Minería de Datos

U1 Introducción

Héctor Maravillo

Section 1

Data Mining

Data Mining

Data Mining is defined as the process of discovering **patterns** in data [Witten & Frank, 2005].

Data mining is the **use of machine learning and statistical analysis** to uncover patterns and other valuable information from **large data sets** [IBM].

Data mining is the non-trivial extraction of implicit previously unknown, and potentially **useful information** from data [Frawley et al., 1991]

Data and patterns

Data are set of facts (such as numbers, words, measurements, observations or just descriptions of things), and **pattern** is an expression to **describe a subset** of the data or a **model** applicable to the subset.

Extracting a pattern also designates **fitting a model** to the data; **finding structure** from data; or, in general, making any **high-level description** of a set of data.

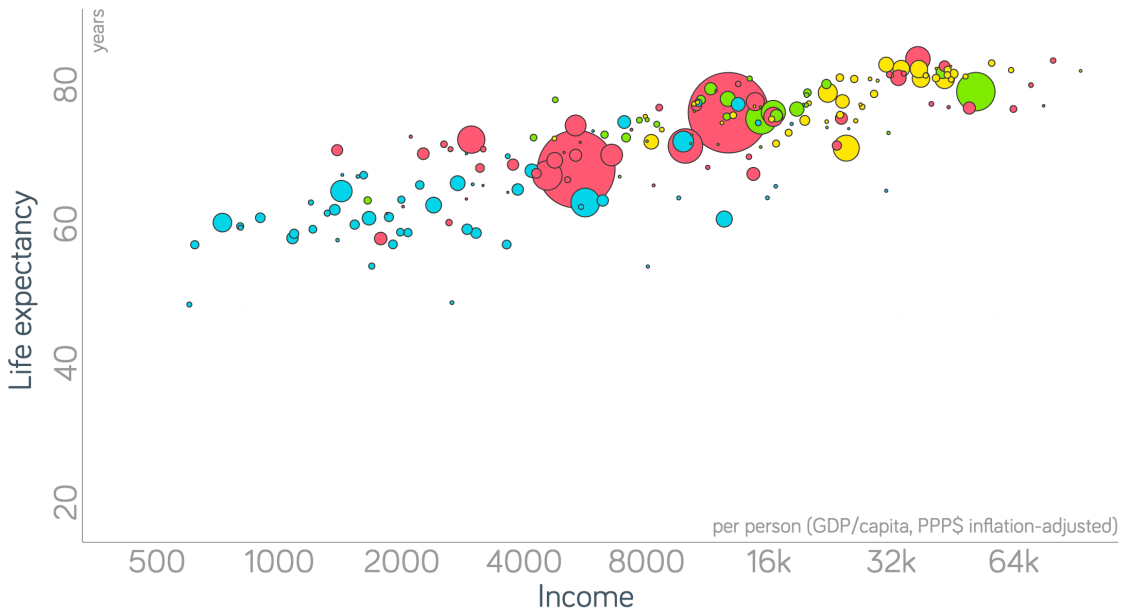

Figure: Example of data and patterns

Data Mining

Historically, the notion of **finding useful patterns** in data has been given a variety of names, including:

- Data Mining
- Knowledge extraction
- Information discovery
- Data archaeology
- Data pattern processing

Many people treat data mining as a synonym for another popularly used term, **Knowledge Discovery from Data**, or **KDD**. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.

Knowledge Discovery in Databases (KDD)

- The term **data mining** has mostly been used by statisticians, data analyst, and the management information systems communities.
- The concept **Knowledge Discovery in Databases (KDD)** was coined at the first KDD workshop in 1989 by Piatetsky-Shapiro in 1991. It has been popularized in the AI and machine-learning fields.

Knowledge Discovery in Databases (KDD)

KDD refers to the **overall process** of discovering useful knowledge from data, while **data mining** is only a **particular step** in this process. Data Mining is the **application** of specific algorithms for extracting patterns from data.

The basic problem addressed by the KDD process is one of mapping **low-level data** (which are typically too voluminous to understand and digest easily) into other form that might be more **compact** (for example, a short report), more **abstract** (for instance, a model), or more **useful** (for example, a predictive model for estimating the value of future cases).

Knowledge Discovery in Databases (KDD)

Why do we need KDD?

- The traditional method of turning data into **knowledge** relies on **manual** analysis and interpretation. It consists fundamentally of one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products.
- As data volumes **grow** dramatically, this type of manual data analysis is becoming completely **impractical** in many domains. Databases are increasing in size in two ways:
  - The number N of records or objects in the database.
  - The number d of fields or attributes to an object.

The KDD Process

The **process** of KDD consists of an iterative sequence of the following steps:

1. **Data integration**: where multiple data source may by combined.
2. **Data cleaning**: To remove noise and inconsistent data.
3. **Data selection**: where data relevant to the analysis task are retrieved from the database.
4. **Data transformation**: where data are transformed or consolidated into forms appropriate for mining.
5. **Data mining**.
6. **Pattern evaluation**.
7. **Knowledge presentation**.

Steps 1 to 4 are different forms of **data preprocessing**, where the data are prepared for mining.

The KDD Process

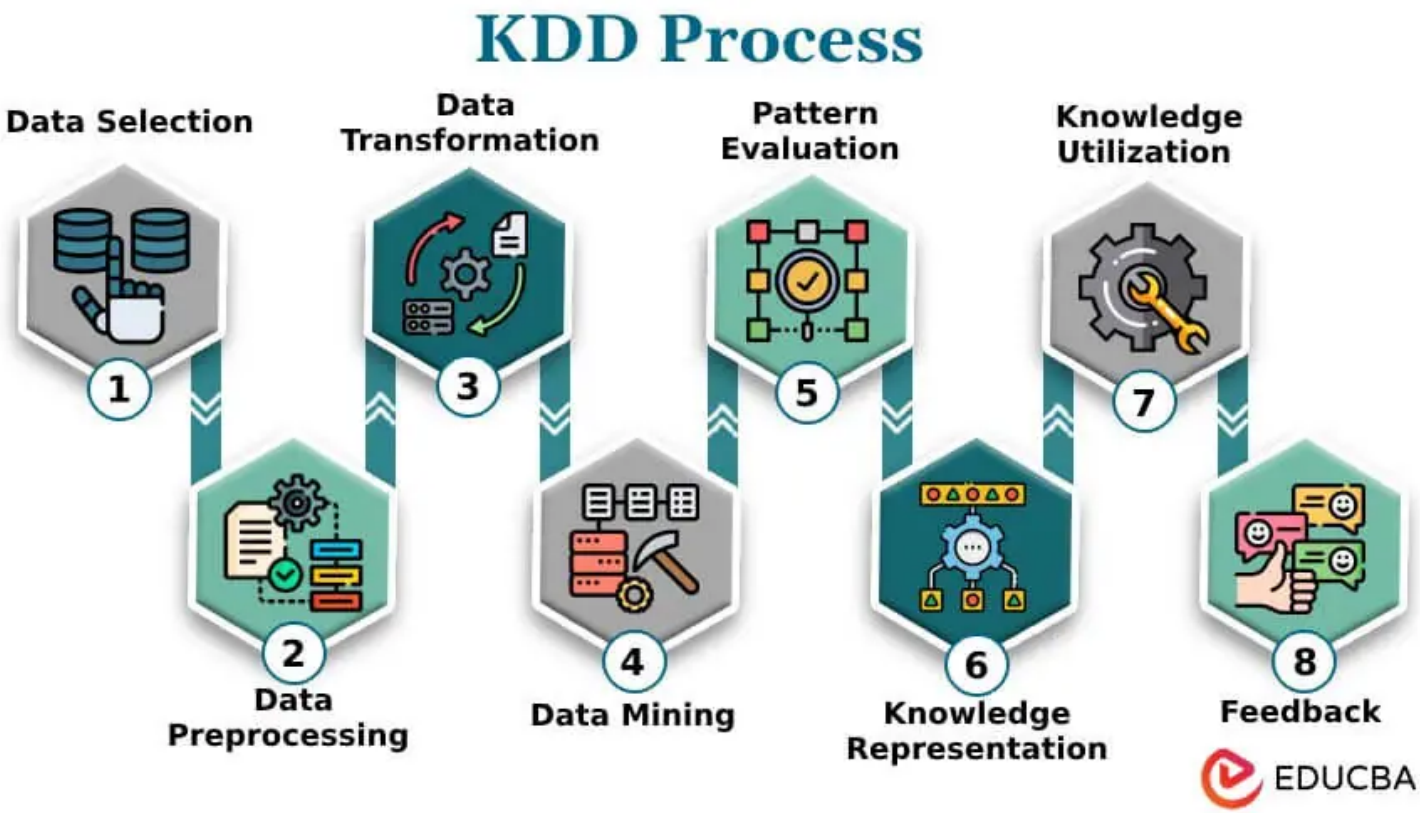

Figure: Stages of Knowledge Discovery

Data-mining goals

We can distinguish two types **goals**:

- **Verification**: The system is limited to verifying the user's hypothesis.
- **Discovery**: The system autonomously finds new patterns. This goal can be subdivided into:
  - **Prediction**: where the system finds patterns for predicting the future behavior of some entities.
  - **Description**: where the system finds patterns for presentation to a user in human-understandable form.

Data Mining methods

- **Classification**: is learning a function that **maps** (classifies) a data item into one of several **predefined classes**.

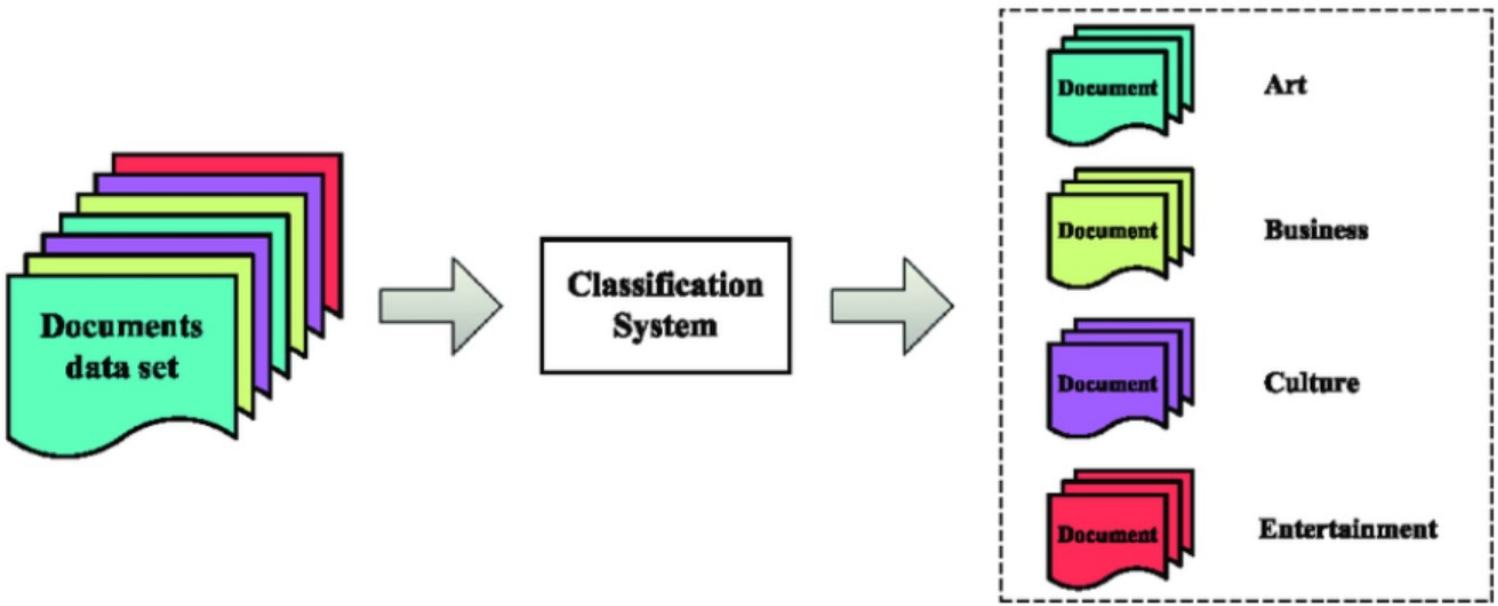

Figure: Example of classification: spam detection

Data Mining methods

- **Regression** is a learning a function that maps a data item to a **real-valued** prediction variable.

Figure: Example of regression: financial forecasts

Data Mining methods

- **Dependency modeling** consists of **finding a model** that describes significant **dependencies** between variables.

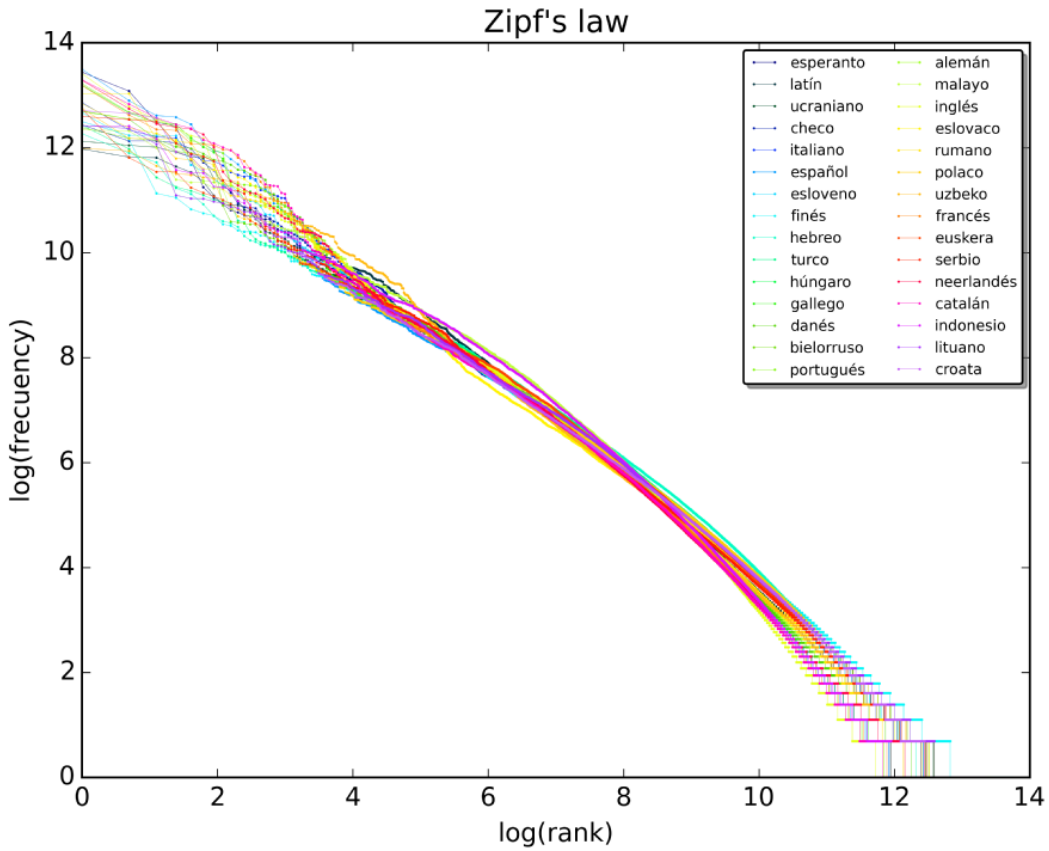

Figure: Example of dependency modeling: power laws in languages.

Data-Mining methods

- **Summarization** involves methods for finding a **compact description** for a subset of data (for instance: summary rules, multivariate visualization techniques, and the discovery of functional relationships between variables).

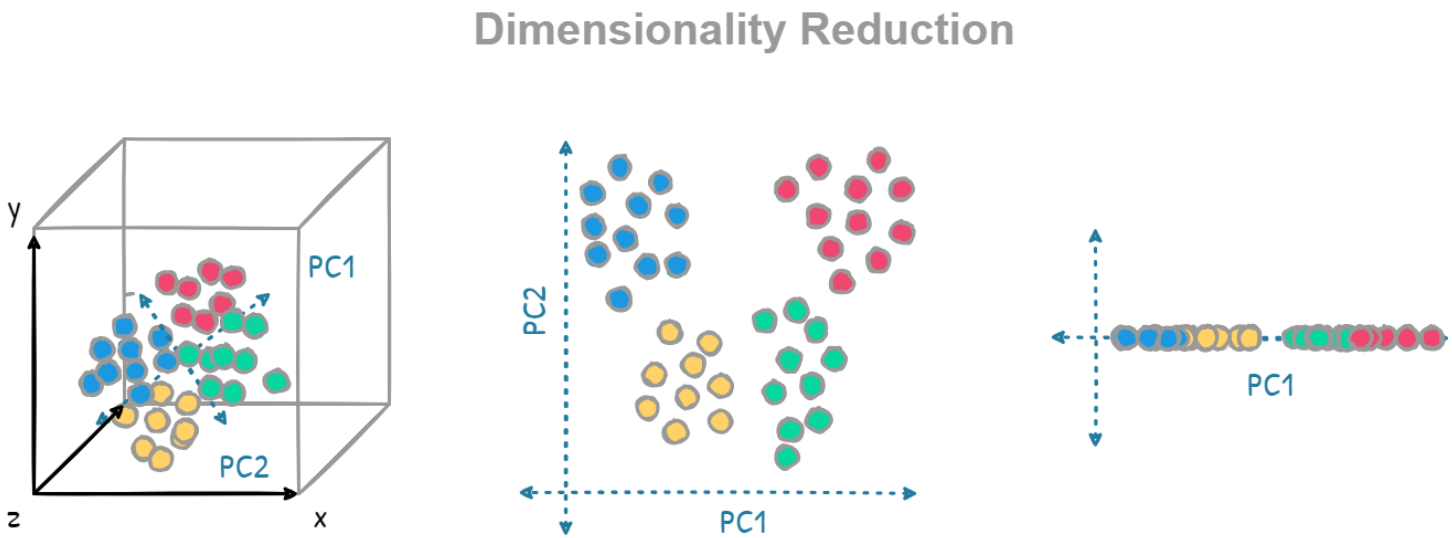

Figure: Example of summarization: data visualization and dimensionality reduction.

Data Mining methods

- **Clustering** is a common descriptive task where one seeks to **identify** a finite set of categories or clusters to describe the data.

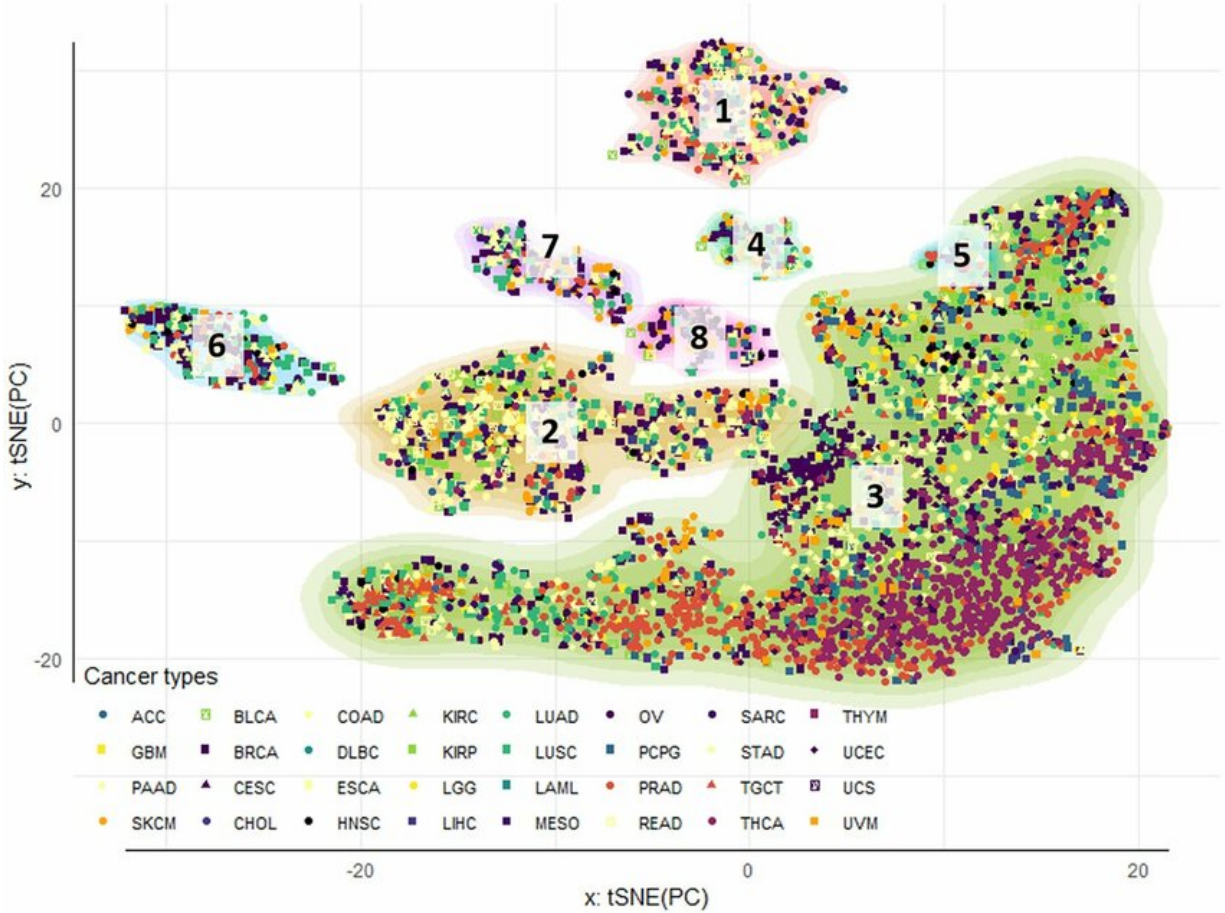

Figure: Example of clustering: cancer detection

Major issues in Data Mining

1. Mining of diverse input **sources**.
2. Incorporation of **background information**.
3. Presentation and **visualization** of data mining results.
4. Handling **noisy** or incomplete data.
5. **Efficiency** and scalability of data mining algorithms.
6. Parallel, distributed, and incremental mining algorithms.

Section 2

References

References

- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996) From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, 17(3): 37–54.
- Han, J. & Kamber, M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006.
- Witten, I. & Frank, E. *Data Mining. Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann Publishers, 2005.