



INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Lina María Castro Córdoba
Segundo semestre 2025

TAREAS Y PENDIENTES

- Reposición clase 7-agosto
 - Miércoles 6-agosto
 - 9 - 11 a.m. **Pendiente respuesta sala**
- Revisión artículos Boletín Economía
- Posibles casos de estudio



POSIBLES CASOS DE ESTUDIO

Modelo predictivo sobre la condición de pobreza de la población colombiana

- **Datos:** Base del SISBEN.
- **Objetivo:** predecir la condición de pobreza de una persona para facilitar decisiones objetivas sobre la entrega de subsidios focalizados.
- **Tipo de análisis:** Machine learning supervisado (clasificación).
- **Modelos:** Logistic Regression, Random Forest, XGBoost.
- **Variables:** Edad, sexo, tamaño del hogar, jefatura de hogar, Nivel educativo, asistencia escolar, analfabetismo, Tipo de vivienda, material paredes/piso/techo, servicios públicos, Condición laboral, ingresos estimados, tipo de ocupación, afiliación a salud, ubicación (depto., municipio, zona rural/urbana).
- **Aplicación:** Mejorar la focalización de subsidios y programas sociales.

Modelo predictivos sobre el nivel de ingresos laborales de una persona según sus características sociodemográficas y laborales

- **Datos:** GEIH.
- **Objetivo:** estimar los ingresos de una persona con base en sus características sociodemográficas y laborales.
- **Tipo de análisis:** Machine learning supervisado (regresión).
- **Modelos:** Regresión lineal, Random Forest Regressor, XGBoost Regressor.
- **Variables:** Edad, sexo, etnia, estado civil, Nivel educativo, Rama de actividad económica (CIIU), sector (público/privado), tipo de contrato, afiliación a salud/pensión, tamaño empresa, Departamento, ciudad principal, zona urbana/rural, Ingreso laboral mensual.
- **Aplicación:** Simulación de ingresos para nuevas políticas públicas (ej. efectos de aumentar el nivel educativo). Detección de brechas salariales por región, género, ocupación. Orientación laboral o predicción de ingresos por carrera u oficio.

POSIBLES CASOS DE ESTUDIO

Segmentación de consumidores colombianos con base en sus patrones de gasto

- **Datos:** Encuesta Nacional de Presupuesto de los Hogares (ENPH) – DANE
- **Objetivo:** Identificar grupos homogéneos de consumidores según su comportamiento de gasto.
- **Tipo de análisis:** Machine learning no supervisado.
- **Modelos:** Clustering (K-Means, DBSCAN).
- **Variables:** Porcentaje del ingreso destinado a: Alimentación, Transporte, Educación, Salud, Entretenimiento, Vivienda, Comunicaciones, Nivel educativo del jefe del hogar, Número de personas por hogar, Zona geográfica (urbano/rural), Estrato socioeconómico, Ingreso mensual.
- **Aplicación:** Generar perfiles útiles para estrategias de marketing o desarrollo de productos financieros. Se pueden analizar patrones por departamentos o ciudades.

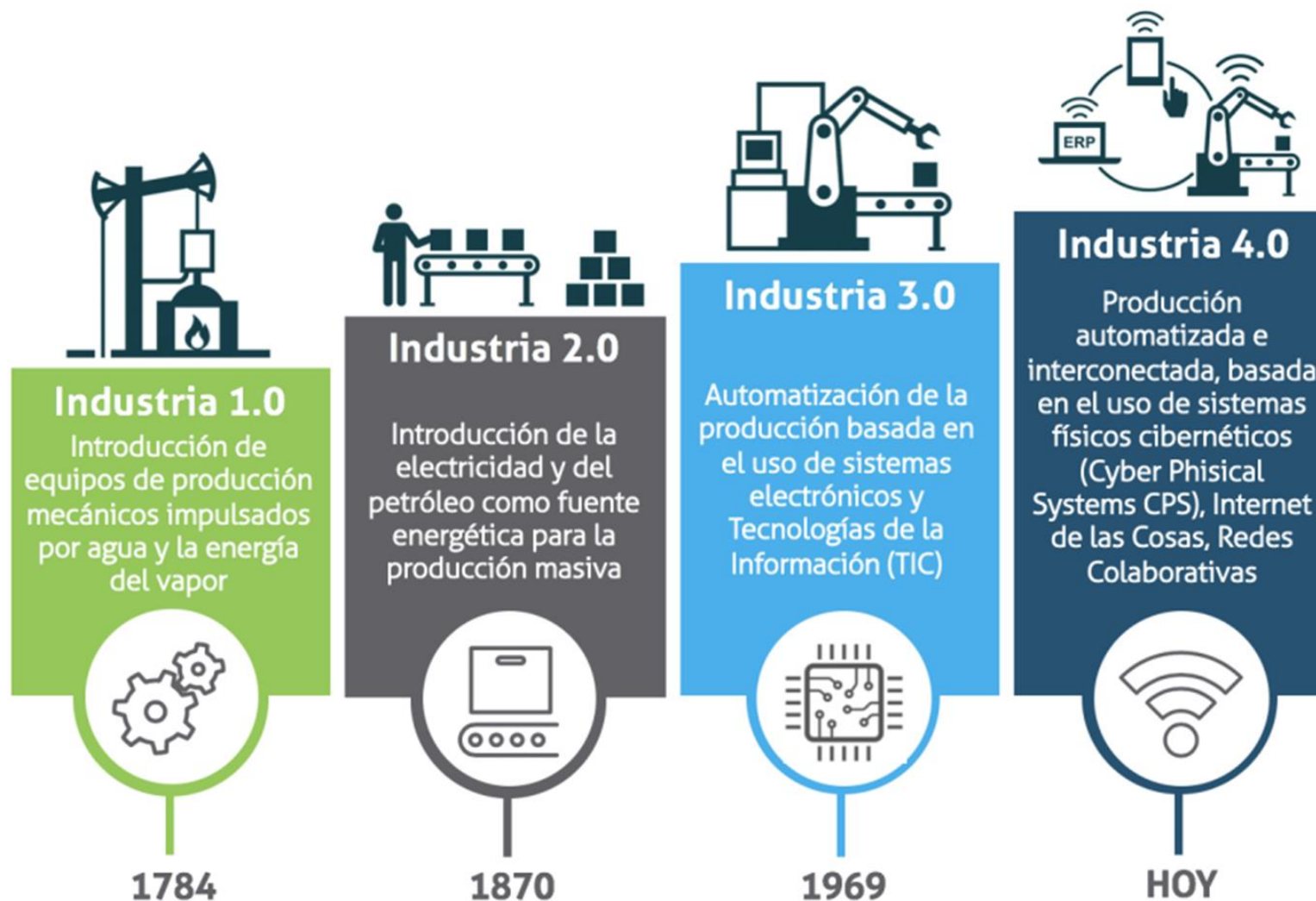
Modelo predictivo de acceso a asistencia técnica

- **Datos:** Censo Nacional Agropecuario.
- **Objetivo:** Predecir qué características influyen en la probabilidad de que un productor reciba asistencia técnica.
- **Tipo de análisis:** Machine learning supervisado (clasificación).
- **Modelos:** Logistic Regression, Random Forest, XGBoost.
- **Variables:** Ubicación, tipo de cultivo, tamaño del predio, nivel educativo, acceso a crédito, tenencia de la tierra.
- **Aplicación:** Focalización de programas de apoyo a los campesinos.

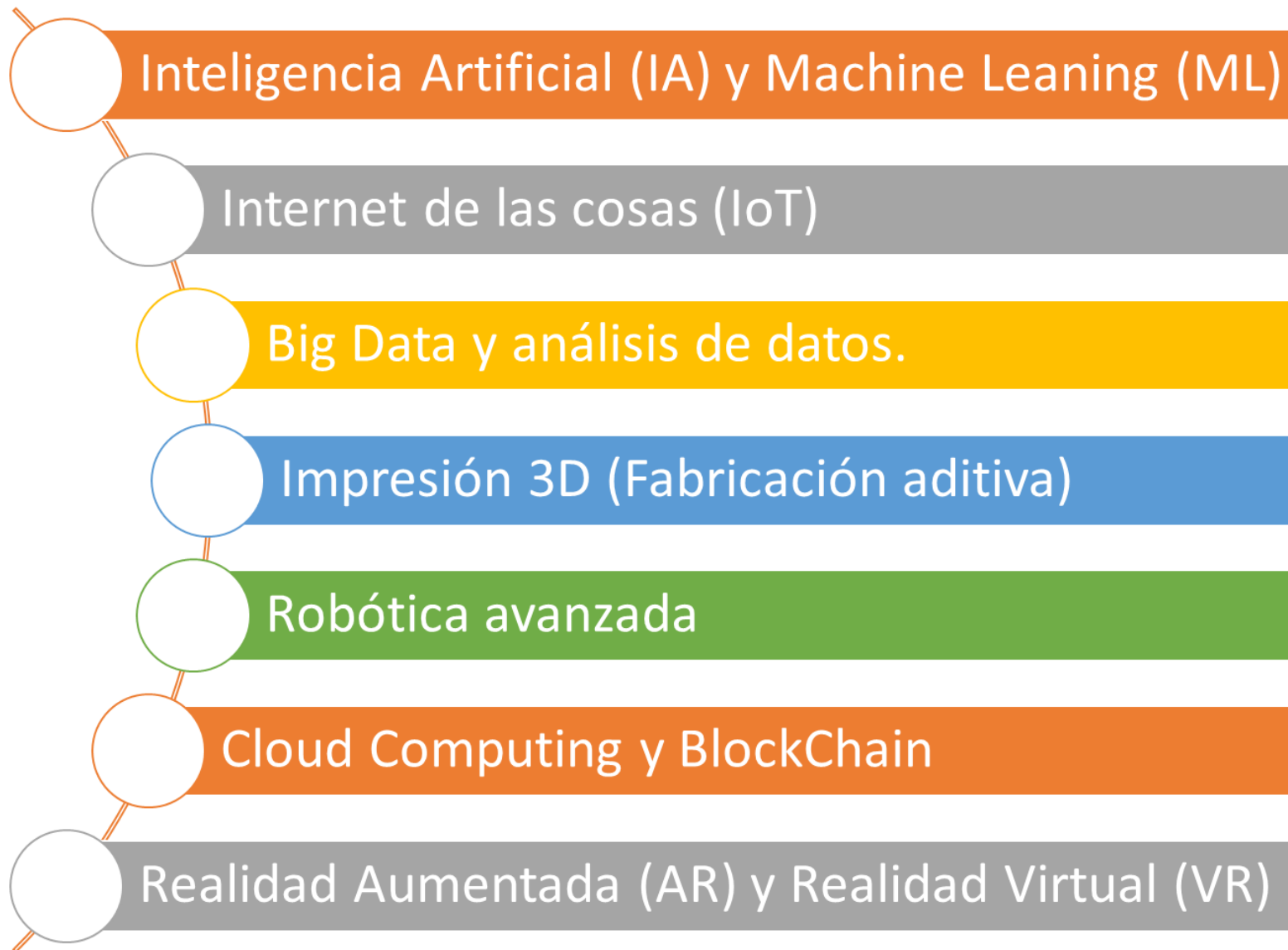


REVOLUCIÓN 4.0

EL MUNDO HA PASADO POR DIFERENTES AVANCES TECNOLÓGICOS



TECNOLOGÍAS CLAVES DE LA REVOLUCIÓN 4.0



PRINCIPIOS DE LA REVOLUCIÓN 4.0

Interconexión e
interoperabilidad.

Automatización y
robótica.

Personalización
masiva.

Agilidad,
flexibilidad,
tiempo real.

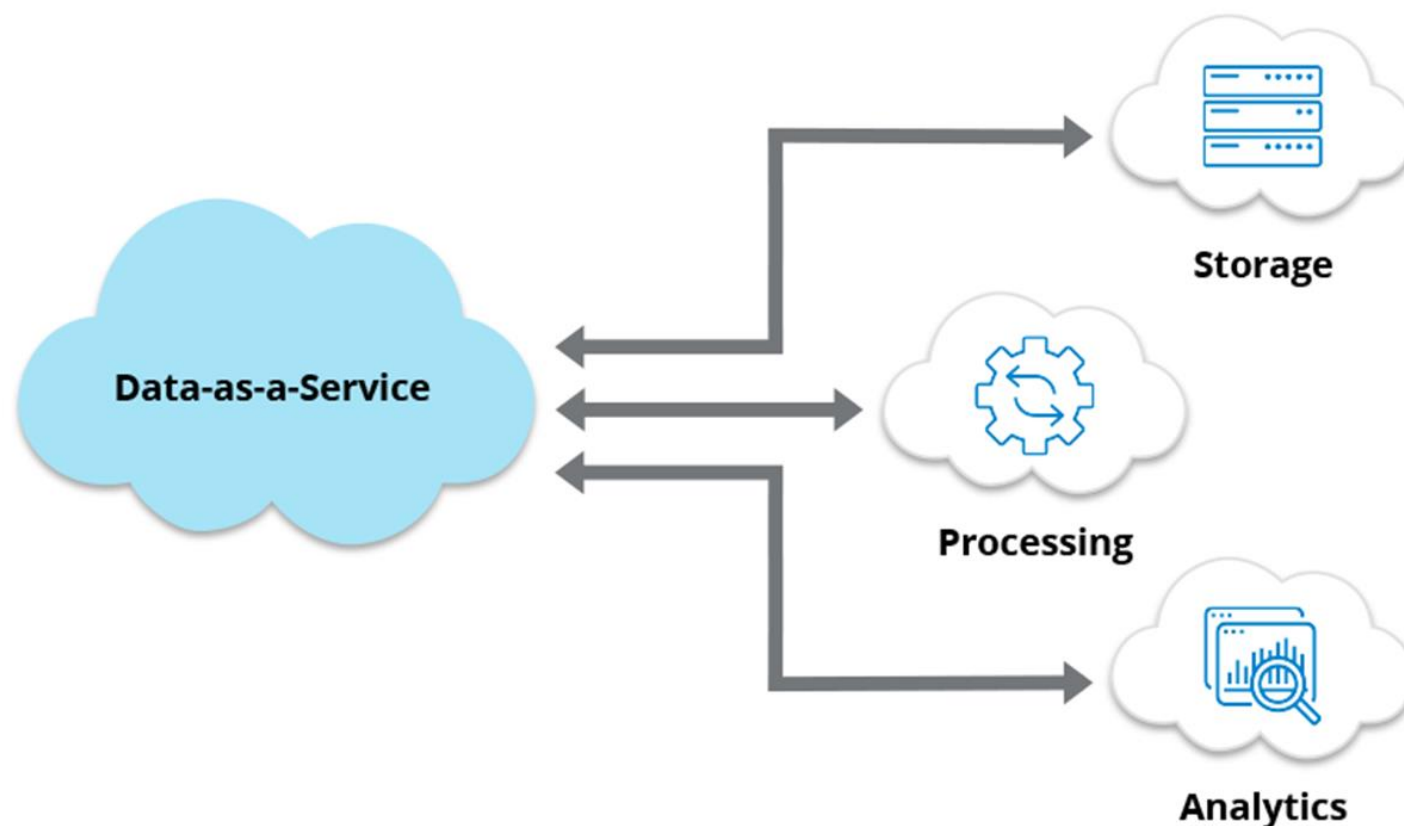
Integración de IA
en todo.

Convergencia de
tecnologías

Reducción de
barreras
geográficas.

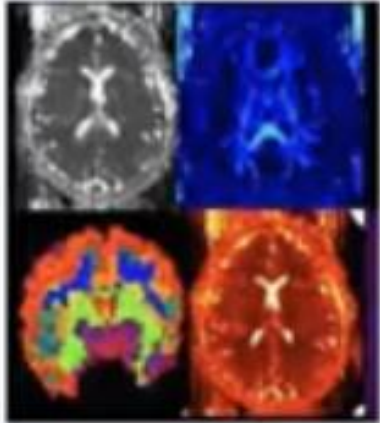
LOS DATOS JUEGAN UN ROL FUNDAMENTAL

La data es el insumo principal para todas las tecnologías de la Revolución 4.0. El análisis y la interpretación de grandes volúmenes de datos permiten **mejorar la toma de decisiones, optimizar recursos y crear productos y servicios personalizados.**



Tomado de: <https://hazelcast.com/glossary/data-as-a-service/>

DATA AGE: LA DIGITALIZACIÓN DEL MUNDO



The data-driven world will be **always on**,
always tracking, **always monitoring**, **always listening** and
always watching - **because it will be always learning.**



Fuente: Reinsel, D.; Gantz, J.; Rydning, J.; (2018) "The Digitization of the World: From Edge to Core", IDC White Paper, #US44413318

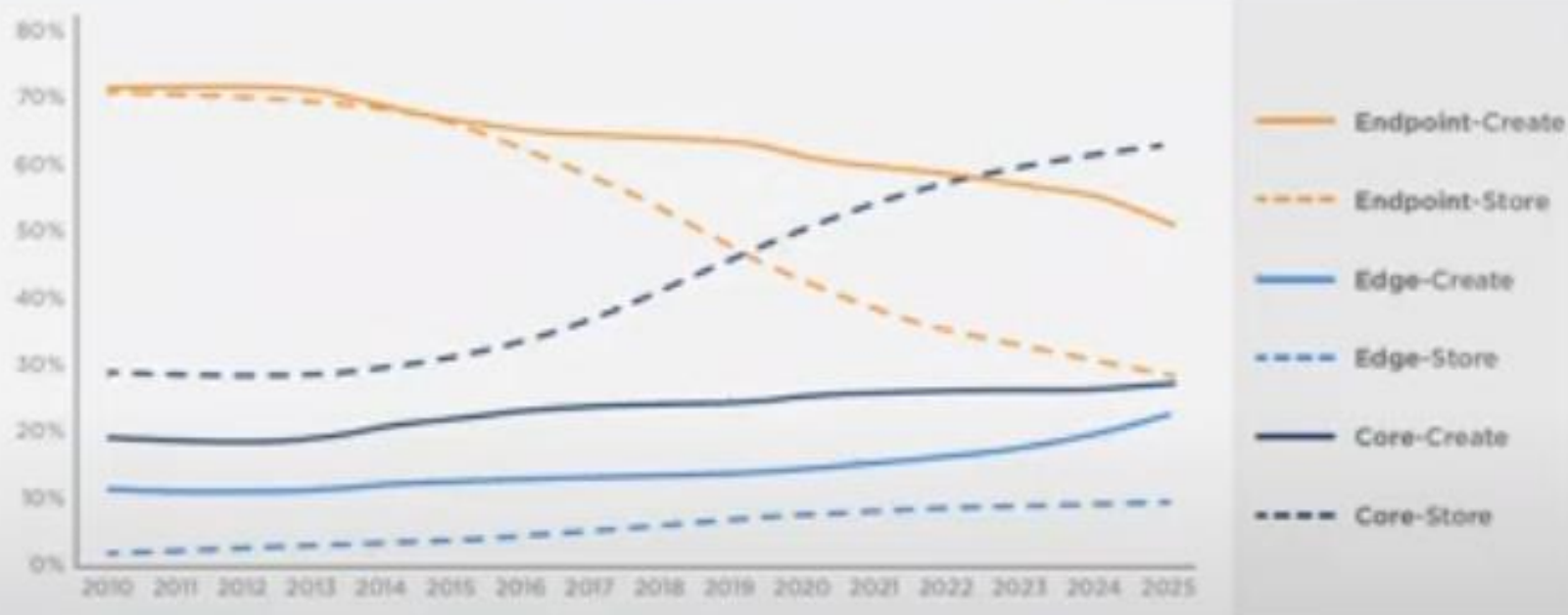
DATA AGE: LA DIGITALIZACIÓN DEL MUNDO



Fuente: Reinsel, D.; Gantz, J.; Rydning, J.; (2018) "The Digitization of the World: From Edge to Core", IDC White Paper, #US44413318

DATA AGE: LA DIGITALIZACIÓN DEL MUNDO

¿Dónde se generan y almacenan estos datos?

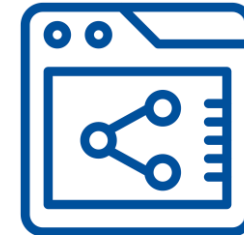


➔ Mayor volumen, variedad, velocidad y acceso a datos

Fuente: Reinsel, D.; Gantz, J.; Rydning, J.; (2018) "The Digitization of the World: From Edge to Core", IDC White Paper, #US44413318

LA COMBINACIÓN DE DATOS E IA TRANSFORMA LOS NEGOCIOS

- **Personalización masiva:** Las empresas pueden usar los datos y la IA para ofrecer productos y servicios adaptados a las preferencias individuales de los clientes. Ejemplo: **Netflix** y sus recomendaciones de contenido basadas en el historial de visualización.
- **Optimización de la cadena de suministro:** **Amazon** usa la IA y Big Data para gestionar inventarios y predecir la demanda en tiempo real, mejorando la eficiencia operativa.
- **Mejora en la toma de decisiones:** Las empresas usan la IA para analizar datos en tiempo real y tomar decisiones más rápidas y precisas. Ejemplo: **Tesla** optimiza sus vehículos mediante actualizaciones de software en tiempo real.



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.



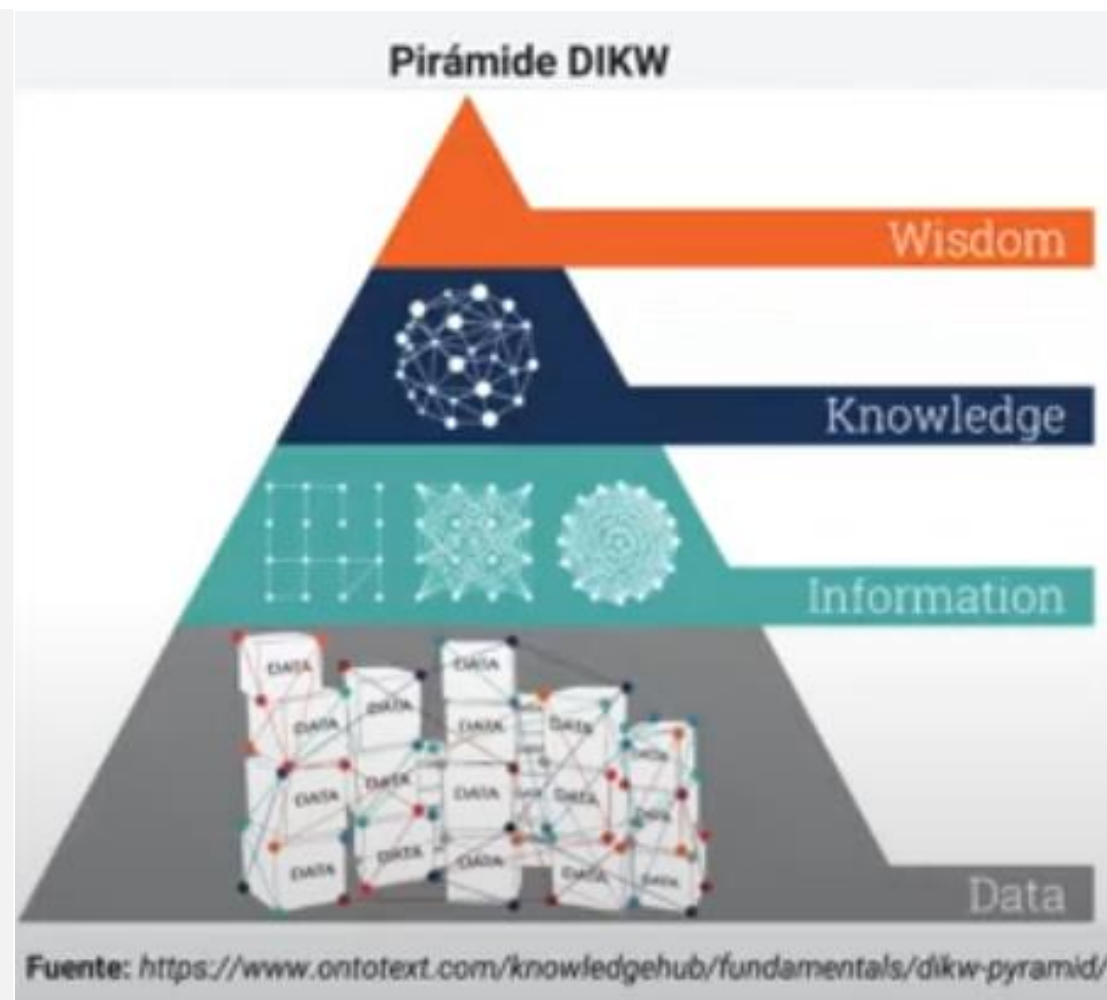
CIENCIA DE DATOS

- 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.

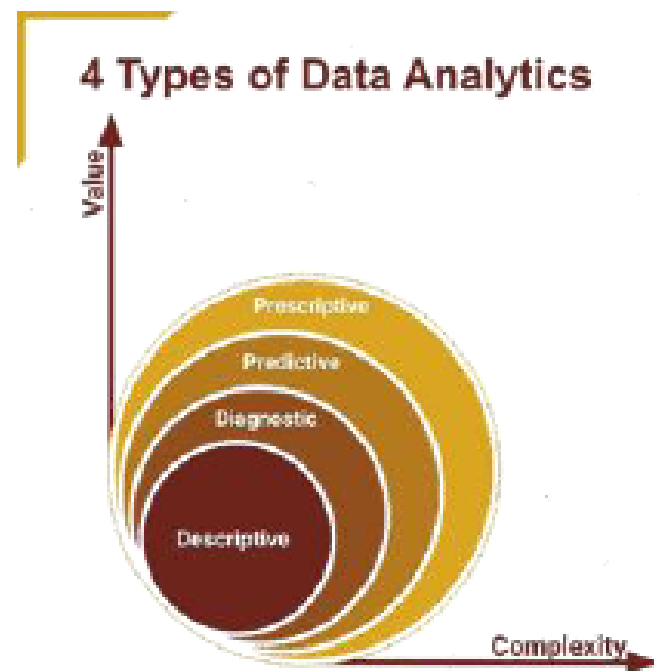


▶ DATOS VS. INFORMACIÓN

- **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



es.gupta.link/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):

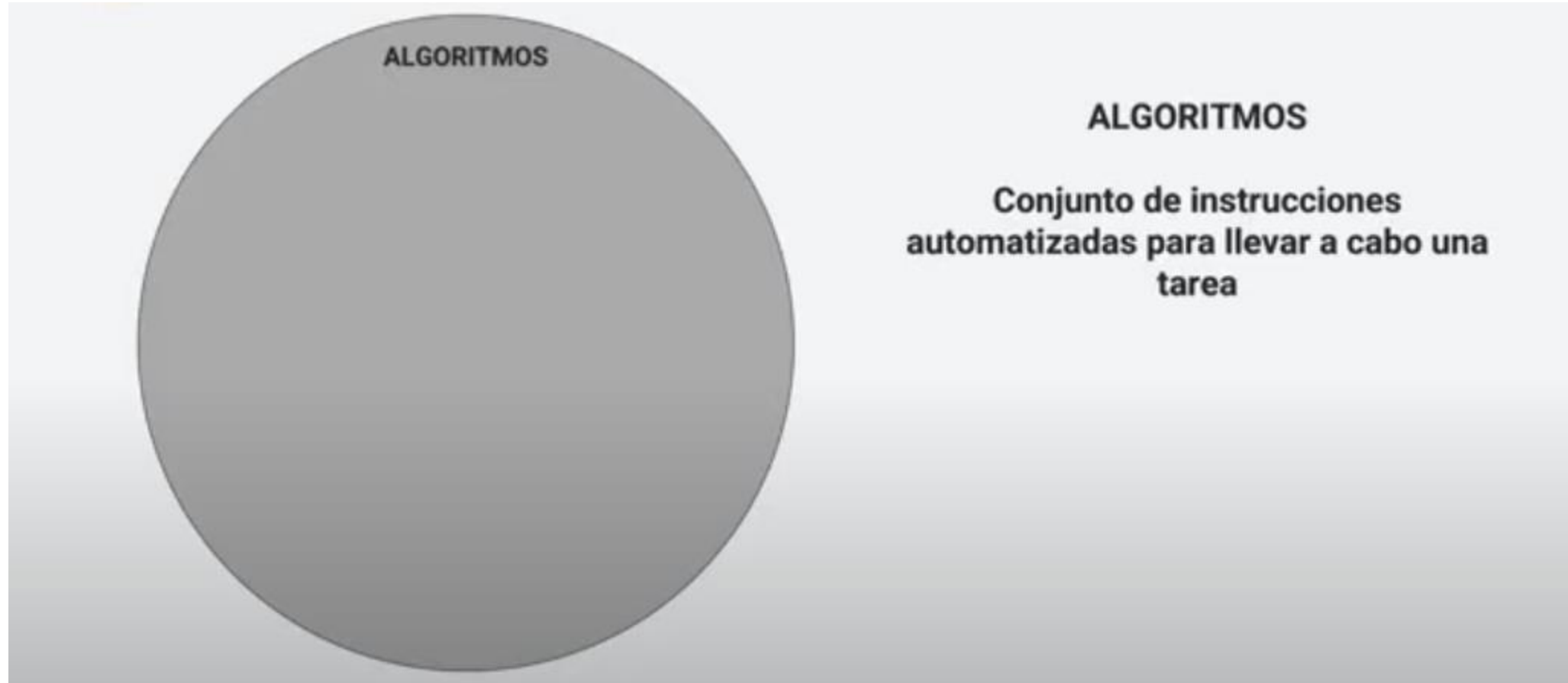
1. **Pregunta:** ¿Qué pasará? (Futuro)
2. **Métodos:** Modelado predictivo y estadístico
3. **Propósito:** Utiliza datos históricos y técnicas estadísticas para predecir eventos y comportamientos futuros.

4.Prescriptivo (Prescriptive):

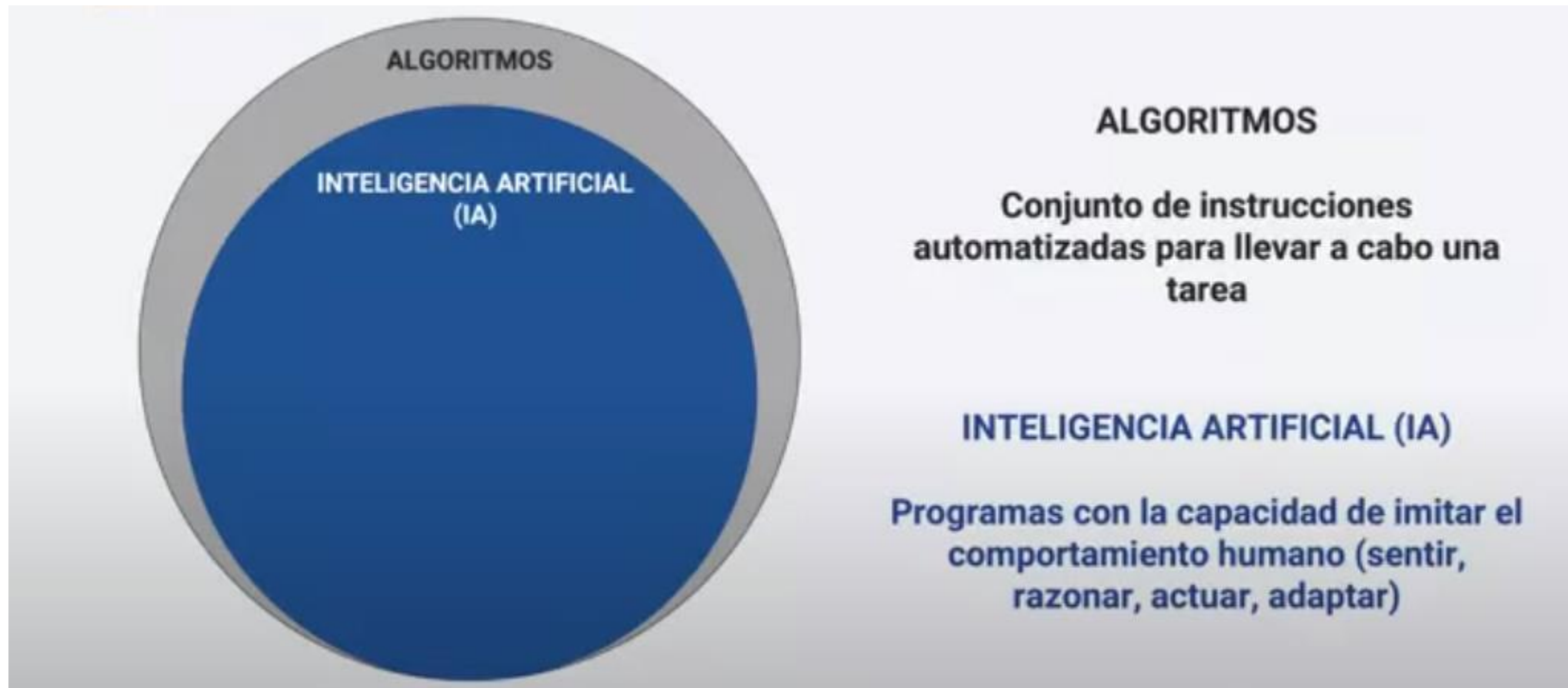
1. **Pregunta:** ¿Qué debo hacer? (Futuro)
2. **Métodos:** Optimización y pruebas aleatorias
3. **Propósito:** Proporciona recomendaciones sobre acciones a tomar para alcanzar objetivos deseados basándose en análisis predictivos y optimización.

Futuro

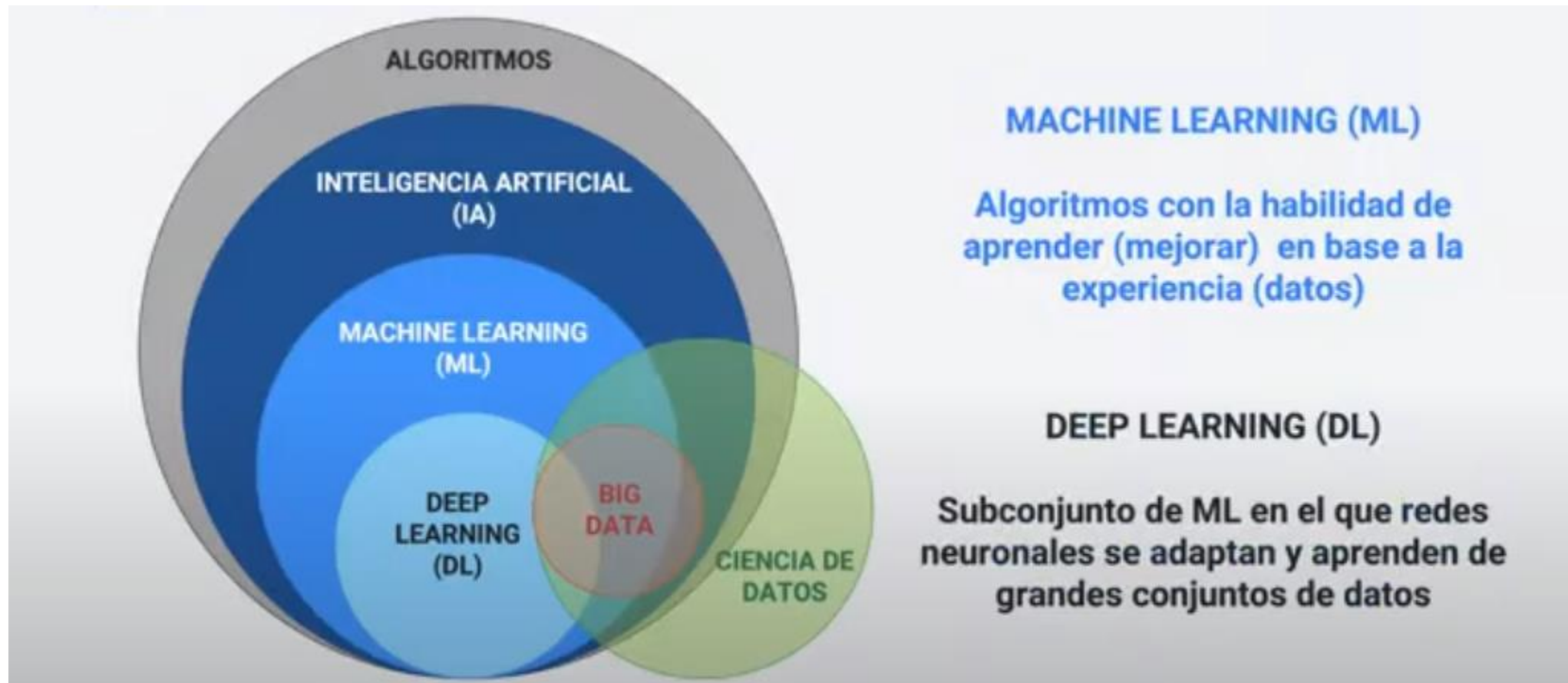
CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL



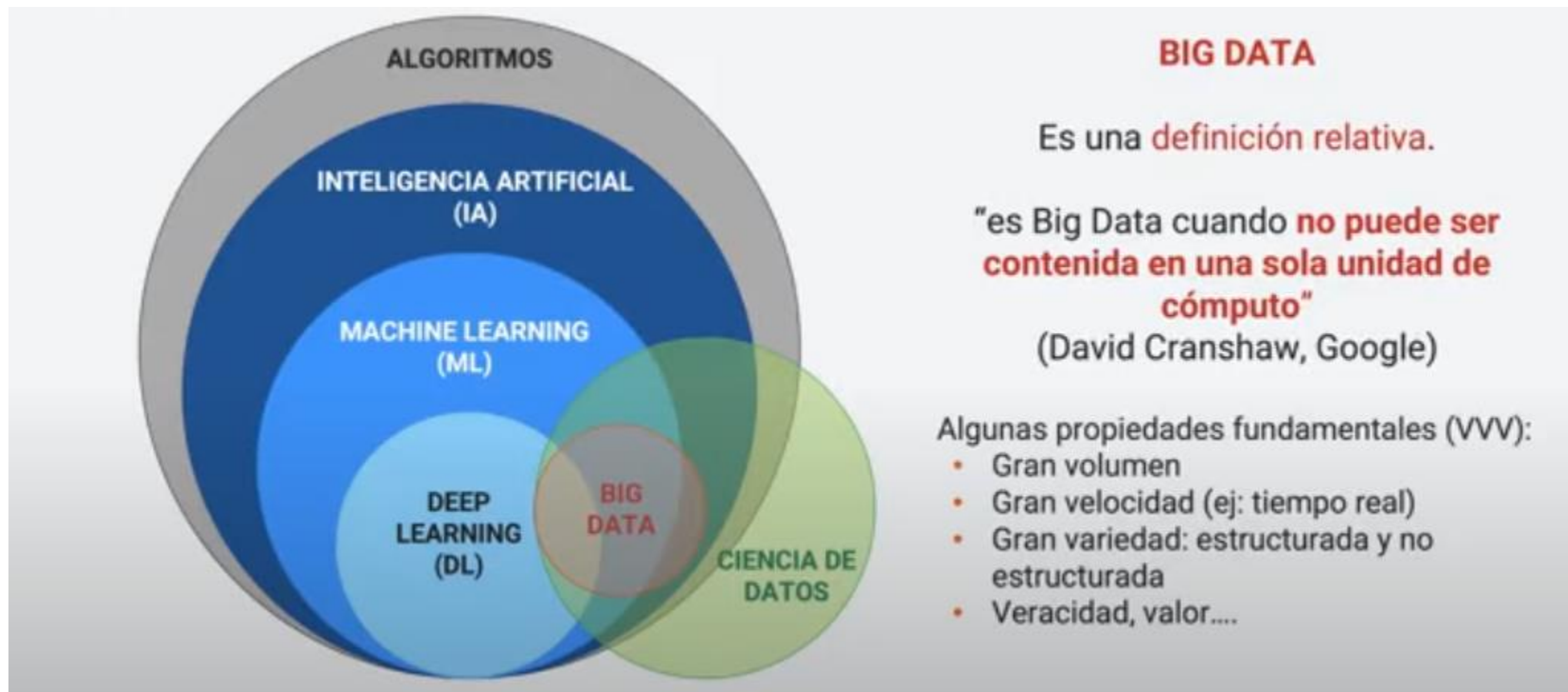
CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL



CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

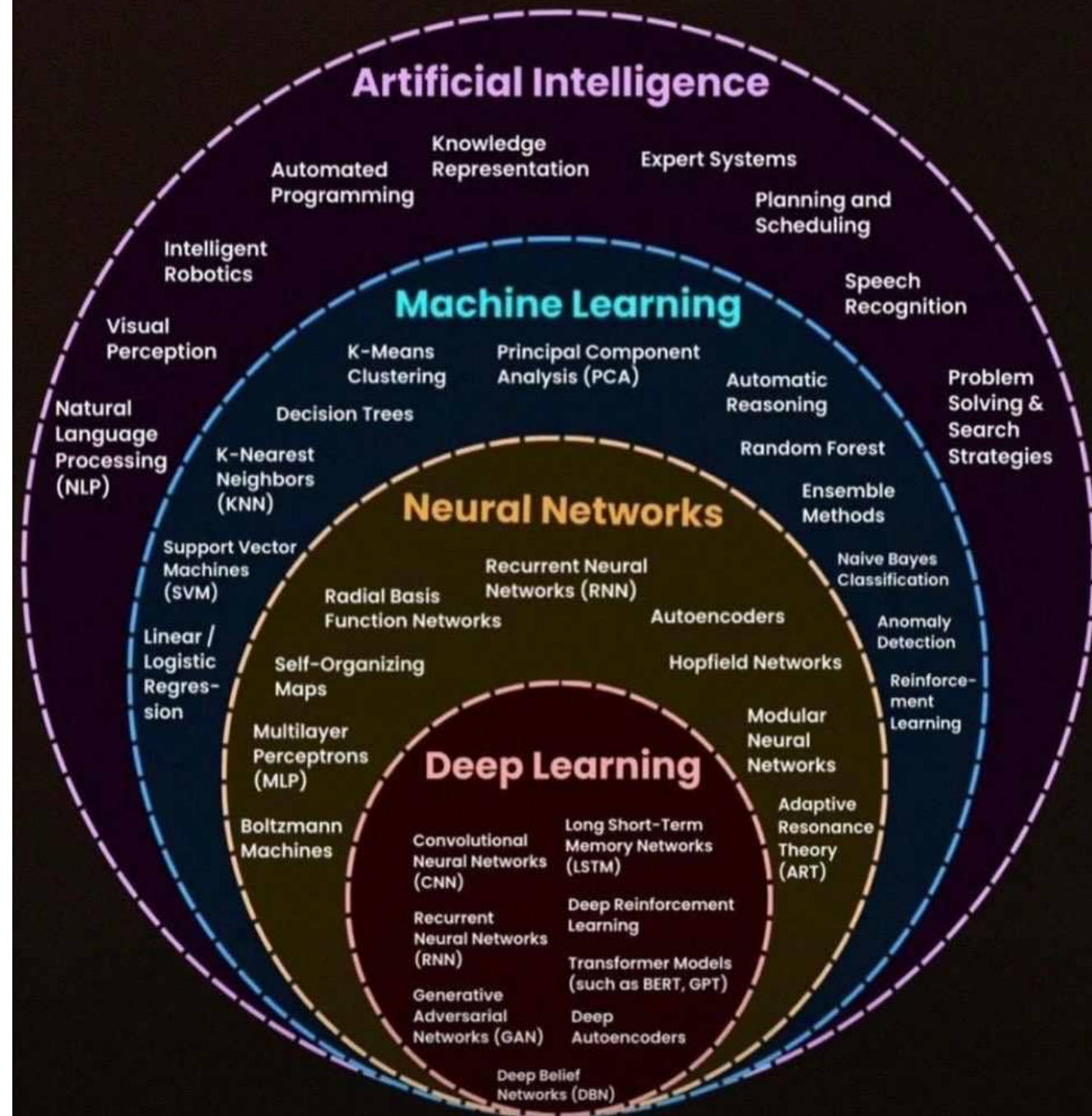


CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL



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.

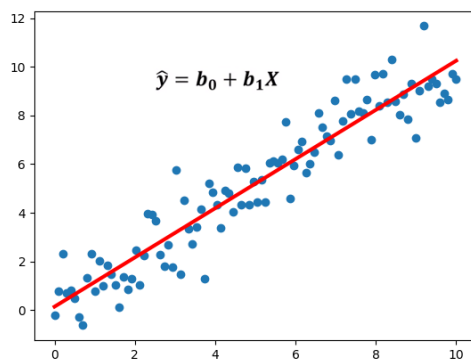
APRENDIZAJE NO SUPERVISADO

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

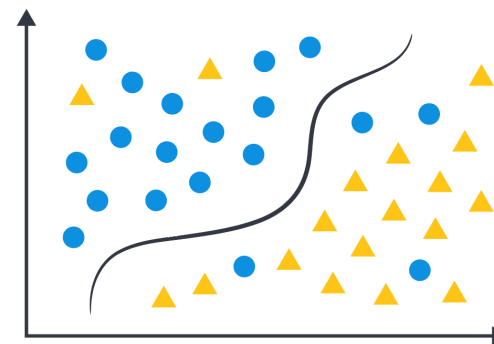
REGRESIÓN

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.

REGRESIÓN LOGÍSTICA

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

KNN

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

ÁRBOLES DE DECISIÓN

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

RANDOM FOREST

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

SUPPORT VECTOR MACHINES

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

GRADIENT BOOSTING MACHINES

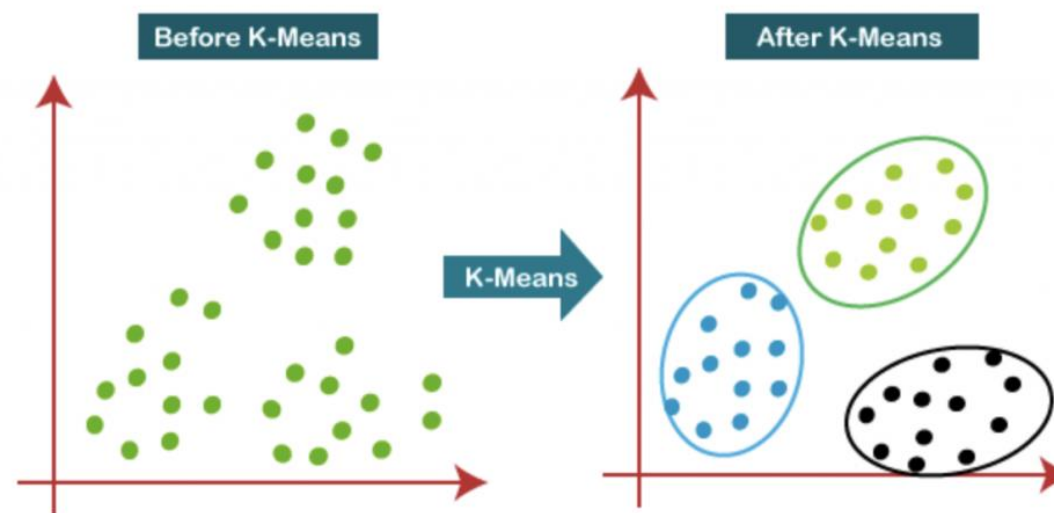
Modelo que construye iterativamente árboles de decisión en serie, cada uno corrigiendo los errores de los anteriores, mejorando el rendimiento en cada iteración.

XGBOOST

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

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.

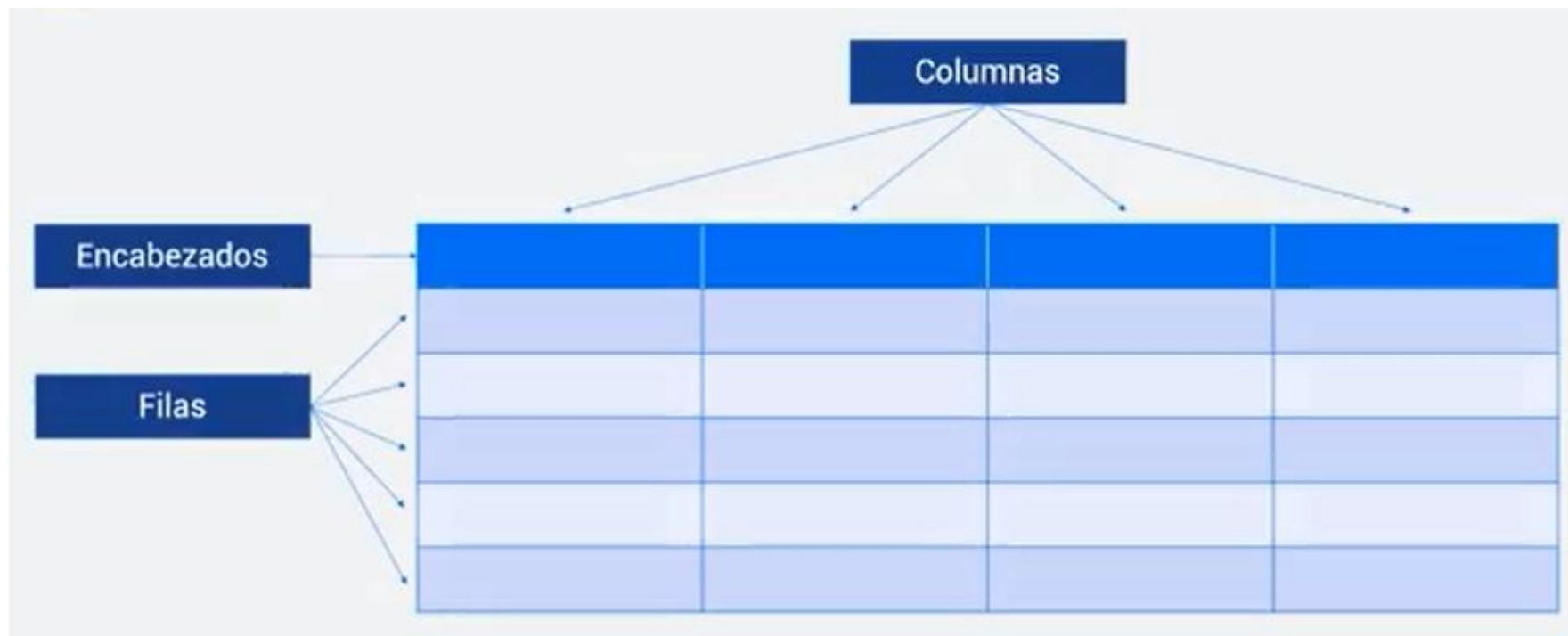
Aplicaciones: Visualización de datos, compresión de datos, preprocesamiento para otros algoritmos.

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.

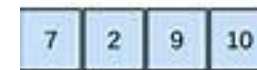


NUMPY (numpy.org)

- 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.

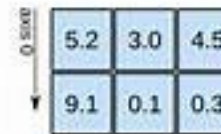


1D array



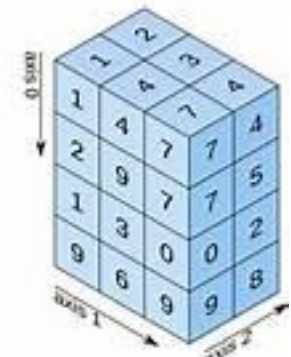
shape: (4,)

2D array



shape: (2, 3)

3D array



shape: (4, 3, 2)

➤ 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.



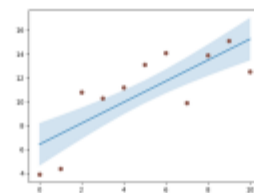
Series			Series			DataFrame	
	peppers			carrots			
0	3	+	0	0	=	0	3
1	2		1	3		1	2
2	0		2	7		2	0
3	1		3	2		3	1
							7
							2

MATPLOTLIB (matplotlib.org)

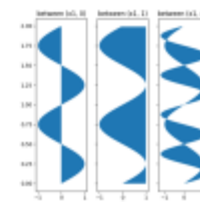
- Es una **biblioteca de Python para crear gráficos y visualizaciones de datos 2D/3D**.
- **Permite generar gráficos variados**, como líneas, barras, histogramas, dispersión, entre otros.
- **Altamente personalizable**, ideal para crear visualizaciones profesionales y adaptadas a distintas necesidades.
- **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).



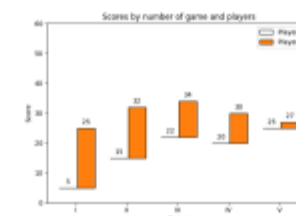
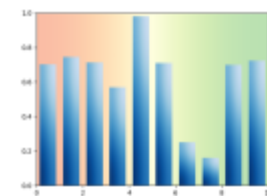
Rellenar entre y alfa



Llenar el área entre líneas



Demostración de relleno de Betweenx



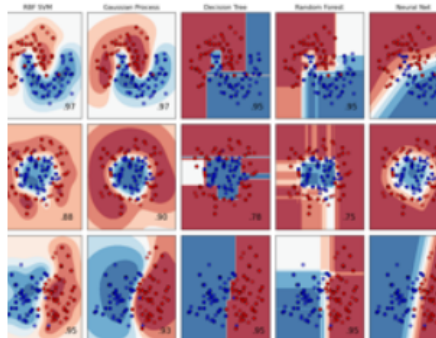
matplotlib

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...](#)

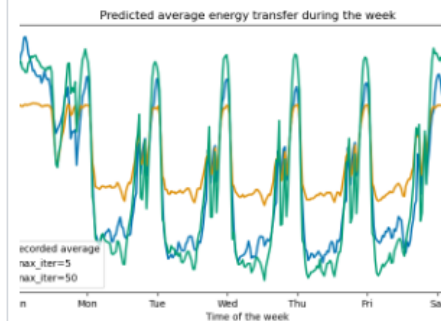


Examples

Dimensionality reduction

Reducing the number of random variables to consider.

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

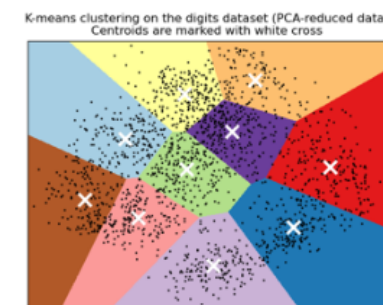


Examples

Model selection

Comparing, validating and choosing parameters and models.

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



Examples

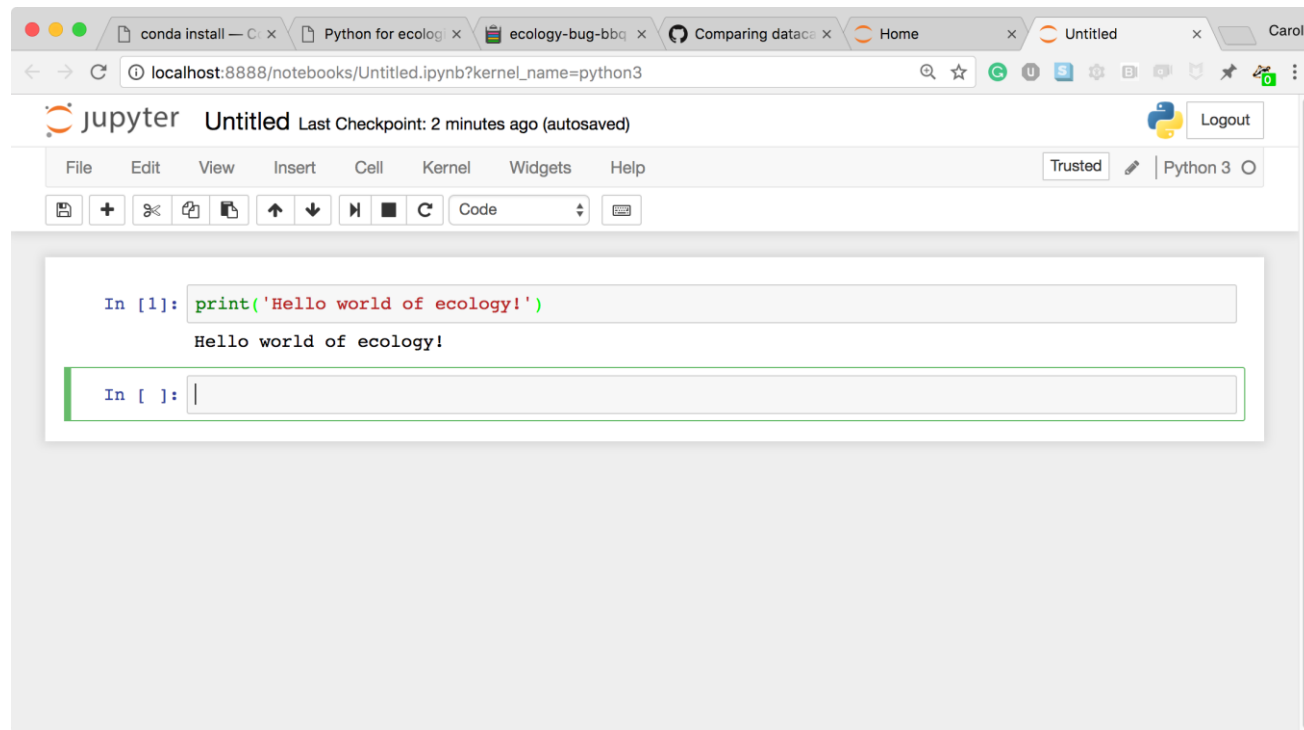
Preprocessing

Feature extraction and normalization.

Applications: Transforming input data

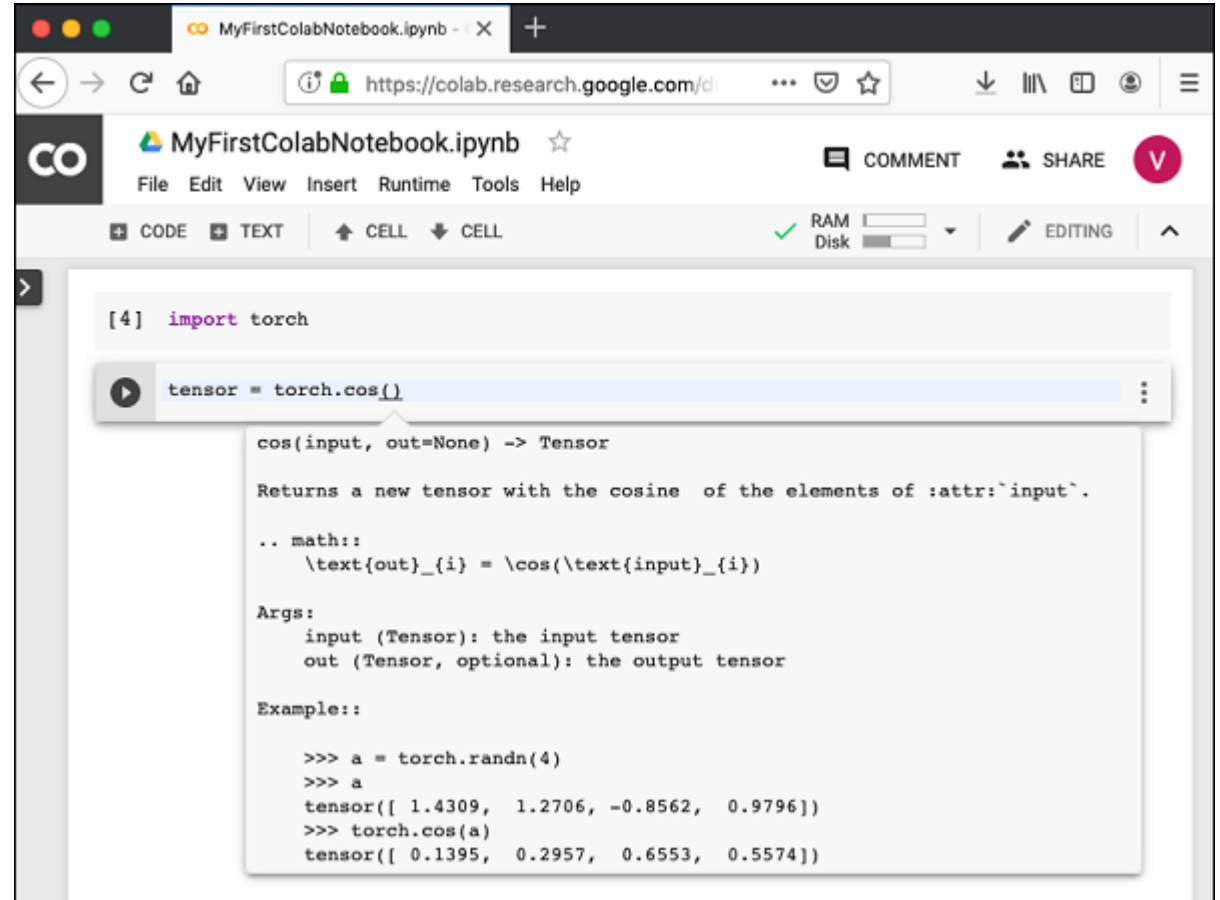
JUPYTER NOTEBOOKS (jupyter.org)

- 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.



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.



```
[4] import torch

tensor = torch.cos()
```

cos(input, out=None) -> Tensor

Returns a new tensor with the cosine of the elements of :attr:`input`.

.. math::

$\text{out}_i = \cos(\text{input}_i)$

Args:

 input (Tensor): the input tensor

 out (Tensor, optional): the output tensor

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.

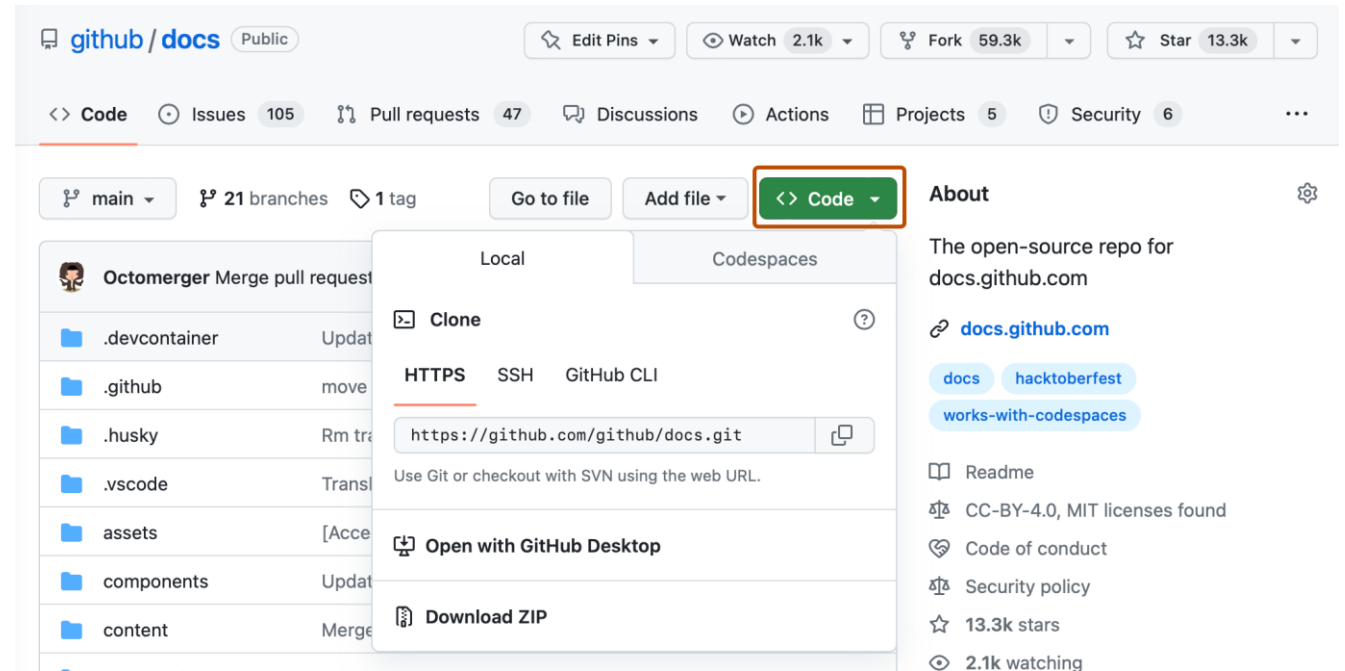


Visual Studio Code

```
21 myVar.printNumbers();
22
23
24 // Solving lexical "This" with .bind()
25 var myVar = {};
26
27 myVar.name = 'pentacode';
28 myVar.numbers = [1,2,3,4,5];
29
30 myVar.printNumbers = function() {
31     any
32     this.numbers.forEach(function(number) {
33         console.log(this.name + ' counts ' + number);
34     }).bind(this));
35 }
36
37 myVar.printNumbers();
38
39 // Solving lexical "This" with ES6 Arrow Function
40 var myVar = {};
41
42 myVar.name = 'pentacode';
43 myVar.numbers = [1,2,3,4,5];
44
45 myVar.printNumbers = function() {
46     this.numbers.forEach((number) => {
47         console.log(this.name + ' counts ' + number);
48     });
49 }
50
51 myVar.printNumbers();
52
53
```

➤ 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 computacionales transparentes.
- Usando modelos y herramientas eficientes de reusabilidad (Ejemplo: software abierto, notebooks, repositorios).