# 1. Introduction

## OLTP - On Line Transaction Processing

Traditional DBMS usage.
These DBs are also called operational DB since they are mainly used in companies
to record/store several stuff(details of certain products etc).
They are mostly relational DB and they tipically have read/write operation that access not too many records.
Characterized also by short transactions which they abide by the ACID properties.
Lastly, the database size is around the megabytes/gigabytes.

## OLAP - On Line Analytical Processing

The main idea behind it it's to analyze the whole data rather than answer specific questions. For example, in a grocery store, we'd want to identify the most critical sector such as the one with the lowest income in a given year, it is thus used for **historical data**, often collected by **operational databases**.
Characterized by access of **million on records** at the same time.
Queries are also more **complex**.
**ACID** properties are less critical.
Database sizes of **terabytes** called **Data warehouse**.

## Data science and Big data

Data whose scale, diversity and complexity require new architectures, techniques, algortihms and alaytics to manage it and extract value and hidden knowledge from it.

## The Vs of Big data

Five properties characterizing big data.

- **Volume:** big data deals with huge data collection, huge *volume*. Those are data regarding social medias, weather forecasting and etc
- **Velocity:** velocity in data production(high data generation rate) and data processing. It used thus for (almost) real time processing.
- **Variety:** Various formats types and structures for data integrated together
- **Veracity**: Data must be of high quality to guarantee a reliable extraction/analysis
- **Value**: Data must be transformed into a *valuable* information. It's supported by domains expert.

## Data science process

1. **Data generation**
2. **Data acquisition**
3. **Data storage**
4. **Data analysis:**
   - Objectives: descriptive/predictive/prescriptive analysis
   - Methods: Statistical analysis, **ML**, **data mining**, text mining, network/graph data mining

## Machine learning and data mining

Non trivial extraction of **implicit**, **previously unknown** and **potentially useful** information from **available data**.
The extraction is automated by certain algorithms.

## Association Rules

**Goal:** extraction of frequent correlations/pattern from a transactional database.

For example, the following *association rule* in a given dataset:
$diapers \rightarrow beer$
where 2% of the transactions contains both items BUT
30% of the transactions containing diapers also container beer.
This is thus an unexpected hidden information that has been extracted from the dataset and that can be used for strategies in companies.

## Classification

**Goal:** prediction of a class label(category) and the definition of an interpretable model.

Training data,which is already labeled/categorized, is used to make a classifier.
This classifier is then utilized to label unclassified data.
This is an example of **supervised** learning

## Clustering

**Goal:** detecting group of similar data objects and exceptions/outliers

Clusters are a collection of data objects with similar aspects.
They can be constructed through **unsupervised** learning.