

# Data Scientist



## UNIDAD DIDACTICA 2

### 2.2 Data Science

Ing. Gary David Guzmán Muñoz

Curso 2025

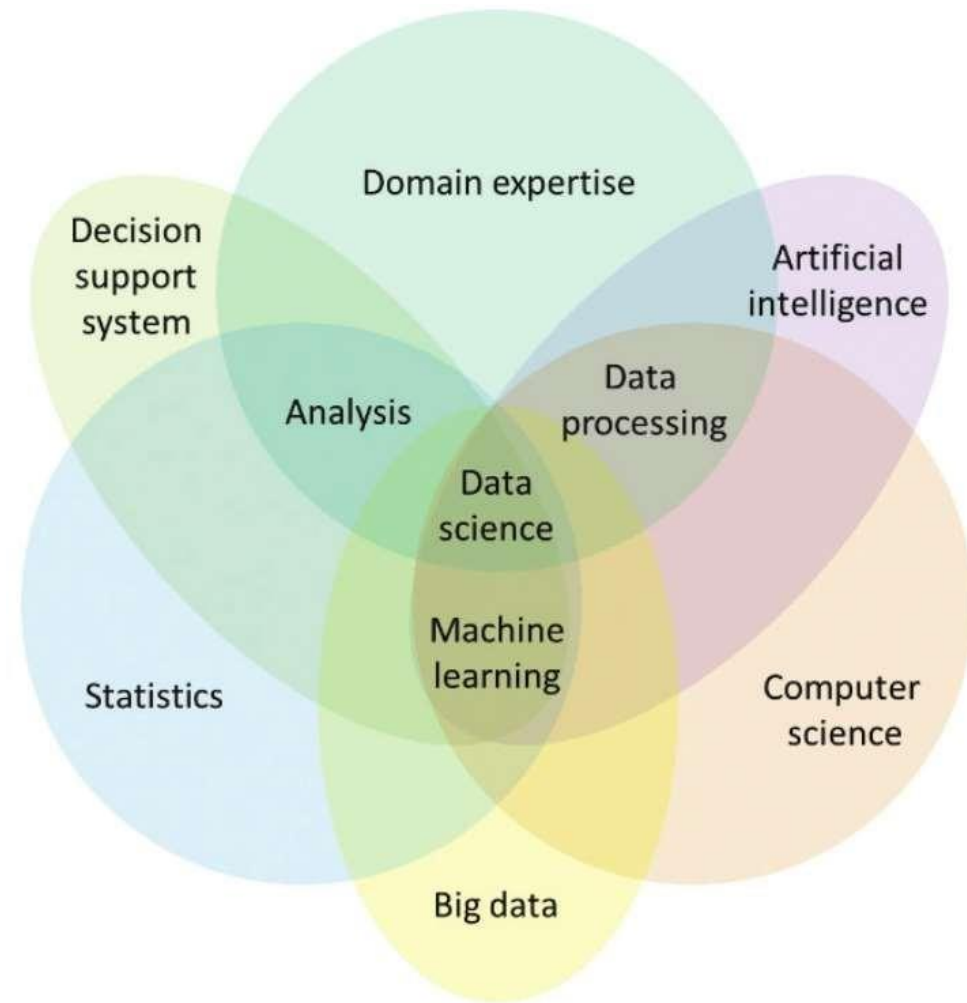
**PREGUNTA**

**¿Qué es la Ciencia de  
Datos o  
Data Science?**

# ¿Qué es la Ciencia de datos o Data Science?

Es un campo interdisciplinar que usa métodos científicos, procesos, algoritmos y sistemas de información para la extracción de información e ideas de una amplia variedad de datos (estructurados y no estructurados).

Engloba áreas desde la estadística pura al machine learning, pasando por la minería de datos (data mining).



# **PREGUNTA**

**¿A qué es debido el  
auge de la ciencia  
de datos?**

# ¿Por qué este auge de la ciencia de datos?

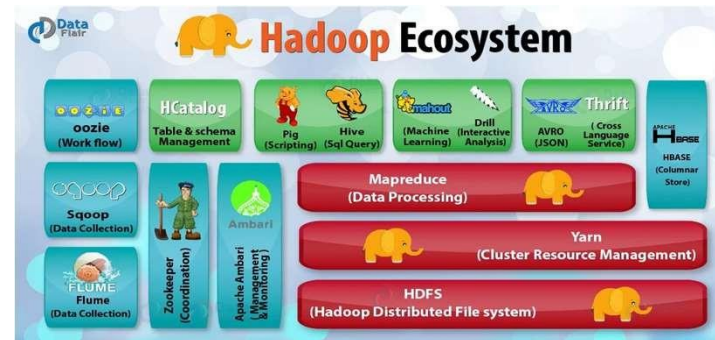
## DISPONIBILIDAD DE DATOS



## HERRAMIENTAS DE ANÁLISIS DE GRANDE DATOS



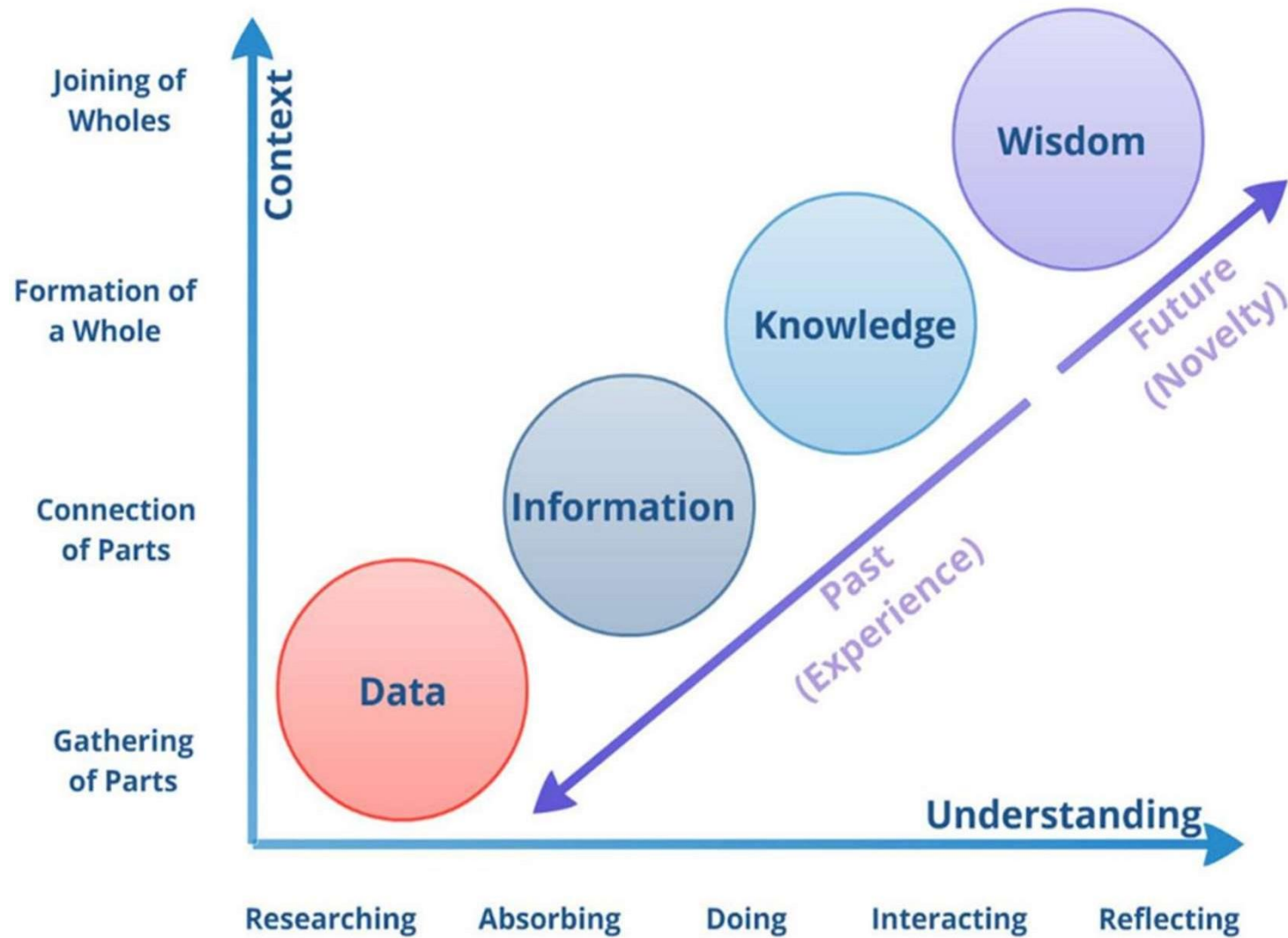
## HERRAMIENTAS DE GESTION DE GRANDE DATOS



## CAPACIDAD DE COMPUTO Y COSTES REDUCIDOS



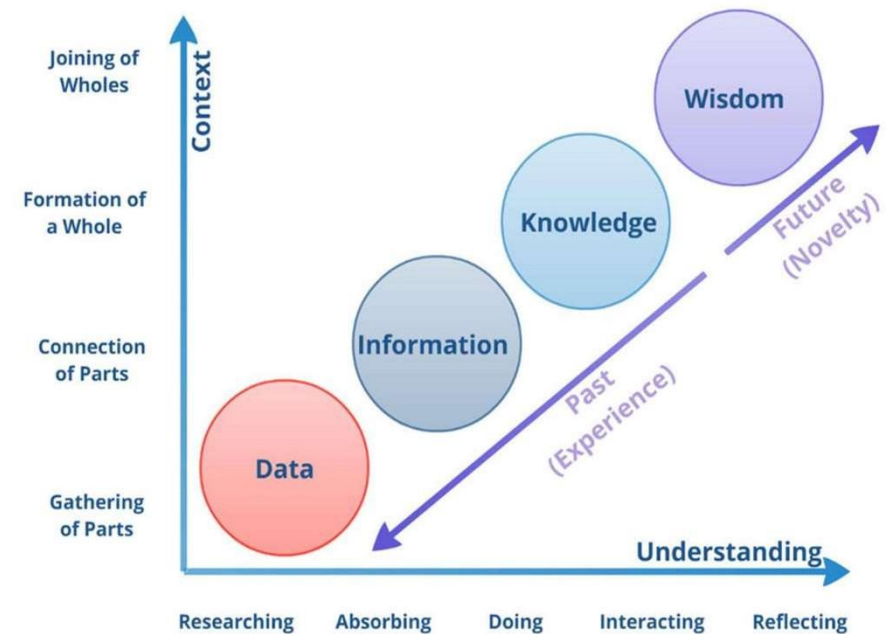
# ¿Qué es la ciencia de datos?



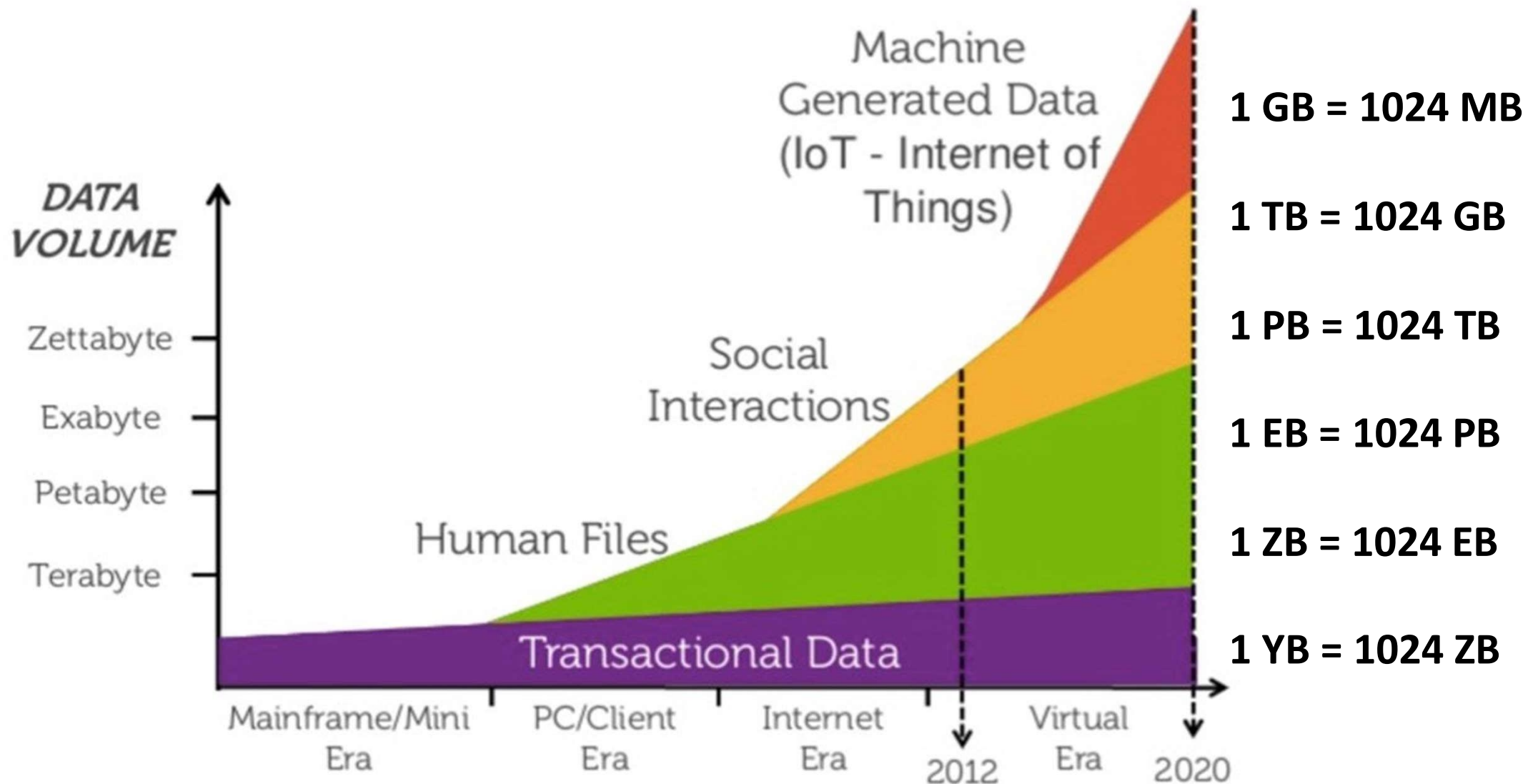
# ¿Qué es la ciencia de datos?

Esto permite:

- Aprovechar información en toda su amplitud
- Proporciona respuestas
- Reducción de costes
- Eficiencia (mejor toma de decisiones)
- Rapidez en la ejecución
- Nuevos productos y oportunidades (relevancias ocultas)
- Reinención y creación de nuevos negocios
- Pro-actividad (Estrategias)

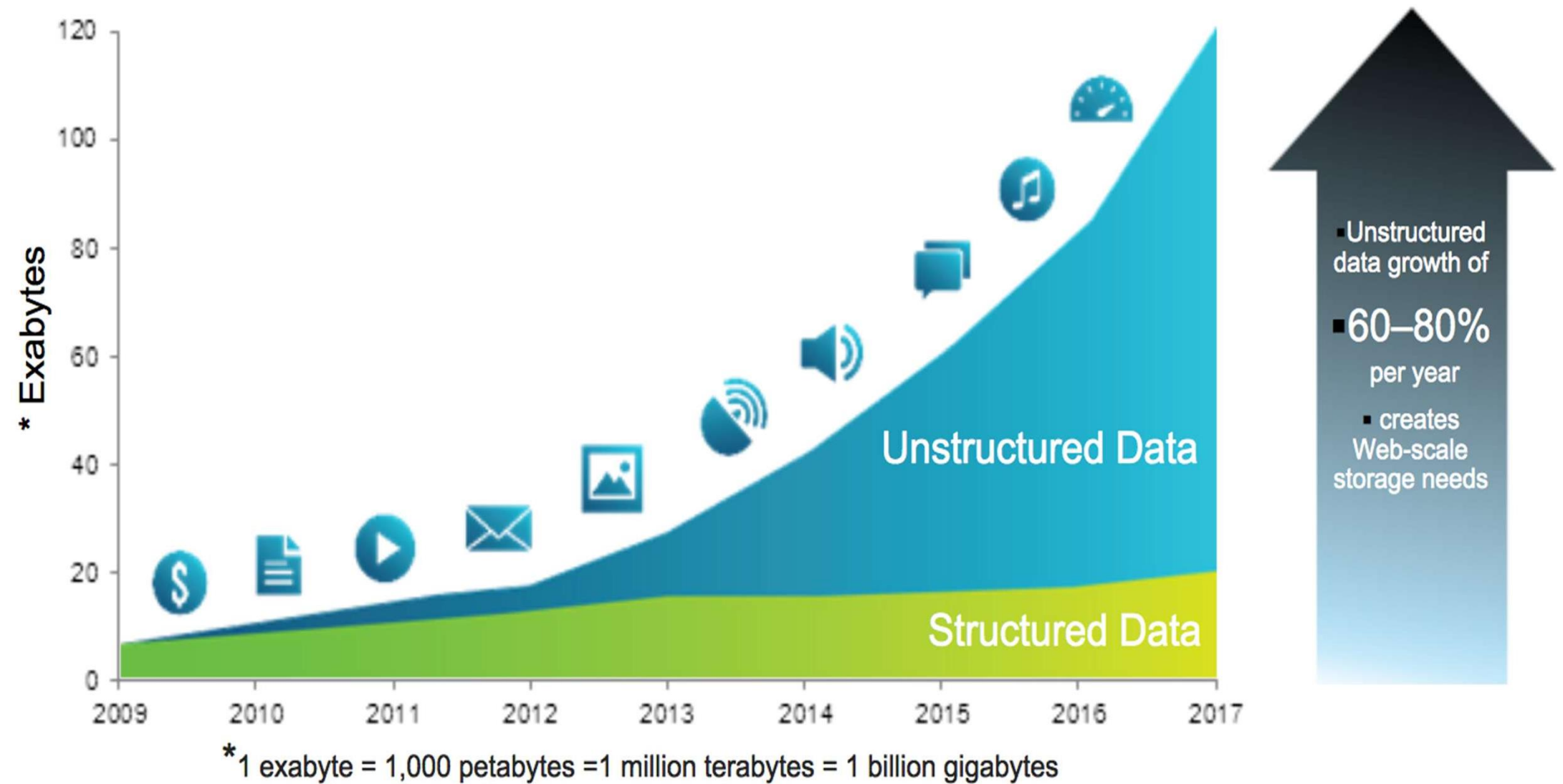


# ¿Qué es la ciencia de datos? - Datos





# ¿Qué es la ciencia de datos? - Datos









# ¿Qué es la ciencia de datos? – Las Vs



La ciencia de datos regida por las Vs del Big Data.

Atendiendo tanto en la cantidad, como en naturaleza y características de los datos recogidos, inicialmente el Big Data se caracterizó por lo que se llegó a conocer como las 3 Vs: **VOLUMEN**, **VARIEDAD** y **VELOCIDAD**. Con el paso del tiempo estas Vs se han incrementado. Un consenso más o menos amplio considera la existencia de 6 Vs:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					


# **PREGUNTA**

**¿Cuáles son los  
principales  
profesionales de la  
ciencia de datos?**

# Profesionales de la ciencia de datos

## DATA ARCHITECT

### THE CONTEMPORARY DATA MODELLER





**Role**  
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

**Languages**  
SQL, XML, Hive, Pig, Spark

**Mindset**  
Inquiring ninja with a love for data architecture design patterns


**Skills & Talents**

- Data warehousing solutions
- In-depth knowledge of database architecture
- Extraction Transformation and Load(ETL), spreadsheet and BI tools
- Data modeling
- Systems development



## DATA ENGINEER

### SOFTWARE ENGINEERS BY TRADE





**Role**  
Develops, constructs, tests and maintains architectures (such as databases and large scale processing systems)

**Languages**  
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

**Mindset**  
All-purpose everyman


**Skills & Talents**

- Database systems (SQL & NO SQL based)
- Data modeling & ETL tools
- Data APIs
- Data warehousing solutions



## DATA SCIENTIST

### AS RARE AS UNICORNS




**Role**  
Cleans, massages and organizes (big) data

**Languages**  
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

**Mindset**  
Curious data wizard


**Skills & Talents**

- Distributed computing
- Predictive modeling
- Story-telling and visualizing
- Math, Stats, Machine Learning



## BUSINESS ANALYST

### CHANGE AGENT




**Role**  
Improves business process as intermediary between business and IT

**Languages**  
SQL

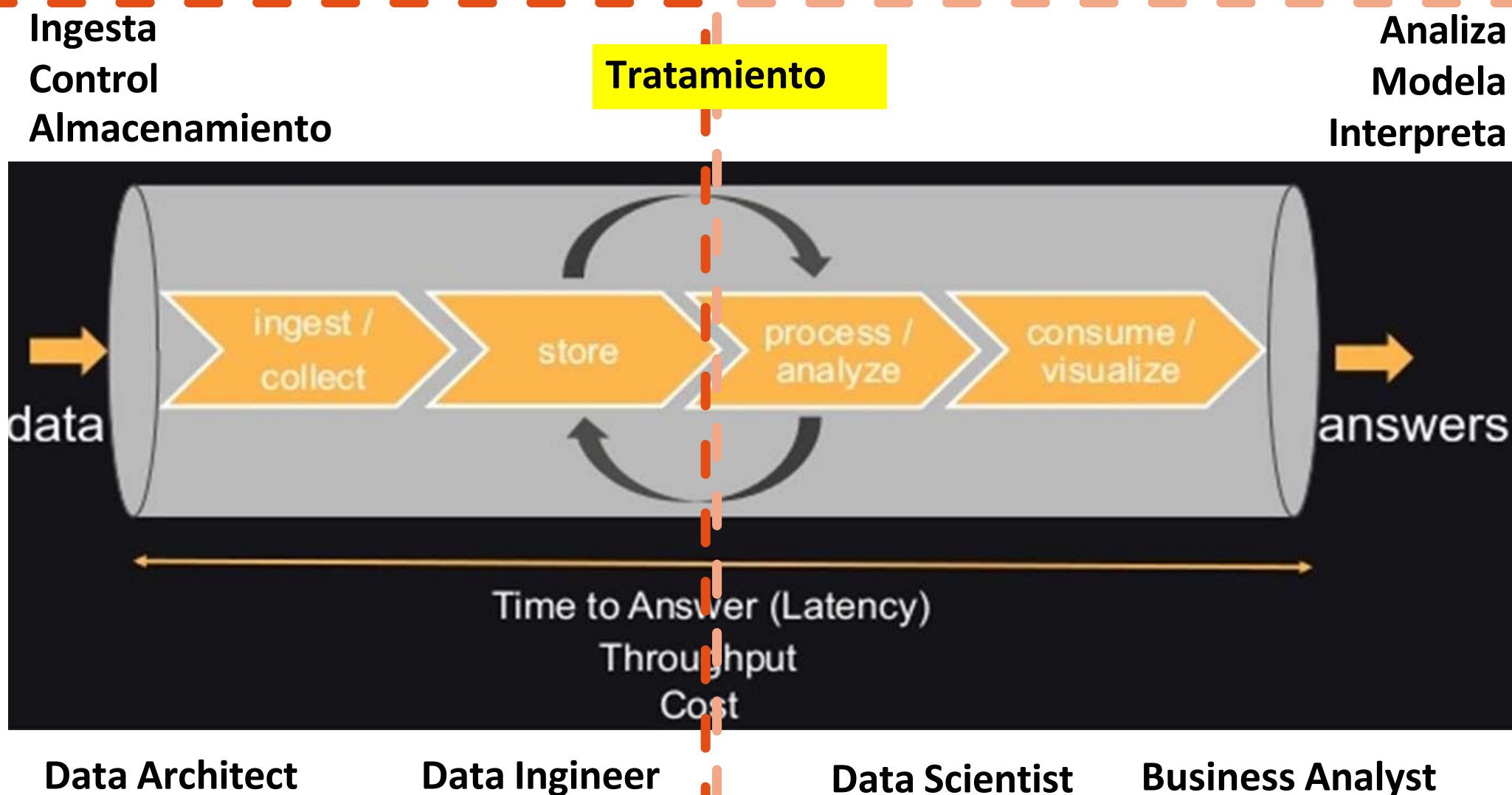
**Mindset**  
Resilient project juggler

**Skills & Talents**

- Basic tools (e.g. MS Office)
- Data visualization tools (e.g. Tableau)
- Conscious listening and storytelling
- Business Intelligence understanding
- Data modeling



# Profesionales de la ciencia de datos

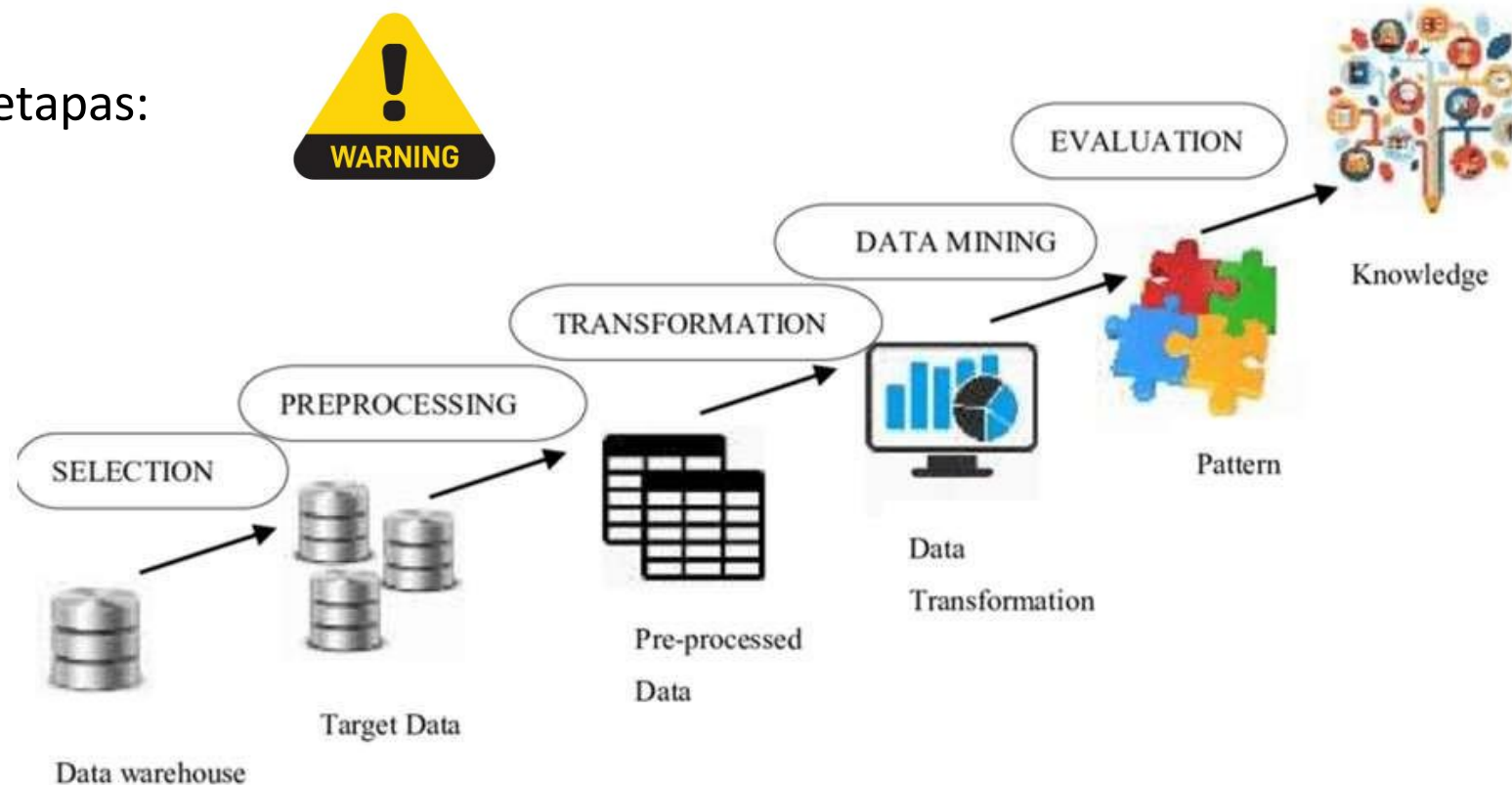


# Ciencia de Datos y KDD

**KDD (Knowledge Discovery in Databases)** es el primer modelo aceptado en la comunidad científica (1996) que establece las etapas principales de un proyecto de explotación de información (Datos).

Consta principalmente de 7 etapas:

- Selección objetivos
- Selección de datos
- Preprocesado
- Transformación
- Minado de datos
- Evaluación
- Conocimiento





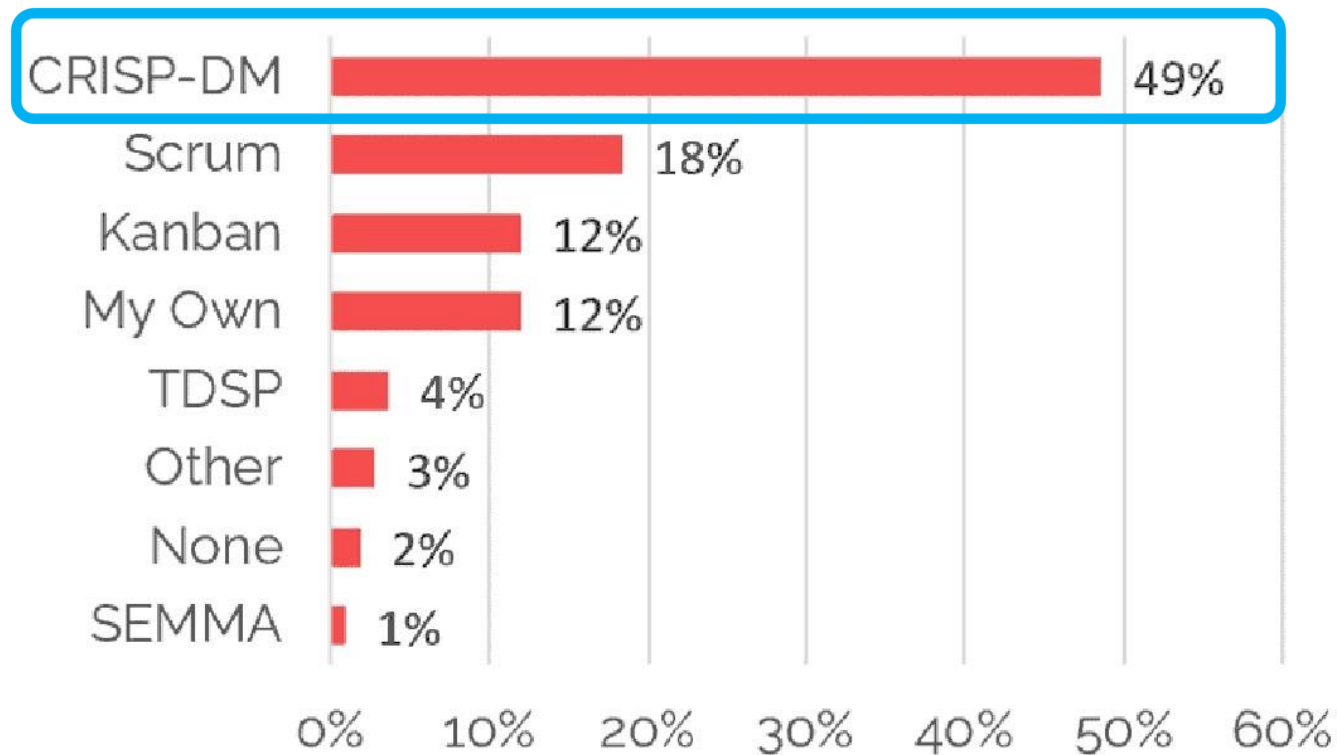
# KDD - ¿En que consiste?

A grandes rasgos el KDD:

- es un **modelo genérico** de análisis de datos
- es un **modelo iterativo e interactivo** para obtener conocimiento a partir de los datos
- requiere muchas **decisiones no automatizadas**
- **establece la toma de decisiones a grandes rasgos** (no profundiza en la descripción de tareas).

# KDD como base a otros modelos

De este modelo surgen modelos más específicos.





# Recursos y Bibliografía

- Iofullstack repository.

**Disponible en línea:**

<https://github.com/iofullstack/data-scientist/tree/main/Web-Mining>

- REF 1 - what\_is\_data\_science.pdf