

MINERÍA DE DATOS

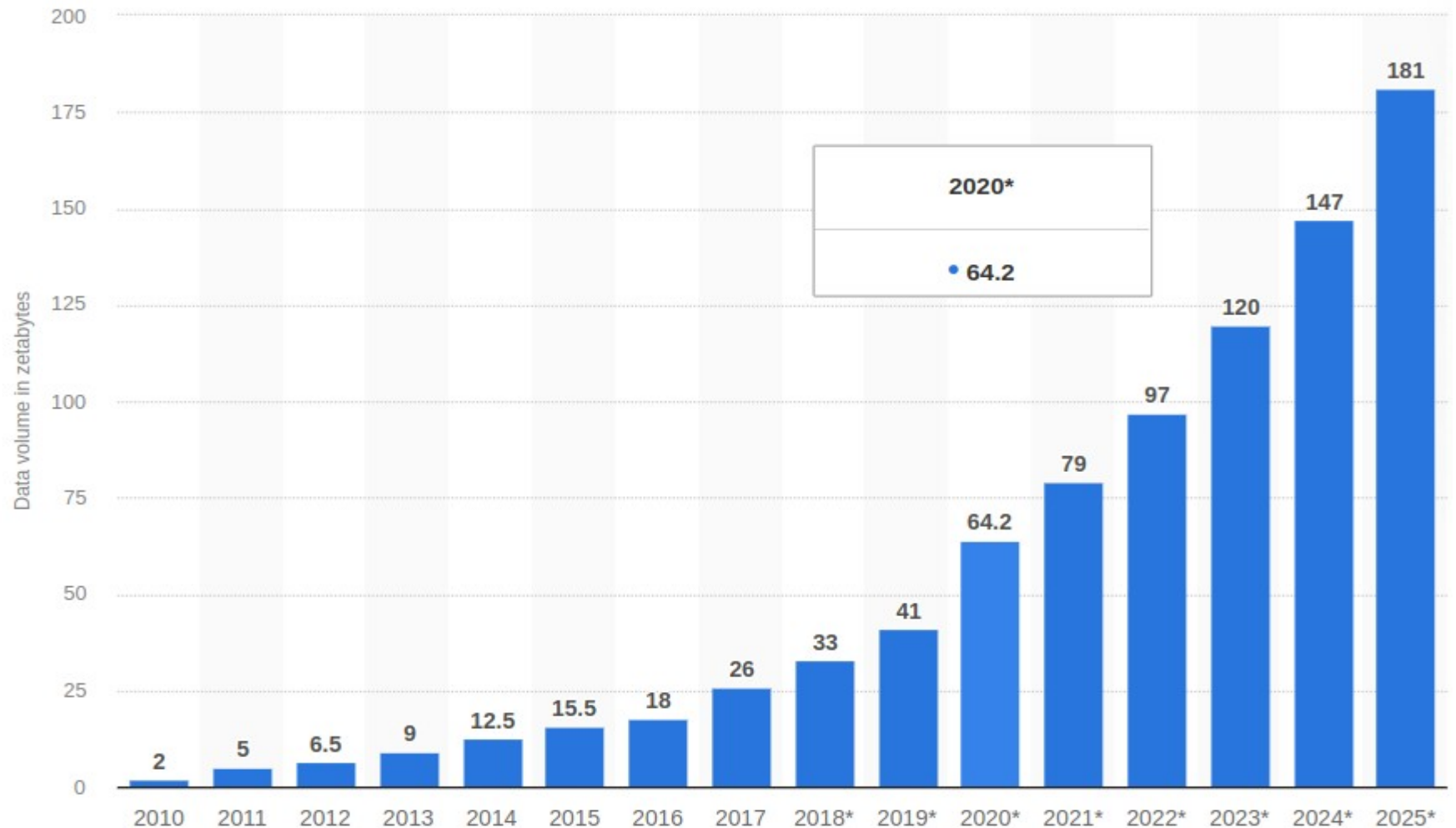


Oswaldo.solarte@correounivalle.edu.co

AGENDA

1. Introducción
2. Business Intelligence & Data warehouses
3. Data Mining

EXPLOSIÓN DE LOS DATOS

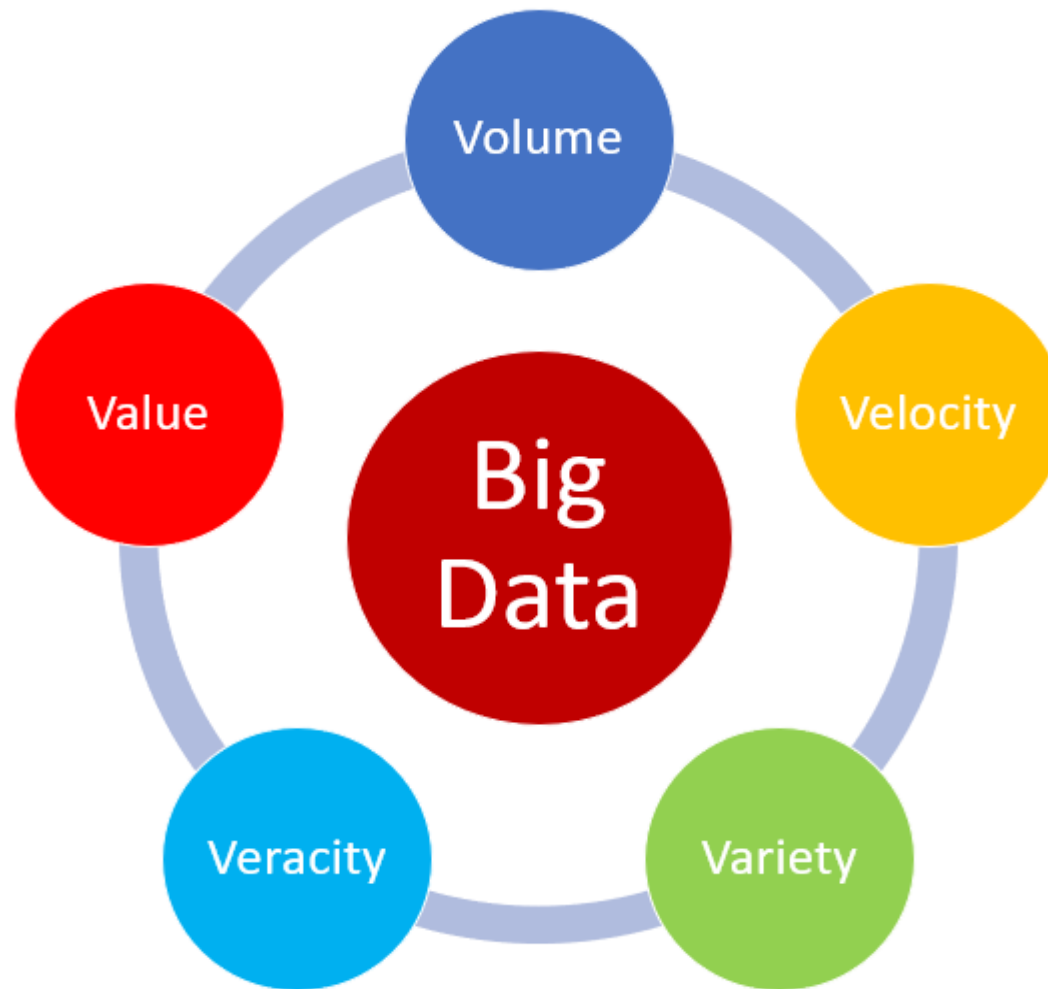


EXPLOSIÓN DE LOS DATOS





Qué es Big Data?



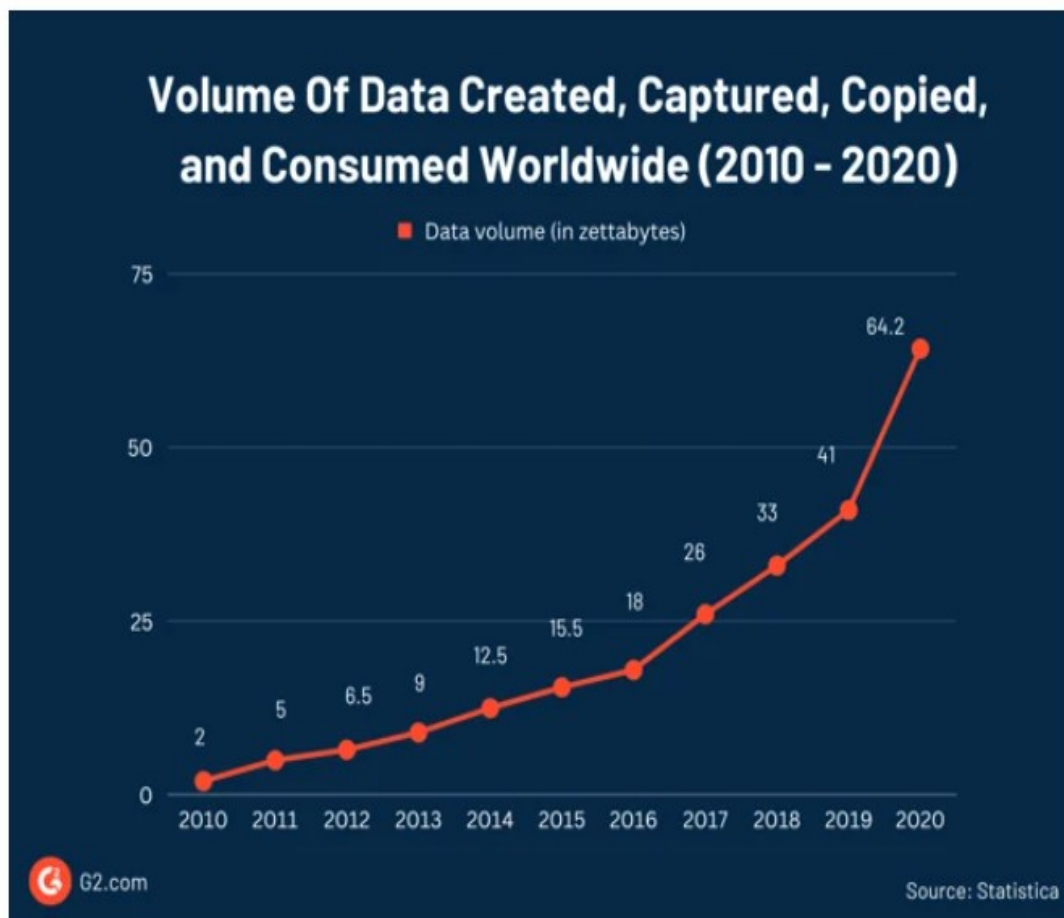
- Diferentes grados de **complejidad, ambigüedad** en los datos
- No pueden ser procesados utilizando **tecnologías tradicionales**

VOLUMEN

97 zettabytes

the estimated volume of data created worldwide in 2022.

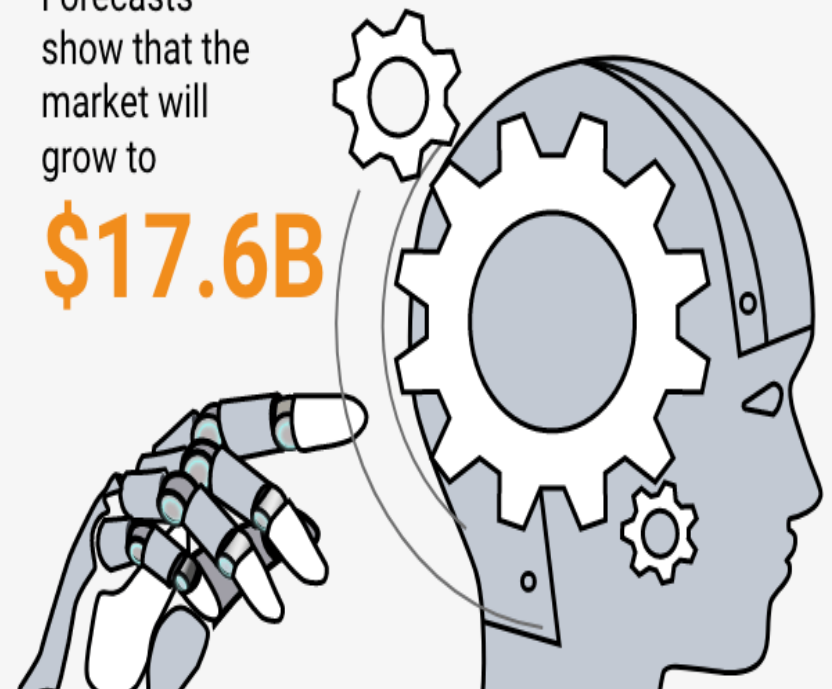
Source: **Statista**



Size of the BI and analytics software application market

Forecasts show that the market will grow to

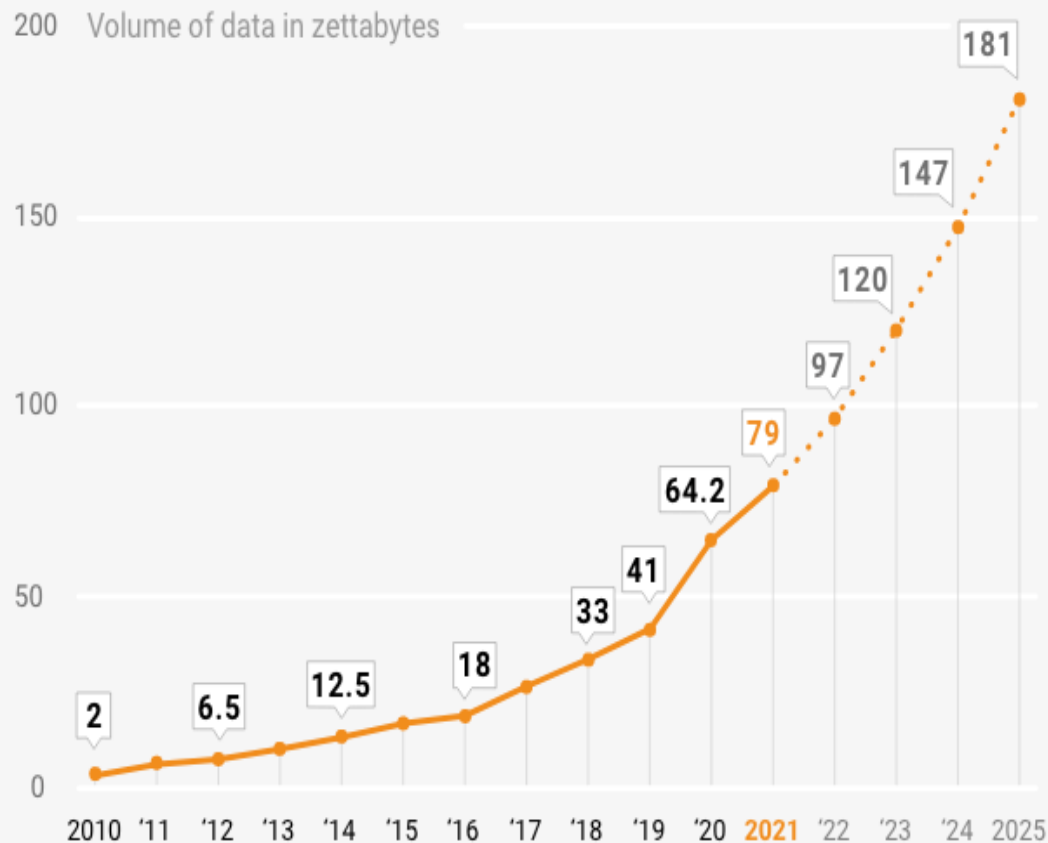
\$17.6B



VOLUMEN



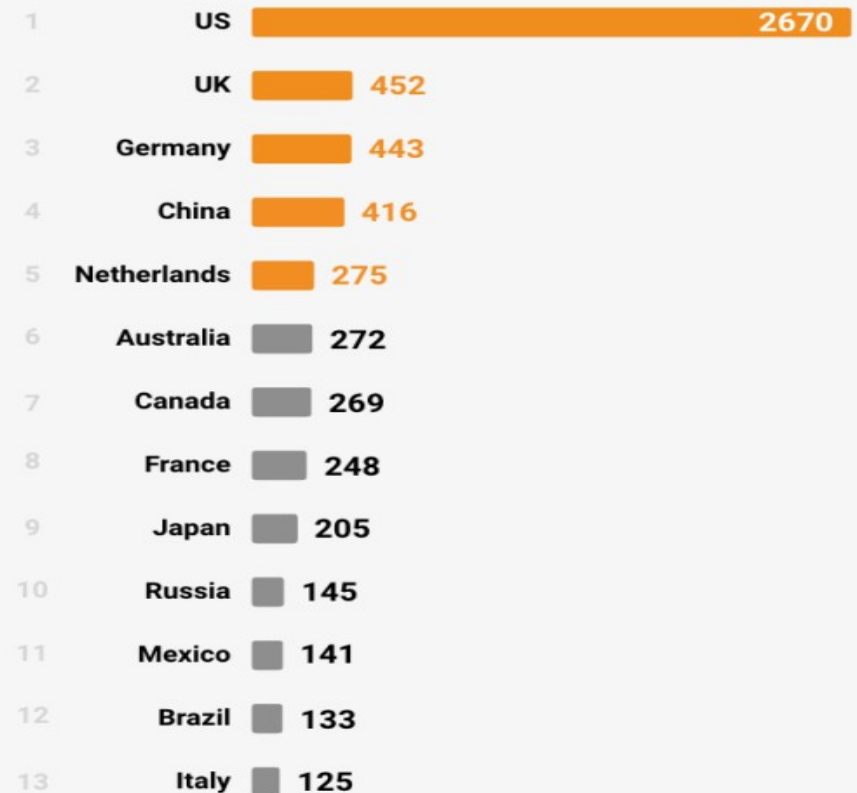
The volume of data generated, consumed, copied, and stored is projected to exceed 180 zettabytes by 2025



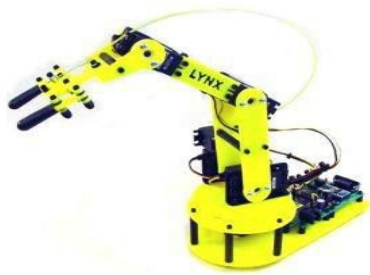
Number of data centers worldwide in 2021



In 2021, the United States is the country with the most data centers (2670) in the world, followed by the UK (452) and Germany (443)



VARIEDAD

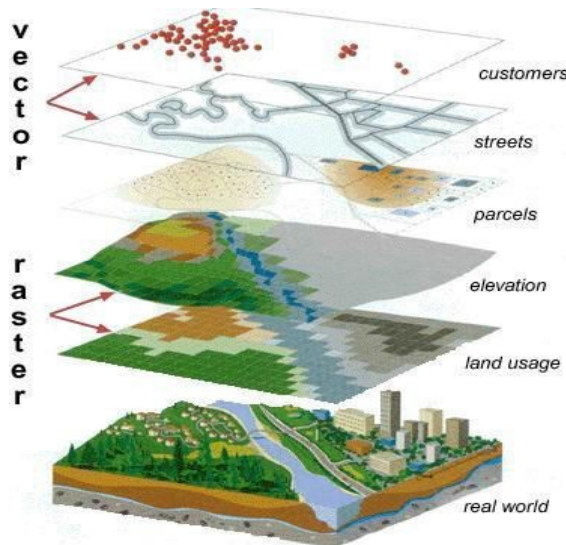


Machine data

Social media data



Geo spatial Data



Email data



DATA MINING



- ❑ Análisis y la exploración de grandes volúmenes de datos para **descubrir patrones significantes** (automática o semi automáticamente)
- ❑ La meta: mejorar procesos de ventas, marketing y en general la relación con los clientes.

DATA MINING



Data mining, also known as knowledge discovery in data (**KDD**), is the process of **uncovering patterns** and other valuable information from large data sets.

Evolution of **data warehousing technology** and the growth of **big data**, adoption of data mining techniques has rapidly accelerated over the last couple of decades, assisting companies by transforming their **raw data** into **useful knowledge**

<https://www.ibm.com/topics/data-mining>

Qué es minería de datos



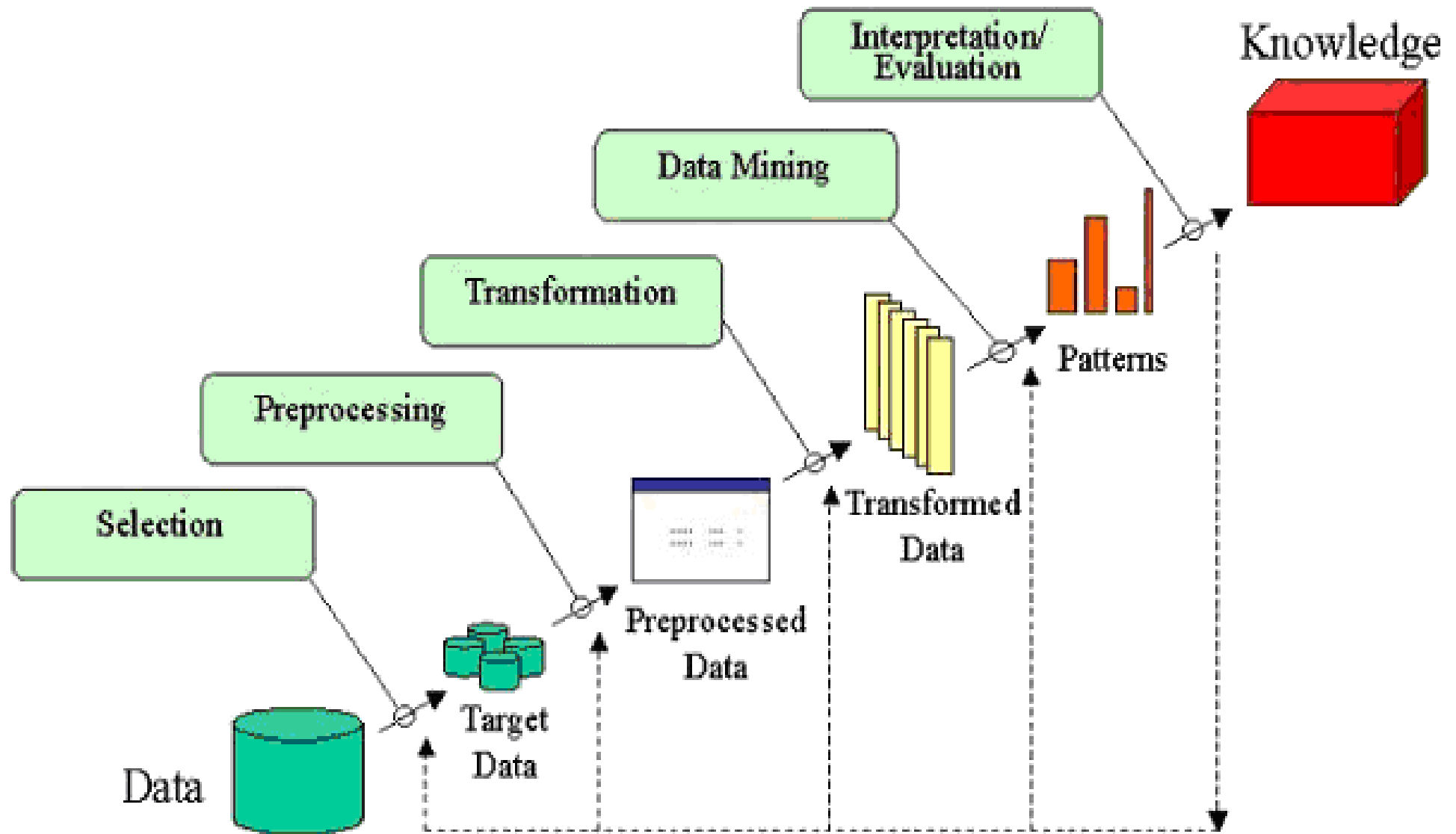
Minería de datos: descubrimiento de conocimiento a partir de datos.

KDD: knowledge discovery from databases

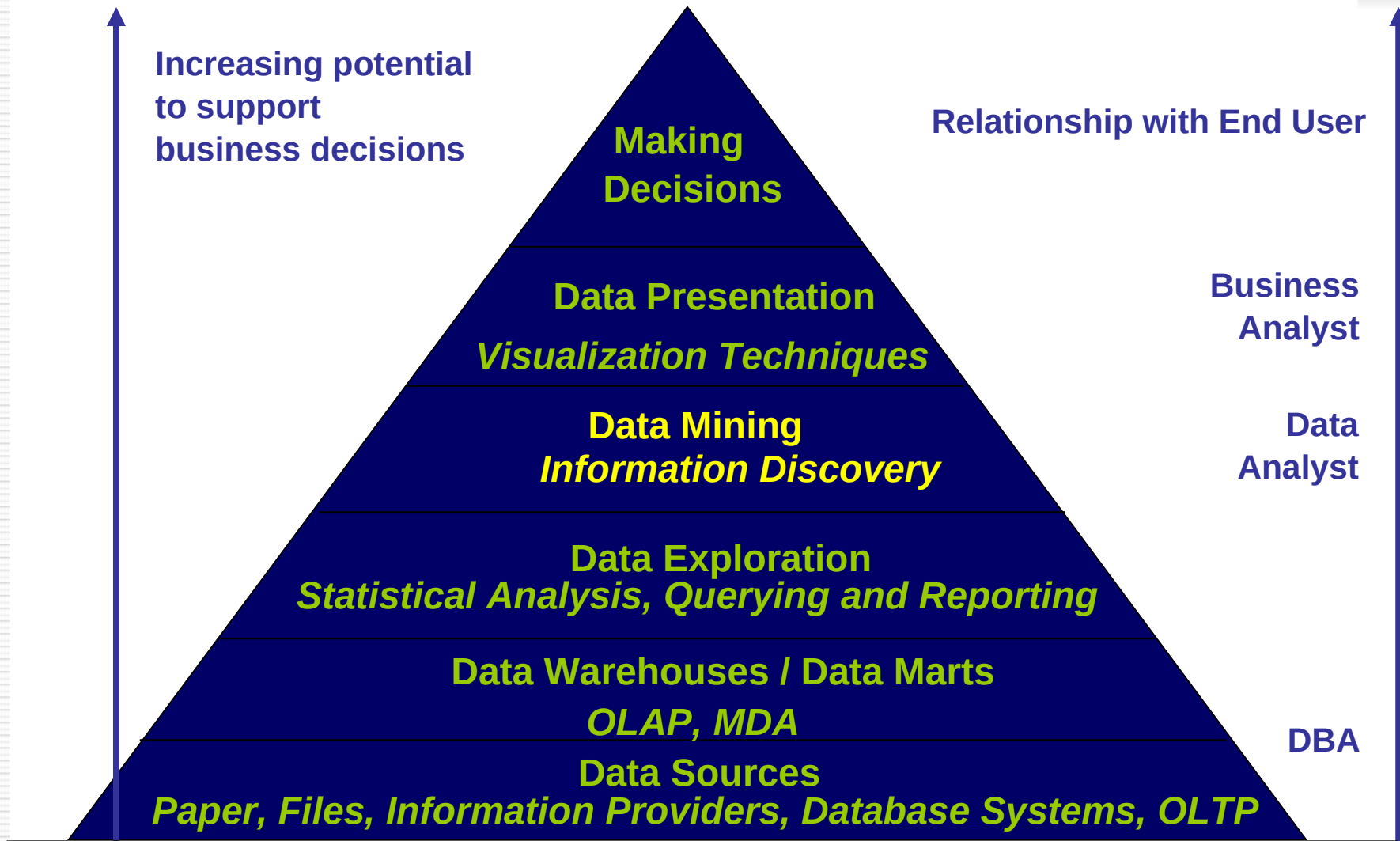
Extracción de datos, patrones o conocimiento de una gran cantidad de datos.

Los datos extraídos son no triviales, implícitos, previamente desconocidos y potencialmente útiles

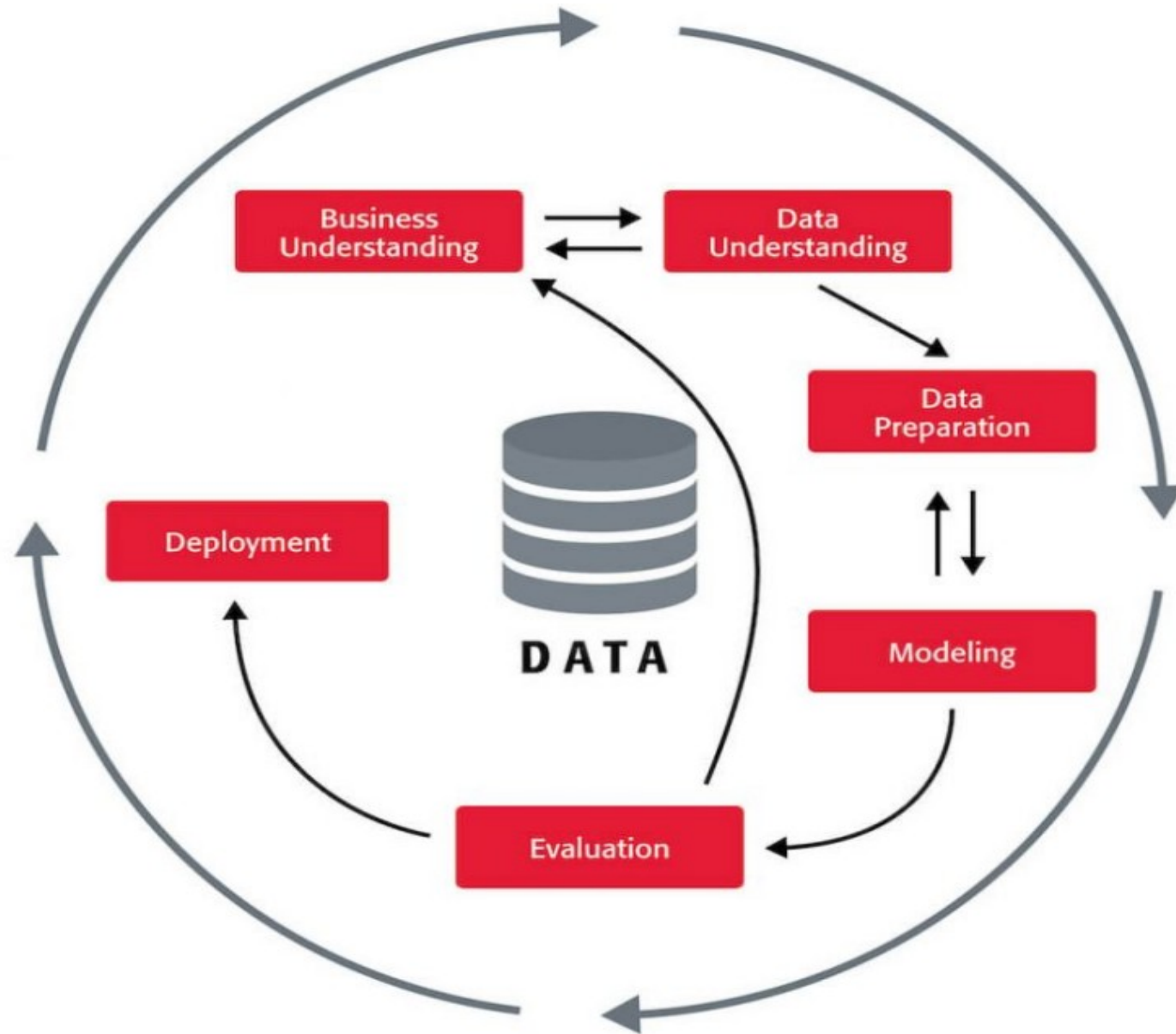
Data Mining es un proceso



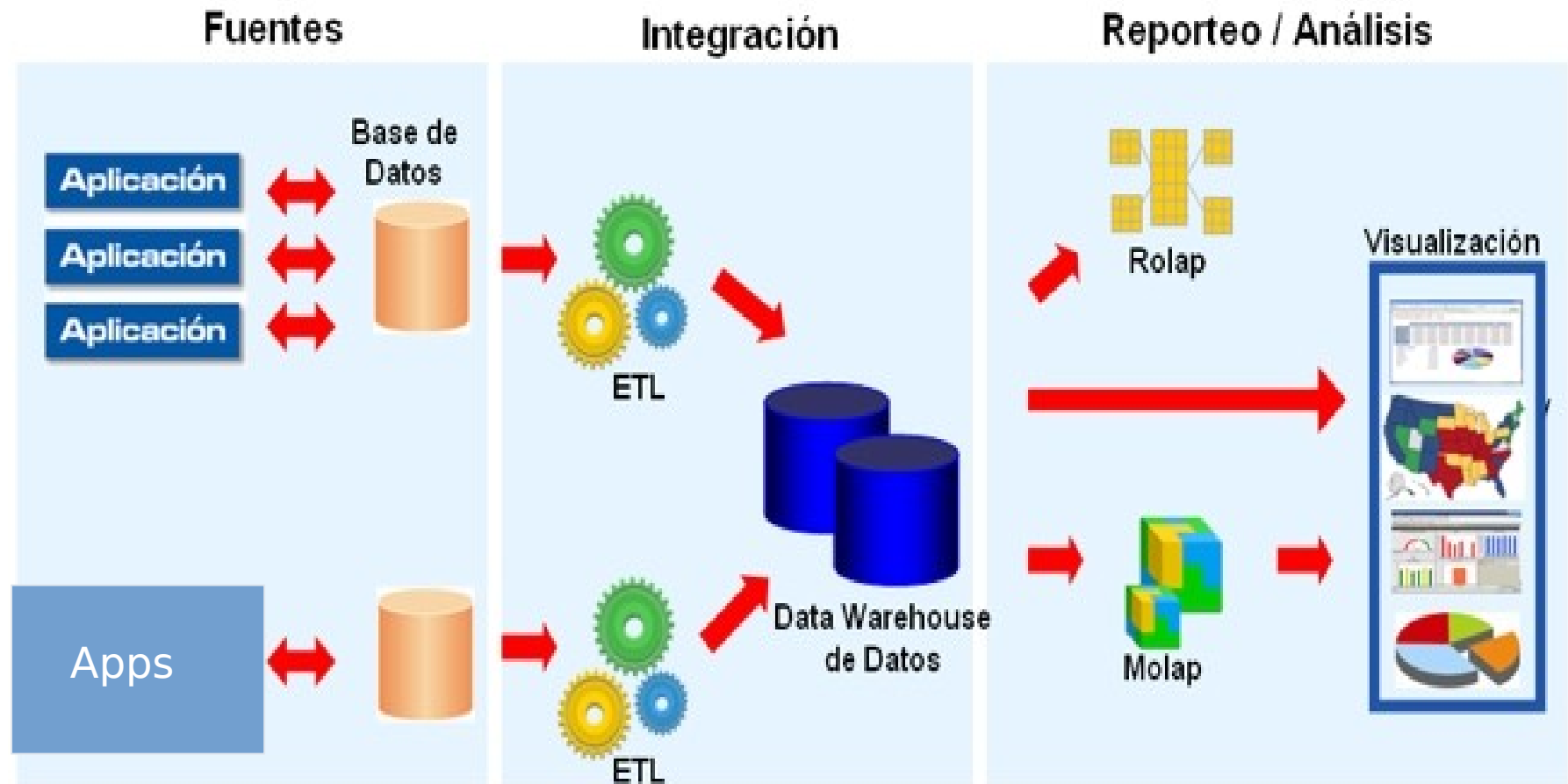
DATA MINING & BI



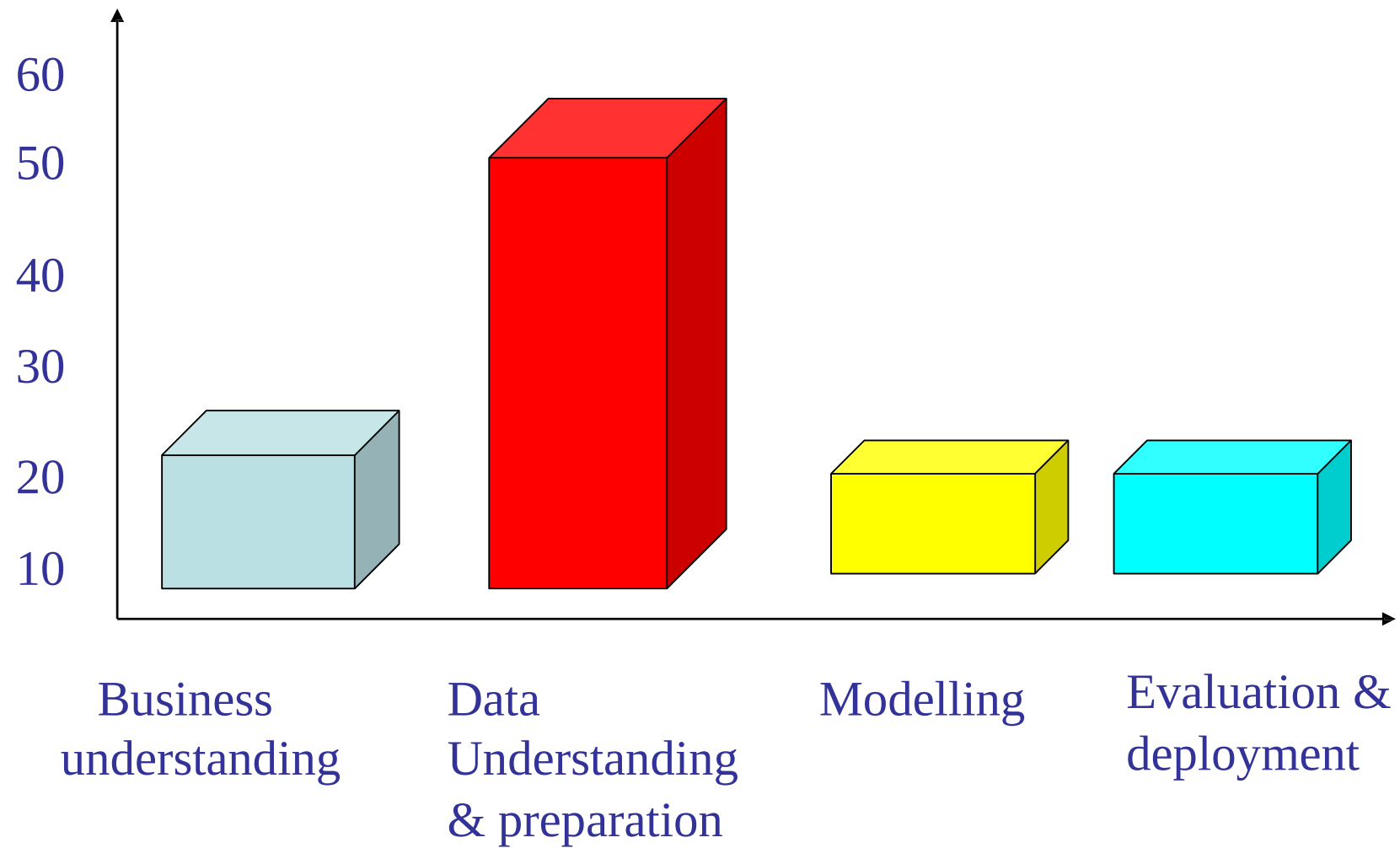
Metodologia CRISP-DM



DATA WAREHOUSE



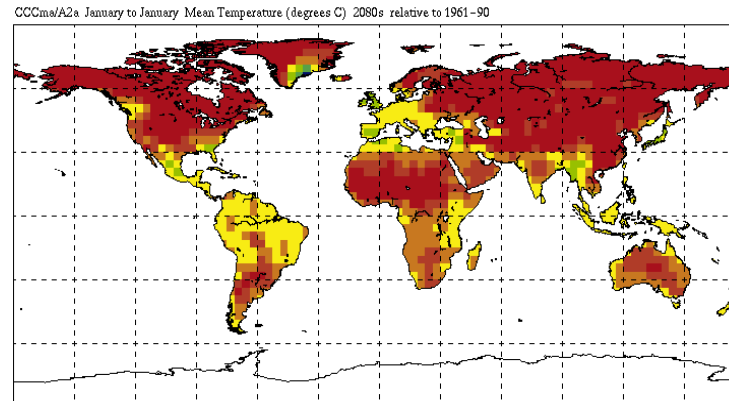
ESFUERZO ASOCIADO



Grandes oportunidades para resolver problemas de gran impacto en la sociedad



Improving health care and reducing costs



Predicting the impact of climate change



Finding alternative/ green energy sources



Reducing hunger and poverty by increasing agriculture production

Data Mining proporciona inteligencia



- ❑ Bases de datos proporcionan los datos
- ❑ Necesidad: explorar datos y encontrar patrones, reglas (entender qué está pasando)
- ❑ Predecir que pasará
- ❑ Se requieren: técnicas y herramientas para extraer el máximo beneficio de los datos.

DATA MINING



¿Cómo nos ayuda?

- ¿Quiénes son nuestros clientes fieles?
 - Clientes que dejarán la compañía.
 - ¿Dónde localizo la próxima sucursal?
 - ¿Cuáles serán mis niveles de venta?
 - ...
- ❑ Las respuestas están en los datos. Técnicas de *data mining* pueden ayudar a encontrarlas

DATA MINING



¿Por qué ahora?

- ❑ Las técnicas existentes
- ❑ convergencia de una serie de factores:
 - Cantidad de datos
 - Datos integrados (data warehouse)
 - Más capacidad de cómputo de los computadores
 - Competencia feroz

DATA MINING



Importante

- ❑ La promesa de Data Mining : encontrar los patrones
- ❑ Hallarlos no es suficiente
- ❑ Los patrones se tienen que entender y valorar
- ❑ El entendimiento de los patrones facilitan actuar
- ❑ se transforman en valor para la compañía.

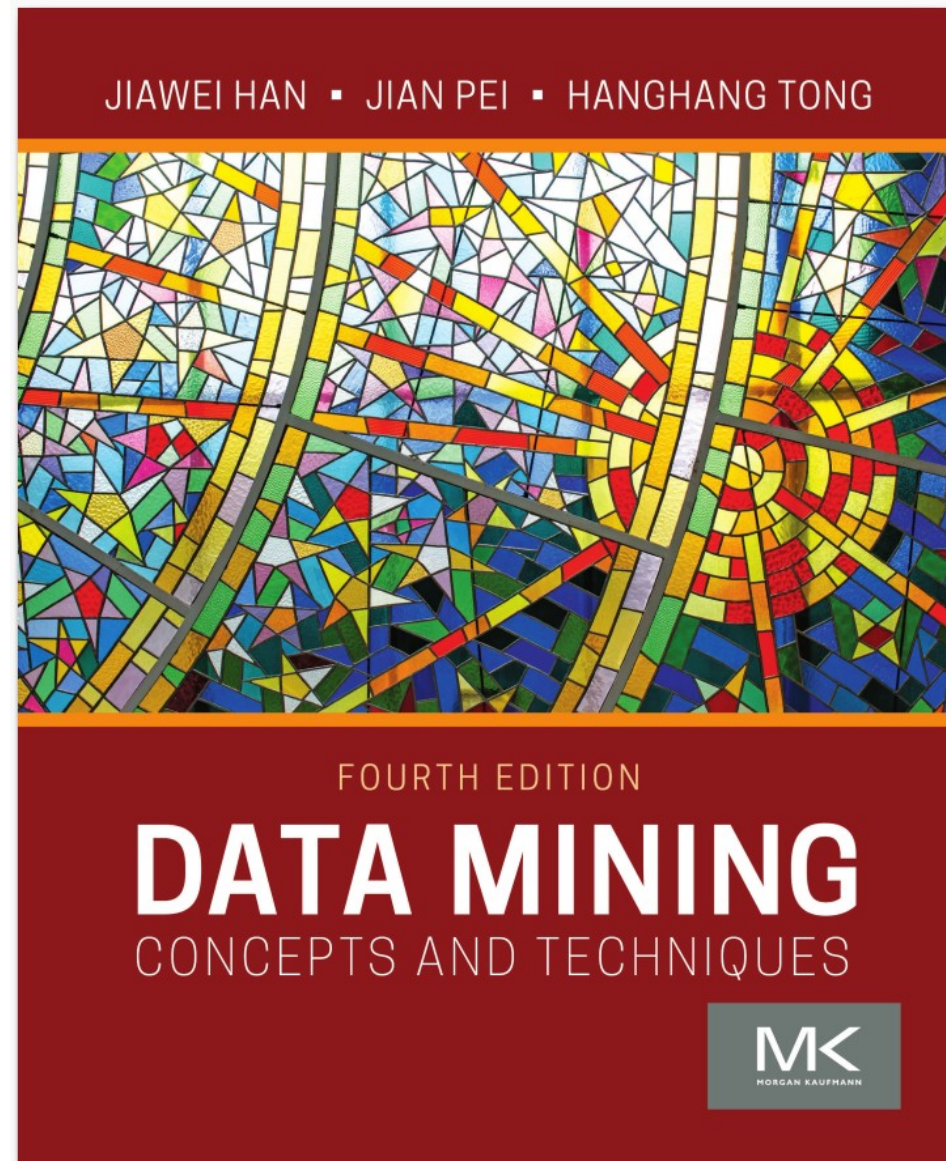
Datos transformados en **Conocimiento** que permita actuar en forma rápida y eficiente

Importante

- La promesa de Data Mining : «Encontrar patrones en los datos»



Libro



<http://hanj.cs.illinois.edu/>


<http://dm1.cs.uiuc.edu/>

Overview

The database group at MIT conducts research on all areas of database systems and information management. Projects range from the design of new user interfaces and query languages to low-level query execution issues, ranging from design of new systems for database analytics and main memory databases to query processing in next generation pervasive and ubiquitous environments, such as sensor networks, wide area information systems, personal databases, and the Web.

Professor Madden offers a class in [Database Systems \(6.830\)](#).

Projects



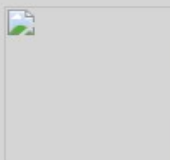
Intel Science and Technology Center in Big Data

In the [Big Data ISTC](#), our mission is to produce new data management systems and compute architectures for Big Data. Together, these systems will help people process data that exceeds the scale, rate, or sophistication of current data processing systems. We are working to demonstrate the effectiveness of these solutions on real applications in science, engineering, and medicine, making our results broadly available through open sourcing.

DataHub

In the [DataHub](#) project, we are building an experimental hosted platform (GitHub-like) for organizing, managing, sharing, collaborating, and making sense of data. The hosted platform provides easy to use tools/interfaces for:

- Managing your data (ingestion, curation, sharing, collaboration)
- Using others' data (discovering, linking)
- Making sense of data (query, analytics, visualization)



CarTel

In [CarTel](#), we are building a system for managing data in the face of intermittent and variable connectivity. We are focusing, in particular, on automotive applications that involve high-rate sensing of road, traffic, and infrastructure conditions. The two key technologies we are developing are CafNet, a carry-and-forward network stack, and a distributed, signal-oriented, priority-driven query processor.

People

Faculty

- [Sam Madden](#)
- [Mike Stonebraker](#)

Administrative Assistant

- Sheila Marian

Ph.D.

- Firas Abuzaid
- Leilani Battle
- Anant Bhardwaj
- Rachel Harding
- Albert Kim
- Yi Lu
- Oscar Moll
- Anil Shanbhag
- [Rebecca Taft](#)
- [Manasi Vartak](#)

Research Staff

- Albert Carter (Staff Programmer, Big Data @ MIT)
- [Stavros Papadapolous](#) (ISTC Researcher, and Visting Researcher)
- [Nesime Tatbul](#) (ISTC Researcher, and Visting Researcher)

Postdoc

- [Raul Castro-Fernandez](#)
- [Holger Pirk](#)

M.Eng

Evangelos Taratoris

Alumni

- [Daniel Abadi](#) (Yale)
- [Ziawasch Abedjan](#)
- Peter Bailis (Stanford University)
- [Magdalena Balazinska](#) (U. Wash.)
- Joshua Blum
- Daniel Bruckner (PhD Student, UC Berkeley)
- [Alvin Cheung](#) (U. Washington)
- [Philippe Cudre-Mauroux](#) (U. of Fribourg, Switzerland)
- [Carlo Curino](#) (Microsoft Research)
- [Jennie Duggan](#) (Northwestern University)
- [Aaron Elmore](#) (University of Chicago)
- [Jakob Eriksson](#) (U. Illinois at Chicago)
- [Stavros Harizopoulos](#)
- [Alekh Jindal](#) (Microsoft Research)
- [Evan Jones](#)
- [Barzan Mozafari](#) (U. Michigan, Ann Arbor)
- [Ryan Newton](#) (U. of Indiana)
- Arvind Thiagarajan (Twitter, Inc.)
- Michael Farry (Charles River Analytics)
- Miguel Ferraria
- [Thomer Gil](#)
- David Goehring (UC Berkeley)
- [Lewis Girod](#)
- [Michael Gubanov](#)
- George Huo (Google)
- [Edmond Lau](#) (Quora)
- Umberto Malesci
- Adam Marcus

CS246: Mining Massive Data Sets

Winter 2024

Logistics

- **Lectures:** are on Tuesday/Thursday 3:00-4:20 PM PDT in person in the NVIDIA Auditorium.
- **Lecture Videos:** are available on [Canvas](#) for all the enrolled Stanford students. You can also check our past [Coursera MOOC](#).
- **Public resources:** The lecture slides and assignments will be posted online as the course progresses. We are happy for anyone to use these resources, but we cannot grade the work of any students who are not officially enrolled in the class.
- **Contact:** Students should ask *all* course-related questions on [Ed](#), where you will also find all the announcements. For external enquiries, personal matters, or in emergencies, you can email us at cs246-win2324-staff@lists.stanford.edu.
- **Academic accommodations:** If you need an academic accommodation based on a disability, you should initiate the request with the [Office of Accessible Education \(OAE\)](#). The OAE will evaluate the request, recommend accommodations, and prepare a letter for faculty. Students should contact the OAE as soon as possible since timely notice is needed to coordinate accommodations.

Instructor



Jure Leskovec

Course Assistants



Ethan Allavarpu (Head
TA)



Aditya Agrawal

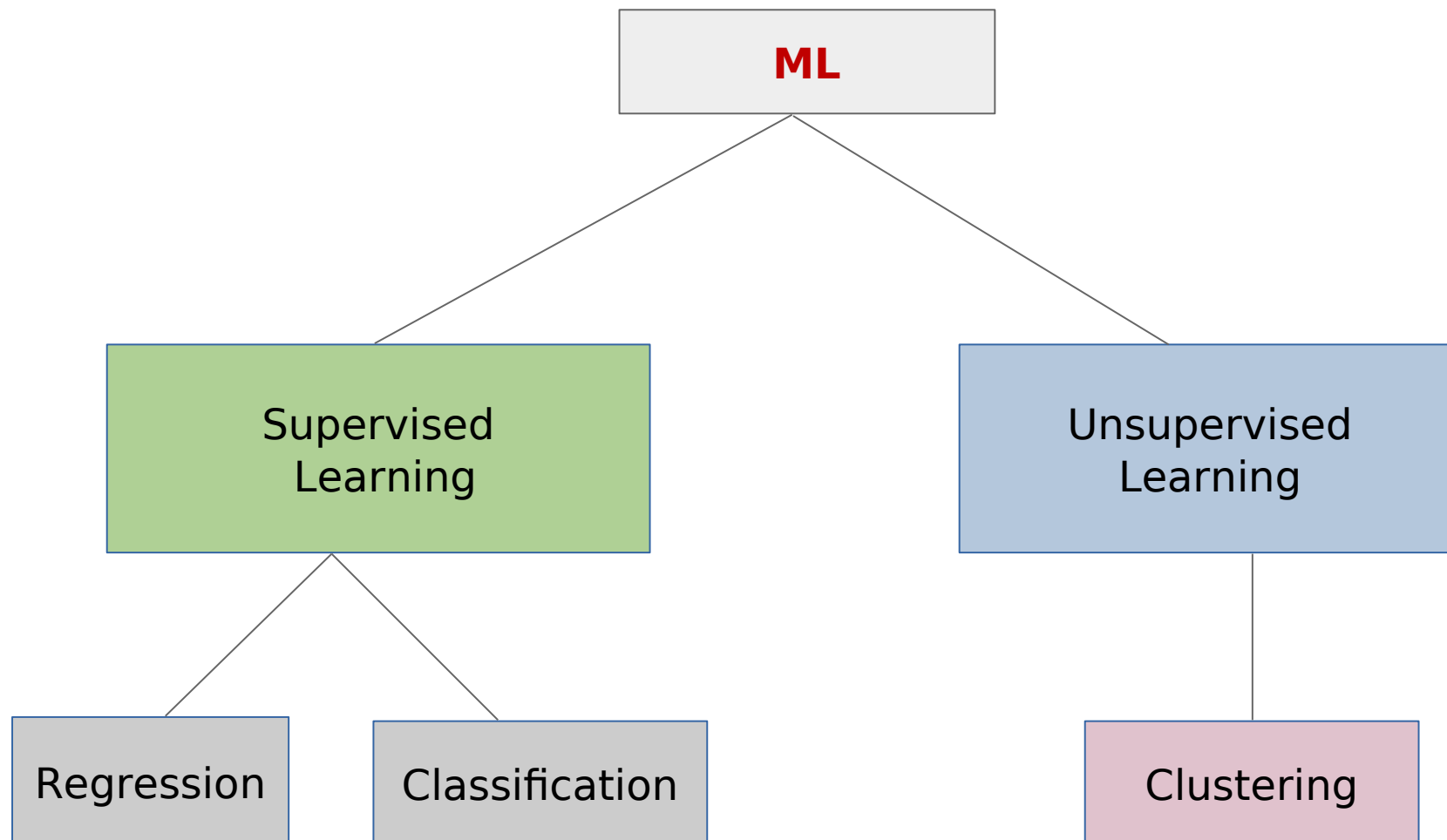


Spencer Siegel

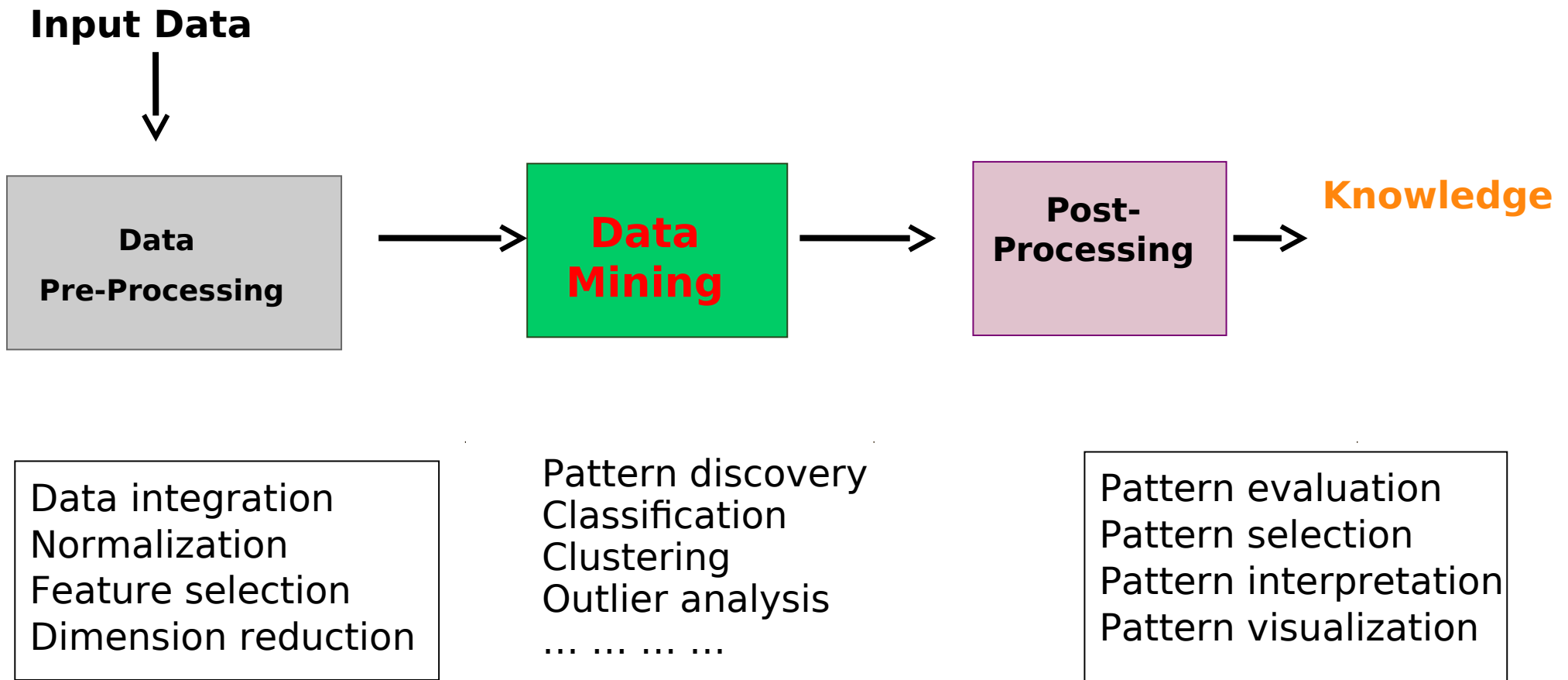
MACHINE LEARNING



Campo de la **Inteligencia Artificial** que usa algoritmos que tienen la capacidad de identificar (**Aprender**) patrones en datos masivos y elaborar **predicciones**.

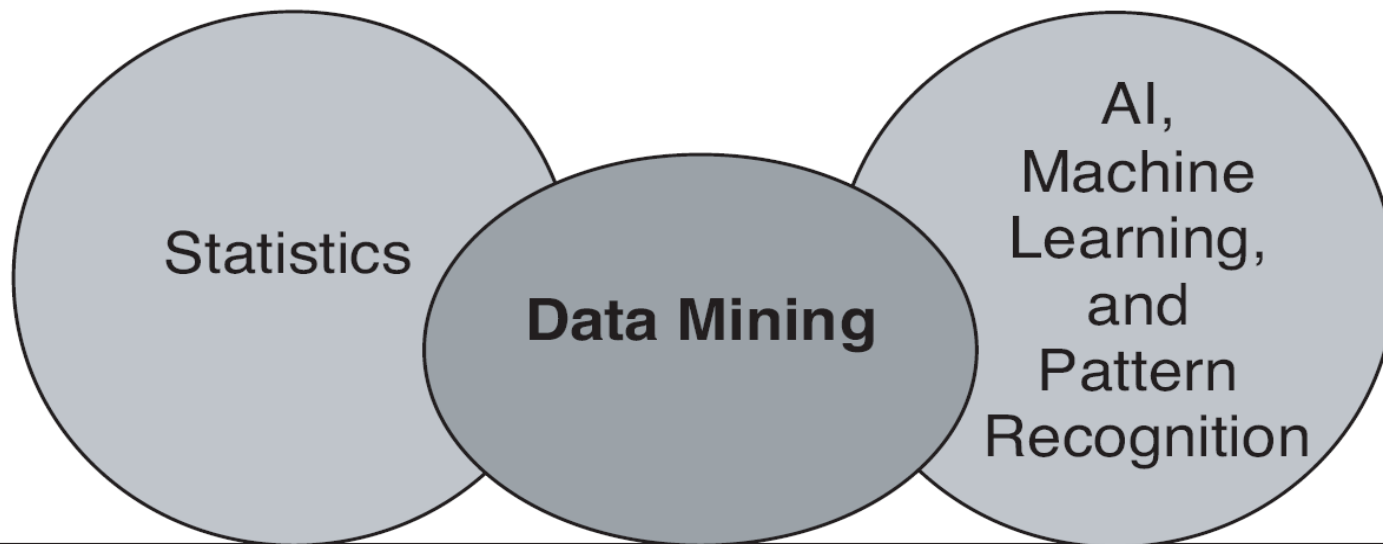


Data Mining: Machine learning View



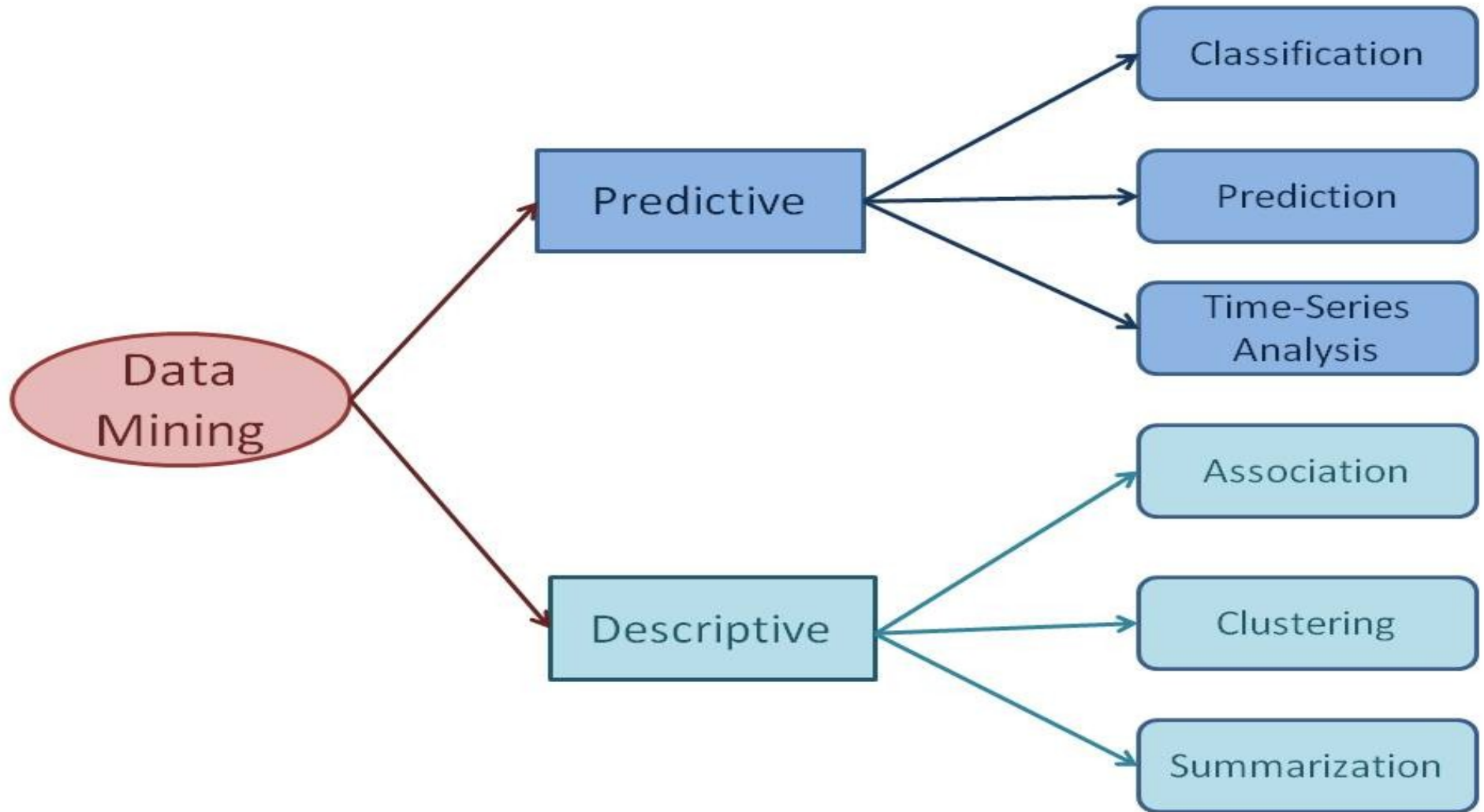
Data Mining y otras áreas

- Extrae ideas de otras áreas como machine learning/AI, pattern recognition, estadística, bases de datos



Database Technology, Parallel Computing, Distributed Computing

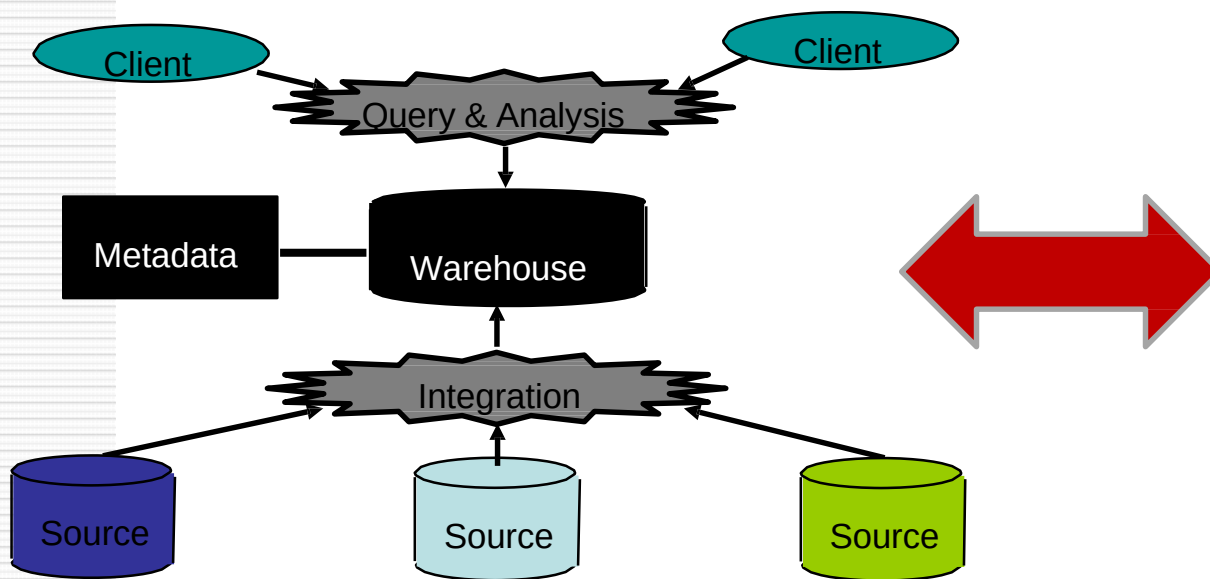
Data Mining: Tareas



Data mining es un proceso



□ Data warehousing



□ Data mining

