Alma Mater Studiorum University of Bologna

Artificial Intelligence

Machine Learning and Data Mining (MLDM)

Course Notes

Author: Simone Reale

Academic Year 2023/2024

Contents

Introduction and main concepts

This course is organised into two big topics which we will go deeper on later. These topics are:

- **Data Mining**, in which we will focus on the *data* side, studying the enabling technologies which have been developed for other purposes, but can positively influence the success of data mining processes;
- **Machine Learning**, in which we will focus on the techniques that support *data–driven decisions*, including learning models and algorithms which allow to extract actionable patterns from data.

First of all we have to say that *Data*, *Data Mining* and *Machine Learning* have many differences. While the last two techniques share different concepts, Data exist independently from them, but they need Machine Learning and Data Mining to infer interesting and actionable insights. Especially in the last years these techniques have acquired much more importance with the diffusion of Big Data.

Now we focus on some concepts that represent the basis of all the things that will come after:

Data: a collection of raw value elements;

Information: the result of collecting, interpreting and organising data (relationships between items, context and meaning)

Knowledge: putting together information in order to recognise patterns, according to the needs of the system;

Insight: all the things that we can infer from knowledge on the basis of what is our goal.

When an event in the *real world* changes the state of the enterprise, one of the events below happens:

- a *transaction* is executed, i.e. a business event that changes or modifies data stored in an information system (*database*);
- a *signal* is collected by the infrastructure and stored, which is the reading of a measure provided by a sensor.

The concept of **business** in this case is referred to a **business process**, i.e. a set of activities that will achieve an *organization goal*, once completed. Outside these two events data can also come from *external subjects*.

OLTP (On-Line Transaction Processing) It is a class of software programs capable of supporting transaction-oriented applications and data storage. It is designed to record the daily routine transactions necessary to run the business.

ERP (Enterprise Resource Planning) It is a software system that integrates and manages key business processes of all departments within a single software product into a single, unified platform. It operates in or near real time and provide a common database, which supports all the applications.

MIS (Management Information Systems) They are standardized reporting systems built on existing OLTP, which work by gathering, processing, managing and delivering important information to support decision-making and management activities within an organization. It is used in working envirnments to generate performance indicators of any type.

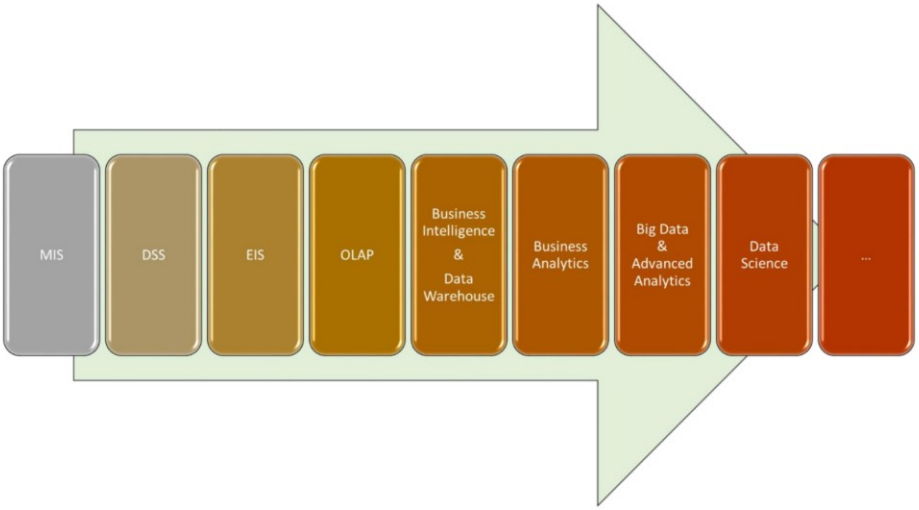


Figure 1.1: Increasing insights

DSS (Decision Support Systems) They are interactive and computerized information systems that aid individuals and organizations in decision-making by providing data analysis, modeling, and decision-making tools to support a wide range of complex and unstructured decisions.

EIS (Executive Information Systems) They are computer-based information systems tailored for senior executives, offering a user-friendly interface that provides consolidated and summarized information from various data sources to support strategic decision-making and organizational performance monitoring.

OLAP (On-Line Analytical Processing) It is a computer-based approach that allows users to interact with and analyze data from multiple dimensions, providing a flexible and intuitive way to explore, query, and report on large datasets. It uses algorithms and data structures specifically designed to ease operations like selections, projections, column exchanges, etc.

BI (Business Intelligence) It is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making (Forrester Research).

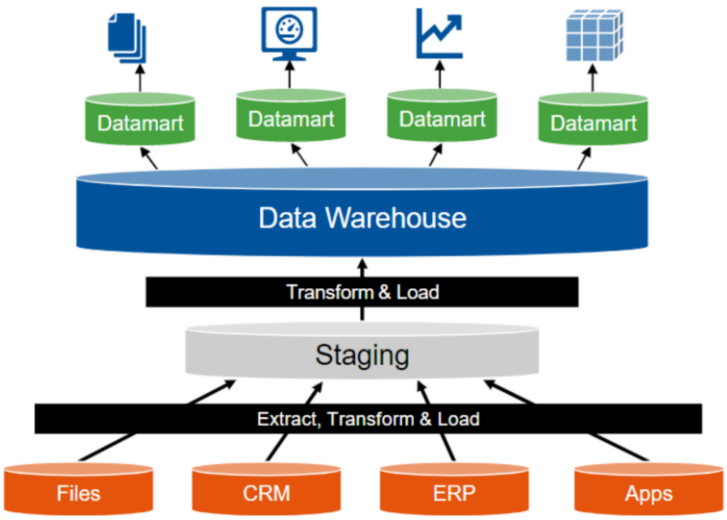


Figure 1.2: BI architecture

Analytics Structured decisions driven by data and there are different types:

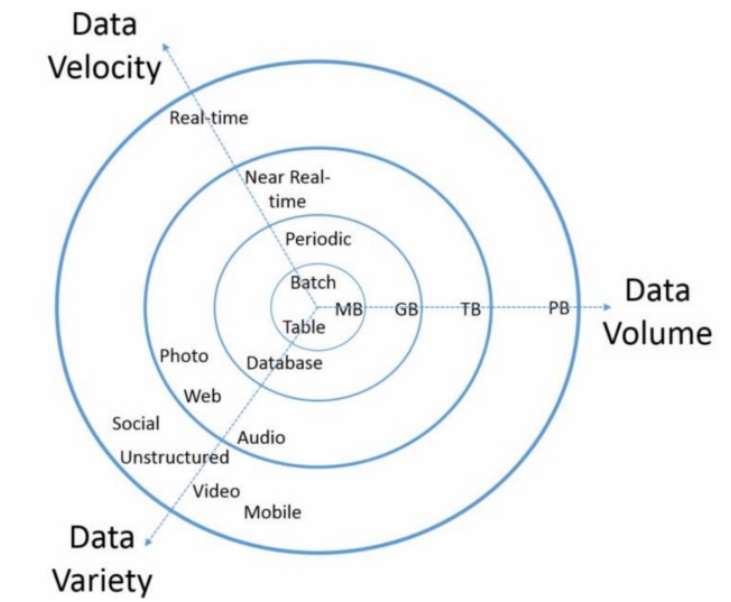
- **descriptive** - understand data
- **diagnostic** - understand causes
- **predictive** - calculate the most probable value of a variable in a future time, given the history of a set (sequence) of variables
- **prescriptive** - suggest actions to be taken to obtain the desired effect.

Cloud Computing Cloud computing refers to the delivery of computing services, including storage, processing power, and applications, over the internet.

It is typically categorized into three main service models:

- **Software as a Service (SaaS):** Delivers software applications over the internet on a subscription basis. Users can access and use software applications running on providers' cloud infrastructure.
- **Platform as a Service (PaaS):** Offers a platform that allows users to develop, run, and manage applications without dealing with the complexities of infrastructure. Then applications are deployed on the provider's cloud infrastructure.
- **Infrastructure as a Service (IaaS):** Consumer can use computing resources within provider's infrastructure upon which they can deploy and run arbitrary software, including OS and applications.

Big Data Big Data are a collection of data sets so large and/or complex and/or fast changing that they are difficult to process using traditional DBMSs or traditional data processing applications.



Their taxonomy is divided into data that are:

- **Structured:** relational tables, spreadsheet (or data which could easily fit in them);
- **Unstructured:** does not have an associated data model (video, audio, pictures, etc.);
- **Semi-structured:** there is some structure, perhaps data refer to different structures (XML, JSON, etc.).

Data Mining

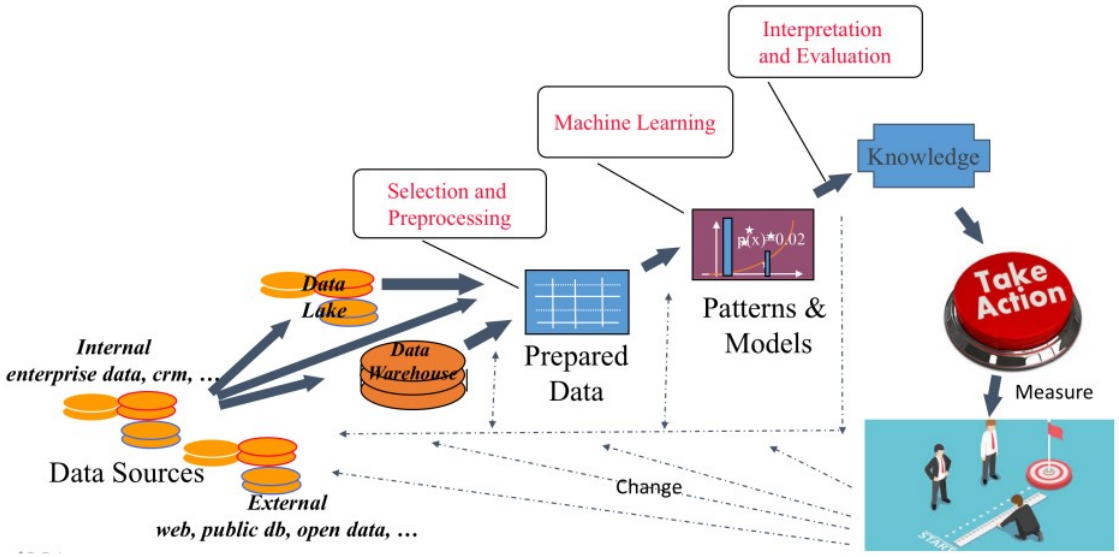


Figure 2.1: The Data Mining Process

Data Mining is the process of discovering patterns, trends, correlations, or meaningful information from large datasets. It involves the application of various techniques from *statistics*, *machine learning*, and *database systems* to extract valuable knowledge from raw data.

Business Intelligence and Data Warehouses

Business Intelligence (BI) represents a key concept in the field of Data Mining and can be described as the process of:

- transforming raw data into useful information to support effective and aware business strategies
- capturing the business data and getting the right information to the right *people*, at the right *time*, through the right *channel*.

There are different definition that has been provided during the years, but there are two of them in particular that we can highlight:

> *Business intelligence (BI) is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.* - ***Gartner***

> *Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making.* - ***Forrester Research***

One of the main tools to support BI is the ***Data Warehouse (DWH)***, which is a type of *Decision Support System (DSS)* and can be seen informally as an optimized repository that stores information for the decision-making process. With the huge and increasing number of information that companies have to manage in order to find relevant business strategies DWHs answer to the necessity of more sophisticated solutions than classical operational databases.
The main advantages are the following:

- they provide the ability to manage sets of historical data;
- they provide the ability to perform multidimensional analysis accurately and rapidly;
- they are based on a simple model that can be easily learned by its users;
- they are the basis for indicator-calculating systems.

Machine Learning