Introduction to Data Mining

Data Mining & Neural Networks

Dr. Wilmer Garzón

Director, Master's Program in Data Science
Department of Computer Engineering

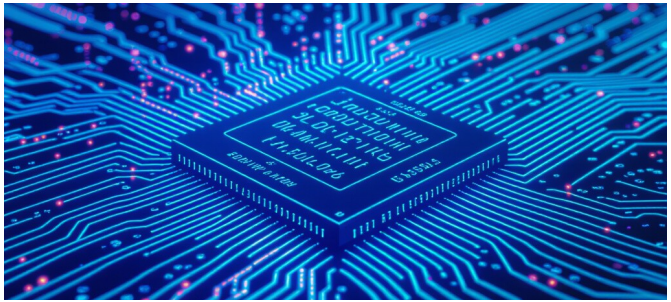
Escuela Colombiana de Ingeniería
Universidade da Coruña

2025

Agenda

- What is Data Mining?
- Key Concepts and Tasks
- Real-World Applications
- Challenges in Data Mining
- Case Studies and Examples
- Conclusions
- Q&A

What is Data Mining?



Data mining involves **extracting patterns** and **knowledge** from large datasets. It is a crucial part of the **Knowledge Discovery in Databases (KDD)** process and utilizes **statistics**, **machine learning**, and **database systems**.

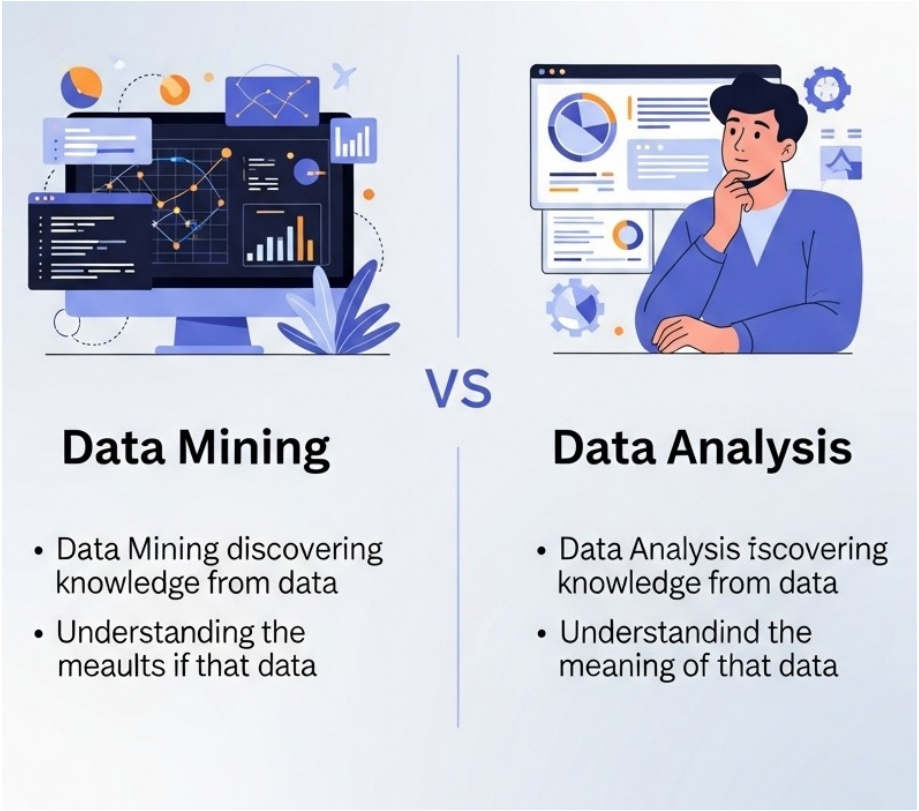
Key Components

Data mining focuses on identifying valuable insights from data, leveraging advanced techniques to analyze and interpret complex datasets.

Data Mining vs. Data Analysis

Data Mining (DM)

Data Mining is **predictive** and **prescriptive**, concentrating on discovering **hidden patterns** and making future predictions based on data.



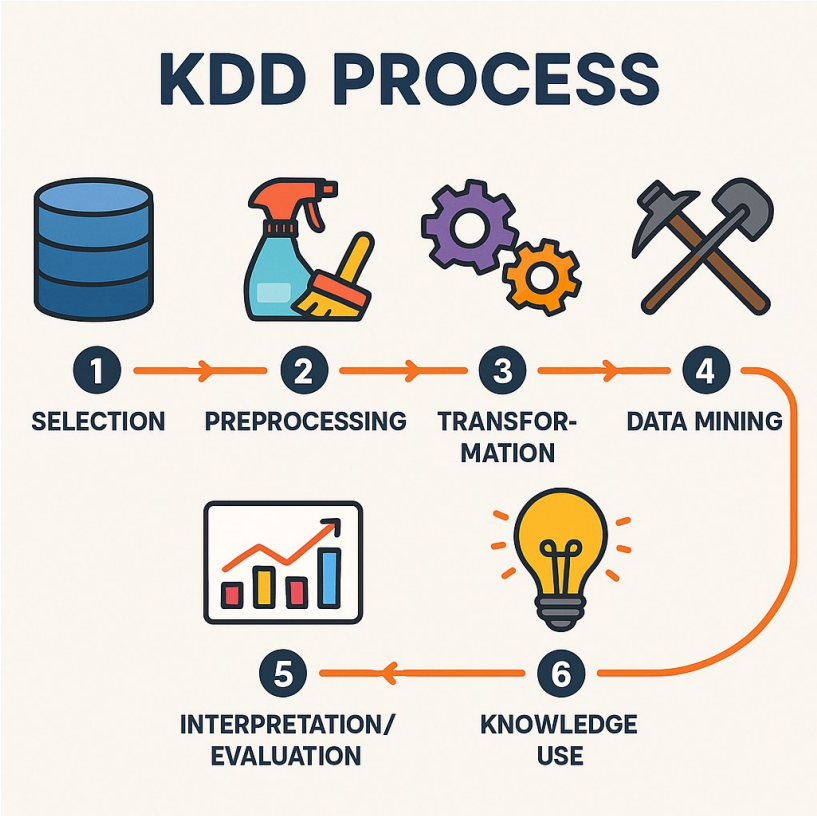
Data Analysis (DA)

Data analysis is primarily **descriptive** and **diagnostic**, focusing on summarizing past data and understanding trends.

The KDD Process

KDD: Knowledge Discovery in Databases, a systematic approach to extract useful knowledge from data.

- Data selection
- Data preprocessing
- Data transformation
- Data mining
- Interpretation and evaluation

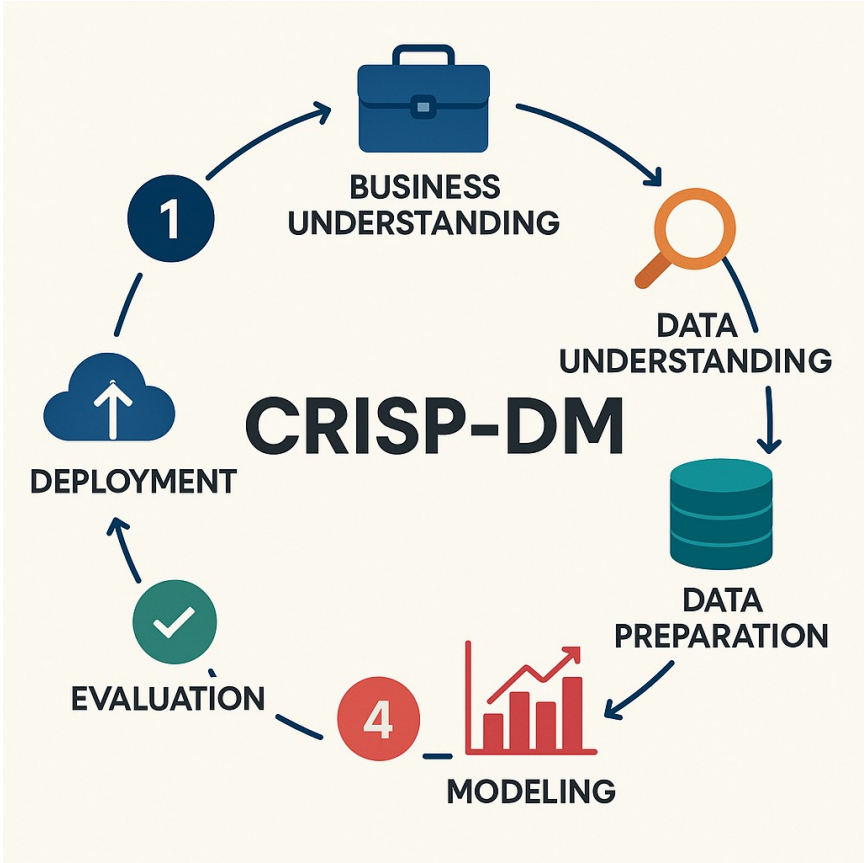


CRISP-DM Methodology

- **Cross-Industry Standard Process for Data Mining.**

The methodology consists of six phases:

1. **Business Understanding –** Define objectives and requirements from a business perspective.
2. **Data Understanding –** Collect and explore data to identify quality issues and insights.
3. **Data Preparation –** Clean, transform, and format data for modeling.
4. **Modeling –** Apply data mining techniques and build predictive models.
5. **Evaluation –** Assess model performance and ensure it meets business goals.
6. **Deployment –** Implement the model in a real-world environment for decision-making.



Key Data Mining Tasks

- Classification
- Clustering
- Regression
- Association rule learning
- Anomaly detection

Classification

- Assign data to predefined categories
- Example: **Spam vs. non-spam emails**
- Algorithms:
  - **Decision trees,**
  - **SVM,**
  - **kNN**

Classification Examples

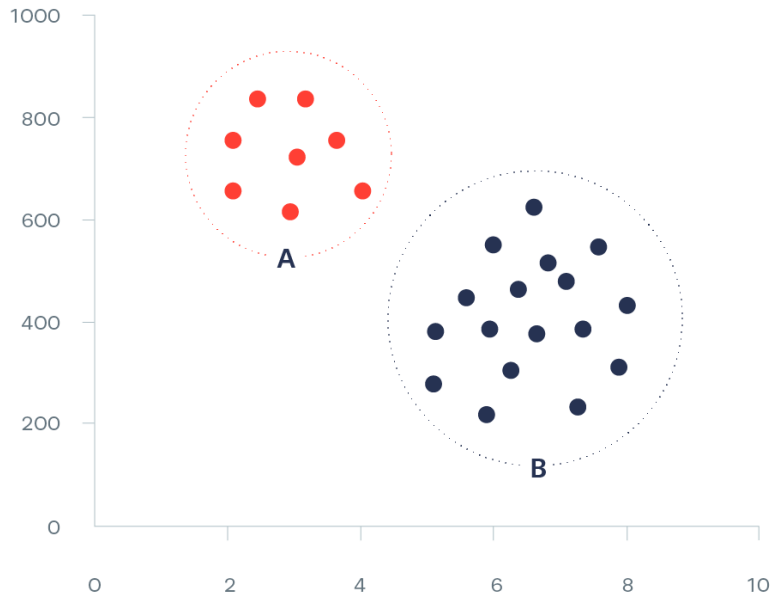
TYPES	EXAMPLES	ORGANIZATIONS
Public Data Data meant for public consumption.	• Government publications • Publicly available research	• Government agencies • Nonprofit organizations • Research institutions • Media outlets
Internal data Data for exclusive enterprise use.	• Employee information • Intellectual property • Operational data	• Research and development organizations • Private corporations and businesses
Confidential data Sensitive data needing protection from unauthorized access.	• Trade secrets • Legal documents • Internal investigations	• Government intelligence agencies • Law firms
Restricted data Data with additional access limitations beyond what is considered confidential due to legal obligations.	• Healthcare records • Government classified information • Trade agreements and contracts	• Research and development organizations • Healthcare institutions
Private data Personal data belonging to an individual.	• Personal identifiers • Contact information • Biometric data	• Hospitals, clinics, and healthcare providers • E-commerce companies • HR departments
Critical data Data vital for business operations and strategic objectives.	• Infrastructure and system configurations • Emergency response plans • Customer databases	• Energy and utility companies • Financial Institutions • Technology Companies
Regulatory data Data subject to legal or regulatory requirements.	• Personal data • Customer financial information • Patient health records	• Banks • Hospitals, clinics, and healthcare providers

Regression

- Predict continuous values
- Example: **Predicting house prices**
- Algorithms: **Linear regression**, **random forest regression**

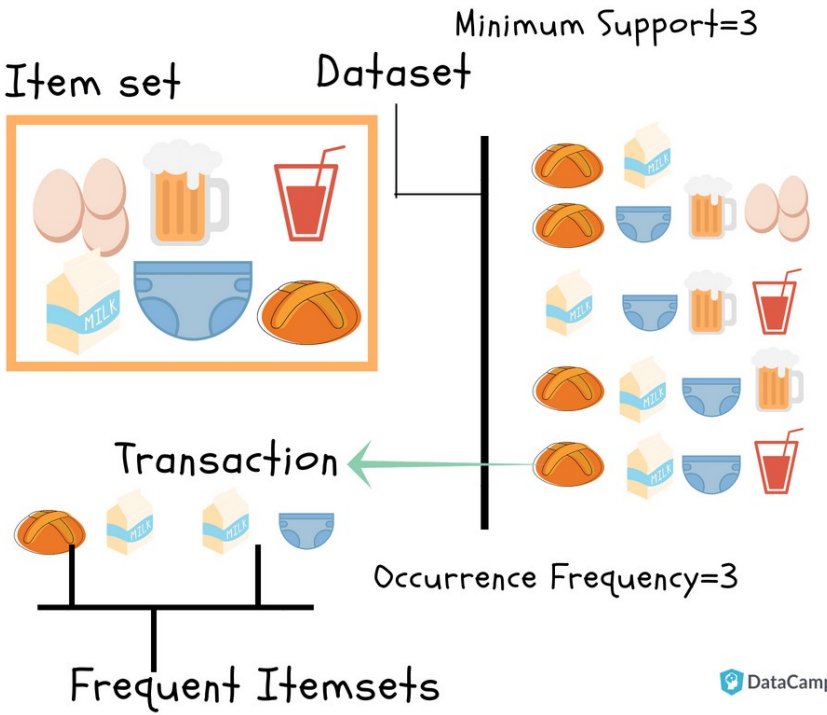
Clustering

- Group similar data without labels
- Example: **Customer segmentation**
- Algorithms:
  - **K-means**,
  - **DBSCAN**,
  - **Hierarchical clustering**



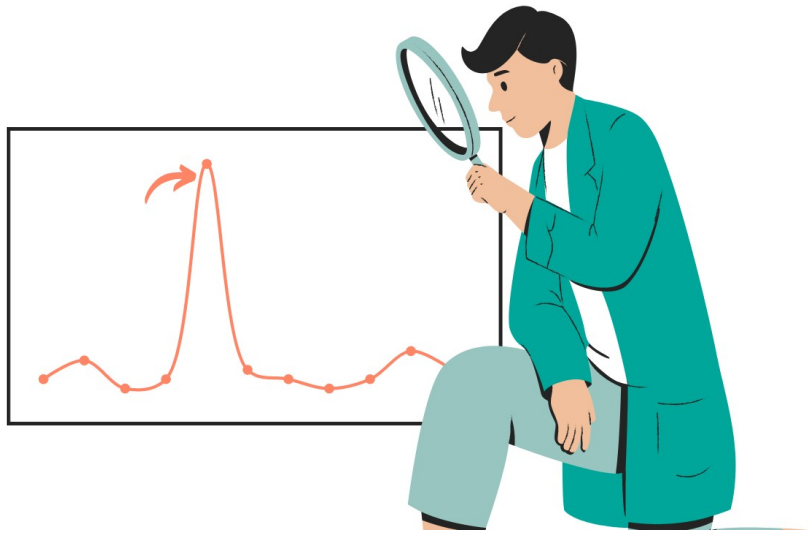
Association Rule Learning

- Discover relationships between variables
- Example: **Market basket analysis**
- Output: “If **X**, then **Y**” rules



Anomaly Detection

- Identify rare or unusual data points
- Example: **Fraud detection**
- Techniques: **Isolation forest**, **statistical methods**



SUMMARY OF DATA MINING TASKS

PREDICTIVE TASKS

Classification

Regression

Time Series Forecasting

Time Series Forecasting

DESCRIPTIVE TASKS

Clustering

Association Rule Mining

Anomaly Detection

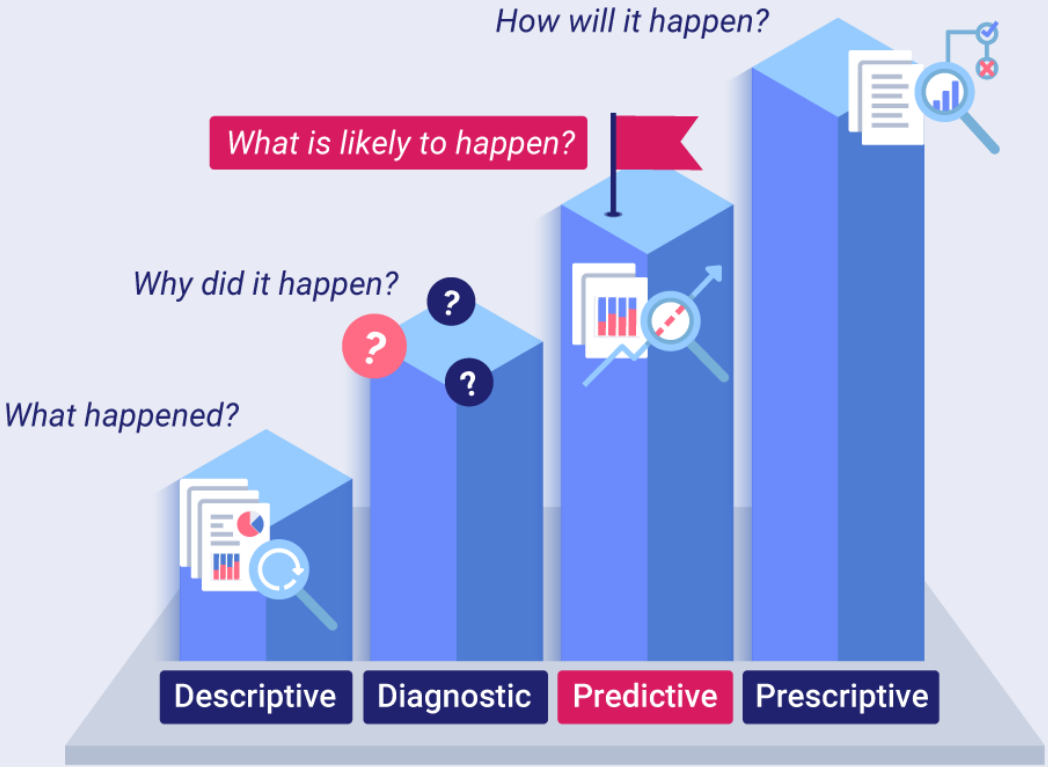
Summarization

Data Science: A Comprehensive Overview

LONGBING CAO, University of Technology Sydney, Australia

Descriptive analytics	Refers to the type of data analytics that typically uses statistics to describe the data used to gain information, or for other useful purposes.
Predictive analytics	Refers to the type of data analytics that makes predictions about unknown future events and discloses the reasons behind them, typically by advanced analytics.
Prescriptive analytics	Refers to the type of data analytics that optimizes indications and recommends actions for smart decision-making.
Explicit analytics	Focuses on descriptive analytics typically by reporting, descriptive analysis, alerting, and forecasting.
Implicit analytics	Focuses on deep analytics, typically by predictive modeling, optimization, prescriptive analytics, and actionable knowledge delivery.
Deep analytics	Refers to data analytics that can acquire an in-depth understanding of why and how things have happened, are happening, or will happen, which cannot be addressed by descriptive analytics.

Types of Advanced Analytics:
Predictive Analytics



Tools for Data Mining

- **Python** (pandas, scikit-learn, matplotlib)
- **R**
- **Weka**
- **RapidMiner**
- **SQL-based tools**

Application Areas Overview

- Healthcare
- Finance
- Retail
- Telecommunications
- Education
- Social media

Healthcare Applications

Healthcare analytics benefit for providers and payers

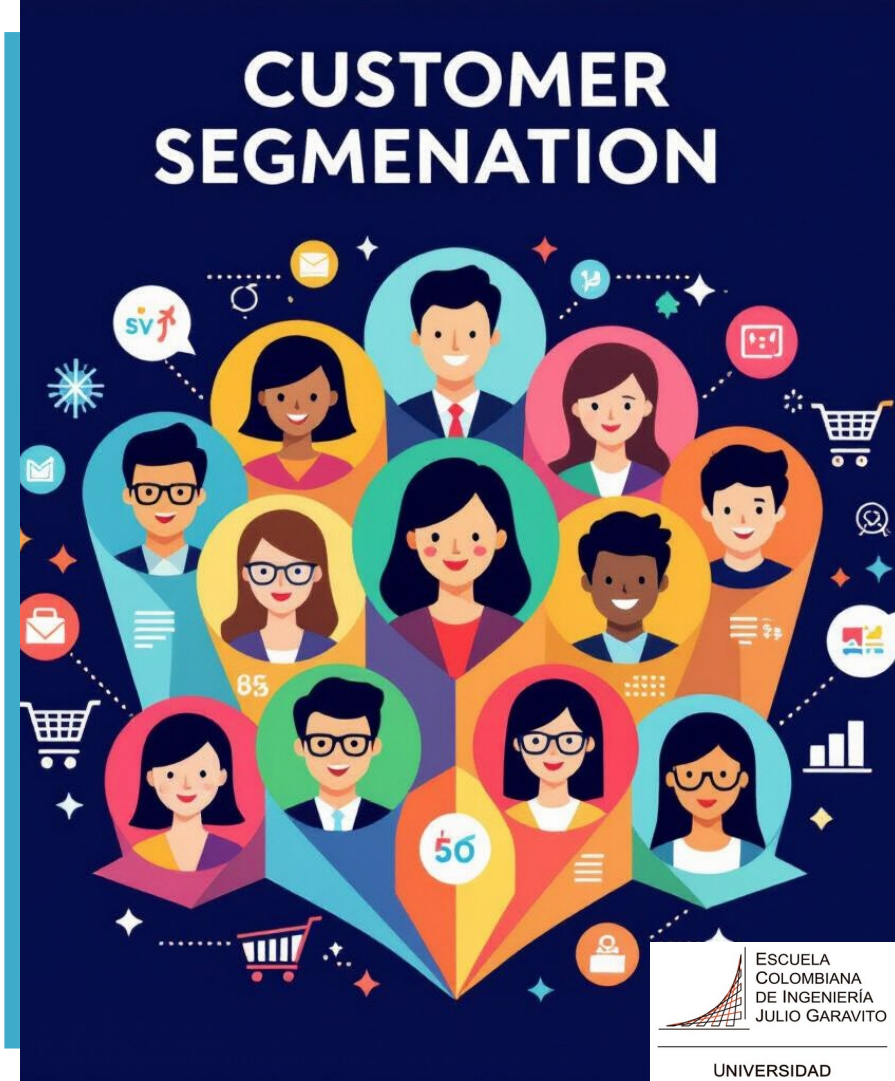
- **Disease prediction** helps in identifying potential health issues early
- **Patient risk profiling** assesses individual health risks effectively
- **Treatment effectiveness analysis** evaluates how well treatments are working and improving outcomes

Finance Applications

- **Fraud detection** helps identify suspicious activities quickly and effectively
- **Credit scoring** evaluates an individual's creditworthiness based on various factors
- **Investment trend analysis** examines market patterns to guide informed decisions

Retail Applications

- **Customer segmentation** is crucial for targeted marketing efforts
- **Product recommendations** enhance user experience and drive sales
- **Inventory optimization** ensures efficient stock management and reduces costs



Telecommunications Applications

- **Churn prediction** helps identify at-risk customers
- **Network optimization** enhances performance and efficiency
- **Customer behavior analysis** provides valuable insights for strategy development

Education Applications

- **Student performance prediction** helps educators identify at-risk students early
- **Dropout risk analysis** provides insights into factors contributing to student attrition
- **Curriculum improvement** focuses on enhancing educational content and delivery methods

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

UNIVERSIDAD

Social Media Applications

- **Sentiment analysis** provides insights into public opinion
- **Trend detection** identifies emerging patterns and shifts
- **Influencer identification** pinpoints key figures driving conversations

Government and Public Sector

- **Crime pattern analysis** helps identify trends and hotspots
- **Resource allocation** ensures efficient use of available assets
- **Policy impact evaluation** assesses the effectiveness of implemented strategies

Manufacturing and Industry

Predictive maintenance helps prevent unexpected equipment failures

Quality control ensures products meet high standards and specifications

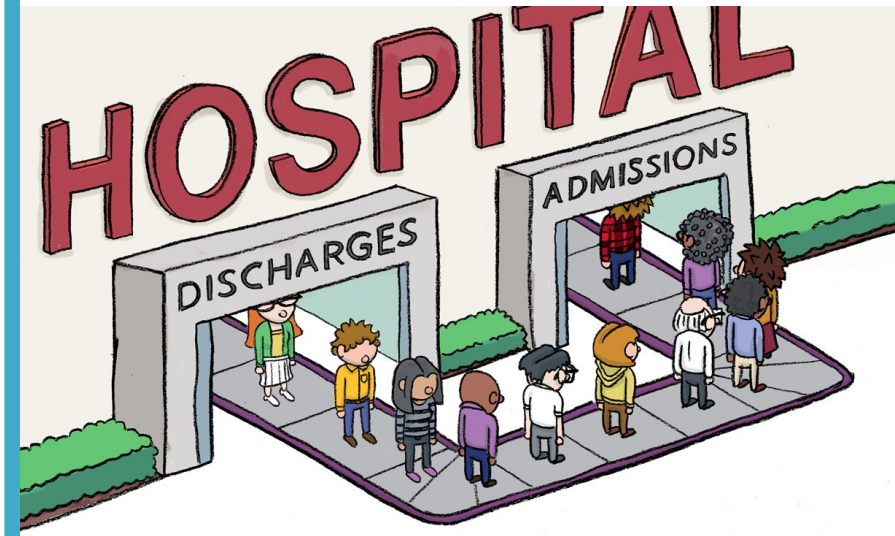
Supply chain optimization streamlines processes for greater efficiency and effectiveness

Real-World Example – Retail

- **Diapers and beer correlation**: This intriguing relationship highlights unexpected consumer behavior
- **Strategic product placement**: Thoughtful positioning can significantly influence purchasing decisions.
- **Increased sales through association rules**: Leveraging these insights can lead to remarkable sales growth.

Real-World:
Example – Healthcare

- **Predicting hospital readmissions** effectively helps in managing patient care
- **Reducing costs and improving care** leads to better healthcare outcomes
- **Based on patient history and treatment data** ensures personalized healthcare solutions

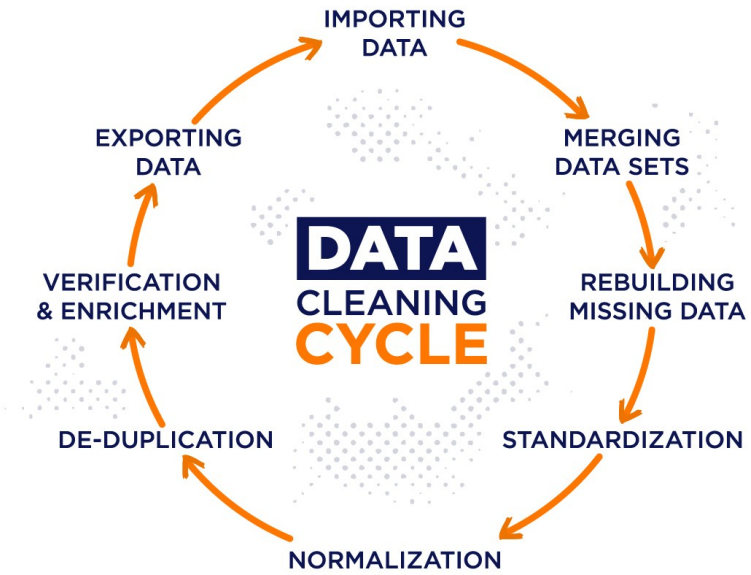


Real-World: Example – Finance

- **Detecting credit card fraud** effectively and efficiently
- **Real-time anomaly detection** helps identify suspicious activities
- **Machine learning models in action** provide powerful insights and predictions

Challenges in Data Mining

- **Data quality** is essential for accurate insights
- **Scalability** ensures systems can grow with demand
- **Privacy and ethics** are crucial for user trust
- **Interpretability** helps users understand complex data
- **Dynamic data** allows for real-time updates and analysis

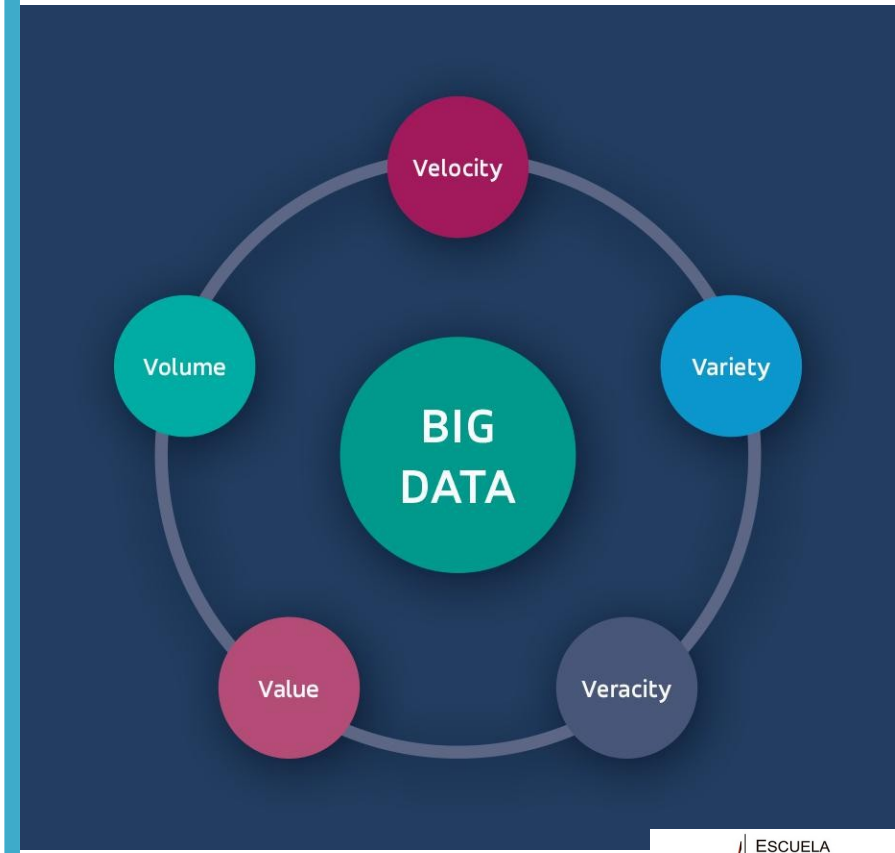


Data Quality Issues

- **Missing values** can lead to inaccurate results
- **Noisy or inconsistent data** complicates analysis and interpretation
- **Data cleaning is essential** for ensuring high-quality datasets and reliable outcomes

Scalability and Big Data

- **Handling large volumes of data**
- **Need for distributed computing**
- **Tools**: Hadoop, Spark

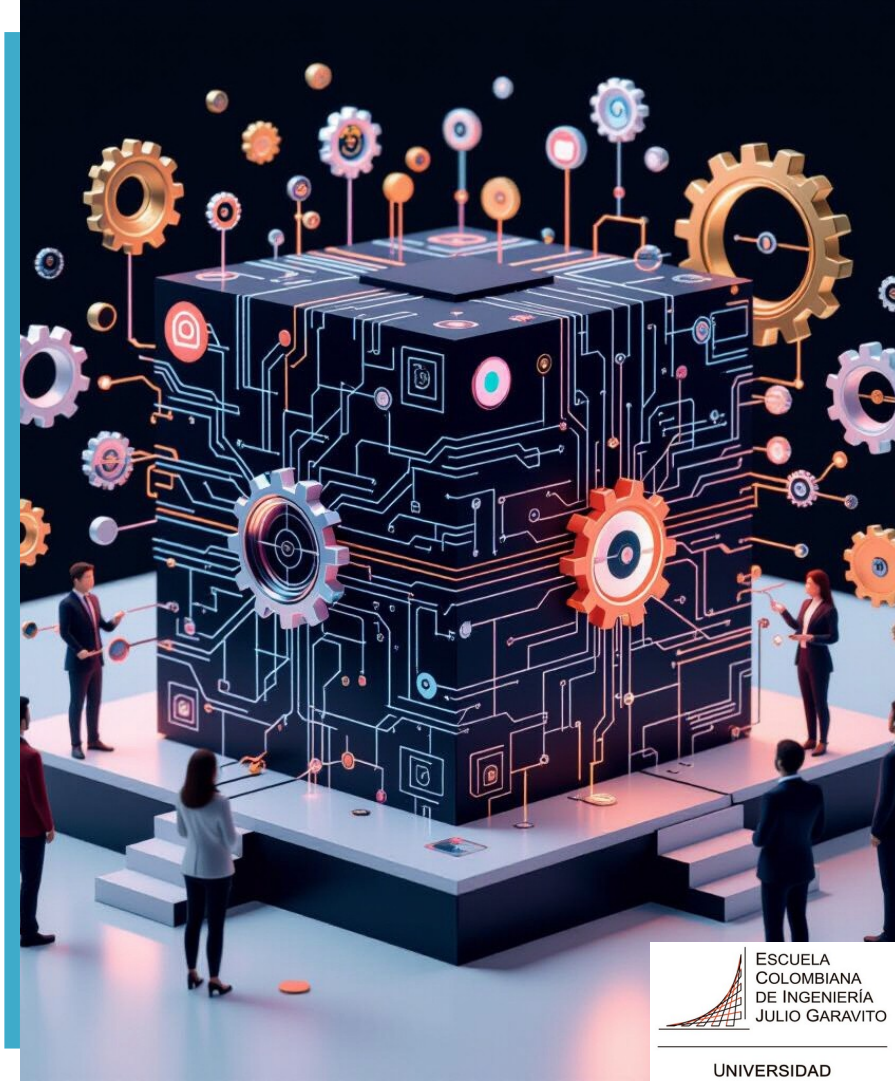


Privacy and Ethics

- **Responsible data use** is essential for maintaining trust in technology.
- **GDPR and data protection laws** ensure user privacy and security.
- **Bias and fairness in algorithms** must be addressed to promote equity in outcomes.

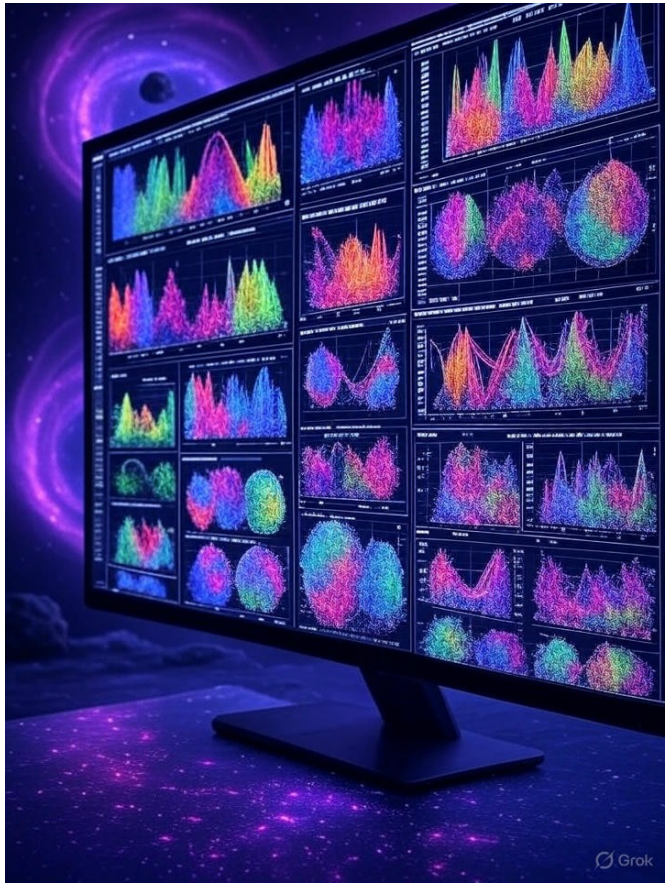
Interpretability and Trust

- Complex models can often be black boxes, making them difficult to interpret
- There is a growing need for explainable AI to enhance transparency
- Stakeholder understanding is absolutely key for successful implementation



Summary

- **Data mining reveals hidden insights** that can significantly enhance decision-making processes.
- **Broad applications across industries** demonstrate its versatility and effectiveness in various fields.
- **Challenges must be addressed** to ensure ethical practices and data integrity in all mining efforts.



INTER NATIONAL SUMMER SCHOOL
UNIVERSIDADE DA CORUÑA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

UNIVERSIDAD