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*.

Figure 1.1: Increasing insights

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.

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).
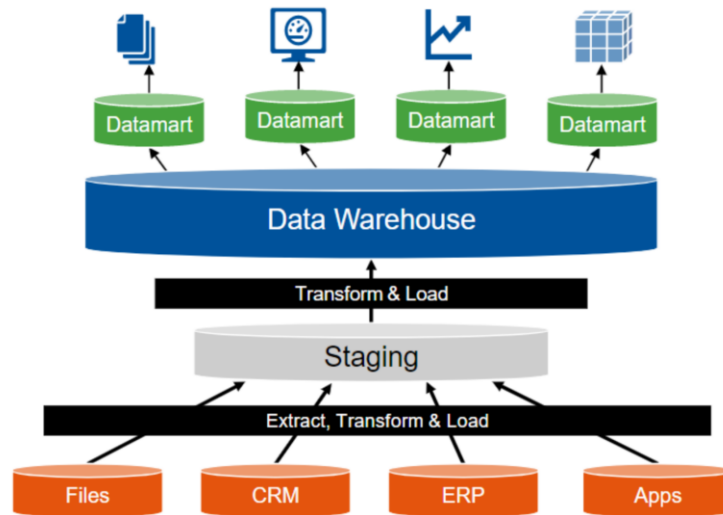

Figure 1.2: BI architecture

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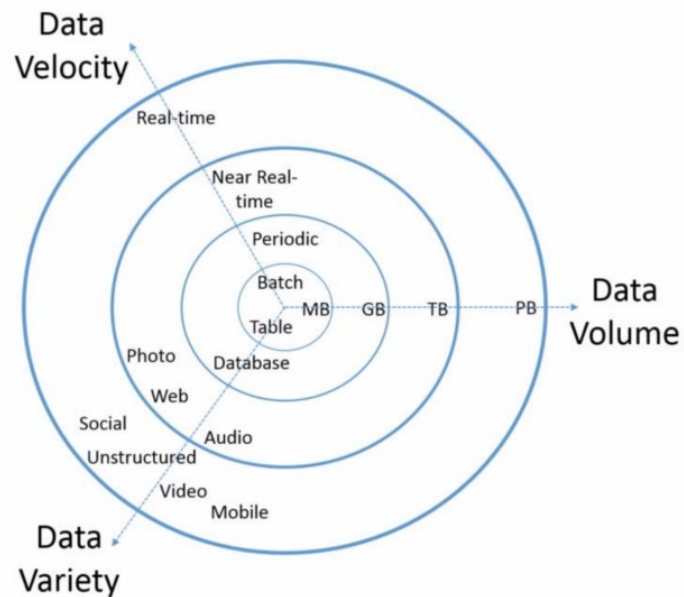
Big Data Big Data are a collection of data sets so large and/or complex and/or fast changing that they are difficult to process using traditional DBMSs or traditional data processing applications.

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 or pre-defined data model (video, audio, pictures, etc.);

- **Semi-structured:** there is some structure, perhaps data refer to different structures (XML, JSON, etc.).

Data Mining
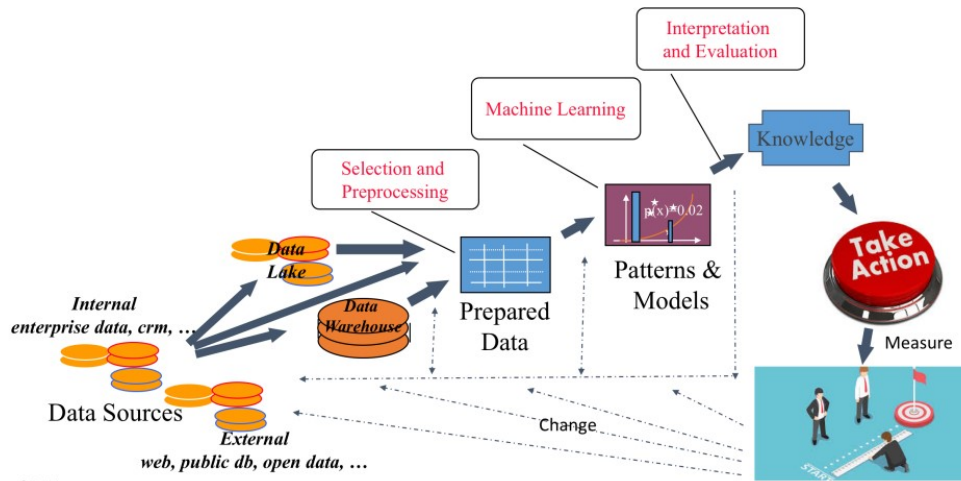
Figure 2.1: The Data Mining Process

    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)   ***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**

Data Warehouse (DWH) 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 run 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.

More formally, we can say that a **Data Warehouse (DWH)** is a specialized database that stores large volumes of historical data and facilitates the analysis and reporting of that data to support *decision-making processes*. It provides several key features:

- **Subject-Oriented:** A data warehouse is designed to focus on specific subjects or domains relevant to the enterprise's operations, such as customers, products, sales, etc. By organizing data around these core concepts, it enables analysts and decision-makers to gain insights into various aspects of the business.

- **Integration and Consistency:** One of the primary functions of a DWH is to integrate data from multiple disparate sources, including transactional databases, spreadsheets, and external systems. This integration ensures that data from different sources is harmonized and provides a unified view across the organization. Consistency in data representation and structure is maintained to ensure accuracy and reliability in analysis.

- **Evolution Over Time and Non-Volatility:** A crucial aspect of a data warehouse is its ability to capture and track changes in data over time. Historical data are preserved, allowing for the analysis of trends and patterns spanning various time periods. Unlike transactional databases where data may be constantly updated or overwritten, data in a data warehouse are non-volatile. Once committed, the data remains static, read-only, and preserved for future reporting and analysis.

In summary, a **DWH** serves as a centralized repository for historical data, providing a comprehensive and consistent view of the organization's information assets. By offering subject-oriented, integrated, and non-volatile data, it empowers decision-makers with valuable insights for strategic planning, performance analysis, and informed decision-making.

6

Data Mart (DM) A **Data Mart (DM)** serves as a specialized subset or aggregation of the data stored within a primary DWH. Unlike the comprehensive nature of a DWH, a **DM** contains a focused set of information tailored to *meet the needs of a specific business area*, corporate department, or category of users.

One of the key roles of **DMs** is to act as building blocks during the incremental development of DWHs. Rather than attempting to construct an entire DWH in one go, organizations often adopt an iterative approach, creating smaller, more *targeted DMs* that address immediate business needs. As the organization's analytical requirements evolve, additional DMs can be added or expanded upon, gradually contributing to the development of a comprehensive DWH architecture.

Moreover, **DMs** help to address the users' queries more efficiently. By tailoring the data content and structure to align with the analytical needs of particular departments or user categories, DMs facilitate more efficient and focused analysis. This granularity enables users to access and analyze relevant data without being overwhelmed by the vast volume of information typically found in a primary DWH.

Furthermore, **DMs** often offer improved performance compared to primary DWHs. Due to their smaller size and targeted scope, DMs can deliver faster query response times and better overall system performance. By focusing on specific subsets of data, data marts reduce the complexity of queries and minimize the processing overhead associated with accessing and retrieving information.
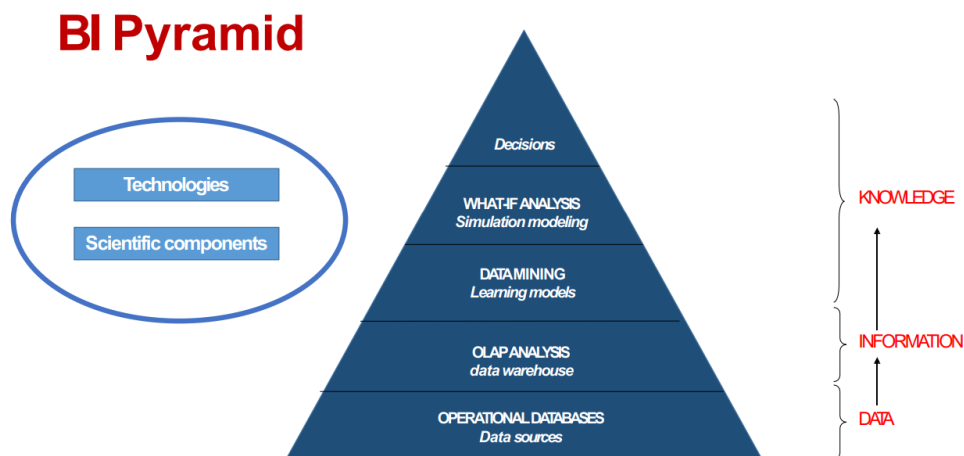
Online Analytical Processing (OLAP)

First of all, data are organised in **tables**, which are defined as collections of related, homogeneous data arranged into a row-column format, which includes:

- **Rows:** the various components stored in the table about a *specific individual*;

- **Columns:** the type and the meaning of a *specific component* of the individuals represented in the table;

- **Key:** a column or a set of columns whose values allow to distinguish *univocally* the rows of the table.

On-Line Analytical Processing (OLAP) allows users to interactively navigate the data warehouse information exploiting the multidimensional model, providing a flexible and intuitive way to explore, query, and report on large datasets. Typically,

the data are analyzed at different levels of aggregation, by applying subsequent **OLAP operators**, each yielding one or more different queries.

In an **OLAP Session** the user can scout the multidimensional model choosing the next operator based on the outcome of the previous ones. In this way, the user creates a navigation path that corresponds to an analysis process for facts according to different points and at different detail levels.

$$Product \longrightarrow Sub\text{-}Category \longrightarrow Category$$

The **OLAP operators** are the following:

- **Roll-up:** causes an increase in data aggregation and removes a detail level from a hierarchy by collapsing the rows that have a feature in common.

			2015		2014	
			Jan-15	Feb-15	Jan-14	Feb-14
Category	**Type**	**Product**				
Food and Beverages	Dairy products	White milk	90	90	60	80
		Chocolate milk	60	80	70	70
		Yogurt XY	20	30	30	35
	Beverages	Cola	20	10	35	30
		Orange Juice X	50	60	60	45

	2015		2014	
	Jan-15	Feb-15	Jan-14	Feb-14
Type				
Dairy products	170	200	160	185
Beverages	70	70	95	75

- **Drill-down:** is the complement to the roll-up operator; it reduces data aggregation and adds a new detail level to a hierarchy (e.g., from category to subcategory).

	2015	2014
Type		
Dairy products	370	345
Beverages	140	170

			2015		2014	
			Jan-15	Feb-15	Jan-14	Feb-14
Category	**Type**	**Product**				
Food and Beverages	Dairy products	White milk	90	90	60	80
		Chocolate milk	60	80	70	70
		Yogurt X	20	30	30	35
	Beverages	Cola	20	10	35	30
		Orange Juice X	50	60	60	45

- **Slice-and-dice:** the *slicing* operation reduces the number of cube dimensions after setting one of the dimensions to a specific value (e.g., category = 'Food and Beverages'); the *dicing* operation reduces the set of data being analysed by a selection criterion.
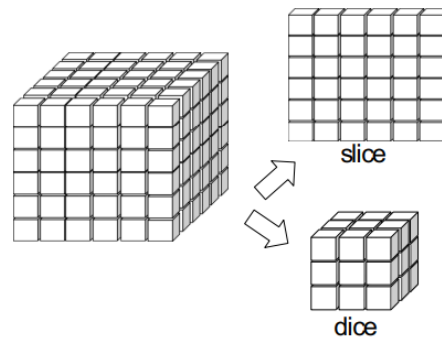

- **Pivot:** implies a change in layouts, aiming at analysing a group of data from a different viewpoint.

	2015	2014
Type		
Dairy products	370	345
Beverages	140	170
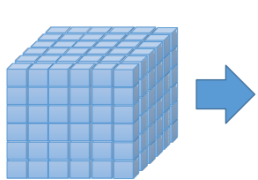
Type	Year	Quantity sold
Dairy products	2015	370
Dairy products	2014	345
Beverages	2015	140
Beverages	2014	170

- **Drill-across:** allows to create a link between concepts in interrelated cubes, in order to compare them.
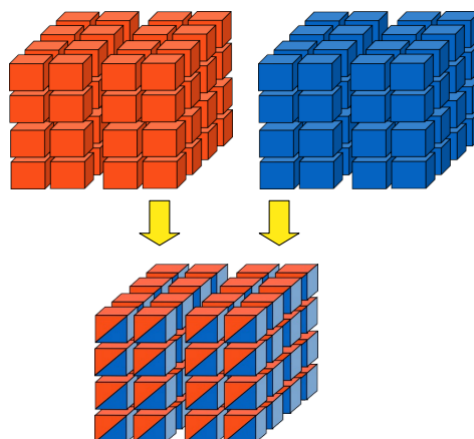
- **Drill-through:** switches from multidimensional aggregate data to operational data insources or in the reconciled layer.

Extraction, Transformation and Loading (ETL)

The ***ETL (Extract, Transform, Load)*** process is a crucial component of data warehousing and business intelligence. It involves **extracting** data from various sources, **cleansing** them, **transforming** them into a consistent format, and **loading** them into a target destination, such as a data warehouse, where they can be analyzed and queried effectively. Let's analyse the single phases.

Extraction

In the **extraction phase**, data (*structured* or *unstructured*) is collected from multiple disparate sources, including databases, flat files, APIs, and external systems. The methods used for this process vary depending on the source systems and the nature of the data:

- **Static extraction:** retrieving all the data from the source. Used to populate the data warehouse for the first time.

- **Incremental extraction:** updating the data warehouse regularly only with the data that have been modified. In this case data comes with an associated *timestamp* and *triggers* (related to change transactions for relevant data).

- **Real-Time extraction:** continous stream of data.

Cleansing

This phase is about all those procedures to improve the **quality** of the retrieved data, by standardizing it and correcting **mistakes** like *duplicate or missing data, unexpeted use of some fields, impossible or wrong values* and **inconsistences** due to *different practices used* or *typing mistakes*. Each type of problem requires different techniques for its solution. We can distinguish three main techniques:

Figure 2.2: Format discrepancies

- **Dictionary-based techniques:** they are used to check the correctness of the attribute values based on *lookup tables* and *dictionaries* to search for abbreviations and synonyms. We can apply these techniques if the domain is known and limited. These techniques are suitable for typing mistakes and format discrepancies [Figure 2.2, 2.3].

Figure 2.3: Inconsistences between correlated attributes

- **Approximate merging:** we use this technique when we need to merge data coming from different sources and we don't have a common key to identify matching tuples:

  - *Approximate join* (DESCRIZIONE E IMMAGINE)
  - *Similary functions* (DESCRIZIONE E IMMAGINE).

- **Ad-hoc algorithms:** custom algorithms based on specific business rules. (IMMAGINE)

Transformation

Loading

Data Warehouse Architectures

Conceptual Modeling: The Dimensional Fact Model (DFM)

Machine Learning