INFORMATION TECHNOLOGIES FOR DATA ANALYSIS

Lecture

Topics:

- Business Intelligence
- OLAP
- Data warehouses
- Exploratory data analysis
- Data mining

Time: 2 hours

Co-funded by the European Union

TET – The Evolving Textbook
Project no: 2022-1-SI01-KA220-HED-000088975

Łukasz Paśko, Rzeszów University of Technology

Introduction

Two main groups of information systems supporting the organization's activities

Record and operational systems

Keeping records of economic events, supporting the company's current operations

Information and analytical systems

They focus on analyzing data and processing it into information useful for decision-making

Record and operational systems

Primarily designed to **enhance daily business operations**, such as accounting, order processing, payment settlements, and warehouse management.
A key challenge is ensuring efficient information retrieval for management purposes:
the system's workload from handling daily transactions can hinder its capacity for analytical tasks like report generation and business analytics.

NOT INTEGRATED

- Basic systems frequently fail to fulfill all user requirements;
- Data within the organization is stored across diverse, distributed systems: data inconsistency and conflicts across different databases, systems, or applications;
- A lack of system integration often hinders access to comprehensive management information.

INTEGRATED

- Excessive information and system complexity can overwhelm users;
- The vast quantity of data resources makes it challenging for individuals to fully understand or utilize their content;
- Data processing methods are primarily designed for recording transactions rather than performing analytical tasks.

Introduction

Two main groups of information systems supporting the organization's activities

Record and operational systems

Keeping records of economic events, supporting the company's current operations

Challenges arise in accessing and presenting management information in a clear and concise format. These systems offer limited informational capabilities, such as generating only basic reports.

Information and analytical systems

They focus on analyzing data and processing it into information useful for decision-making

1ST GENERATION

2ND GENERATION

Information and analytical systems

Two generations of the systems

BUSINESS INTELLIGENCE

1ST GENERATION

- Basic information and analytical systems;
- Designed to handle specific reporting and analysis tasks;
- Common tools include report generators, spreadsheets, data visualization software, statistical tools, and specialized applications;
- The data sources: transaction system databases and/or manually entered user data.

2ND GENERATION

- Advanced information and analytical systems incorporate a broad range of applications and technologies to collect, integrate, organize, filter, analyze, and clearly present information from multiple sources, tailored to specific business areas;
- These systems utilize data warehouse technology, along with automated processes for extracting, integrating, and loading data from source databases into the warehouse;
- They also employ tools for multidimensional data analysis (OLAP) and data mining.

! Currently, the demand for information obtained through cross-sectional analyzes and unusual queries is significantly increasing.

Business Intelligence

Definition and features

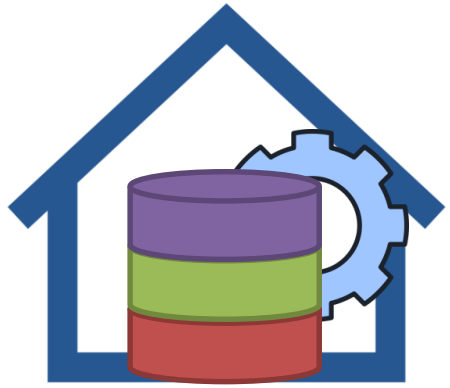
Information and analytical systems that process both internal enterprise data and external data to address the comprehensive informational and analytical needs of the organization and its environment, built on data warehouse technology and utilize tools for multidimensional analysis and data exploration.

They employ more advanced processing methods compared to transaction systems.

Their technical and software infrastructure is distinctly separate from transaction systems, utilizing data warehouse technologies and advanced tools for multidimensional analysis and data mining.

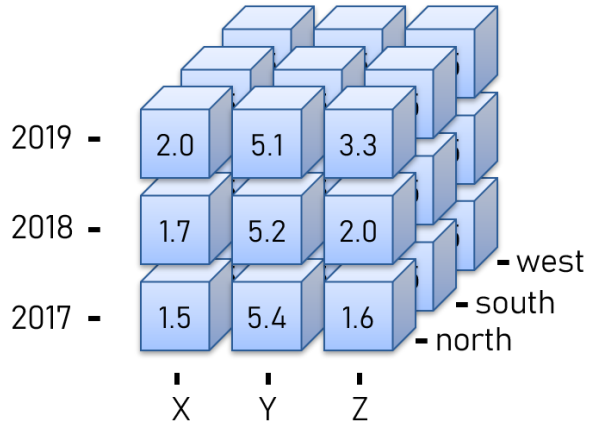
Components of BI systems

Data warehouse software: includes pre-configured programs for data retrieval, cleansing, transformation, access, and database structure creation.



Basic reporting and ad hoc queries: require no advanced analytical skills or training, enabling beginners to generate and use reports.

OLAP tools: provide an environment for multidimensional data analysis.



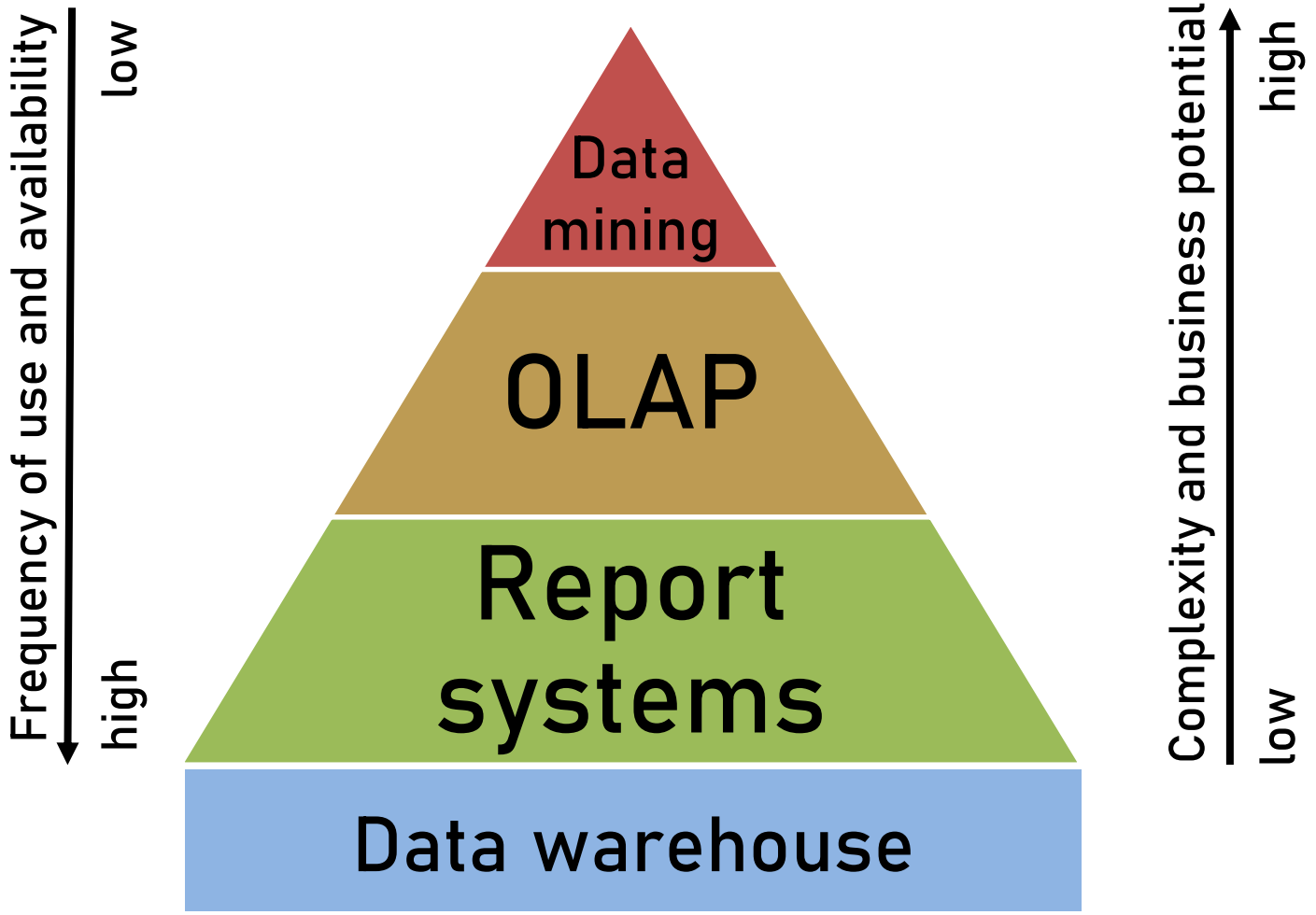
Exploration tools (data mining): apply advanced techniques to uncover hidden relationships, patterns, and trends in data.

Tools for informing management: offer an intuitive, visual interface for displaying trends, dependencies, and identifying issues and opportunities.

Analytical tools of BI systems

! The simpler the tool and methods, the greater the number of potential users and the frequency of use.

! Integrating all tools into a unified system creates a synergy, resulting in an intelligent decision-making environment.

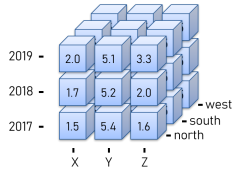


Layers of BI systems

1st layer
Integration and storage

ETL

Data warehouse



OLAP aggregates

2nd layer
Analytical processing

Basic analytical tools

Advanced analytical tools

Analytical software

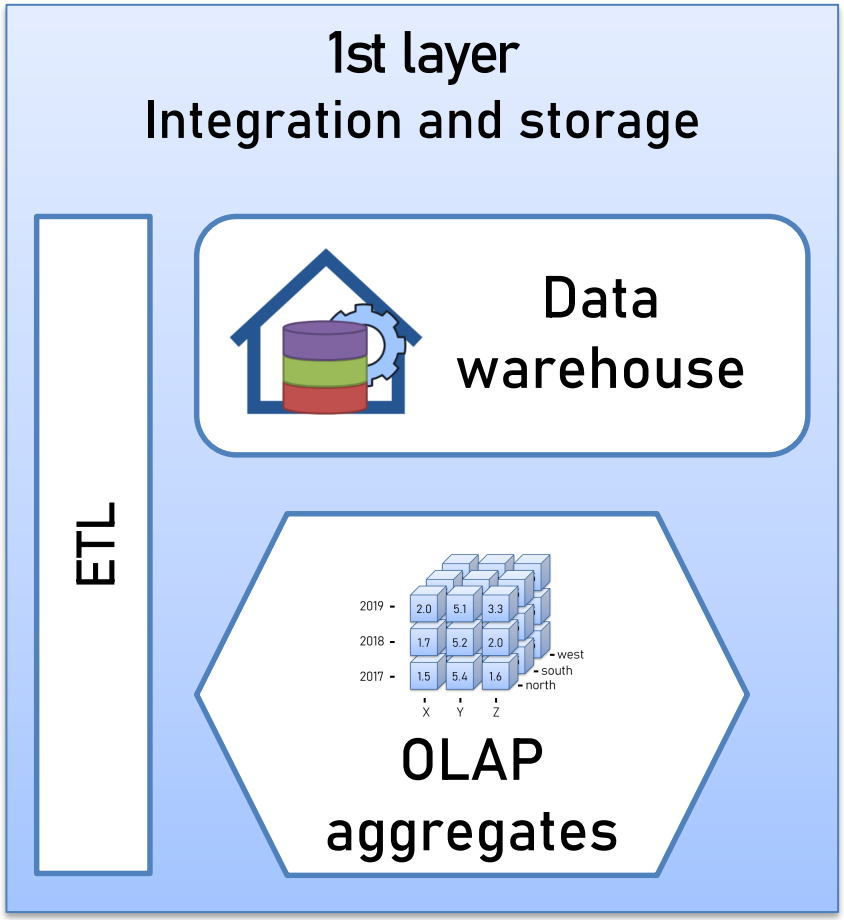
3rd layer
Sharing results

Information portals

Automatic distribution

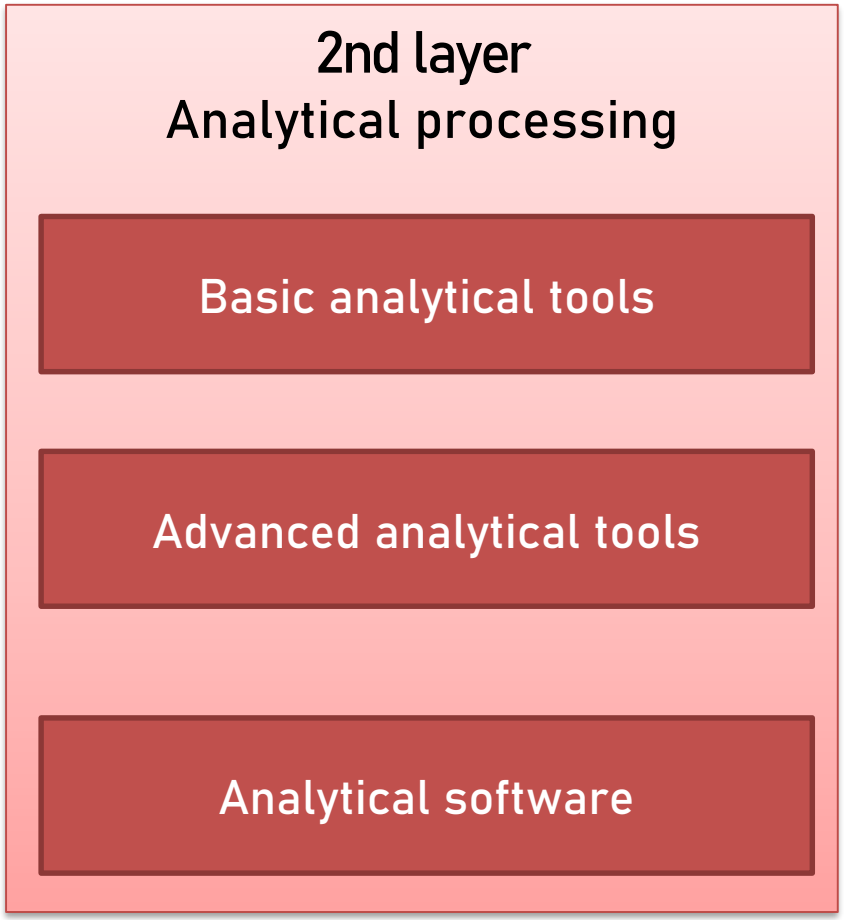
4th layer Administration

1 st layer of BI systems



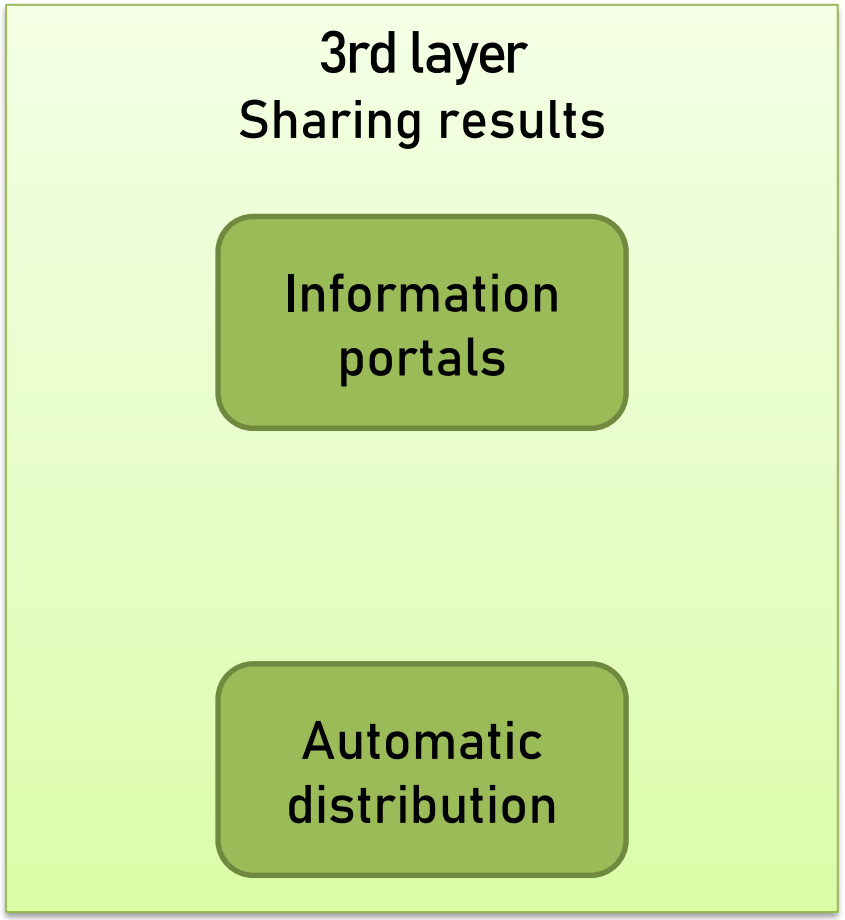
- ETL tools (Extraction, Transformation, Loading) handle the processes of extracting, transforming, and loading data into the data warehouse. They integrate data from multiple sources, ensuring high-quality and consistent data for the analytical tools in the second layer.
- The data warehouse stores both raw and aggregated data generated through the ETL process, with aggregated data either stored in the warehouse database or in separate files with specialized multidimensional structures.
- OLAP aggregations involve tools and specific multidimensional structures designed for storing aggregated data.

2nd layer of BI systems



- Basic reporting and visualization tools include report generators, wizards, query languages, spreadsheets, and OLAP tools for multidimensional analysis.
- Advanced analytical tools involve exploring databases with numerical data (data mining) and unstructured text data (text mining).
- Dedicated analytical software are:
  - Field-specific: targeting a particular business area (e.g., logistics),
  - Problem-specific: focusing on one or more detailed methods (e.g., financial analysis),
  - Industry-specific: addressing issues from a particular industry,
  - Comprehensive: supporting overall enterprise management,
  - Additionally, modules that enhance and extend the functionality of ERP systems.

3rd layer of BI systems



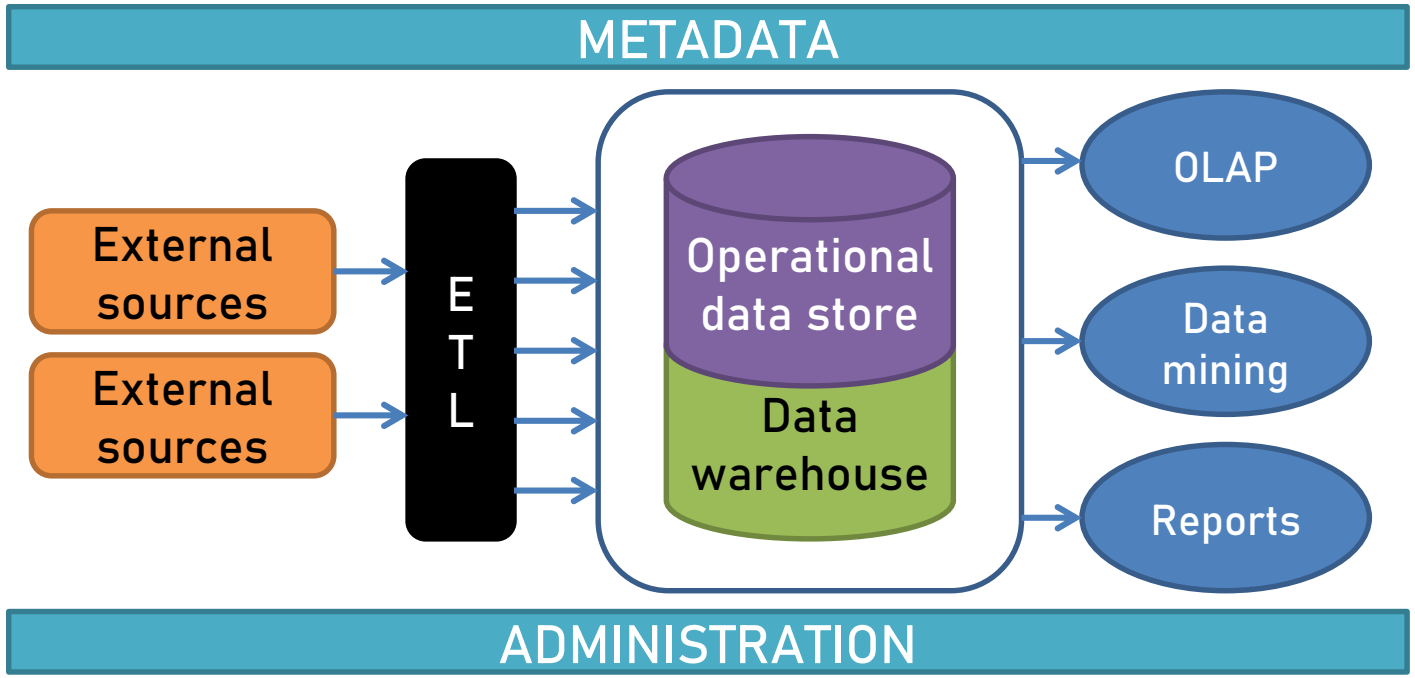
- File servers are used to store and share analysis results with decision-makers.
- Information portals operate within the organization's intranet.
- Automated tools for distributing information include email, instant messaging, and wireless communication.

4th layer of BI systems

- It spans across all three core layers, with various administrative tasks carried out in each layer:
  - Tools for managing data access and the metadata repository,
  - Content management for the metadata repository,
  - Tools for monitoring ETL and analytical process performance,
  - Configuration tools,
  - Personalization tools.

ETL: Extraction, Transformation, Loading

ETL tools are intended to extract data from source systems, apply necessary transformations, and load it into the appropriate locations within the data warehouse. They utilize information stored in the metadata repository and typically include additional features for documenting transformation methods or managing ongoing processes.



!

The ETL process is the process of **feeding the data warehouse**.

Experts note that designing and developing the ETL process accounts for 60-70% of the time needed to build a BI system.

ETL: Extraction, Transformation, Loading

Extraction → Transformation → Loading

- Identifying data sources,
- Selecting data that meets specific criteria,
- Extraction utilizes relevant metadata stored in the repository and interfaces that allow access to the systems from which the data is retrieved.

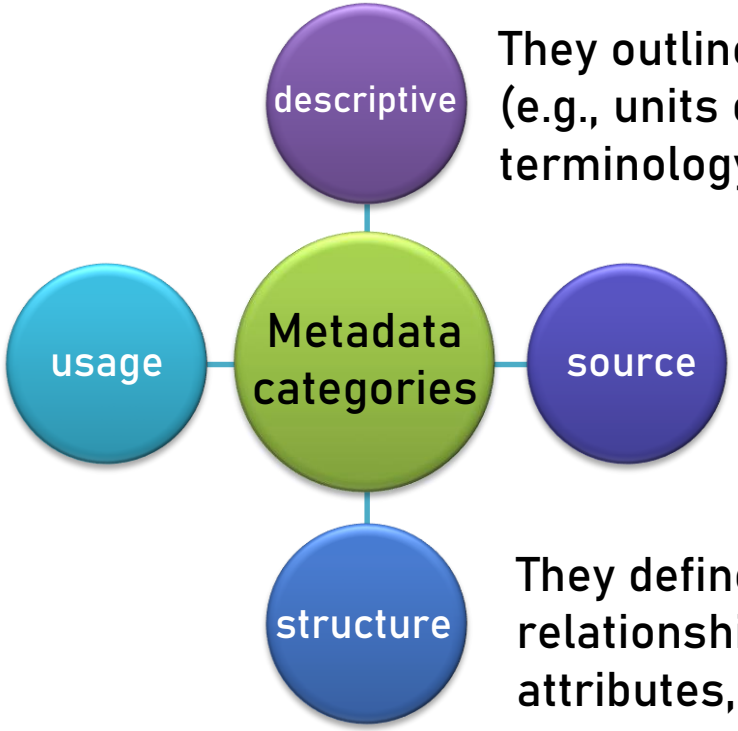
- **Data cleaning**: correcting erroneous records by removing inconsistencies that occur during data collection.
- **Data conversion**: changing data formats (e.g., transforming numerical values into qualitative ones) and standardizing record identifiers when source systems use different identifiers.
- **Semantic integration of data**: eliminating ambiguities in data interpretation by providing a clear description of each attribute in the warehouse.
- **Denormalize data**: links records and tables to enhance the performance of analytical processing.
- **Data sorting and aggregation**: are done based on specific criteria stored in the repository.

Saving data in the warehouse in fact tables and dimension tables.

Metadata repository

Metadata refers to data that defines the characteristics of collected information. It is essential for everyone involved in creating and managing a data warehouse: programmers use it to develop ETL and reporting applications, administrators rely on it to configure the hardware and software environment, and users utilize it to understand the content of the source data.

They specify which entities are authorized to access data, including access methods (access rights and data transformation rules during the process of making data available to users). They also define the guidelines for data integration and conversion during the loading process into the warehouse.



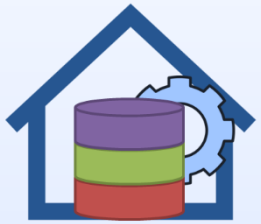
They outline the basic characteristics of data (e.g., units of measurement) and standardize terminology and nomenclature.

They indicate the origin of the data, including the location of sources and the connections between sources.

They define the structure of data sets, the relationships between tables and their attributes, and how data is connected.

Data warehouse

In a more specific context, a data warehouse is a **database that stores selected and organized data**, making it easily accessible and usable for decision-making.



In a broader sense, a data warehouse is **an IT architecture that manages the acquisition and organization of data required to support decision-making.**

It includes a database with systems for extracting data from various sources and processes that transform it into a format suitable for analysis.

Information in a data warehouse

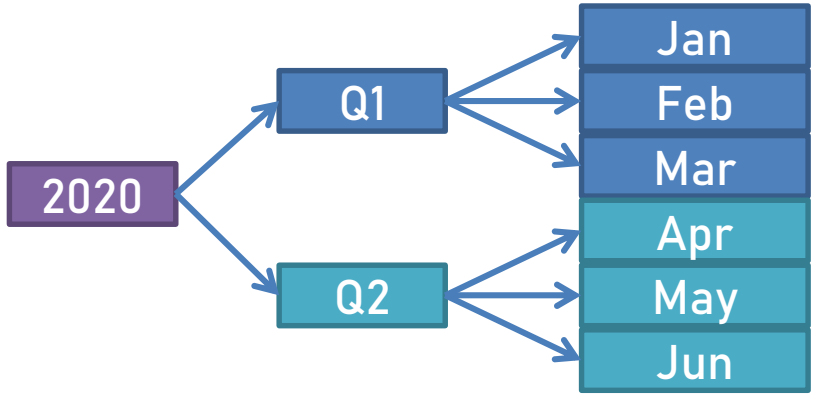
		Examples
Facts	They relate to the occurrence of specific events in the real world (e.g., business operations within an organization) and form the core of the analysis.	sales value, number of pieces, number of failures, number of complaints
Descriptions	They define the dimensions (or 'categories') through which actual data is analyzed and specify the areas where the data should be aggregated.	time, product, service, geographic region, customer, distribution channel, staff
Aggregates of facts	Aggregated data, stored at multiple levels, allows for detailed analysis—from general to specific. The purpose of storing these aggregates is to improve the speed of user query responses.	total sales values for weeks, months, quarters, years
Metadata		

Information in a data warehouse

An example of dimension an its hierarchy

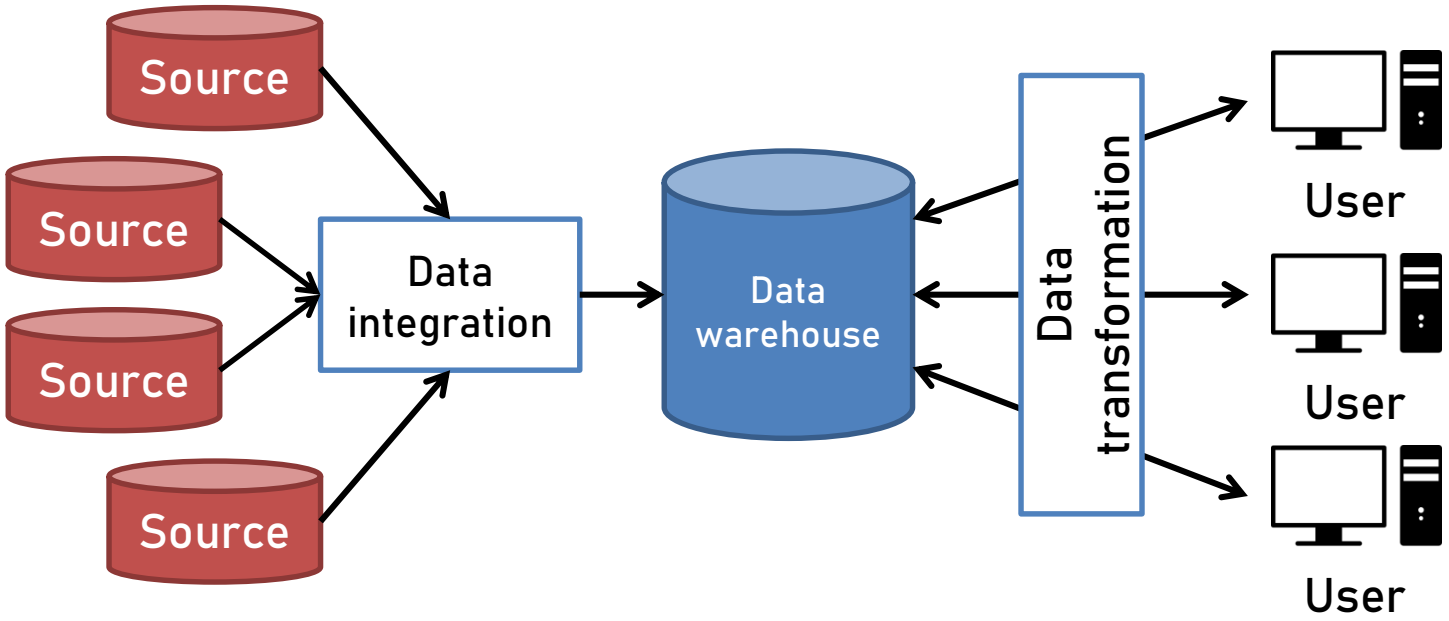
Dimension	Dimension attribute	Attribute element
Time	Year	2018, 2019, 2020, …
	Quarter	I, II, III, IV
	Month	January, …, December

Hierarchy in the TIME dimension



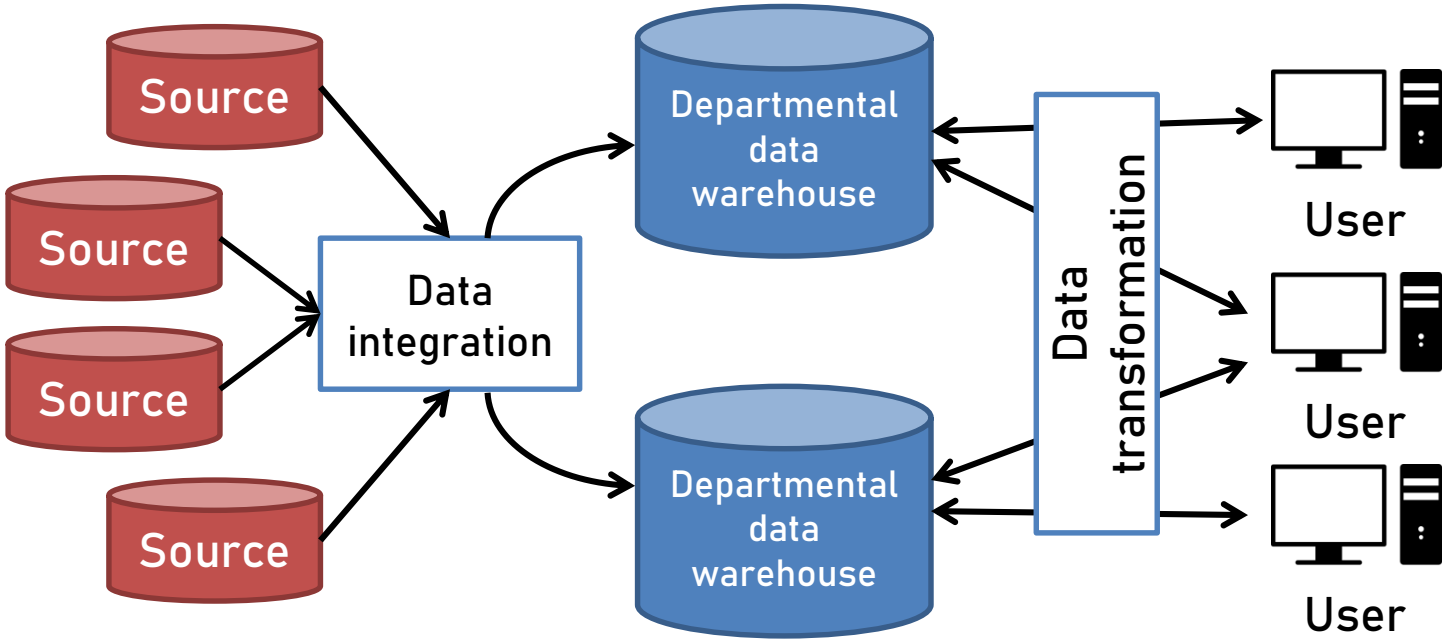
Data warehouse architectures (1)

General architecture of the warehouse



Data warehouse architectures (2)

A system based on departamental data warehouses

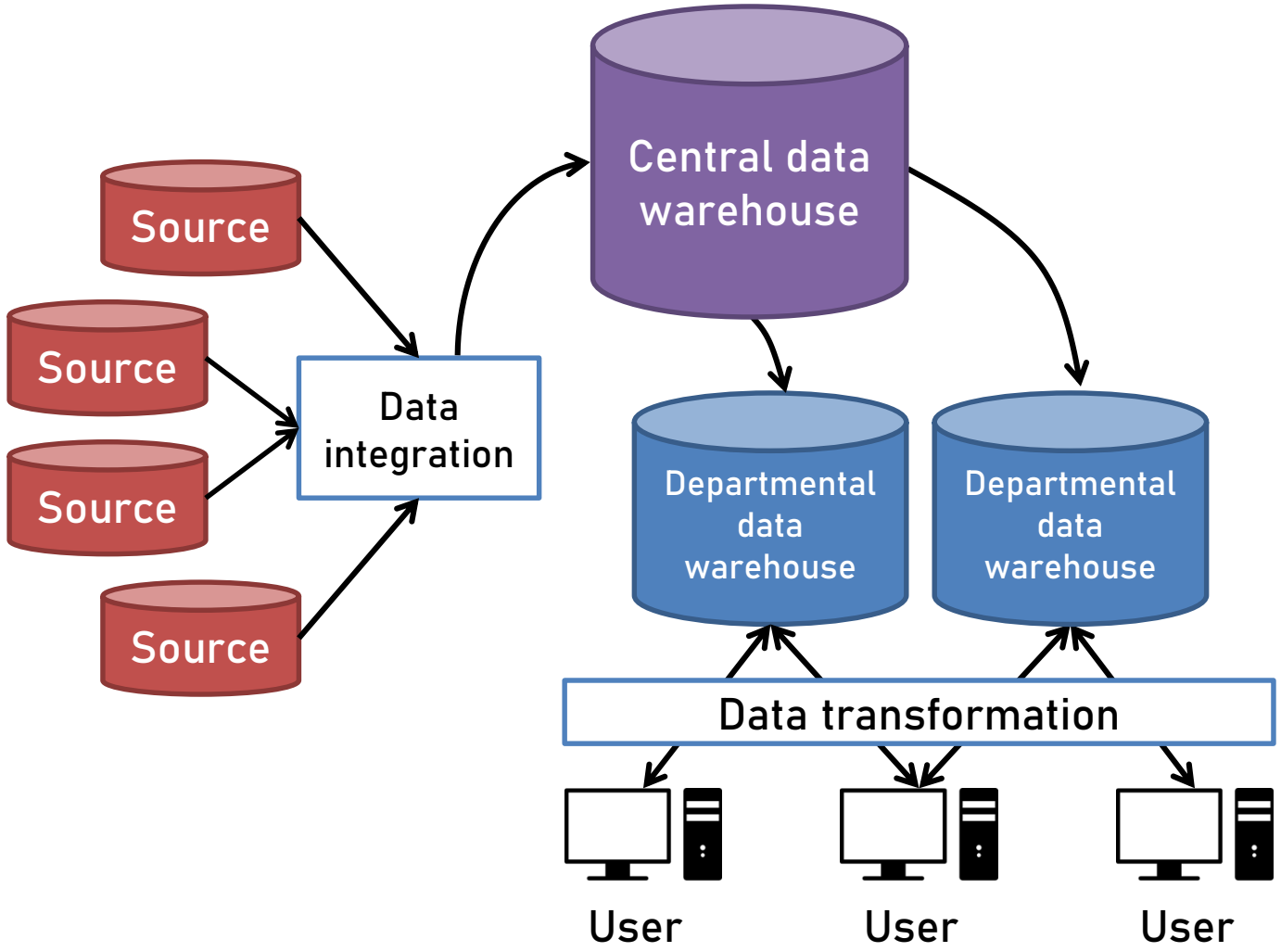


Mini warehouse (departmental data warehouse or Data Mart)

A data warehouse focused on a specific theme, created to address the information requirements within a particular area (e.g., marketing, finance, production, sales).It targets specific business challenges within one department of the company.

Data warehouse architectures (3)

A system based on a central warehouse and departamental warehouses



Central data warehouse

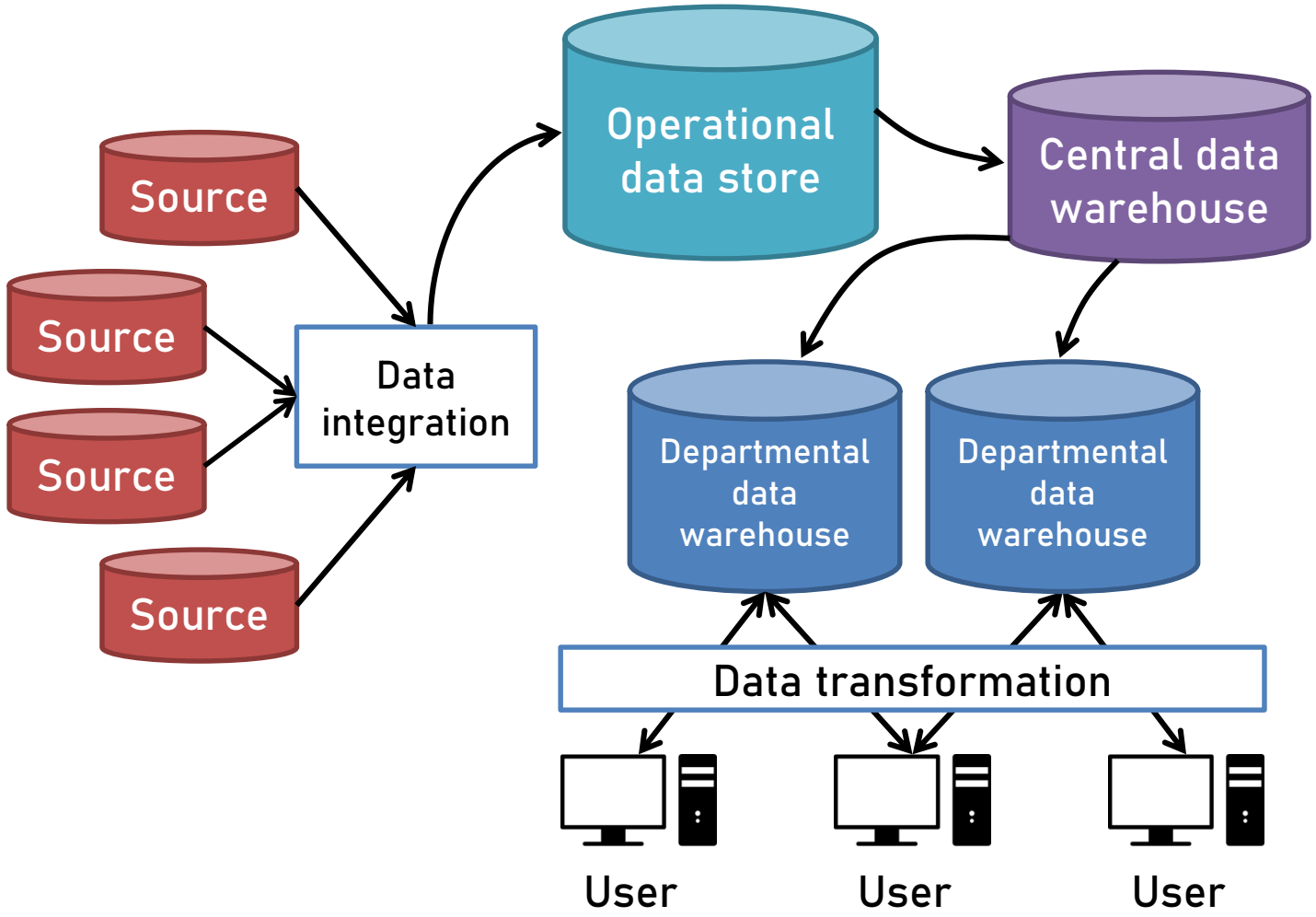
Two main purposes:

1. To perform comprehensive analyses for the entire organization
2. To establish a single central source of consistent data, which distributes information to specialized data warehouses.

It facilitates the standardization of business terminology across departmental warehouses through a central metadata repository.

Data warehouse architectures (4)

A system based on an operational data warehouse, a central warehouse and departamental warehouses



Operational Data Store

- It holds all up-to-date information about the company, reflecting its current status, and must be regularly updated.
- Its purpose is to offer quick access to detailed and current data (e.g., the debt status of a particular client).
- It is not suitable for analytical queries, as it does not contain archived data, and queries to the ODS focus only on current information.

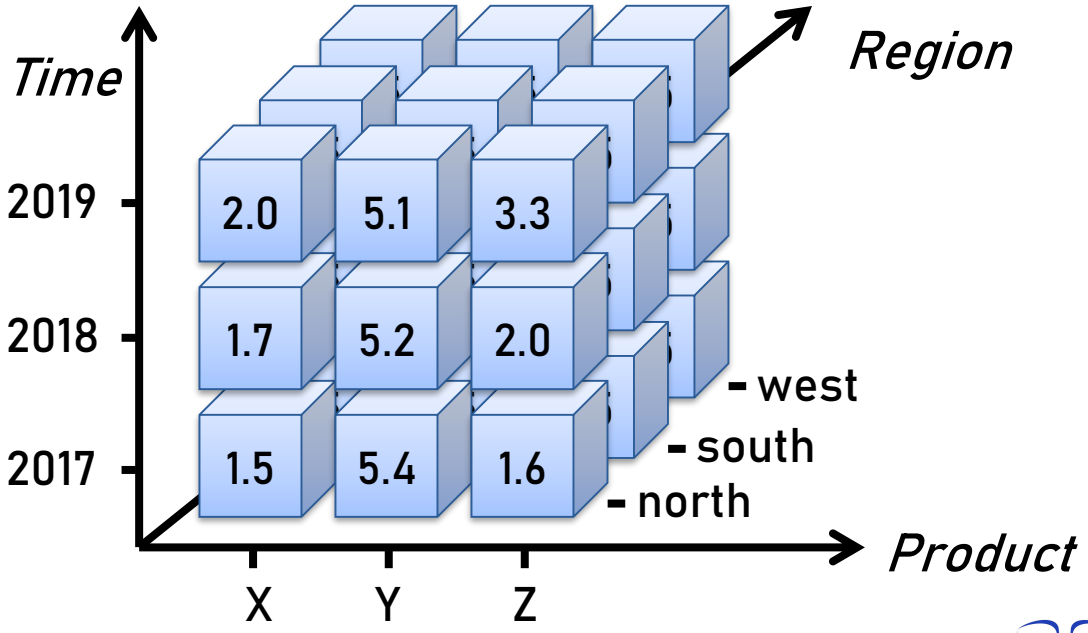
OLAP

On Line Analytical Processing

OLAP is a software technology that enables analysts to explore data by quickly and easily accessing various perspectives of information. These perspectives are derived from raw data and represent the organization's dimensions in a user-friendly way.

The goal of OLAP is to allow users to conduct thorough data analyses by providing rapid access to multidimensional views of the organization.

Data cube with sales value

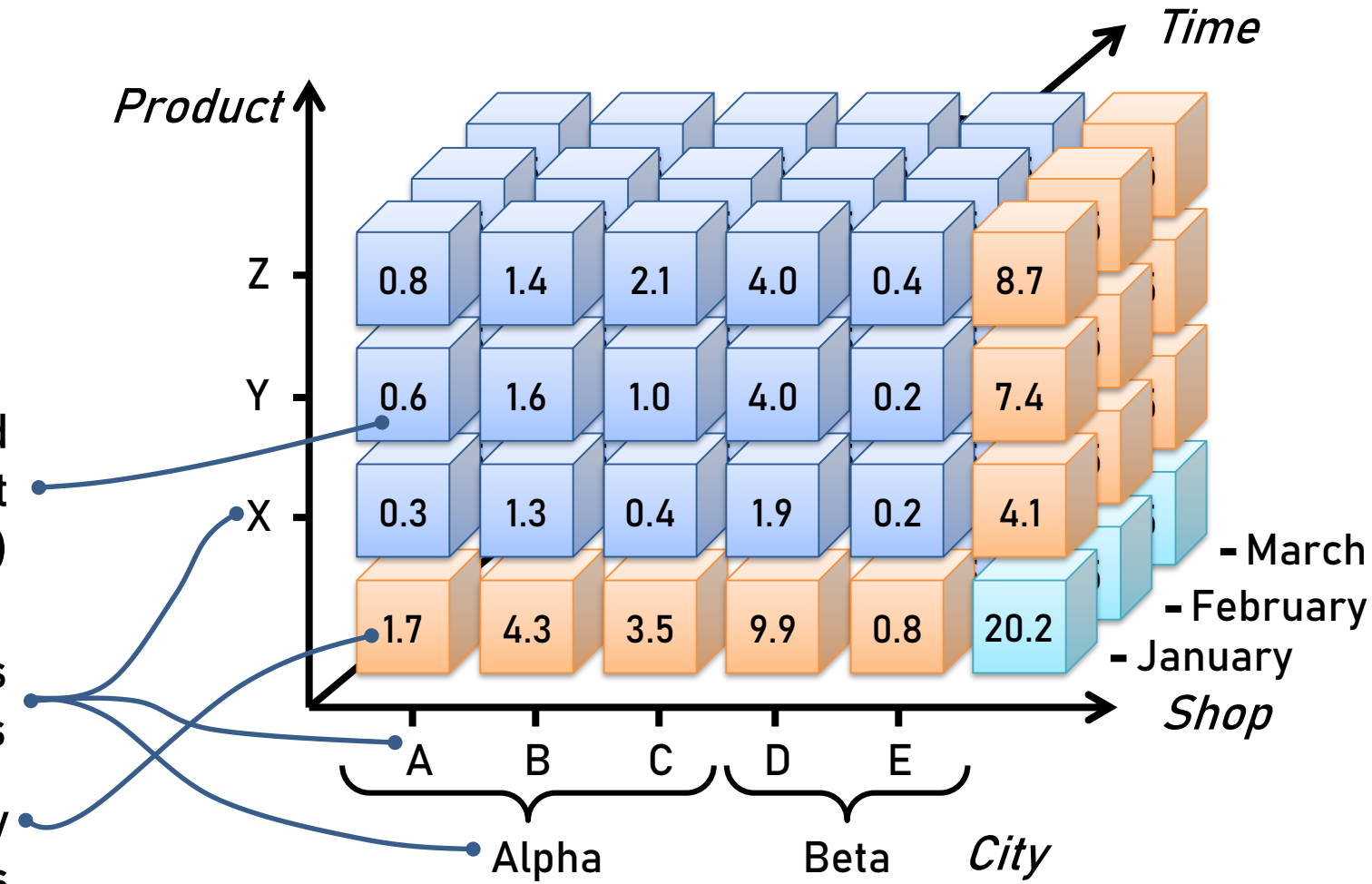


! Data for OLAP is presented in the form of multidimensional data cubes (three or more dimensions).

OLAP data cube

! Cubes overcome the limitations of relational databases and extend the "two-dimensional world" of spreadsheet tables.

A data cube with sales value and summaries



Cube elements are aggregated (by dimensions) measurement values (e.g. sales)

Row and column headings are dimension attributes

Cubes also contain auxiliary summaries

Knowledge Discovery from Databases

Knowledge Discovery from Databases – KDD

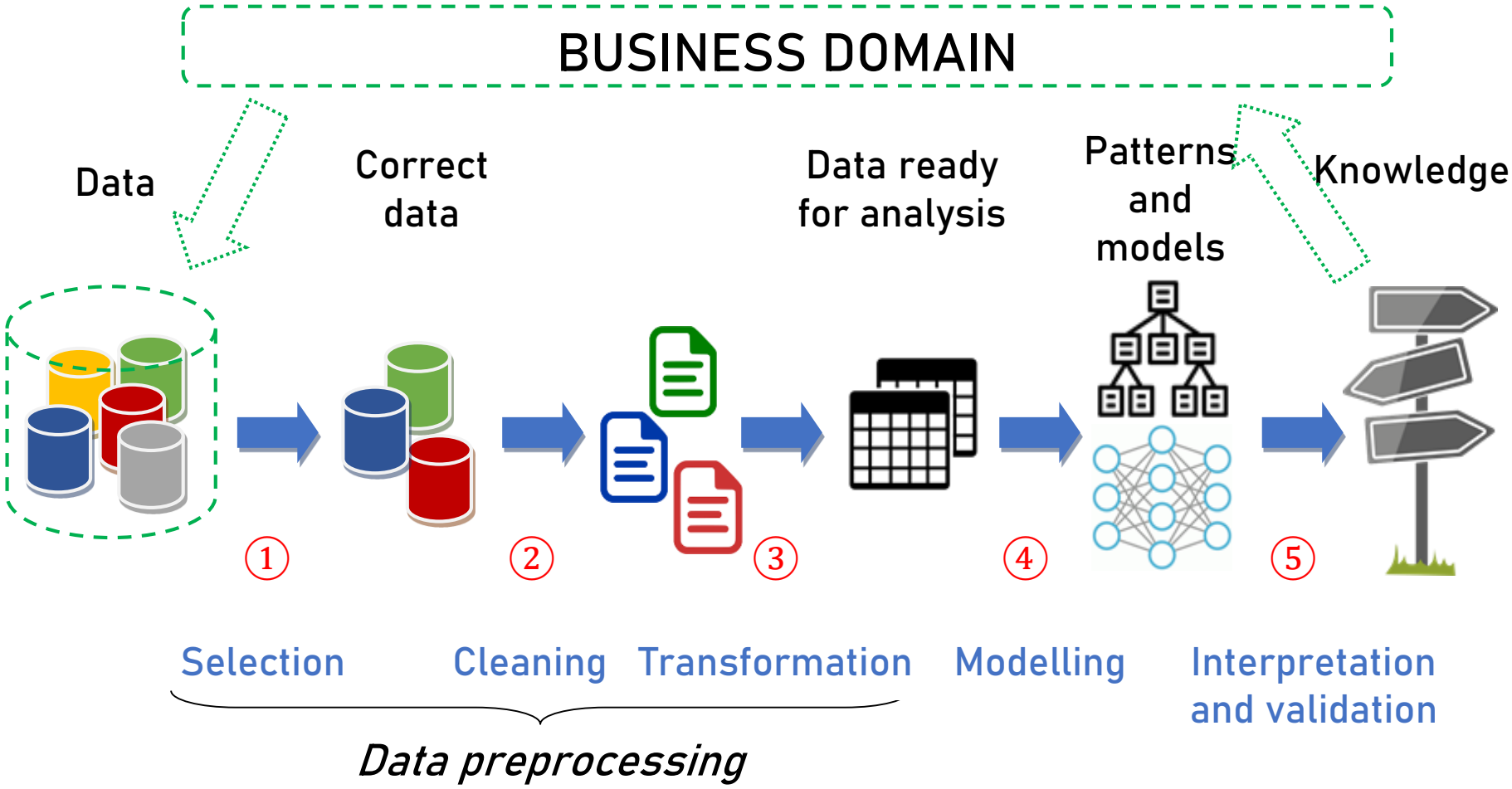
The process aims to thoroughly analyze data, beginning with a clear understanding of the problem, followed by data preparation, application of appropriate models and analyses, and their subsequent evaluation.

The goal of KDD is to extract information that is hidden due to the large volume of data, converting this information into actionable knowledge that can, among other things, support decision-making.

CRISP-DM stages (Cross-Industry Standard Process for Data Mining – one of the KDD varieties):

1. Understanding the business domain where the data comes from
2. Detailed understanding of the data,
3. Data preparation,
4. Creation of models,
5. Evaluation of the results obtained,
6. Implementation of discovered knowledge in the business field.

Knowledge Discovery from Databases



① selecting data relevant to the problem under consideration

② handling incorrect or missing data

③ giving the data proper representation

④ finding patterns

⑤ checking whether the identified phenomena occur only in the analyzed data and how the models deal with new data

Data mining

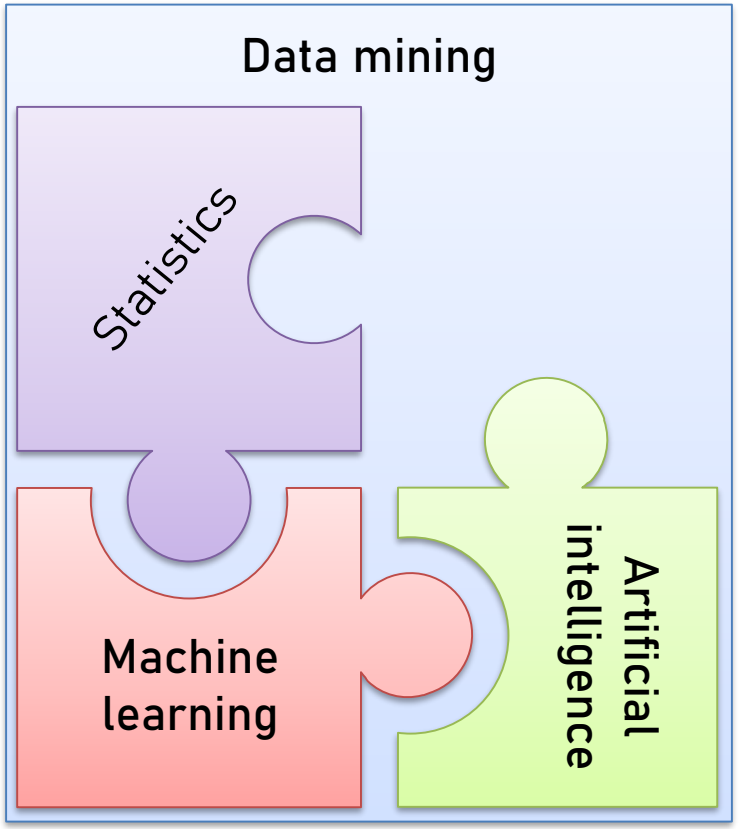
Data mining

Analyzing large datasets to uncover unexpected relationships and present the data in a novel way, making the discovered information both comprehensible and valuable to the user.

The goal of data mining is to identify complex and previously unknown correlations, patterns, and trends hidden within the data.

!

Data mining uses both statistical data analysis and artificial intelligence with machine learning.



Data mining

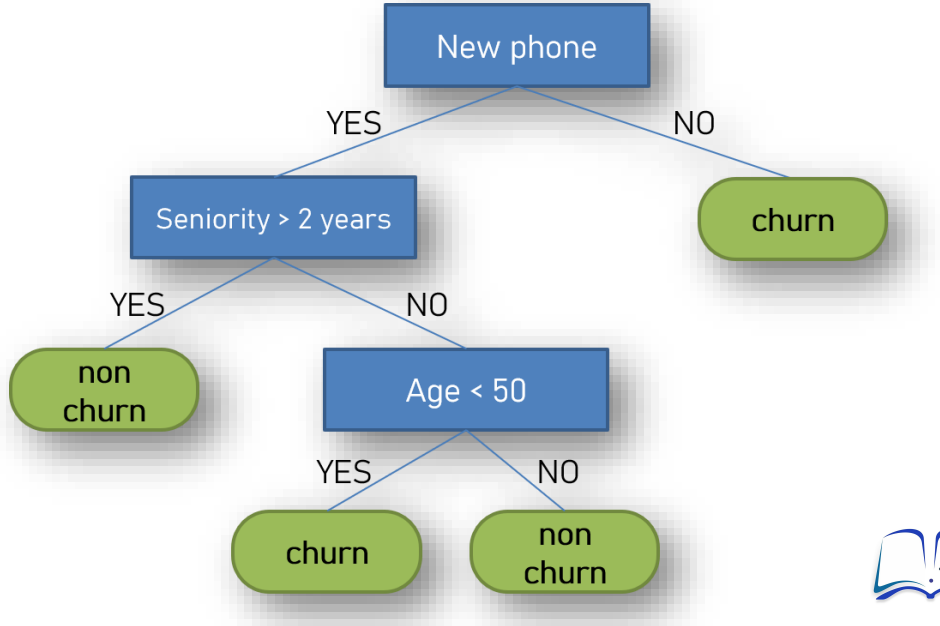
The summaries and dependencies that result from data mining are called **patterns**. Their examples could be:

- Linear or non-linear equations,
- Rules,
- Graphs,
- Tree structures,
- Recursive patterns in time series.

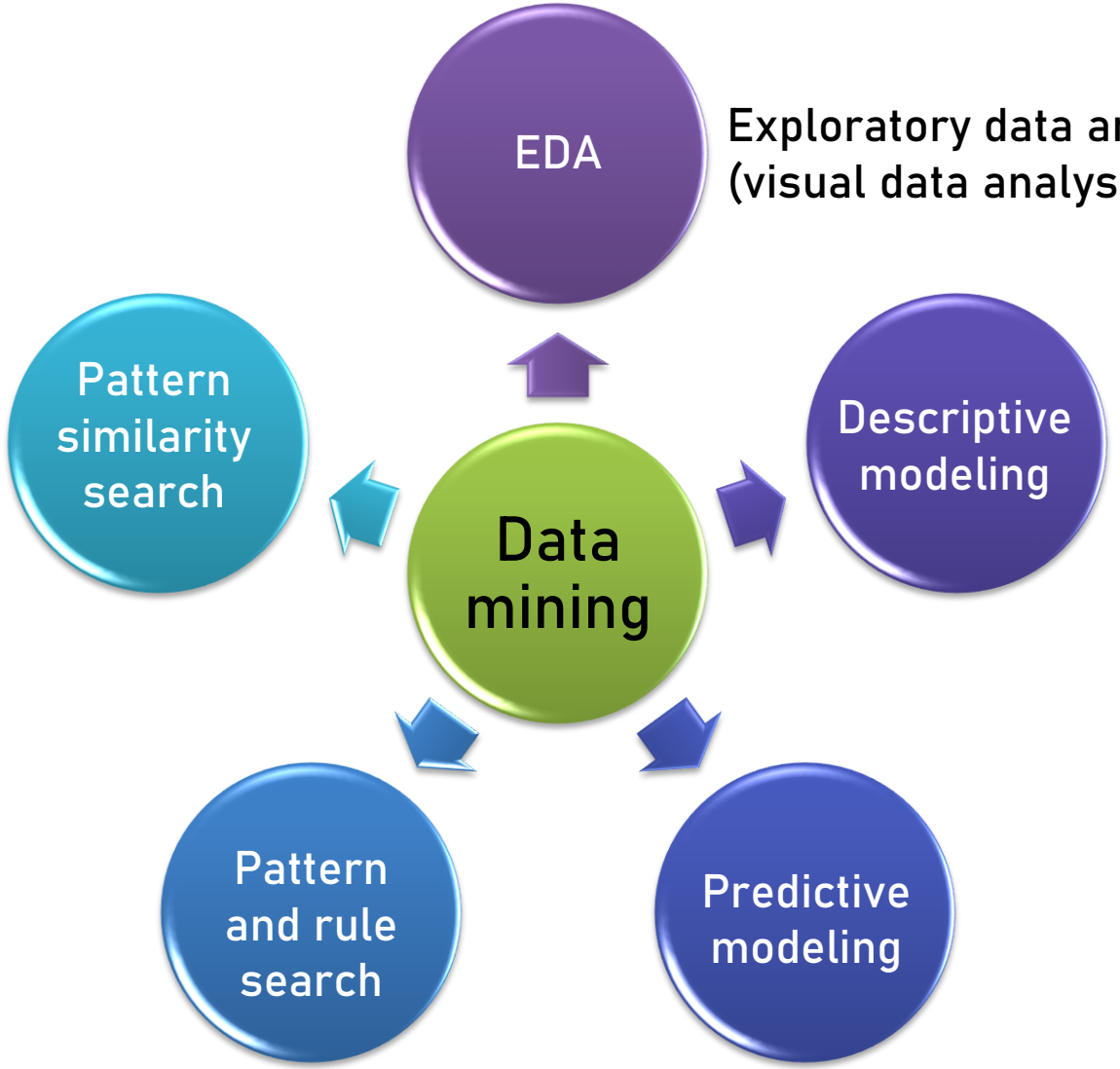
RULE 1: **IF** parameter1 < 50 **i** parameter2 > 10, **THEN** percentage of defects = 3%

RULE 2: **IF** parameter1 >= 50 **i** parameter3 = 3rd_shift, **THEN** percentage of defects = 10%

RULE 3: **IF** parameter1 <= 10, **THEN** percentage of defects = 40%



Types of data mining applications



Exploratory data analysis

EDA involves using visual methods to find certain structures (patterns) in the data that may signal deeper dependencies.

- EDA leverages human skills to interpret patterns through visualizations.
- There is no need to define assumptions beforehand—we begin by exploring the data and then develop hypotheses based on our observations.
- This contrasts with statistical data analysis, where we start with a hypothesis and then apply statistical methods to verify whether the data supports it.

! EDA is particularly valuable and useful when the data are little known and the purpose of the study is not precisely specified.

Techniques used in EDA can visualize:

- Single variables,
- Relationships between two variables,
- More than two variables,
- Multidimensional scaling.

Descriptive modeling

It is used to describe the data under study and includes:

- Models for the overall probability distribution of the data,
- Models that define the relationships between variables,
- Models for partitioning multidimensional data into subgroups.

- Both methods create subsets (groups, segments, clusters) containing elements with similar characteristics.

Segmentation

- The researcher defines the characteristics and number of groups as well as establishes the criteria in advance for classifying an object into a particular group.
- Once the group characteristics are set, the dataset is examined to identify objects that match the criteria and can be assigned to the appropriate group.

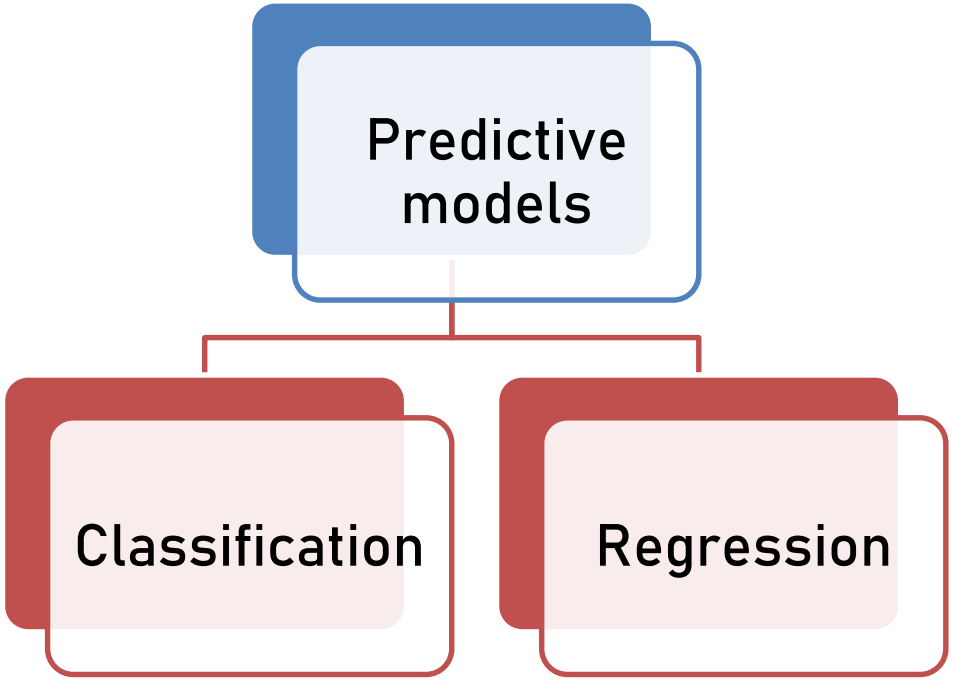
Clustering

- The aim is to **identify natural groups** in the data,
- The clustering technique (algorithm) chosen determines the number of groups and the characteristics used to classify an object into a specific group.
- This involves searching for objects with similar characteristics without predefining the feature sets and their values that would define the groups.

Predictive modeling

It allows predicting the unknown value of the result variable for certain given values of other variables called explanatory variables.

There are two types of predictive models



Predictive modeling

Classification

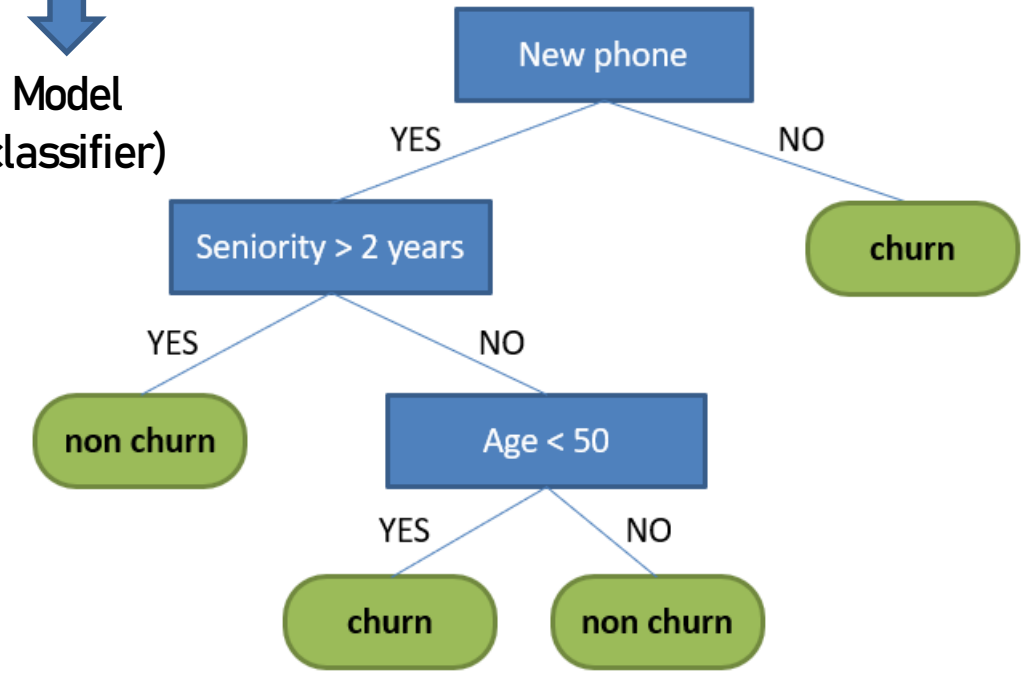
- The result variable has categorical values, i.e. from a finite set of categories.
- **1st STAGE:** building a model based on historical data–the model divides the set of objects into mutually exclusive classes so that objects belonging to the same class are similar to each other in the context of the result variable.
- **2nd STAGE:** using the model built in the 1st stage to classify new objects that were not present in the historical data.

1st stage

New phone	Seniority	Age	Churn?
Yes	5	32	Non churn
Yes	1	20	Churn
…	…	…	…

Historical data

Model (classifier)



2nd stage

New phone	Seniority	Age	Churn?
Yes	12	52	???

Predictive modeling

Regression

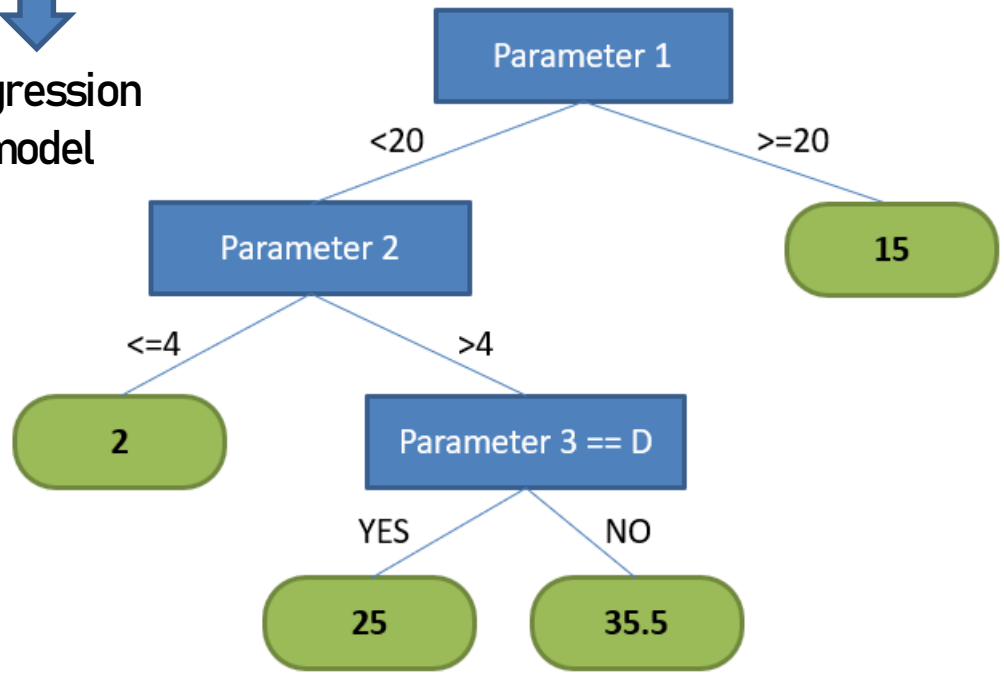
- The result variable has numeric values.
- **1st STAGE:** building a model based on historical data–the model will express the relationship between the values of the resultant variable and the explanatory variables.
- **2nd STAGE:** using the model built in stage 1 to predict the value of the outcome variable for new objects that were not present in the historical data.

1st stage

Param1	Param2	Param3	Result
10	5	A	5
8	1	D	23
…	…	…	…

Historical data

Regression model



2nd stage

Param1	Param2	Param3	Result
14	3	C	???

Discovering patterns and rules

Pattern

- A pattern is a local concept that describes only a certain aspect of the data (as opposed to a model, which describes the entire data set);
- Represents a feature of the data that may be valid for only a few records or a few variables.

- Pattern search algorithms are used, for example, to predict risk, discover the causes of observed phenomena, and identify customers exhibiting similar behavior.
- Pattern search most often concerns discrete data stored in a standard data matrix.

Rule

IF (set of conditions) **THEN** (set of facts)

- The most popular way to describe a pattern is an association rule:

  IF *A*, THEN *B* with probability *p*

- *p* is called the accuracy or confidence of the rule (the probability that *B* is true given that *A* is true)
- A rule's support specifies the portion of all objects for which the left and right sides of the rule are true.

Search by pattern

- Pattern search algorithms aim to identify objects that resemble a given pattern.
- Similarity conditions or measures must be defined to assess whether an object is similar to the reference pattern.

The pattern search task is primarily applied to datasets that include:

- Texts, where a pattern could be a set of keywords or a phrase,
- Images, where a pattern could be a sketch or a description of an image,
- Time series, where a pattern could be a sequence of data points over time,
- Other types of sequential data, where the sequences are not time-dependent.

Thank you for your attention!