What is Data?

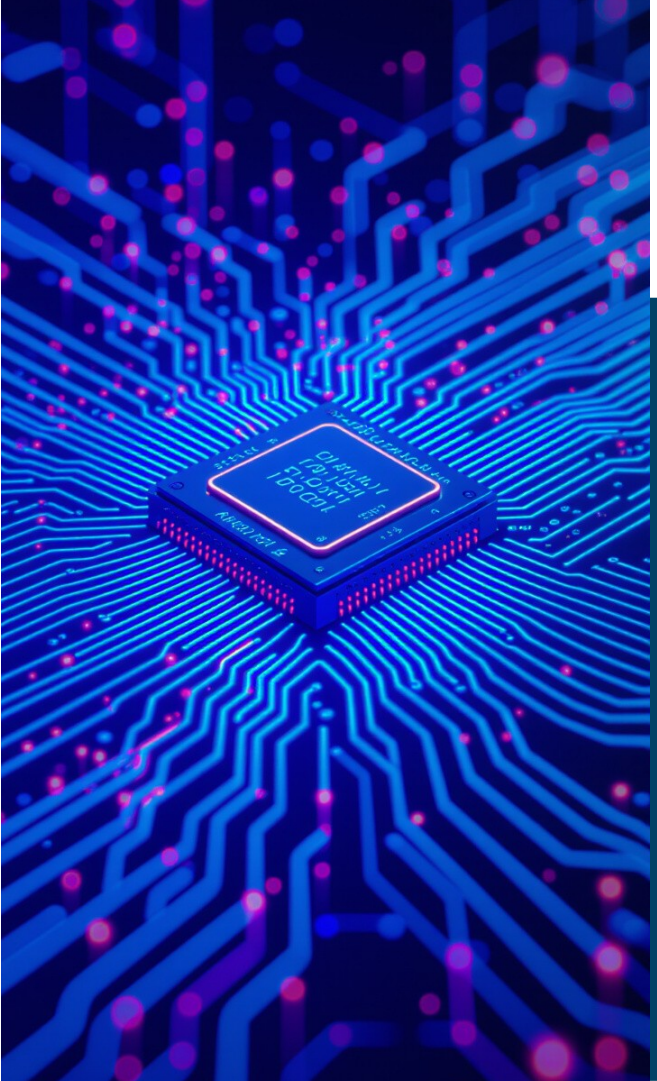
Data Mining & Neural Networks

Dr. Wilmer Garzón

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

Escuela Colombiana de Ingeniería
Universidade da Coruña

2025



What is Data?

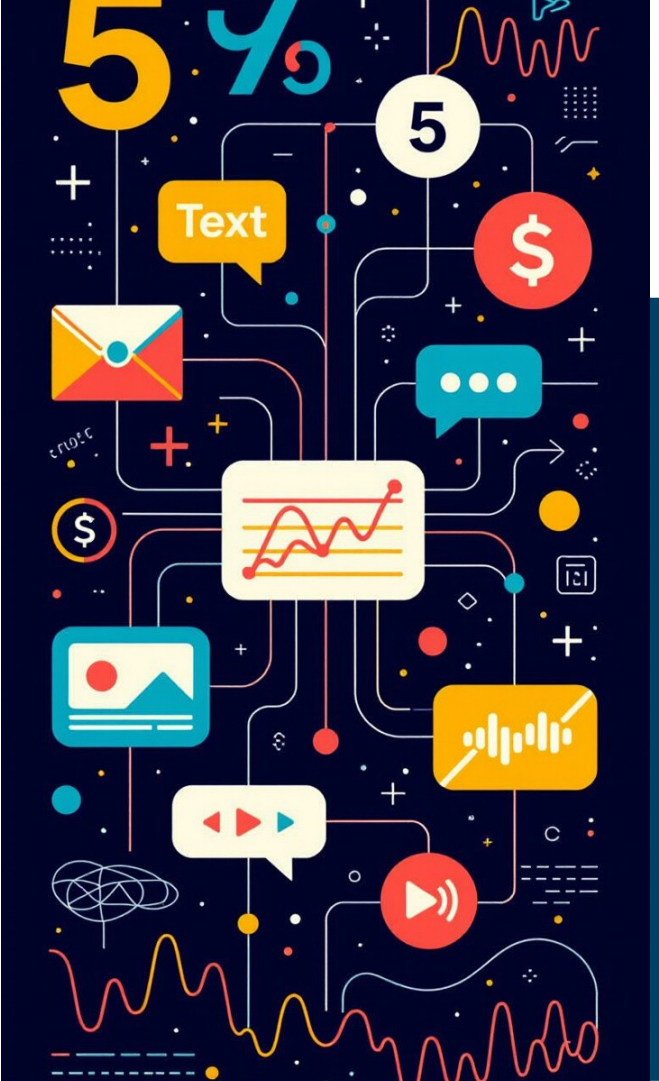
- This presentation introduces the **foundational concept of data**, exploring its definitions, types, and roles in modern computing, especially in data mining and neural networks.
- Understanding what data is and how it is used is the **first step in becoming proficient** in AI, machine learning, and data science disciplines.
- We'll also explore **real-world examples and reflect on how data powers intelligent systems**.

Learning Objectives

By the end of this session, you will be able to:

- Define data and distinguish it from information and knowledge.
- Classify data as structured, semi-structured, or unstructured.
- Identify the key attributes of high-quality data.
- Recognize common sources and representations of data.
- Understand the critical role of data in data mining and neural networks.

Each of these skills is essential for successful work in data science.



What is Data?

- Data refers to unprocessed facts, figures, and symbols that describe objects, events, or conditions.
- It can be numerical, textual, visual, or audio:
  - **42**, **blue**, or an image file are all data
- They have no meaning. It’s only when we analyze and contextualize them that they become useful.
- In AI and ML, data serves as the foundation for learning patterns and making predictions.

From Data to Wisdom – The DIKW Pyramid

The DIKW pyramid outlines the transformation of data:

- **Data**: Raw values, e.g., “35°C”
- **Information**: Contextualized data, e.g., “Today’s temperature is 35°C”
- **Knowledge**: Patterns and understanding, e.g., “It’s hotter than usual in July”
- **Wisdom**: Actionable insight, e.g., “Avoid outdoor work in the afternoon”

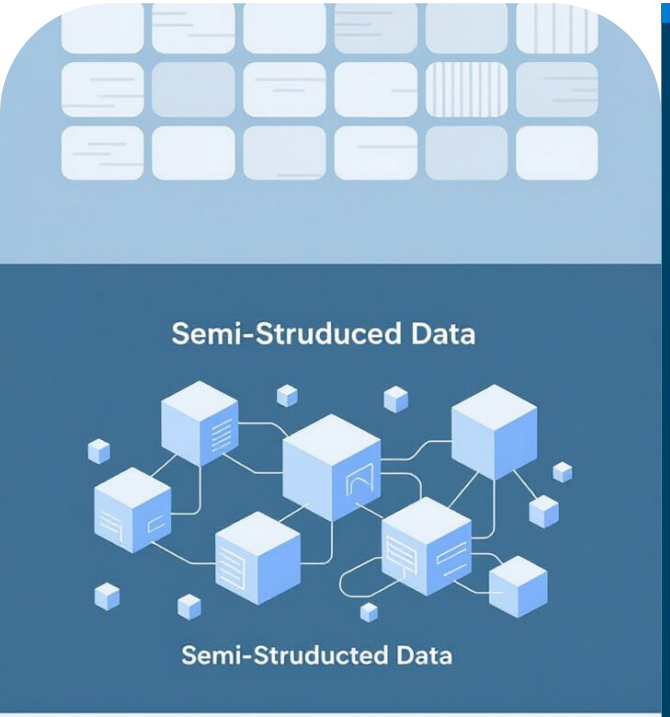
This hierarchy guides how systems convert raw input into valuable decisions.

Real-World Examples of Data

Consider various domains:

- In **healthcare**: Patient vitals (heart rate, blood pressure)
- In **e-commerce**: Purchase history, clicks, and ratings
- In **transportation**: GPS coordinates, traffic data
- In **finance**: Stock prices, transactions, credit scores

Each piece of data, though small on its own, can be combined and processed to reveal valuable insights or automate decisions.

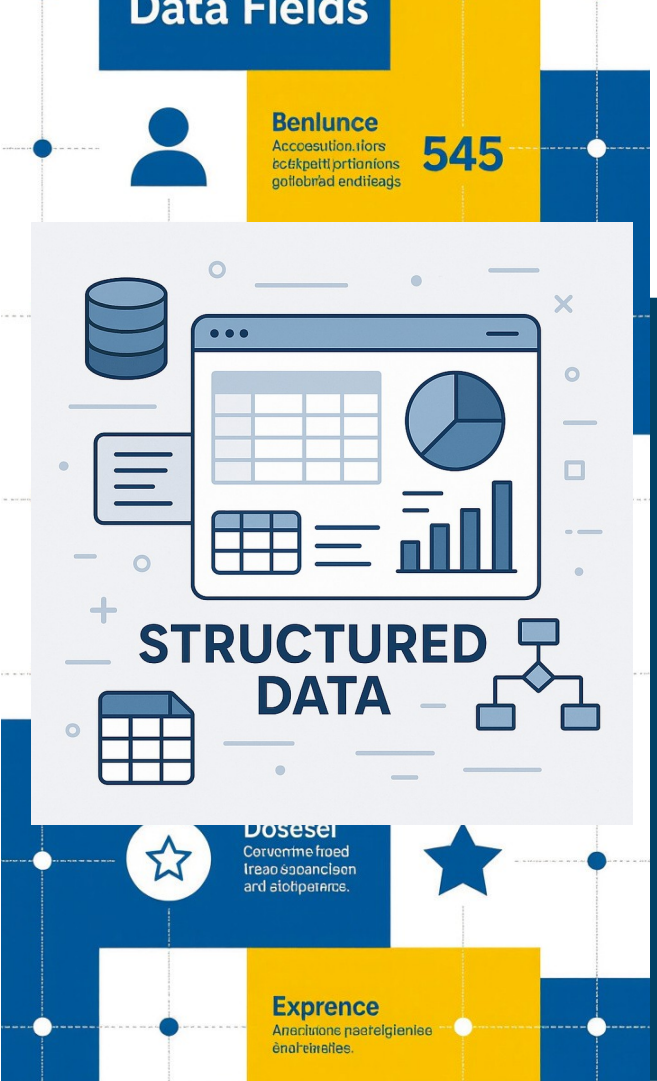


Types of Data

Data is commonly divided into:

- **Structured**: Clearly organized in tables, rows, and columns.
- **Semi-structured**: Partially organized, like XML or JSON.
- **Unstructured**: No specific format, such as images, audio, video.

Understanding these distinctions is crucial because different tools and techniques are needed to store, retrieve, and analyze each type.



Structured Data

Structured data is the easiest to analyze using traditional tools like SQL. It follows a strict schema, enabling fast querying and filtering.

- A student database with fields: name, ID, GPA
- A bank's transactions table

Inventory in a retail system Its predictable format supports relational databases, making it common in business applications.

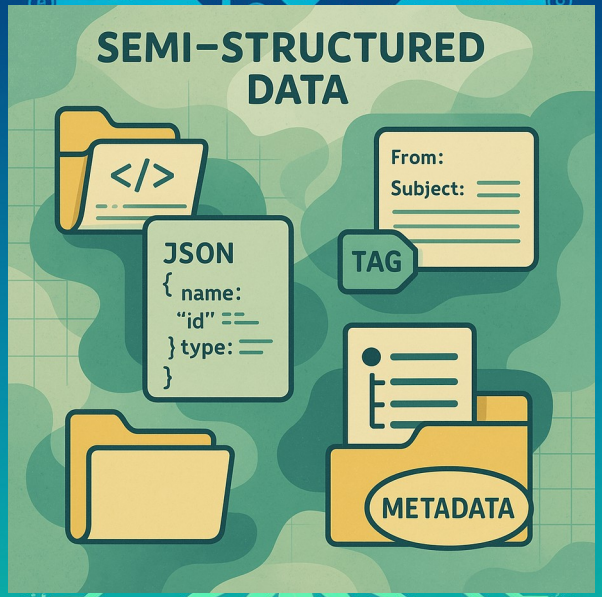
Unstructured Data

Unstructured data lacks a predefined format, making it more difficult to store and analyze.

- Videos and images from social media
- Recorded customer service calls

Text from emails or chat messages Advanced techniques such as natural language processing (NLP) or computer vision are **needed to convert unstructured data into usable forms**.

Semi-structured Data



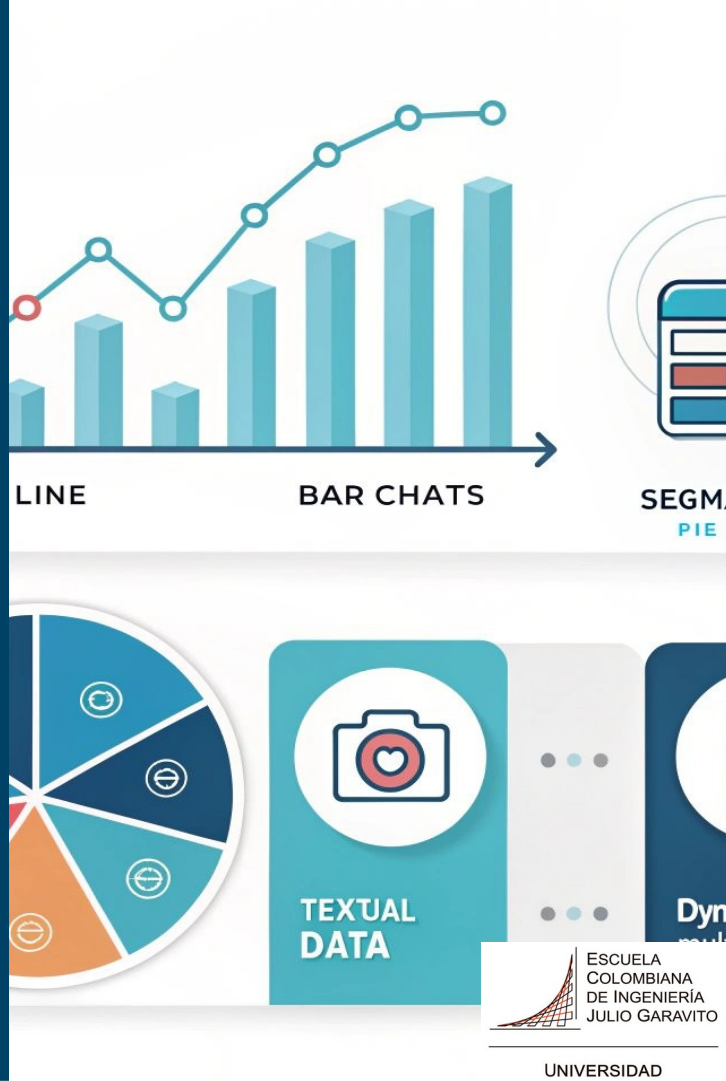
Semi-structured data contains tags or markers but doesn't conform to rigid schemas. It's flexible and commonly used in web and application development.

- JSON objects used in APIs
- XML files for configuration or data exchange This type is especially common in NoSQL databases and systems that require scalability.

Data Types by Value

- **Numerical**: e.g., 98.6°F, 20kg. Enables statistical analysis.
- **Categorical**: e.g., "male", "female", "premium". Useful for grouping.
- **Textual**: e.g., reviews, tweets. Requires NLP to interpret.
- **Multimedia**: e.g., images, audio. Processed with deep learning.

Choosing the correct data type helps in selecting appropriate models and preprocessing techniques in ML.

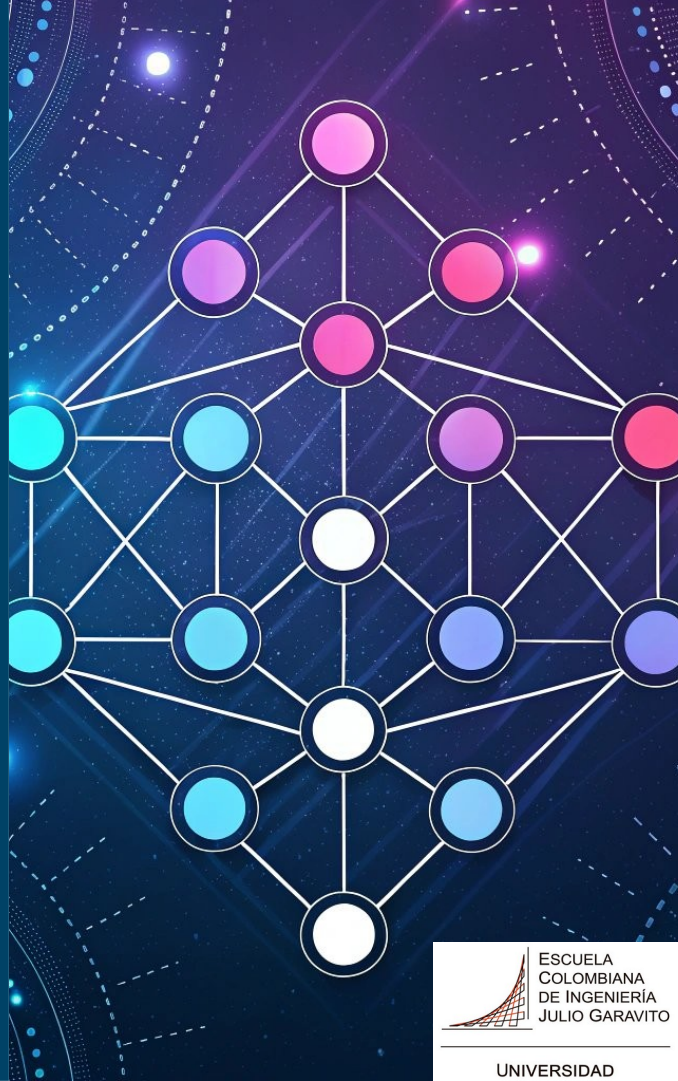


Data in Data Mining

Data mining uses algorithms to **discover patterns in large datasets**. It involves:

- **Classification** (e.g., spam vs. non-spam emails)
- **Clustering** (e.g., customer segmentation)
- **Association rule mining** (e.g., "people who buy X also buy Y")

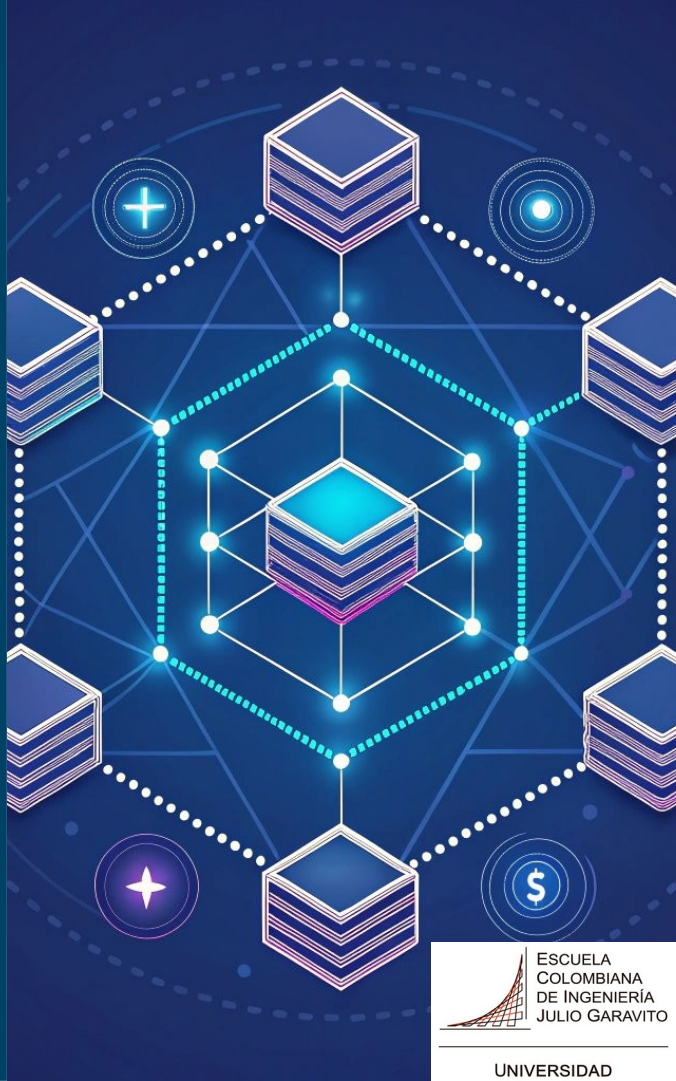
High-quality and well-prepared data enables these algorithms to extract insights efficiently, which can then support decision-making or predictions.



Data in Neural Networks

Neural networks simulate the **brain to learn patterns from data.** They require large, labeled datasets for training, such as:

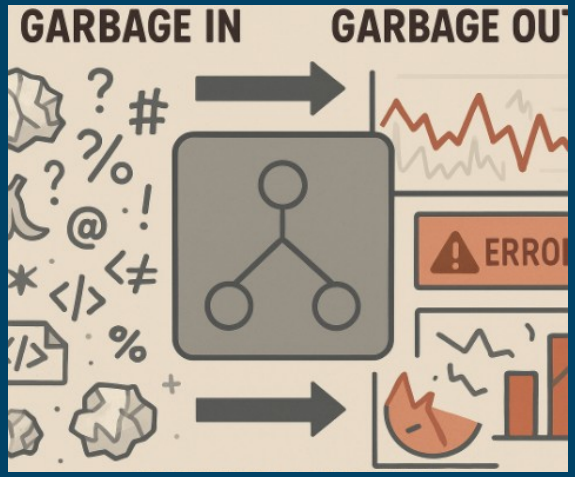
- **ImageNet** for image classification
- **IMDB** for sentiment analysis Each input must be converted into numerical form. Data normalization, encoding, and augmentation are critical steps to ensure convergence and model accuracy.



Importance of Data Quality

Poor-quality data leads to inaccurate models and **bad decisions**. Five key attributes define good data:

- **Accuracy**: Is the data correct?
- **Completeness**: Are values missing?
- **Consistency**: Are values standardized?
- **Timeliness**: Is the data up-to-date?
- **Validity**: Does it follow expected rules?



Common Data Sources

Data comes from multiple domains:

- **Business systems**: Sales, HR, CRM tools
- **Sensors**: IoT, weather stations
- **Web activity**: Search logs, clicks
- **Public datasets**: Kaggle, UCI ML Repository
- **Social platforms**: Twitter API, Reddit datasets

Choosing a reliable and relevant source ensures that models receive meaningful inputs.

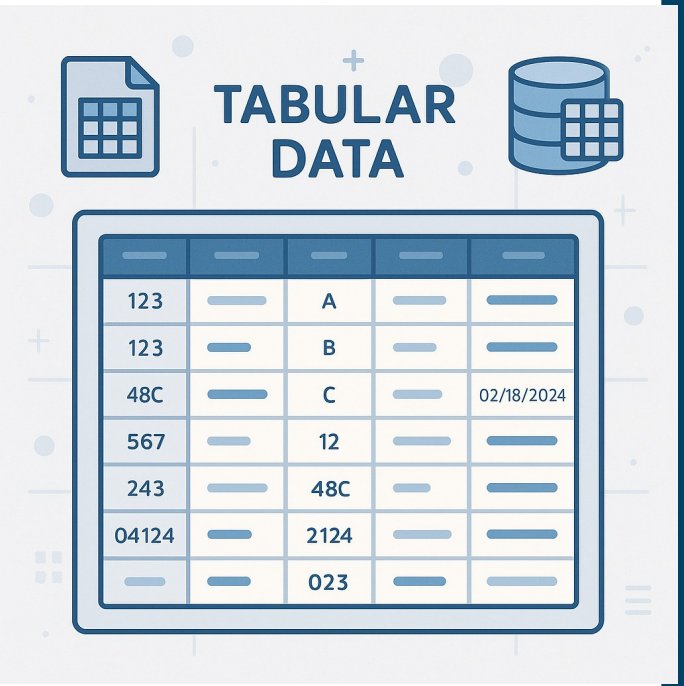
Data Representation

Depending on the application, data must be represented differently: Tabular

- Graphs
- Vectors
- Time series
- Pixel arrays

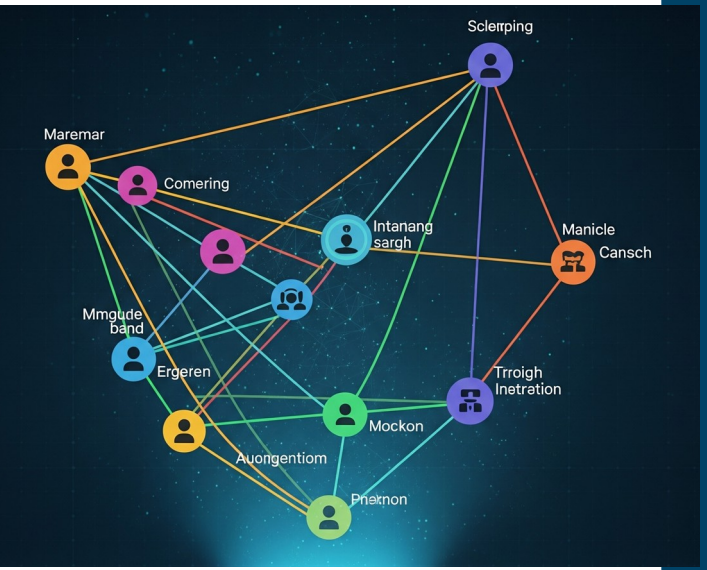
Each representation affects the choice of algorithm and preprocessing pipeline.

Data Representation - Tabular

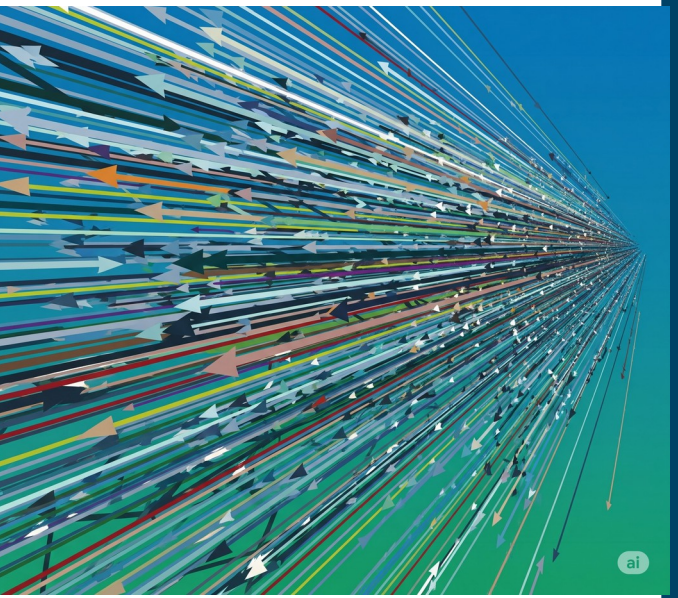


- Structured data organized in rows and columns
- Each row is a record; each column is a feature
- Common formats: CSV, Excel, SQL tables
- Easy to store, query, and analyze
- Used in most business and scientific applications
- Easy for statistical analysis

Data Representation - Graphs

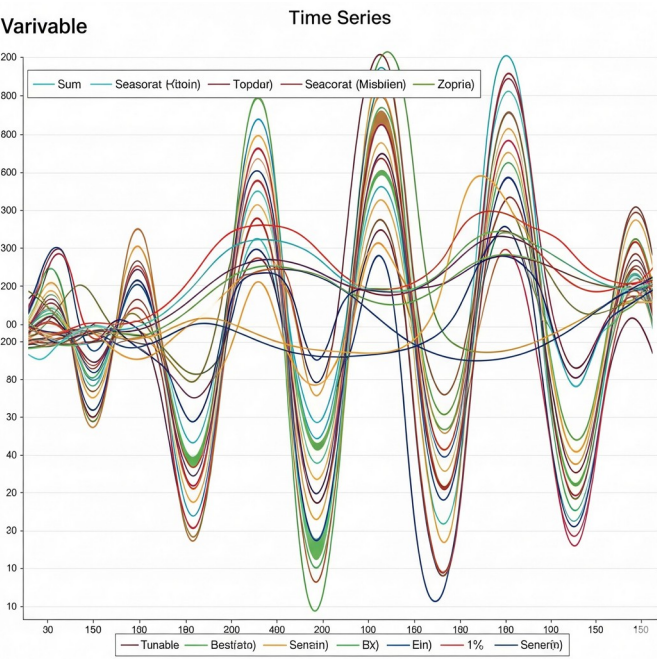


- Represents entities (nodes) and their relationships (edges)
- Ideal for modeling social networks, recommendation systems, knowledge graphs
- Nodes = objects (e.g., users, products)
- Edges = connections (e.g., friendships, purchases)
- Graph databases: Neo4j, Amazon Neptune
- Enables traversal and pattern discovery in connected data



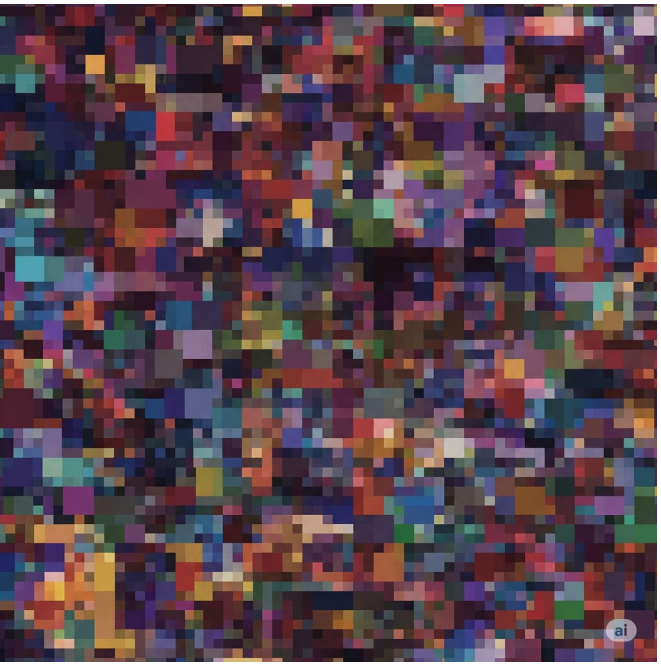
Data Representation - Vectors

- A vector is an ordered list of numerical values
- Represents data in a format suitable for machine learning models
- Each element is a feature (e.g., height, weight, age)
- Used in text embeddings, image representations, user profiles
- Vectors enable computations: distance, similarity, transformations
- Common in algorithms like k-NN, SVM, neural networks
- Used in machine learning and deep learning



Data Representation - Time Series

- A time series is a sequence of data points indexed in time order
- Each observation is associated with a specific timestamp
- Common in finance, weather forecasting, sensor monitoring
- Key properties: trend, seasonality, noise, stationarity
- Used in models like ARIMA, LSTM, Prophet
- Visualization: line charts over time for pattern detection



Data Representation - Pixel Arrays

- Images are represented as pixel arrays, typically 2D or 3D matrices
- Each pixel contains intensity values, e.g., grayscale or RGB channels
- A grayscale image uses one channel (0–255)
- Pixel arrays are input data for computer vision models
- Used in CNNs for image classification, detection, and segmentation

Big Data and the 5 V’s

Big Data is characterized by:

- **Volume**: Terabytes to petabytes
- **Velocity**: Real-time data streams
- **Variety**: Diverse data formats and types
- **Veracity**: Reliability of sources
- **Value**: Insights that drive actions

These dimensions define modern data challenges and opportunities in analytics and AI.

How Data is Collected

Methods include:

- **Manual**: Surveys, forms
- **Automated**: IoT devices, log files
- **Digital tracking**: Clickstreams, session logs
- **External**: APIs, third-party platforms
- **Scraping**: Crawling web content

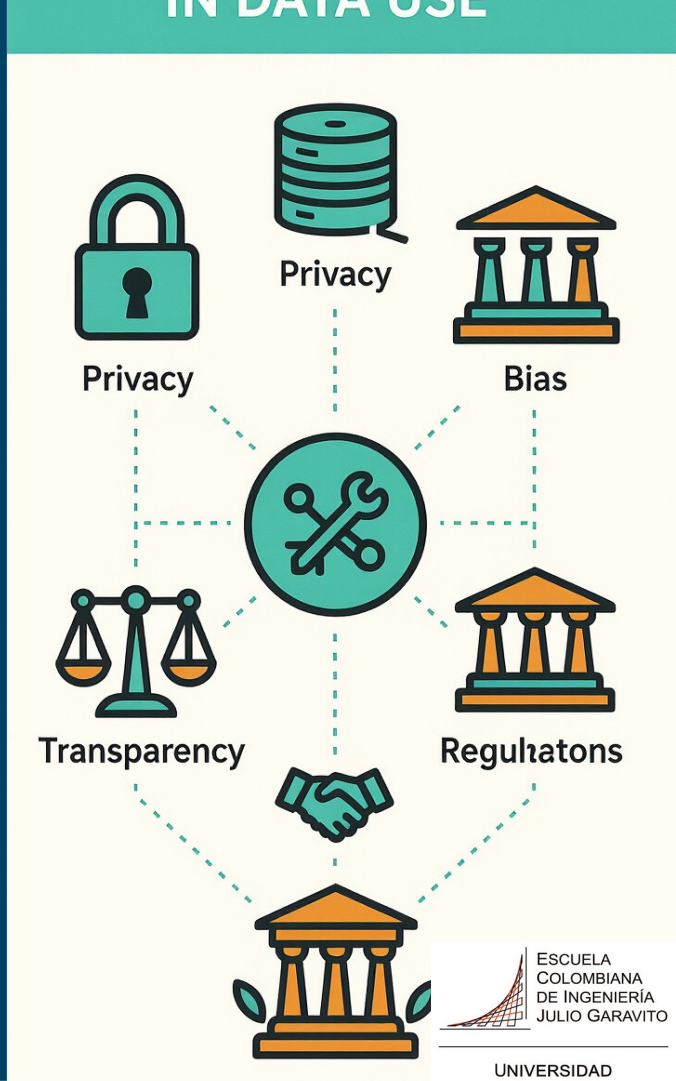
Each method has pros and cons in terms of scale, accuracy, and ethics. Understanding them is essential to managing datasets effectively.

Ethical Considerations

Data use must respect legal and moral boundaries:

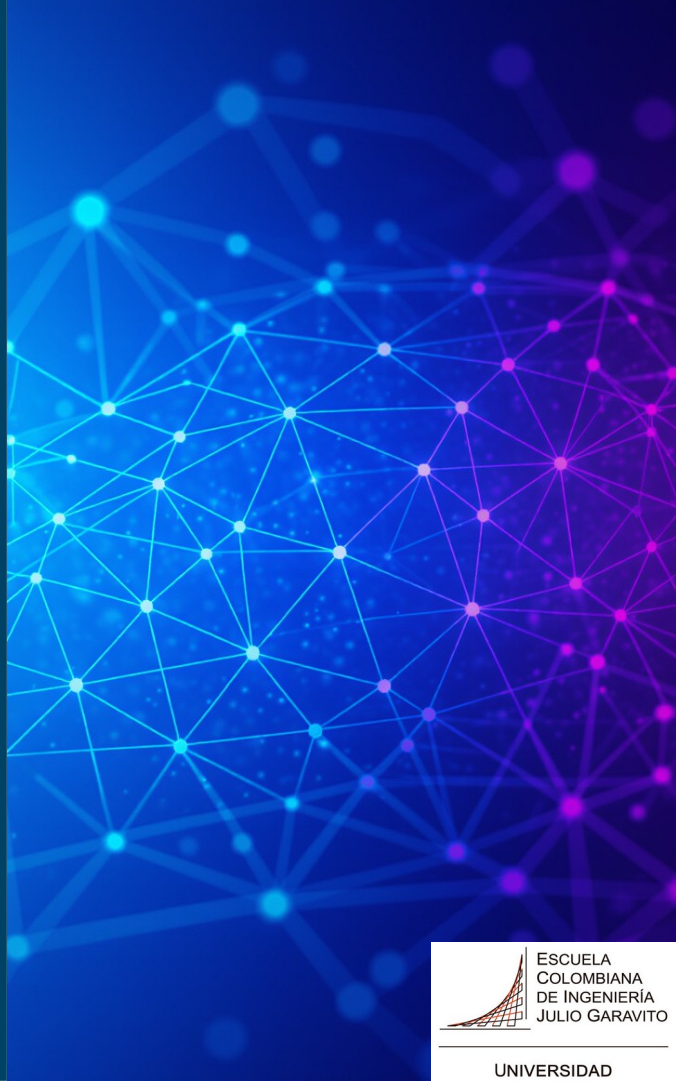
- **Privacy**: Use only data obtained with consent.
- **Bias**: Ensure datasets represent diverse populations.
- **Transparency**: Explain how data is collected and used.
- **Regulations**: Follow rules like **GDPR** or **HIPAA**.

Ethical practices prevent harm, build trust, and ensure fairness in algorithmic systems.



Summary and Key Takeaways

- Data is foundational to all modern intelligent systems.
- Understand what kind of data you are working with and how to represent it.
- The quality of your data influences the accuracy of your results.
- Data mining and neural networks rely heavily on clean, relevant data.
- Always consider the ethical implications of data use.



What’s Next? Data Preprocessing

Next class we’ll focus on how to prepare data:

- **Cleaning**: Remove duplicates, fix errors
- **Transformation**: Normalize, scale, encode
- **Feature engineering**: Extract meaningful attributes
- **Handling missing data**: Imputation or exclusion

Proper preprocessing is key to successful modeling and system performance.

INTER NATIONAL SUMMER SCHOOL
UNIVERSIDADE DA CORUÑA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

UNIVERSIDAD