Data Analysis and Knowledge Discovery

Lecture 1

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

Sergiu Limboi, PhD Teaching Assistant

Motto: From raw data to understanding and decisions.

FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ • BABEȘ-BOLYAI •

50 de ani de Informatica la Cluj

FACULTATEA DE MATEMATICA ȘI INFORMATICA • BABEȘ-BOLYAI •

Π Σ

ROMÂNIA

Introduction into Data Analysis, Data Mining and Knowledge Discovery

AGENDA

- Course organization
- Why this course matters?
- Let’s get to know the audience
- Evaluation
- The Big Picture: From Data to Knowledge
- What is Data?
- What is Data Analysis?
- What is Machine Learning?
- What is Data Mining?
- What is Knowledge Discovery?
- Data science

AGENDA

- The Relevance of Data-Driven Disciplines in 2026
- Where Things Fail in Practice
- Types of Data We Will Handle
- Data Libraries in Python
- Key Takers

Course organization

Faculty of Mathematics and Computer Science

Course organization

- Instructor: Teaching assistant PhD Sergiu Limboi
- Teams channel code: **9dtc90d**
- Course structure:
  - Lectures 2 hours per week
  - Seminars 2 hours every two weeks
- Additional information:
  - Seminar sessions are dedicated to the **presentation, discussion, and evaluation of semester work.**
  - Seminars will be scheduled as needed.

Why this course matters?

Faculty of Mathematics and Computer Science

Why this course matters?

- Data is everywhere. Organizations collect massive amounts of data.
- Data alone has no value.
- Value appears only when:
  - Patterns are discovered;
  - Results are interpreted;
  - Decisions are made.
- Examples:
  - Finance-> risk detection
  - Healthcare-> diagnosis support
  - Marketing-> customer behaviour
  - Science-> hypothesis validation
  - etc.

Let’s get to know the audience

Faculty of Mathematics and Computer Science

Let’s get to know the audience

Go to www.menti.com and enter the code **5924 0715**

or use the QR code

Evaluation

Faculty of Mathematics and Computer Science

Evaluation

- **Evaluation Structure:**
  - Written Exam (during the official exam session): 30%
  - Project and Research Report (completed during the semester): 70%
  - Further details are available in the *Information and Requirements* document.

- To successfully pass the course, students must obtain:
  - **Minimum grade of 5** in the Written Exam
  - **Minimum grade of 5** in the Project and Research Report
  - Both conditions must be fulfilled independently.

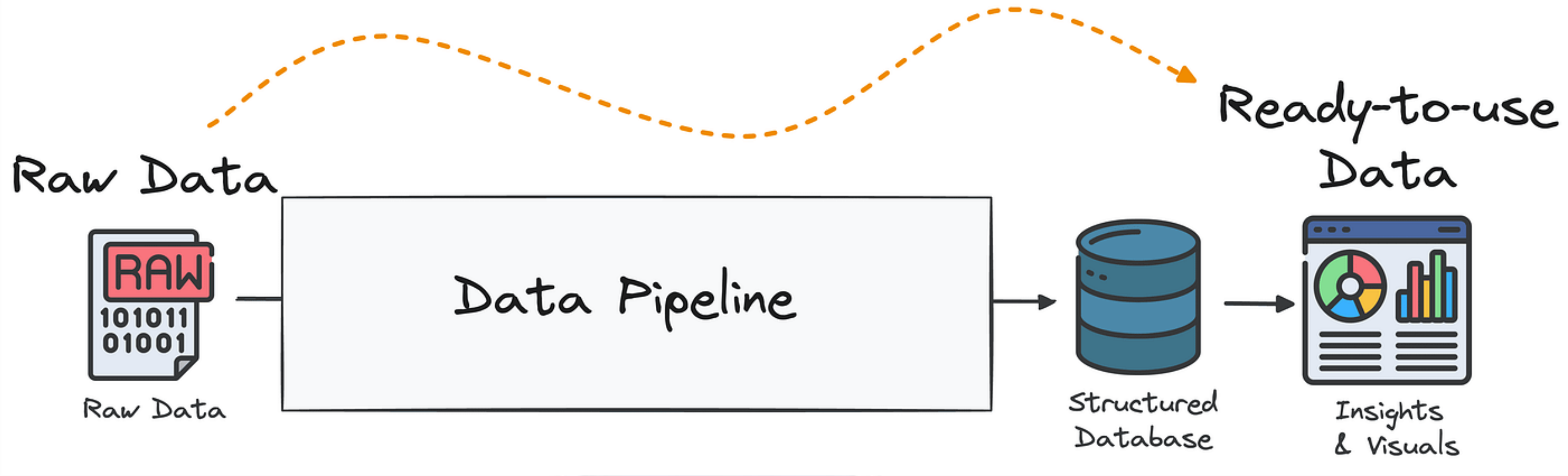
- **Additional Points**
  - **In-class quizzes** conducted throughout the semester may provide bonus points.
  - **Attendance at invited lectures** will be rewarded with additional points added to the final grade.

The Big Picture: From Data to Knowledge

Faculty of Mathematics and Computer Science

The Big Picture: From Data to Knowledge

- Algorithms alone **do not create knowledge** – interpretation does.
- Conceptual pipeline: raw data → processed data → patterns → interpretation → knowledge → decisions



What is Data?

Faculty of Mathematics and Computer Science

What is Data?

- Data = recorded observations
- Data ≠ information
- Data ≠ knowledge
- Information = data plus context and meaning
- Knowledge = information + understanding, interpretation and validation
- Examples:
  - Numbers (prices, temperature)
  - Categories (gender, product type)
  - Text (reviews, comments)
  - Images/ signals
  - etc.

What is Data Analysis?

Faculty of Mathematics and Computer Science

What is Data Analysis?

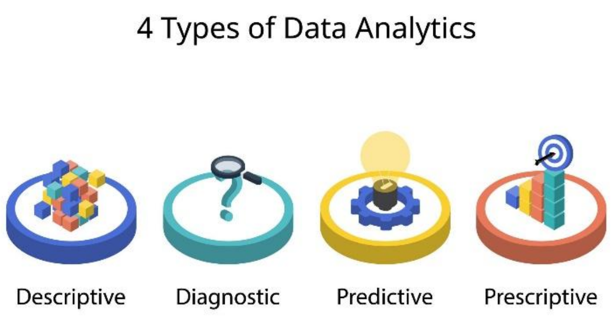
- Process of systematically inspecting, cleaning, transforming and modelling data to discover useful information, draw conclusions, and support decision-making.
- Answers:
  - What happened?
  - How often?
  - What trends exist?
- Key concepts of Data Analysis:
  - Data collection
  - Data cleaning
  - Data transformation
  - Exploratory Data Analysis (EDA)
  - Interpretation

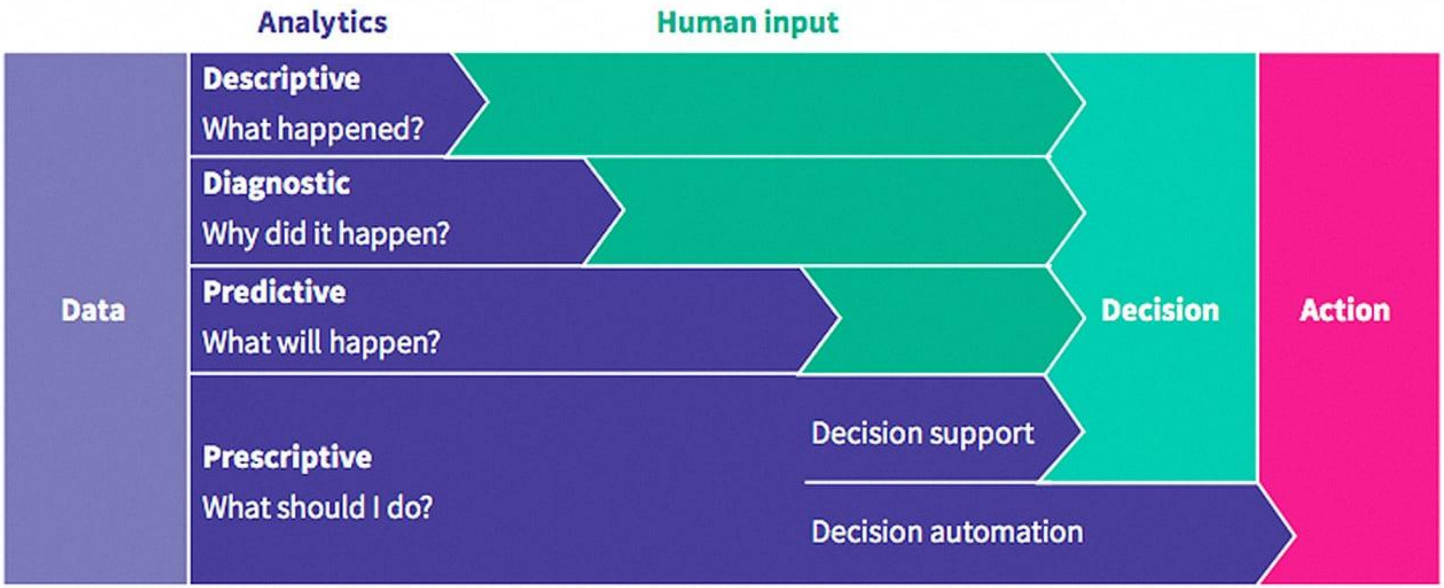
What is Data Analysis?

- Why is Data Analysis Important?
  - Problem-solving;
  - Performance tracking;
  - Informed decision-making.

- The role of a Data Analyst:
  - Data interpretation
  - Reporting
  - Decision support
  - Tool proficiency
  - Collaboration

Types of Data Analysis





Descriptive Analysis (What happened?)

- Summarizes historical data to understand **what has already occurred**.

- **Typical outputs**
  - Averages, totals, percentages;
  - Tables and dashboards;
  - Line charts, bar charts.

- **Examples**
  - Monthly sales report;
  - Average exam score;
  - Number of users per day.

- **Methods**
  - Basic statistics (mean, median, variance);
  - Aggregations (group by, counts);
  - Data visualization.

Diagnostic Analysis (Why did it happened?)

- Investigates **causes and relationships** behind observed outcomes.

- **Typical outputs**
  - Correlations
  - Comparisons between groups
  - Root-cause explanations

- **Examples**
  - Why did sales drop in March?
  - Why did website traffic decrease after the redesign?
  - Why did customer satisfaction decline this quarter?

- **Methods**
  - Correlation analysis
  - Segmentation
  - Hypothesis testing
  - Drill-down analysis

Predictive Analysis (What is likely to happen?)

- Uses historical data to **forecast future outcomes**.

- **Typical outputs**
  - Predictions;
  - Probabilities;
  - Forecasted trends.

- **Examples**
  - What will next month's demand be?
  - Which customers are at highest risk of leaving in the next 3 months?

- **Methods**
  - Regression models;
  - Classification algorithms;
  - Time-series forecasting;
  - It can imply Machine learning models.

Prescriptive Analysis (What should we do?)

- Recommends **actions or decisions** based on predictions and constraints.

- **Typical outputs**
  - Optimal actions;
  - Decision rules;
  - What-if scenarios.

- **Examples**
  - What price should we set?
  - Which customers should receive a discount?

- **Methods**
  - Optimization;
  - Simulation;
  - Business rules + Machine Learning.

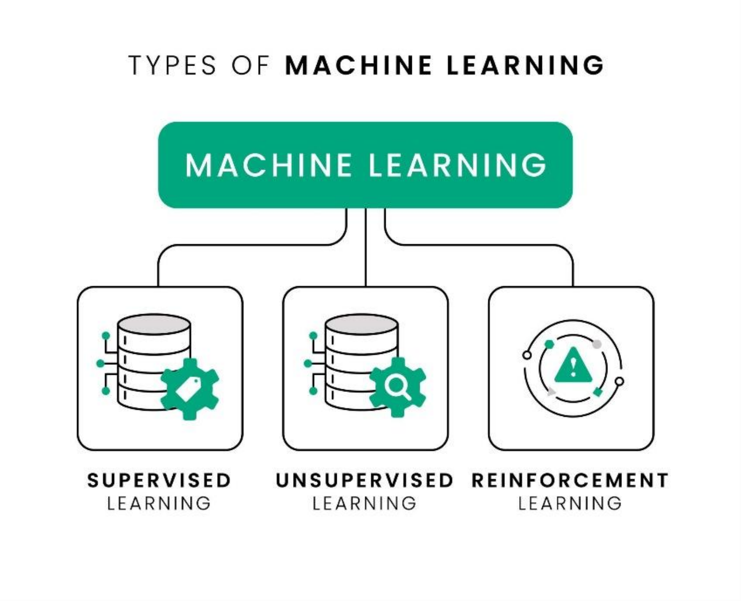
What is Machine Learning?

Faculty of Mathematics and Computer Science

What is Machine Learning?

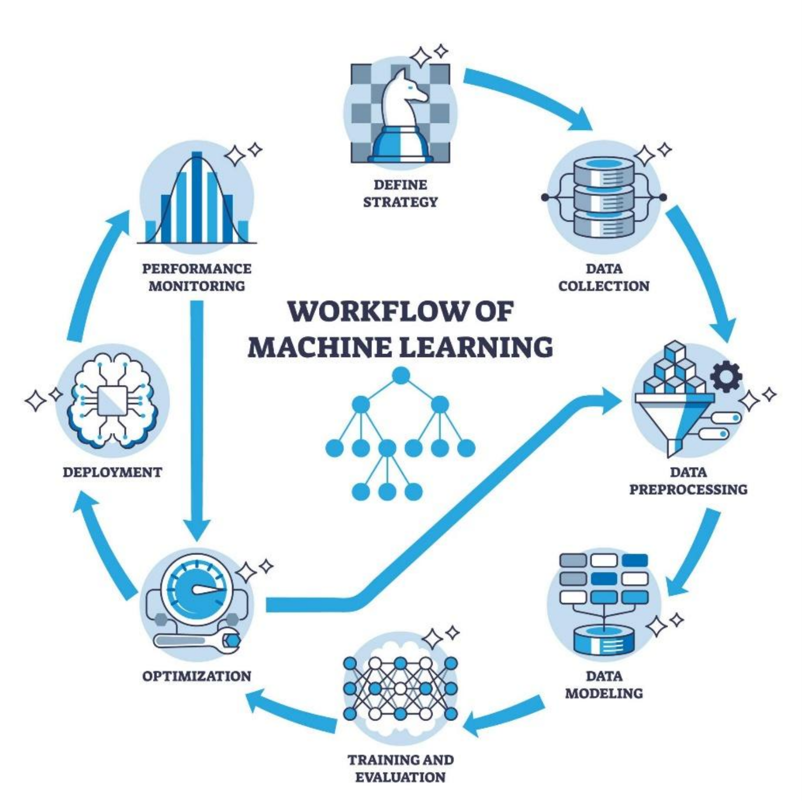
- Machine Learning (ML) is a subset of Artificial Intelligence (AI) that focuses on building systems that **learn patterns from data** and **improve their performance automatically**, without being explicitly programmed with fixed rules.

- Key components of Machine Learn
  - Adaptive learning;
  - Data analysis automation;
  - Iterative algorithms.



What is Machine Learning?

- The role of Machine Learning Professionals:
  - Algorithm development;
  - Data-driven solutions;
  - Cross-disciplinary integration.
  - Optimization & scaling.



What is Data Mining?

Faculty of Mathematics and Computer Science

What is Data Mining?

- **Data Mining** is the process of extracting valuable information from large databases that was previously unknown and using it to make informed business decisions.

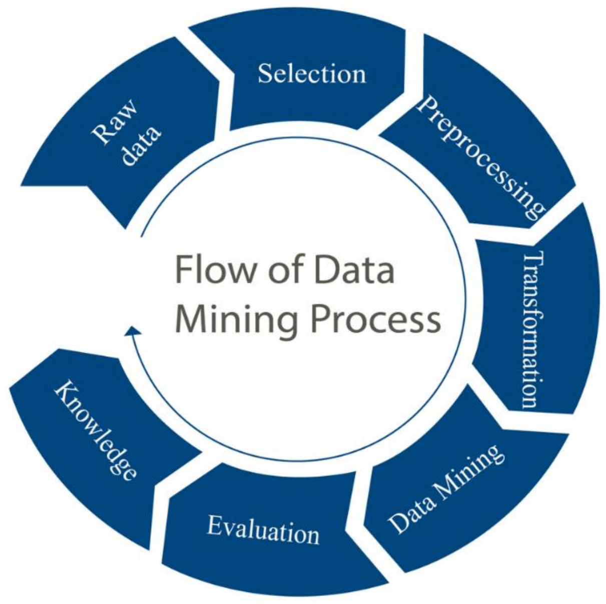
- Why is Data Mining Important?
  - Insight extraction: transforms complex data sets into understandable and actionable information;
  - Decision-making support: helps businesses make data-driven decisions;
  - Pattern recognition : reveals trends and relationships that were previously hidden.

Why Data Mining?

- Fraud detection
- User profile
- Market analysis
- Time-based pattern mining
- Association rules
- House price prediction
- Energy consumption prediction
- Spam detection
- Credit risk detection
- Medical diagnosis
- etc.

What is Data Mining?

- The role of Data Mining professionals:
  - Data transformation
  - Driving innovation
  - Application across fields
  - Communication & decision support



- Data mining professionals don’t just build models — they explain data.

- Data mining does NOT work on raw data

Data Mining vs. Machine Learning

Aspect	Data Mining	Machine Learning
Main Goal	Discover hidden patterns and knowledge from data	Build models that learn and make predictions
Focus	Knowledge extraction	Prediction & automation
Output	Patterns, rules, insights	Trained model
Typical Questions	“What patterns exist in the data?”	“Can we predict future outcomes?”
Techniques Used	Clustering, association rules, anomaly detection	Regression, classification, clustering
Relation	Broader analytical process	Subfield of AI used inside data mining
Human Involvement	More exploratory & interpretative	More algorithm-driven & automated

Data Mining vs. Machine Learning

- **Machine Learning** = set of algorithms that learn from data
- **Data Mining** = process of discovering knowledge from data
- **ML algorithms are tools used inside the data mining process**

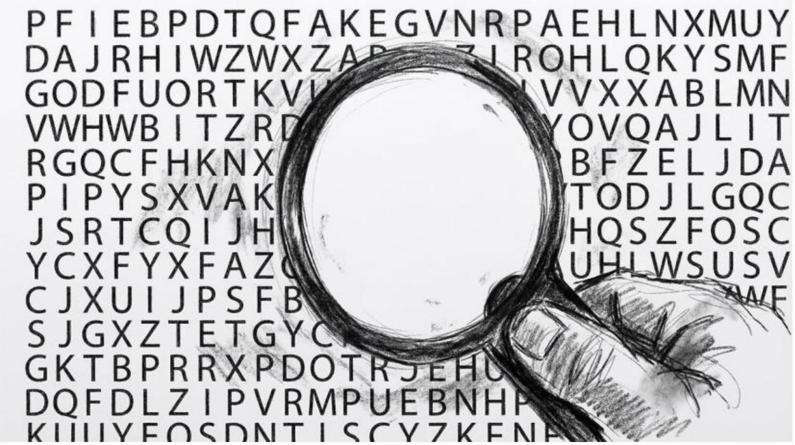
What is Knowledge Discovery?

Faculty of Mathematics and Computer Science

What is Knowledge Discovery?

- **Knowledge Discovery** is the process of identifying valid, novel, and useful patterns in data, transforming raw data into meaningful information.

- **Key Concepts of Knowledge Discovery:**
  - Data selection;
  - Data preprocessing;
  - Data mining;
  - Pattern evolution;
  - Knowledge representation.



What is Knowledge Discovery?

- The Data Analyst acts as the **bridge between raw data and knowledge**, ensuring that discovered patterns are **understandable, valid, and useful.**

- The role of a Data Analysist in Knowledge Discovery:
  - Data preparation;
  - Pattern identification;
  - Result interpretation;
  - Knowledge presentation;
  - Business alignment.

Data Science

Faculty of Mathematics and Computer Science

Data Science

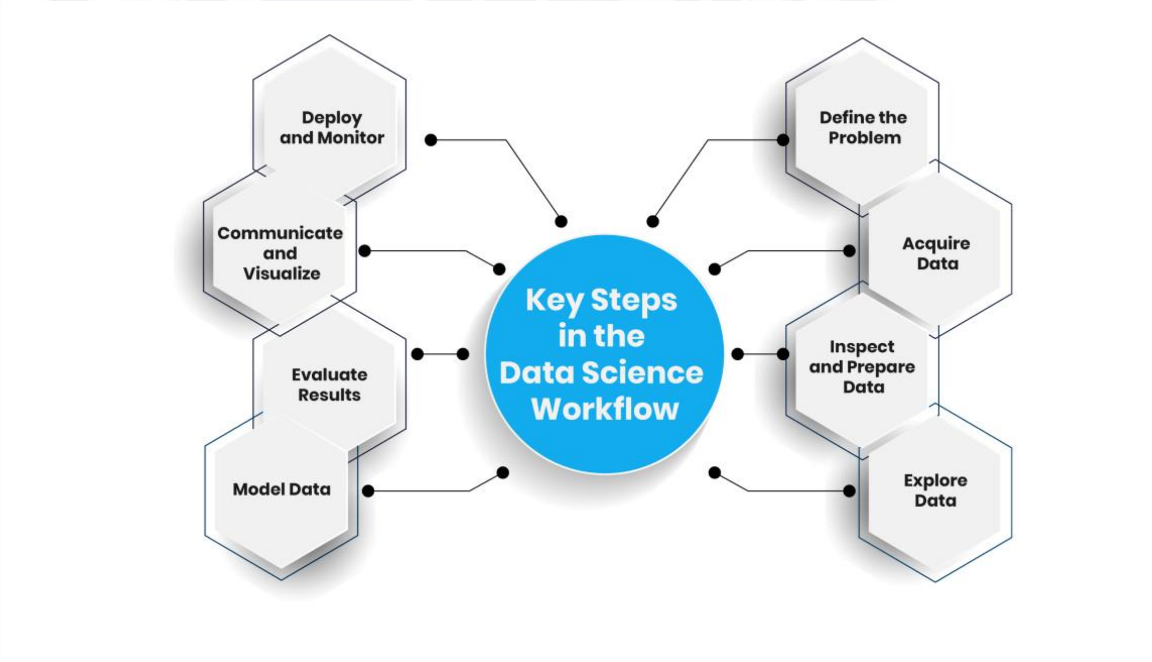
- **Data Science** is a multidisciplinary field that focuses on extracting insights and knowledge from both **structured** and **unstructured data.**

- Data Science integrates:
  - Data Analysis;
  - Data Mining;
  - Machine Learning;
  - Statistics;
  - Domain knowledge;
  - Communication.

- Data Science is the umbrella discipline.

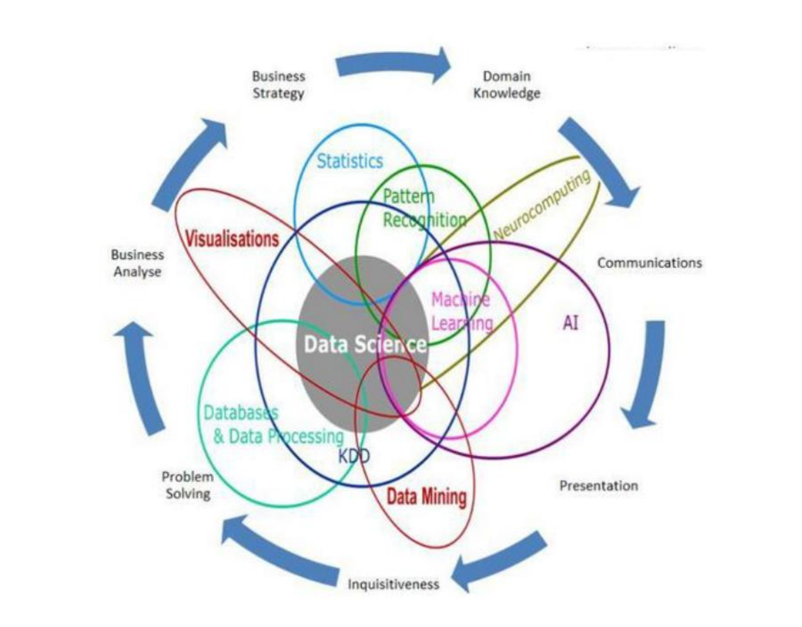
Data Science

- Why is Data Science important?
  - New perspectives;
  - Data preparation;
  - Innovative data capture.



Data Science

- Data Analysis = summarizing and understanding data.
- Machine Learning = training algorithms to predict.
- Data Mining = discovering patterns.
- Knowledge Discovery = interpreting those patterns meaningfully.
- Data Science = combining all of these into a coherent workflow.



Data Science

- The role of data scientists:
  - Developing data products;
  - Extracting insights;
  - Communication & Storytelling
  - Critical thinking
  - Ethical awareness
  - Deployment & Monitoring (Optional)

The Relevance of Data-Driven Disciplines in 2026

Faculty of Mathematics and Computer Science

The Relevance of Data-Driven Disciplines in 2026

- Data-driven disciplines are **core to modern decision-making**

- Data Science is **no longer niche**

- There is **strong and sustained demand** for data scientists and related roles

- **Artificial Intelligence Integration:** The synergy between Data Mining and ML allows for predictive analytics, helping businesses anticipate customer behaviour or detect system anomalies.

The Relevance of Data-Driven Disciplines in 2026

- **Informed Strategic Decision-Making**

- **Ethics, Privacy, and Compliance: as data privacy regulations** tighten, organizations must use Data Mining and Data Analysis responsibly, ensuring compliance while still extracting value.

- **Complexity of Data Structure:** Data is no longer just structured (tables, databases) but increasingly unstructured (text, images, videos) and semi-structured (JSON, XML).

- Data Science is not a trend — it is an infrastructure skill for the modern world.

The Relevance of Data-Driven Disciplines in 2026

- **Jobs:**
  - **Data Scientist;**
  - **Data Analysist;**
  - **Machine Learning Engineer;**
  - **Business Intelligence Analysist;**
  - **AI/ Data Consultant;**
  - **Research Scientist.**

- **Industries Actively Hiring**
  - Finance & Banking (risk, fraud, forecasting)
  - Healthcare (decision support, prediction)
  - Technology & AI companies
  - Marketing & E-commerce
  - Manufacturing & IoT
  - Public sector & smart cities

Where Things Fail In Practice

Faculty of Mathematics and Computer Science

Where Things Fail In Practice

- Dirty Data;
- Biased sampling;
- Wrong evaluation metrics;
- Misinterpretation;
- Overfitting;
- Data leakage.

Types of Data We Will Handle

Faculty of Mathematics and Computer Science

Types of Data We Will Handle

- **Structured data** consists of records (instances) described by a **fixed and predefined set of features**, typically organized in tabular form (e.g., relational database tables, CSV files, Excel spreadsheets).

- **Semi-structured data** refers to data that **does not follow a rigid schema**, where different instances may contain **different sets of attributes**, while still preserving some structural organization (e.g., XML, JSON).

- **Unstructured data** includes data that **lacks a predefined organizational model**, making it unsuitable for direct tabular representation (e.g., free text documents, images, audio recordings, video content).

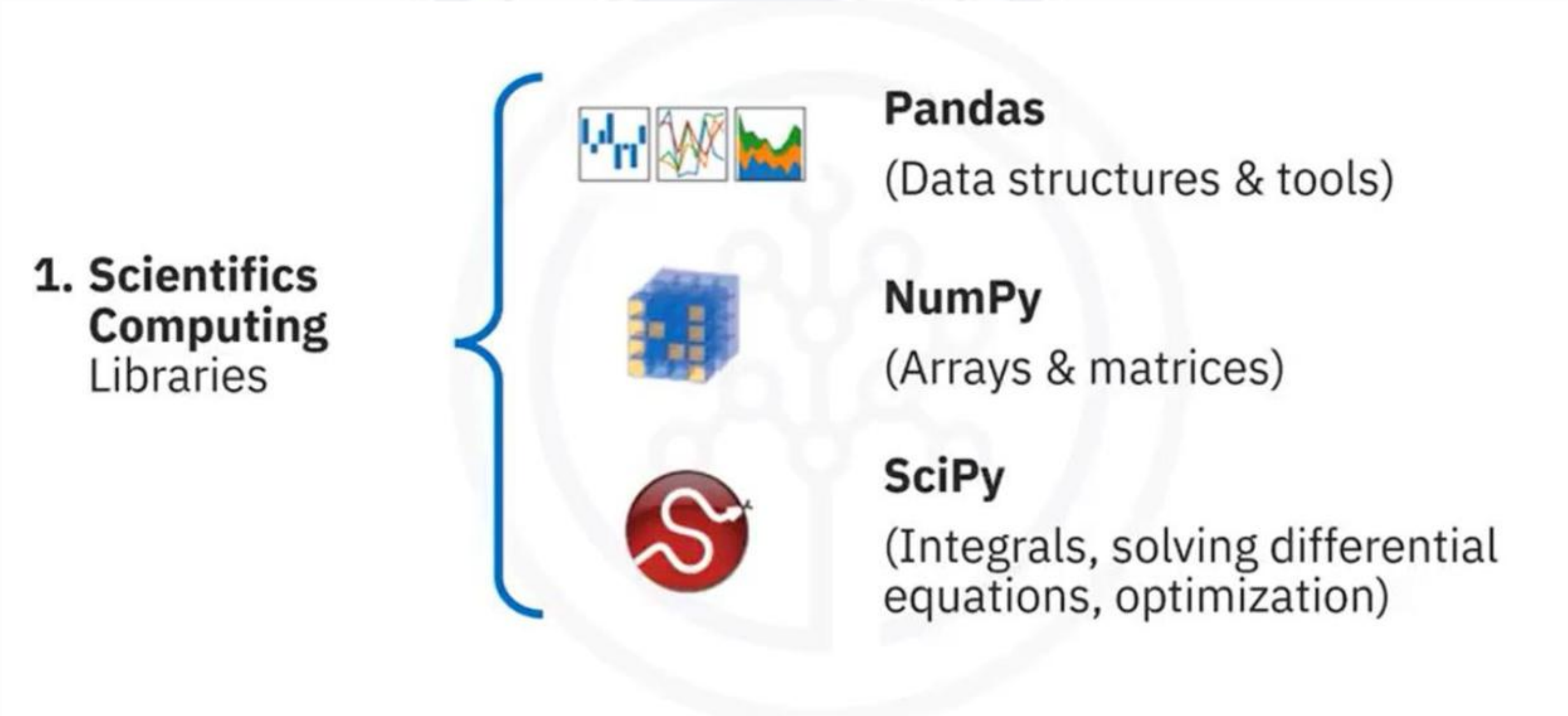
- Small vs Large Data
  - Small data : thousands to millions of rows; fits in memory
  - Large data: millions to billions of records; requires distributed systems; does not fit easily in memory

Data Libraries in Python

Faculty of Mathematics and Computer Science

Data Libraries in Python

- Throughout the semester, all demonstrations and examples will be conducted in Python using Jupyter Notebook.



1. Scientifics Computing Libraries

Pandas

(Data structures & tools)

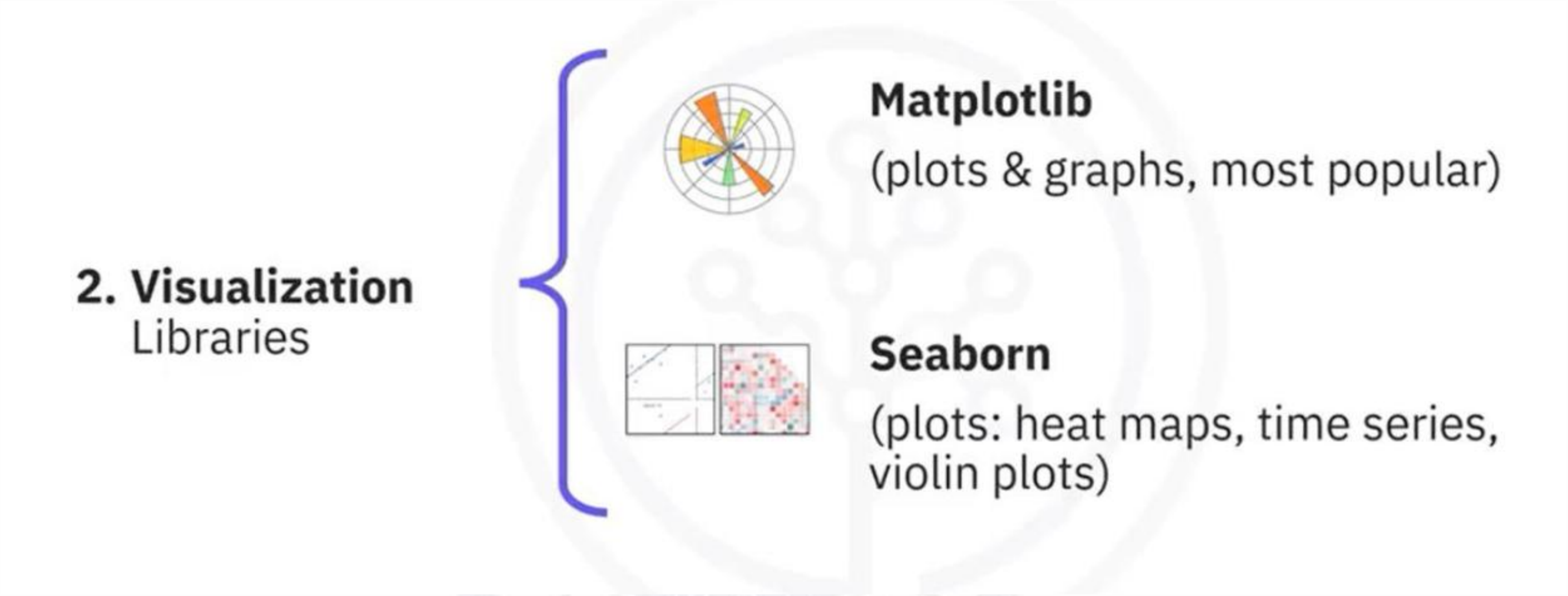
NumPy

(Arrays & matrices)

SciPy

(Integrals, solving differential equations, optimization)

Data Libraries in Python

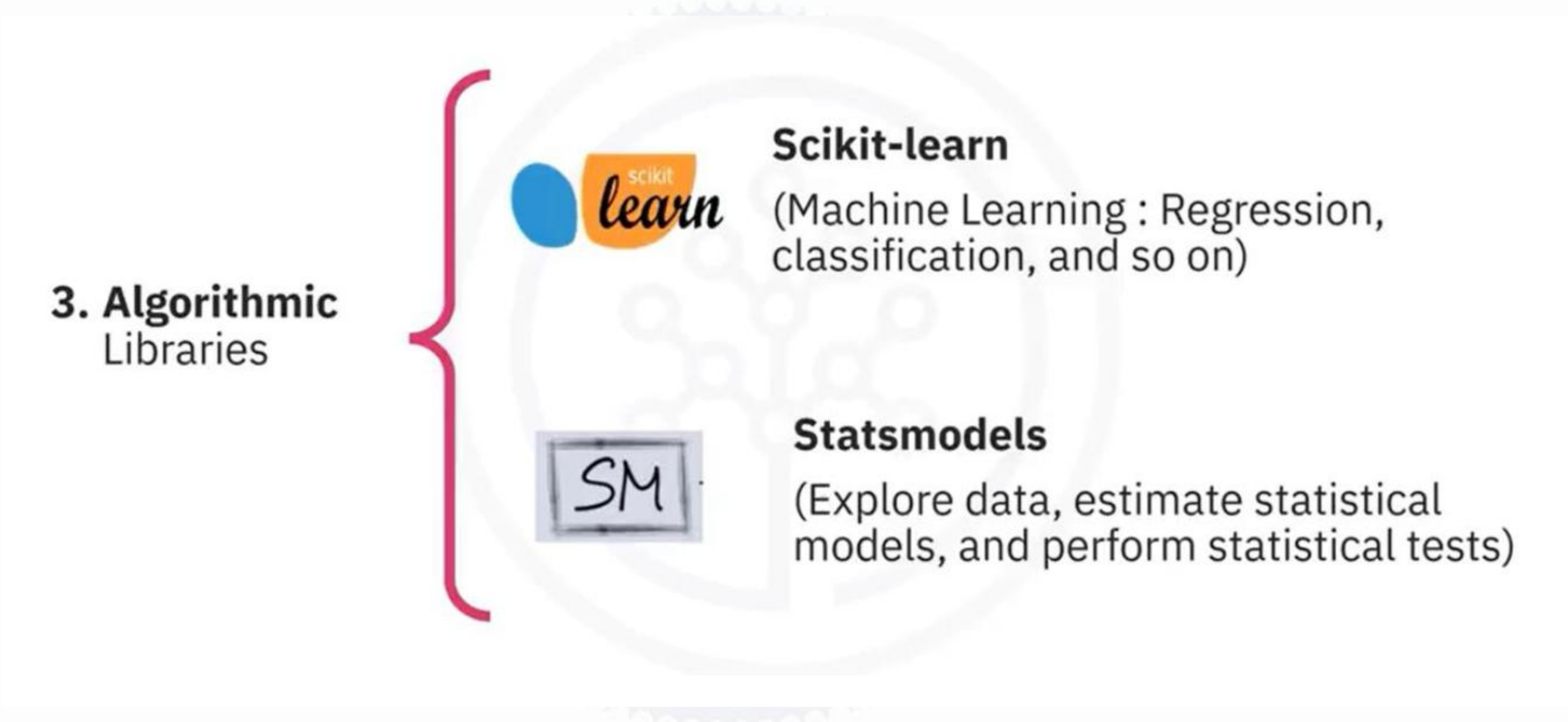


2. Visualization
Libraries

Matplotlib
(plots & graphs, most popular)

Seaborn
(plots: heat maps, time series, violin plots)

Data Libraries in Python



3. Algorithmic Libraries

Scikit-learn

(Machine Learning : Regression, classification, and so on)

Statsmodels

(Explore data, estimate statistical models, and perform statistical tests)

Key Takers

Faculty of Mathematics and Computer Science

Key Takers

- Data ≠ Information ≠ Knowledge
- Data Science is an **ecosystem**, not a single method
- Interpretation and context are essential
- Data Analysis focuses on understanding and summarizing data.
- Data Mining focuses on discovering hidden patterns and structure.
- Knowledge Discovery focuses on interpretation and meaning.
- Machine Learning focuses on learning predictive or decision models.

Thank you for your attention — questions, thoughts, or challenges?

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
BABEȘ-BOLYAI UNIVERSITY

1 Mihail Kogălniceanu Street,
Cluj-Napoca, Cluj, România

www.cs.ubbcluj.ro