMPLO

A background photograph of a person's face, partially obscured by a translucent blue overlay. The overlay features a complex circuit board pattern in white and light blue. In the center, there is a graphic of a central processing unit (CPU) with the letters "Ai" written on it.

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Lina María Castro Córdoba
Primer semestre 2026



CONCEPTOS BÁSICOS

CIENCIA - MÉTODO CIENTÍFICO

- Estudio sistemático de la estructura y comportamiento del mundo físico y natural, mediante la **observación y experimentación**.
- Se basa en la construcción, testeo y comprobación de una hipótesis.
- Busca identificar relaciones causales.



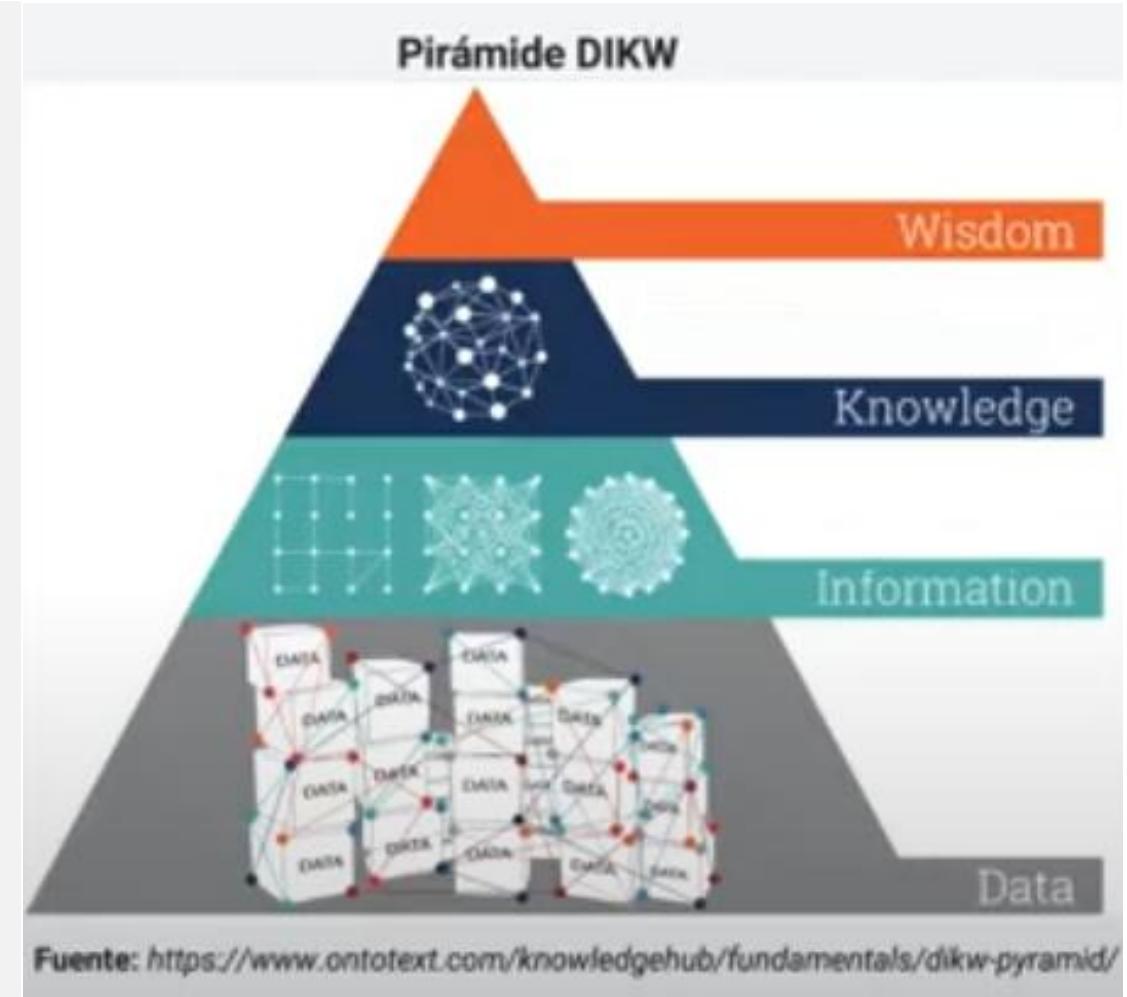
CIENCIA DE DATOS

- Es una disciplina que **combina estadísticas, programación y conocimiento del dominio** para extraer valor de grandes volúmenes de datos.
- **Involucra todo el ciclo de vida del dato**, desde la recolección y limpieza hasta el análisis, modelado y visualización.
- Utiliza herramientas y lenguajes como **Python, R, SQL, pandas, scikit-learn y más**.
- Permite **descubrir patrones, generar predicciones y tomar decisiones basadas en evidencia**.
- Su foco está en **correlaciones y patrones**, mas que en causalidad.
- Transforma los datos en **información y conocimiento**.
- Es **clave en múltiples sectores**, como salud, finanzas, marketing, logística, políticas públicas, entre otros.
- **Relevante en la era digital**, donde los datos son uno de los activos más valiosos para la innovación y la competitividad.

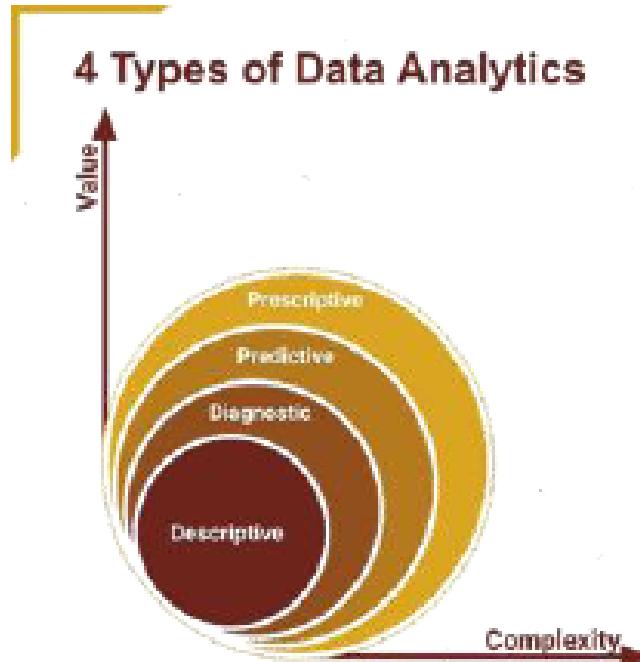


DATOS VS. INFORMACIÓN

- **Datos:** abstracciones o mediciones obtenidas del mundo real.
- **Información:** datos que han sido procesados, estructurados o contextualizados de manera que son significativos para las personas.
- **Conocimiento:** Información que ha sido interpretada y comprendida por una persona, de manera que puede actuar a partir de ella.
- **Sabiduría:** actuar de forma apropiada a partir del conocimiento.



TIPOS DE ANÁLISIS DE DATOS



www.simplilearn.com/data-analytics

1. Descriptivo (Descriptive):

1. Pregunta: ¿Qué pasó? (Pasado)
2. Métodos: Estadísticas descriptivas y Análisis Exploratorio de Datos (EDA), Análisis Ad Hoc.
3. Propósito: Proporciona una comprensión básica de los datos y describe lo que ha ocurrido en el pasado.

2. Diagnóstico (Diagnostic):

1. Pregunta: ¿Por qué pasó? (Pasado)
2. Métodos: Análisis Exploratorio de Datos (EDA), Profundizaciones y descubrimientos.
3. Propósito: Examina los datos para entender las causas subyacentes de los eventos y comportamientos observados.

Pasado

3. Predictivo (Predictive):

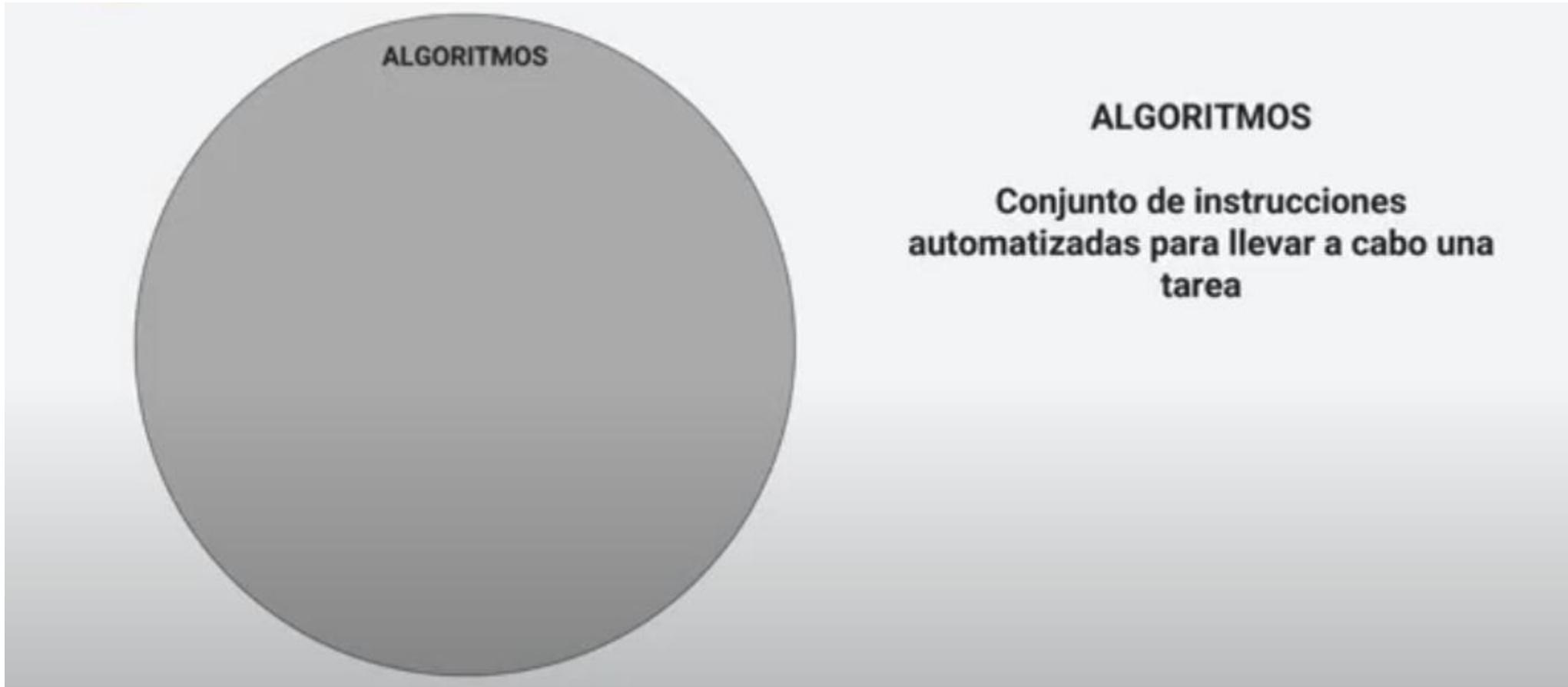
1. Pregunta: ¿Qué pasará? (Futuro)
2. Métodos: Modelado predictivo y estadístico
3. Propósito: Utiliza datos históricos y técnicas estadísticas para predecir eventos y comportamientos futuros.

4. Prescriptivo (Prescriptive):

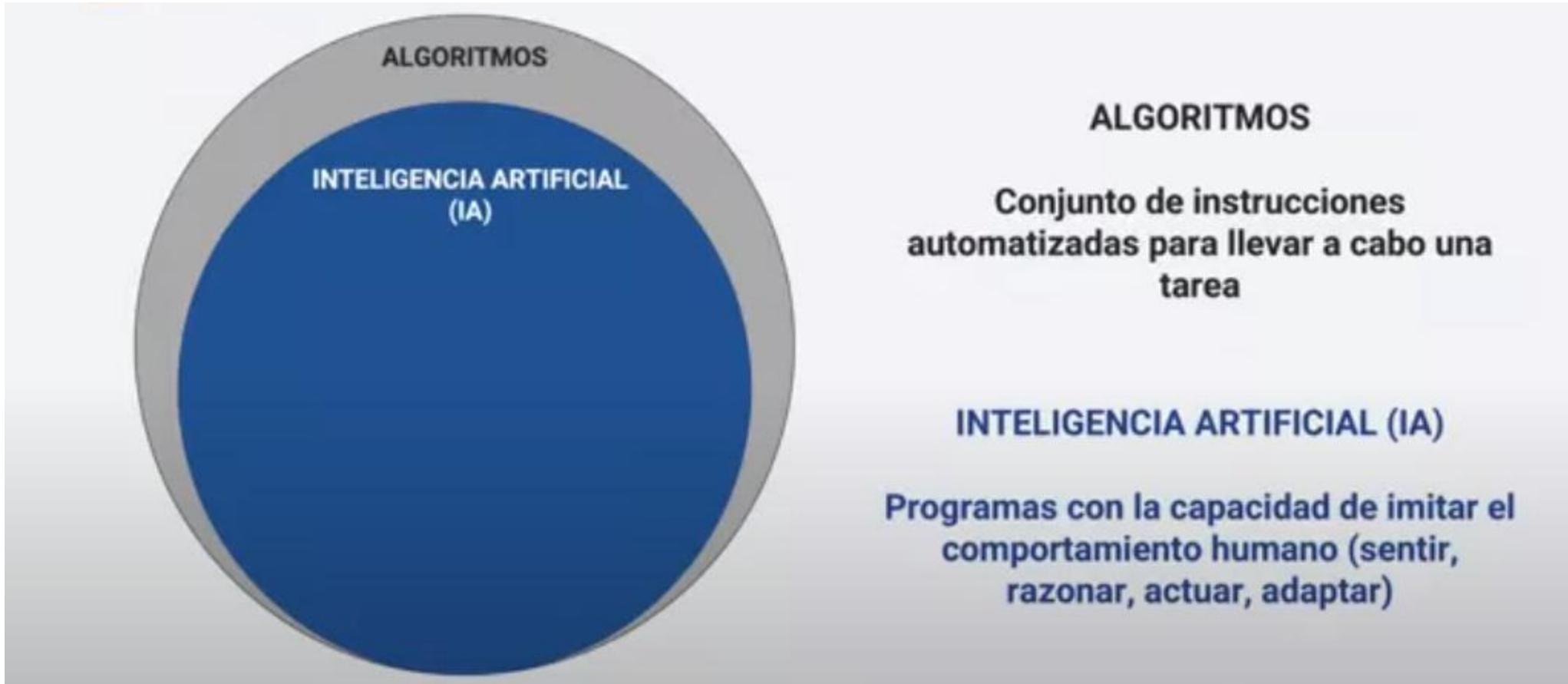
1. Pregunta: ¿Qué debo hacer? (Futuro)
2. Métodos: Optimización y pruebas aleatorias
3. Propósito: Proporciona recomendaciones sobre acciones a tomar para alcanzar objetivos deseados basándose en análisis predictivos y optimización.

Futuro

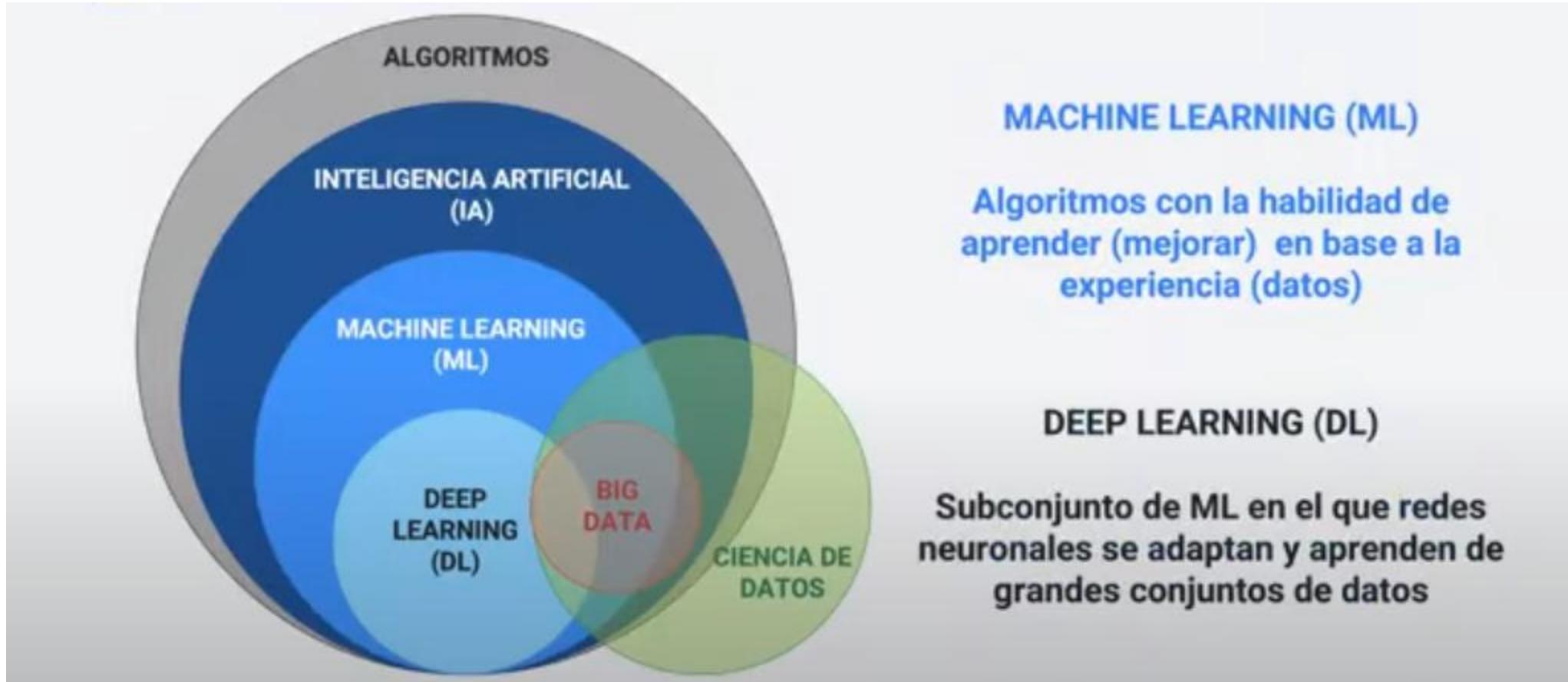
CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL



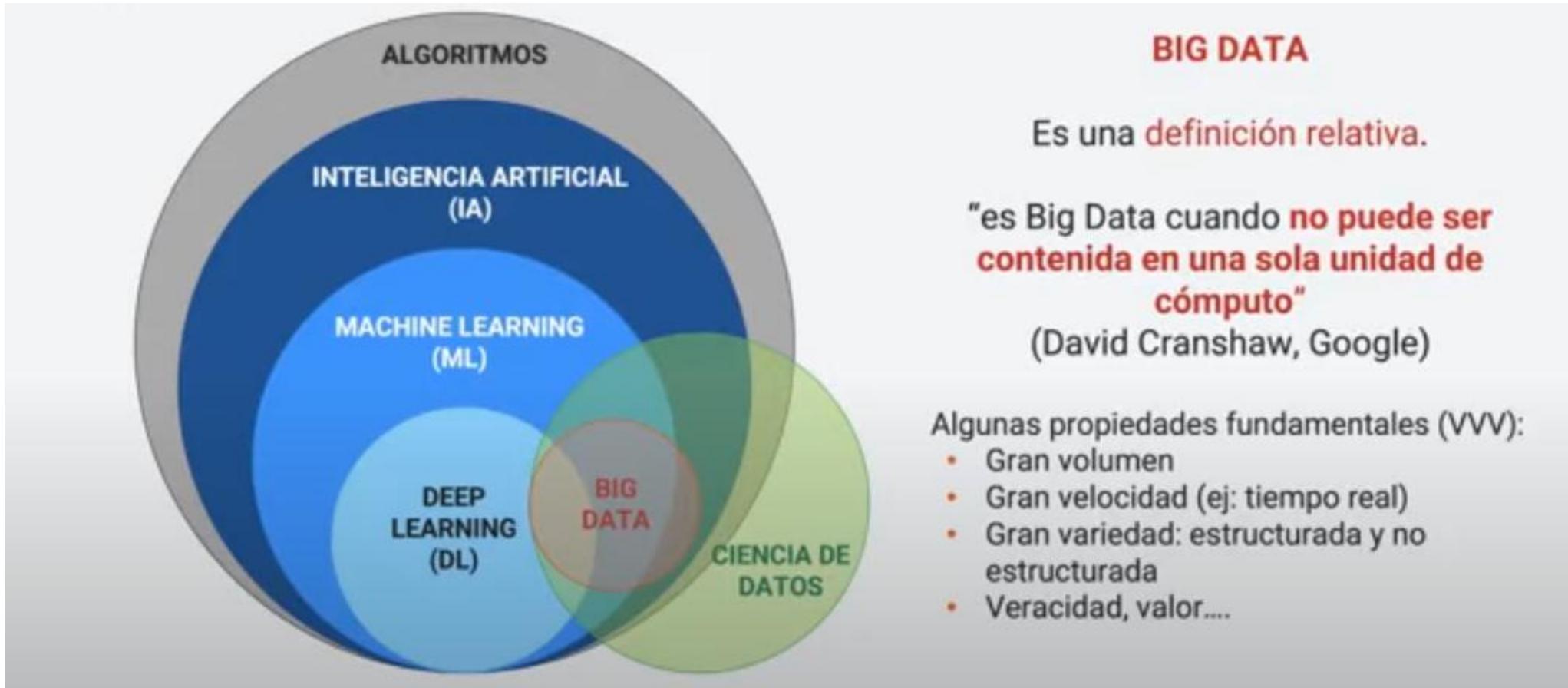
CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL



CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

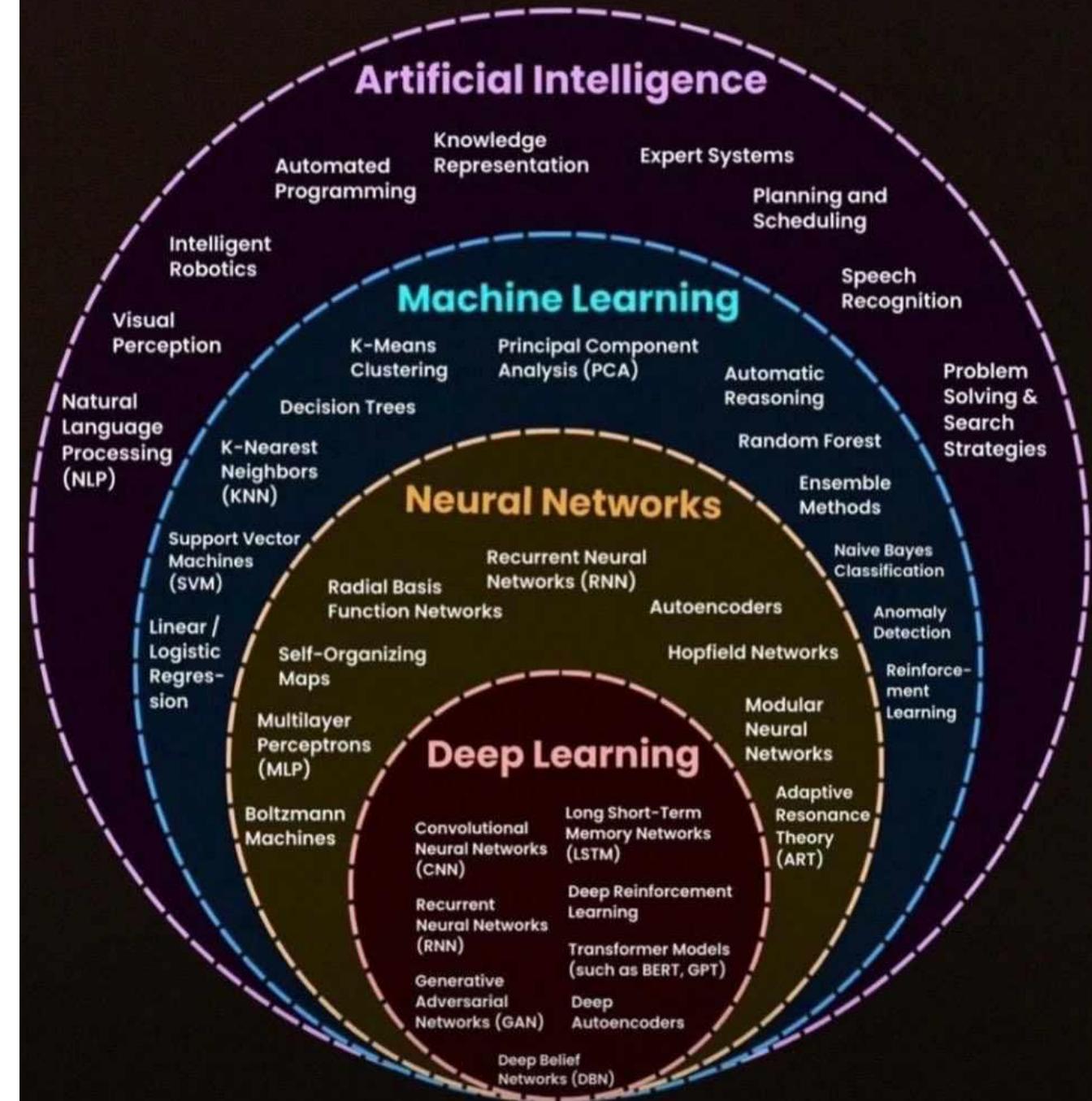


CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL



MACHINE LEARNING

Disciplina científica cuyas técnicas permiten a los ordenadores **aprender de forma automática** a partir de un **conjunto de datos**, de tal forma que sean capaces de hacer predicciones sobre un proceso o describirlo de forma compacta.





MACHINE LEARNING

APRENDIZAJE SUPERVISADO

Algoritmos que permiten inferir, a partir de ejemplos, una función o fórmula matemática que **relacione un conjunto de atributos con una variable de respuesta**. El objetivo es predecir (generalizar) **la respuesta** ante futuras observaciones de los atributos.

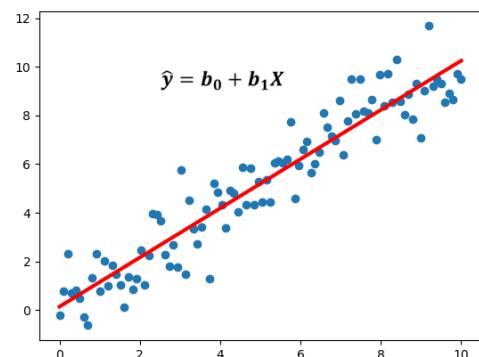
APRENDIZAJE NO SUPERVISADO

Algoritmos que permiten describir cómo están organizados o agrupados un conjunto de datos que **no tienen una variable de respuesta**. El **objetivo** de este método es **conseguir agrupaciones de datos** no detectables a simple vista, **con base en las variables** que describen cada uno de los ejemplos de la muestra.

APRENDIZAJE SUPERVISADO

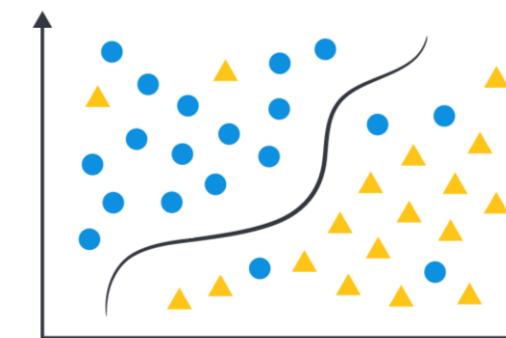
REGRESIÓN

Salario	Nivel de estudios	Sexo	Años de experiencia
\$1.000	Bachillerato	M	2
\$5.500	Doctorado	F	5
\$3.600	Maestría	M	3
\$2.800	Pregrado	F	4



CLASIFICACIÓN

Pobre	Nivel de estudios	Sexo	Salario
SI	Bachillerato	M	\$1.000
NO	Doctorado	F	\$5.500
NO	Maestría	M	\$3.600
SI	Primaria	F	\$500



ALGUNOS ALGORITMOS DE APRENDIZAJE SUPERVISADO

REGRESIÓN LINEAL

Permite predecir el valor de una variable continua como una función lineal de las variables de entrada.

ÁRBOLES DE DECISIÓN

Estructura en forma de árbol que toma decisiones basadas en reglas if-then.

GRADIENT BOOSTING MACHINES

Modelo que construye iterativamente árboles de decisión en serie, cada uno corrigiendo los errores de los anteriores, mejorando el rendimiento en cada iteración.

REGRESIÓN LOGÍSTICA

Permite estimar la probabilidad de que ocurra (o no) un evento como función de las variables de entrada.

RANDOM FOREST

Conjunto de árboles de decisión que se entrena de manera independiente y cuyas predicciones se combinan.

XGBOOST

Es una implementación avanzada y eficiente de Gradient Boosting, optimizada para mayor velocidad y rendimiento.

KNN

Clasifica puntos según la mayoría de votos de sus K vecinos más cercanos teniendo en cuenta las características o variables de entrada.

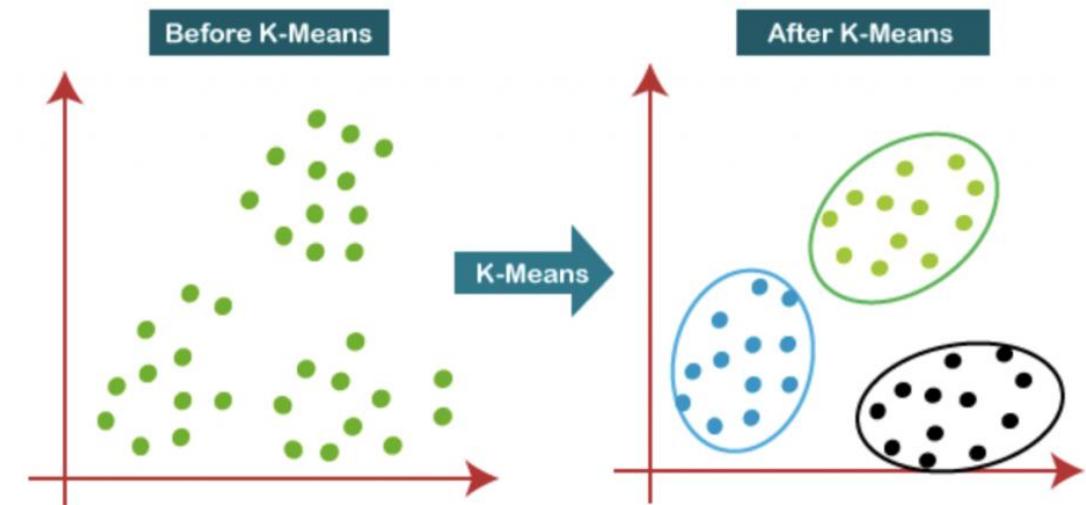
SUPPORT VECTOR MACHINES

Algoritmo de clasificación que encuentra el hiperplano óptimo que separa las clases teniendo en cuenta las características.



APRENDIZAJE NO SUPERVISADO

Cliente	Edad	Sexo	Compras mensuales
Fernando Mora	35	M	\$1.000
Andrea Parra	20	F	\$5.500
Cristina Ramírez	50	F	\$3.600
Juan Gómez	65	M	\$2.800





ALGUNOS ALGORITMOS DE APRENDIZAJE NO SUPERVISADO

K-MEANS

Algoritmo de clustering que agrupa un conjunto de datos en K grupos distintos basados en la distancia. Cada punto de datos se asigna al clúster cuyo centroide está más cercano, y los centroides se recalculan iterativamente hasta que las asignaciones de los puntos no cambian.

Aplicaciones: Segmentación de clientes, análisis de patrones.

PCA

Método de reducción de dimensionalidad que transforma los datos a un nuevo sistema de coordenadas donde las nuevas variables (componentes principales) son combinaciones lineales de las variables originales, ordenadas por la varianza que explican. Se utiliza para simplificar los datos mientras se retiene la mayor variabilidad posible.

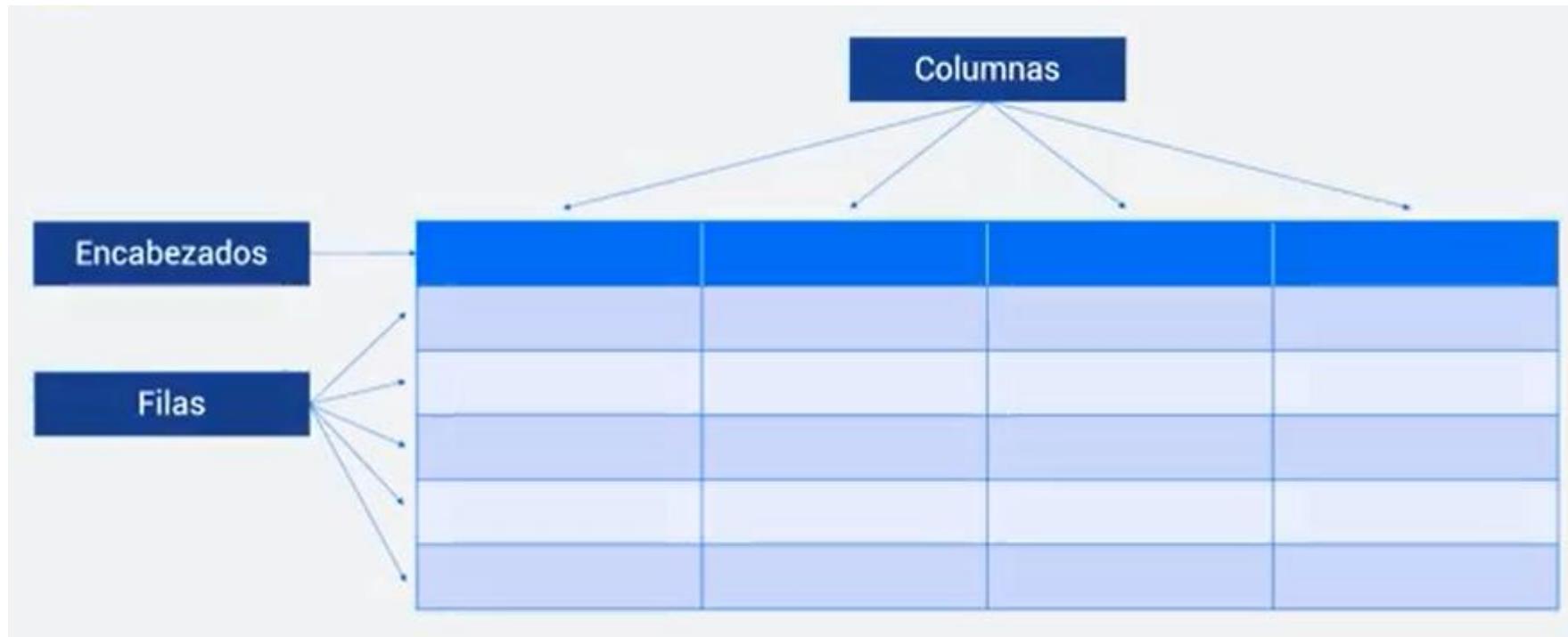
Aplicaciones: Visualización de datos, compresión de datos, preprocesamiento para otros algoritmos.

APRIORI

Algoritmo de reglas de asociación que se utiliza para descubrir relaciones entre variables en grandes conjuntos de datos. Busca conjuntos frecuentes de ítems y genera reglas que describen relaciones entre los ítems.

Aplicaciones: Análisis de cestas de compra, detección de patrones en transacciones, recomendadores de productos.

DATOS ESTRUCTURADOS



Archivos de texto
simple



Planillas de
cálculo



Bases de datos
relacionales

DATOS SEMIESTRUCTURADOS

Los datos semiestructurados cuentan con una organización jerárquica, a partir de llaves y etiquetas semánticas.

Ejemplo formato JSON

JavaScript Object Notation

```
{  
    'nombre': 'Luis',  
    'apellido': 'Perez',  
    'mail': 'luisperez@mail.com',  
    'comprasMes': 178540  
},  
{  
    'nombre': 'Victoria',  
    'apellido': 'Solar',  
    'mail': 'vsolar@mail.com',  
    'compras': {'Cuaderno cuadros':5,'Lapiz grafito':10,'Goma de borrar':8},  
    'telefono':5687654325  
}
```

> DATOS NO ESTRUCTURADOS



No siguen una organización o jerarquía interna clara.



Pueden contener mucha información cualitativa.



Requieren de herramientas de análisis distintas como PLN.

Ejemplos:

- Audios
- Imágenes
- Documentos
- Comentarios en redes sociales



HERRAMIENTAS ANALÍTICAS

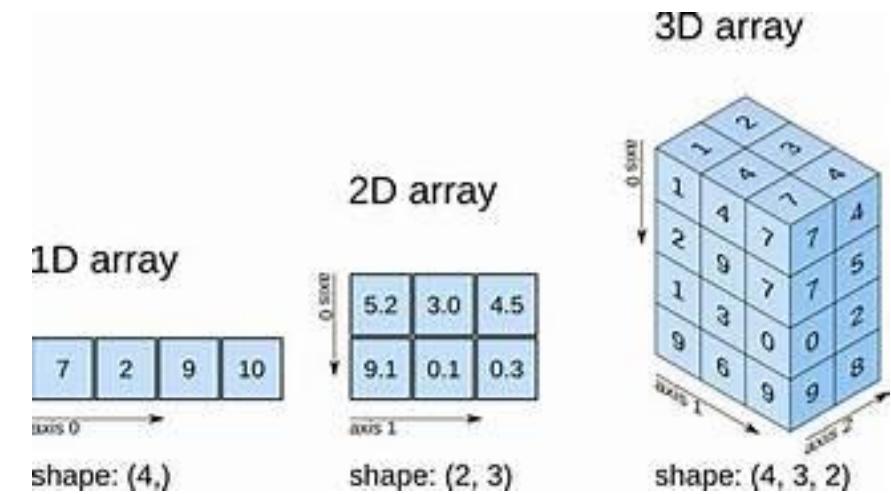
> PYTHON (www.python.org)

- Es un **lenguaje de programación interpretado**, de alto nivel y fácil de aprender, con una sintaxis clara y legible. Los comandos o instrucciones se ejecutan paso por paso.
- Es **multipropósito**, usado en desarrollo web, automatización, análisis de datos, inteligencia artificial, entre otros.
- Cuenta con una **amplia y activa comunidad y miles de bibliotecas**, lo que acelera el desarrollo de proyectos complejos.
- **Relevante en ciencia de datos y machine learning** por su versatilidad, facilidad de integración y herramientas especializadas.



➤ NUMPY (numpy.org)

- Es una **biblioteca fundamental** de Python para el cálculo numérico y científico.
- Proporciona **arreglos multidimensionales (ndarrays)** y funciones eficientes para operar sobre grandes volúmenes de datos.
- **Optimiza el rendimiento** en operaciones matemáticas complejas gracias a su implementación en C.
- **Provee las estructuras**, algoritmos y funciones para la mayoría de las aplicaciones científicas relacionadas con datos numéricos.
- **Es la base de otras bibliotecas** como pandas, scikit-learn y TensorFlow, lo que la hace esencial en ciencia de datos y machine learning.





PANDAS (pandas.pydata.org)

- Es una **biblioteca de Python especializada en la manipulación y análisis de datos estructurados y tablas**.
- Ofrece **estructuras como DataFrame y Series**, que facilitan el manejo de datos tabulares y temporales.
 - **Dataframe**: estructura de datos tabular con etiquetas de filas y columnas.
 - **Series**: objeto tipo arreglo 1-D con etiqueta.
- Permite **leer, limpiar, transformar y exportar datos** de manera eficiente desde múltiples fuentes (CSV, Excel, SQL, etc.).
- **Combina la eficiencia de cálculo de NumPy con capacidades de manipulación de datos típicas de hojas de cálculo** como Excel y bases de datos relacionales.
- Es **fundamental en ciencia de datos y análisis**, por su facilidad de uso y su integración con otras bibliotecas como NumPy, Matplotlib y Scikit-learn.



Series

peppers	
0	3
1	2
2	0
3	1

Series

carrots	
0	0
1	3
2	7
3	2

+

DataFrame

	peppers	carrots
0	3	0
1	2	3
2	0	7
3	1	2

=

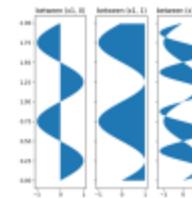
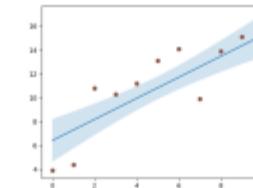


MATPLOTLIB (matplotlib.org)

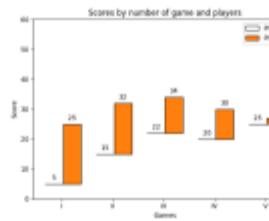
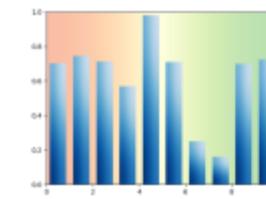
- Es una **biblioteca de Python para crear gráficos** y visualizaciones de datos 2D/3D.
- **Permite generar gráficos variados**, como líneas, barras, histogramas, dispersión, entre otros.
- **Altamente personalizable**, ideal para crear visualizaciones profesionales y adaptadas a distintas necesidades.
- **Relevante en análisis de datos**, ya que facilita la comprensión de patrones y tendencias en los datos.
- **Librería base para otras interfaces de ploteo de alto nivel** como Seaborn (gráficos estadísticos) o Plotly (gráficos interactivos).



Rellenar entre y alfa



Demostración de relleno de Betweenx



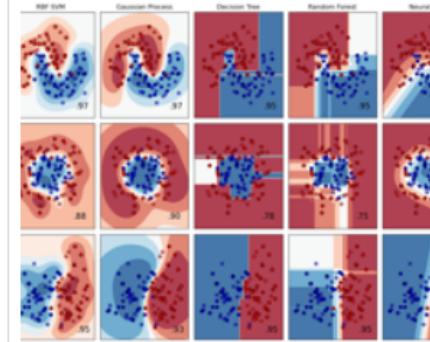
matplotlib

SCIKIT-LEARN (scikit-learn.org)

- Es una **biblioteca de Python** de código abierto para **machine learning** y análisis de datos.
- Ofrece **herramientas simples y eficientes** para tareas como clasificación, regresión, clustering y reducción de dimensionalidad.
- Fácil de usar e integrar con otras bibliotecas como NumPy, pandas y matplotlib.
- Relevante en **ciencia de datos** por su amplia adopción, documentación clara y utilidad en proyectos tanto académicos como industriales.

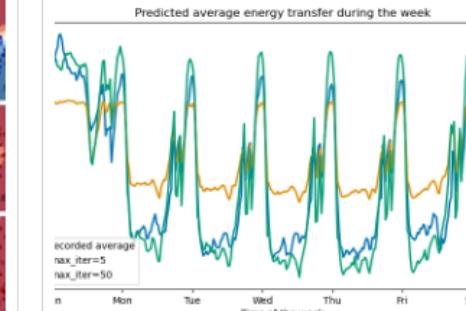


Algorithms: [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [logistic regression](#), and [more...](#)



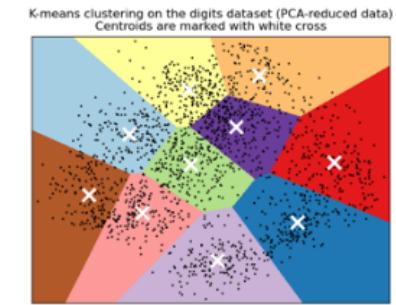
Examples

Algorithms: [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [ridge](#), and [more...](#)



Examples

Algorithms: [k-Means](#), [HDBSCAN](#), [hierarchical clustering](#), and [more...](#)



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Model selection

Comparing, validating and choosing parameters and models.

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data

JUPYTER NOTEBOOKS (jupyter.org)

- Es una **aplicación web** que permite crear y compartir documentos con código en vivo, visualizaciones y texto explicativo.
- **Soporta varios lenguajes de programación** (R, Julia, JavaScript, SQL, C++, tec) aunque es más comúnmente usado con Python para análisis de datos y machine learning.
- **Facilita el desarrollo interactivo**, permitiendo ejecutar bloques de código por separado y ver resultados inmediatos.
- **Ideal para investigación, educación y ciencia de datos**, ya que combina código, gráficos y documentación en un solo entorno.

A screenshot of a Jupyter Notebook interface running in a web browser. The title bar shows multiple tabs: "conda install", "Python for ecolog", "ecology-bug-bbq", "Comparing data...", "Home", "Untitled", and "Carol". The main window displays a notebook cell with the code "print('Hello world of ecology!')". The output of the cell is "Hello world of ecology!". Below the cell is another input field labeled "In []:".

GOOGLE COLAB (colab.research.google.com)

- Es una **plataforma gratuita basada en la nube** que permite escribir y ejecutar código en Python directamente desde el navegador.
- **No requiere instalación**, ya que funciona con notebooks de Jupyter almacenados en Google Drive.
- **Incluye acceso a GPUs y TPUs gratuitas**, lo que facilita el desarrollo de proyectos de machine learning y análisis de datos.
- Ideal para colaborar, ya que **permite compartir notebooks fácilmente** y trabajar en equipo, aunque no en tiempo real.

A screenshot of a Google Colab notebook titled "MyFirstColabNotebook.ipynb". The notebook interface includes a toolbar with file operations, a menu bar with File, Edit, View, Insert, Runtime, Tools, and Help, and a sidebar with tabs for CODE, TEXT, CELL, and EDITING, along with RAM and Disk status. A code cell contains the following Python code:

```
[4] import torch  
  
tensor = torch.cos()  
tensor
```

A tooltip for the `torch.cos()` function is displayed, providing documentation: "cos(input, out=None) -> Tensor", "Returns a new tensor with the cosine of the elements of :attr:`input`.", "... math::", "\text{out}_{\text{i}} = \cos(\text{input}_{\text{i}})", "Args:", "input (Tensor): the input tensor", "out (Tensor, optional): the output tensor", "Example::", and a code example:

```
>>> a = torch.randn(4)  
>>> a  
tensor([ 1.4309,  1.2706, -0.8562,  0.9796])  
>>> torch.cos(a)  
tensor([ 0.1395,  0.2957,  0.6553,  0.5574])
```

VISUAL STUDIO CODE (code.visualstudio.com)

- Es un **editor de código fuente gratuito**, ligero y multiplataforma desarrollado por Microsoft.
- **Soporta múltiples lenguajes de programación**, como Python, JavaScript, HTML, C++, entre otros.
- **Ofrece herramientas integradas** como depuración, control de versiones (Git) y terminal, lo que mejora la productividad del desarrollador.
- Es **altamente personalizable** mediante extensiones, temas y configuraciones para adaptarse a distintos proyectos y flujos de trabajo.



Visual Studio Code

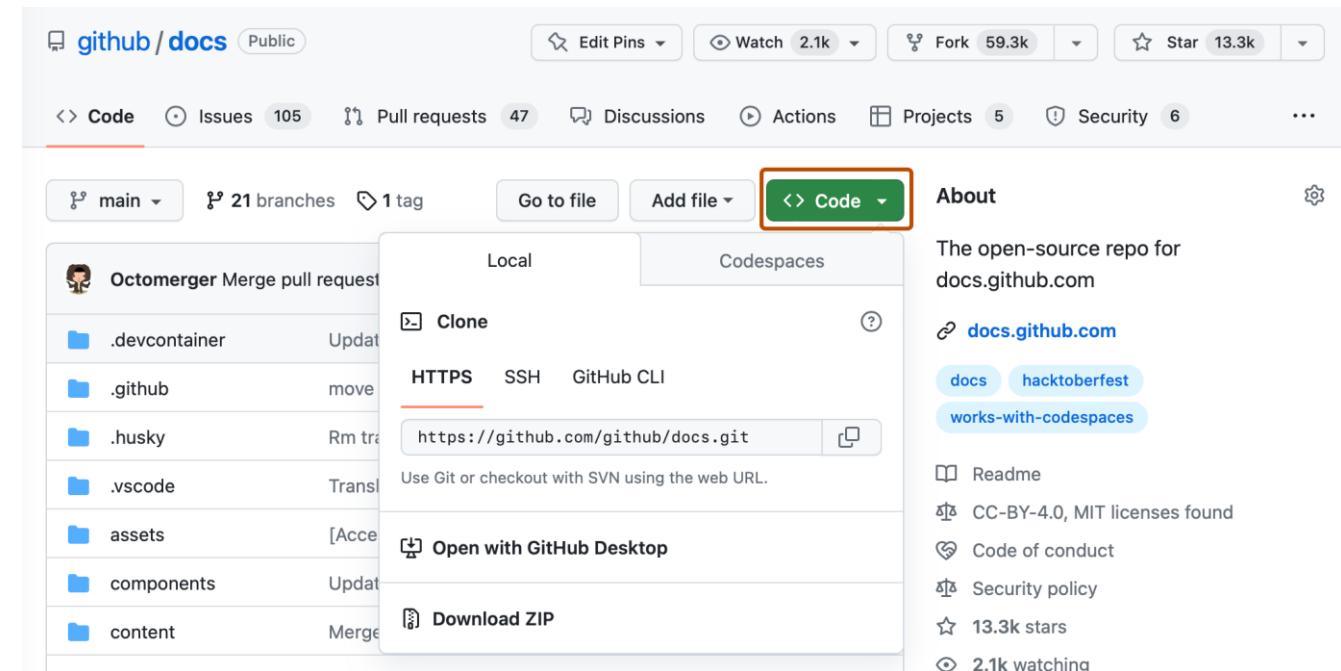
A screenshot of the Visual Studio Code interface. The Explorer sidebar on the left shows a file tree with various projects and files, including '.vscode', 'AdBlockDetector', 'AsyncJs', 'BackOffAlgorithm', 'Docker', 'DockerLEMP', 'ES6', 'GAEvents', 'GitBisect', 'JavascriptModules', 'Jshint', 'Memoization', 'MinimalExpress...', 'Prehooks', 'RedisAndChill', 'Webpack', and 'README.md'. The Editor pane in the center displays a JavaScript file named '01-arrow_functions.js' with the following code:

```
myVar.printNumbers();
// Solving lexical "This" with .bind()
var myVar = {};
myVar.name = 'pentacode';
myVar.numbers = [1,2,3,4,5];
myVar.printNumbers = function() {
    any
    this.numbers.forEach(function(number) {
        console.log(this.name + ' counts ' + number);
    }.bind(this));
}
myVar.printNumbers();
// Solving lexical "This" with ES6 Arrow Function
var myVar = {};
myVar.name = 'pentacode';
myVar.numbers = [1,2,3,4,5];
myVar.printNumbers = function() {
    this.numbers.forEach((number) => {
        console.log(this.name + ' counts ' + number);
    });
}
myVar.printNumbers();
```

The Terminal pane at the bottom right shows the command '01-arrow_functions.js - pentacode'.

GIT Y GITHUB (github.com)

- **Git** es un **sistema de control de versiones** que permite llevar un seguimiento de los cambios en el código y trabajar en equipo sin perder historial.
- **GitHub** es una **plataforma en la nube** que **aloja repositorios Git** y facilita la colaboración entre desarrolladores.
- Permiten **trabajar de forma organizada y segura**, integrando funciones como ramas, revisiones de código y control de versiones.
- **Relevantes para cualquier proyecto de desarrollo**, ya que mejoran la productividad, la colaboración y el manejo de versiones del software.



REPRODUCIBILIDAD CIENTÍFICA

Capacidad de reproducir procedimientos científicos (experimentos, análisis, resultados).

¿Por qué es importante?

- Para repetir un análisis con distintos parámetros.
- Para escalar un análisis y entregarlo a equipo de producción.
- Para crear confianza en los resultados.
- Para compartir el conocimiento.

¿Cómo lo hacemos?

- Disponibilización de datos.
- Automatización de procesos, sin pasos “manuales”.
- Desarrollando procesos computaciones transparentes.
- Usando modelos y herramientas eficientes de reusabilidad (Ejemplo: software abierto, notebooks, repositorios).

TAREA

- Crear usuario en Github e incluirlo en el Excel que se enviará por email.
- Seguir todos los pasos de la Guía práctica: Uso de Python, VS Code y Git en computador personal

