Data Analysis and Knowledge Discovery

Lecture 12

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

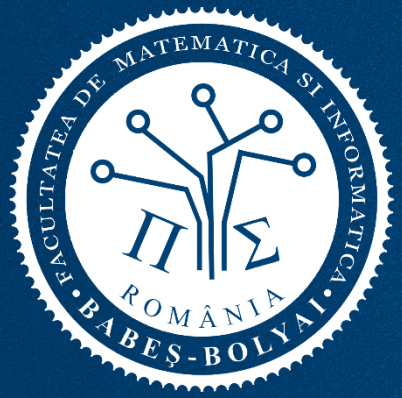
Sergiu Limboi, PhD Teaching Assistant

Motto:“Hidden patterns become valuable when they support real decisions.

Association Rules and Knowledge Discovery in Practice

AGENDA

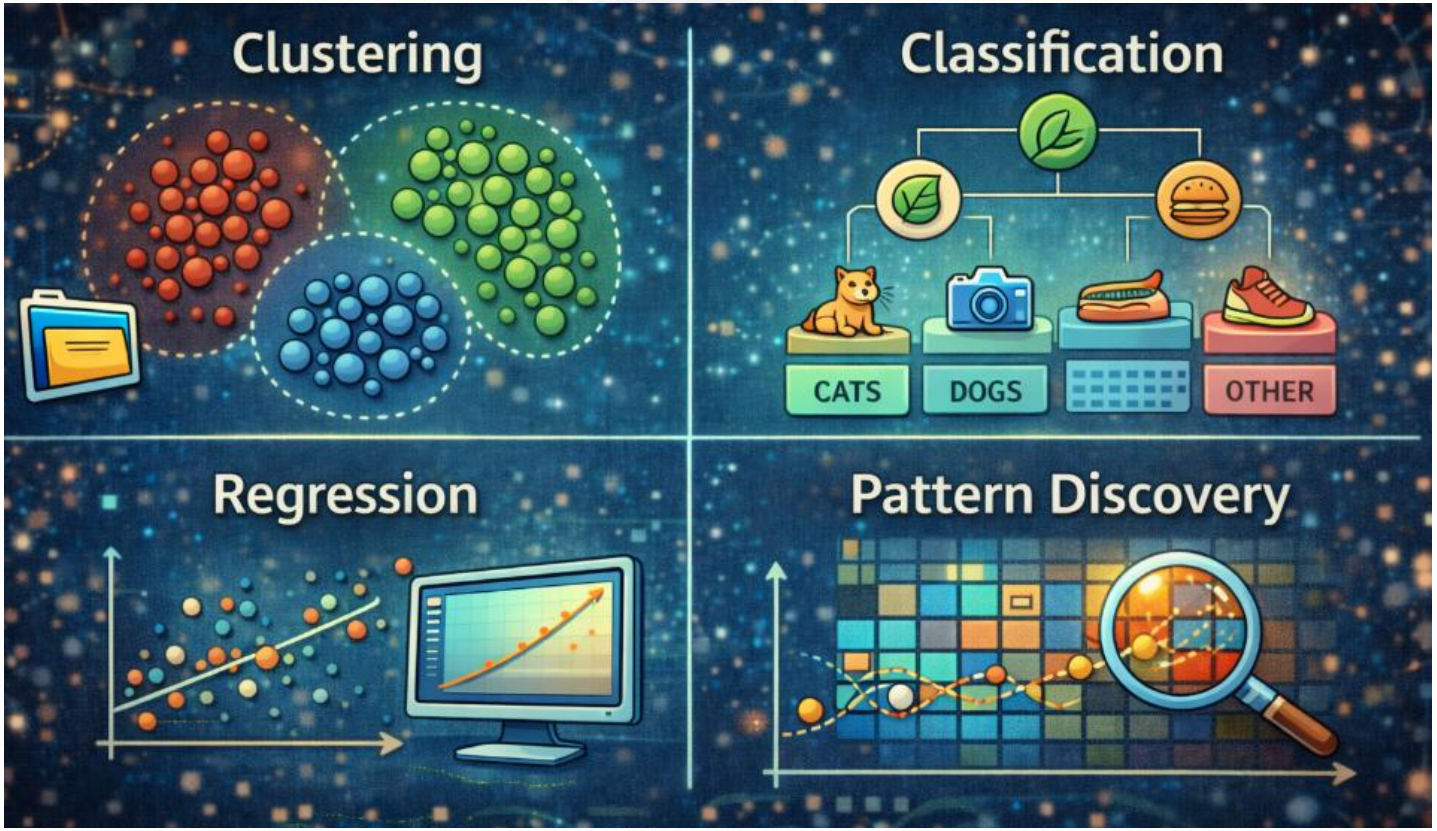
- Pattern Discovery
- Association Rule Mining
- Software Tools for Knowledge Discovery
- Standards in Data Mining
- Exam Details



Pattern Discovery

Faculty of Mathematics and Computer Science

Fundamental Data Mining Tasks



Pattern Discovery

- Goal: Find interesting relationships between variables.

- Instead of predicting a value or assigning a label, pattern discovery focuses on finding hidden structures in datasets.

- The goal is to answer questions such as:
  - Which events occur together?
  - Which behaviors frequently appear?
  - What sequences happen repeatedly?
  - What relationships exist between variables?

Pattern Discovery

- Large datasets often contain complex relationships that humans cannot easily detect.

- For example:
  - A retailer may have millions of transactions.
  - It is impossible for analysts to manually inspect them.

- Pattern discovery algorithms help reveal insights such as:
  - products frequently purchased together
  - browsing behaviors of users
  - sequences of actions before churn
  - relationships between symptoms and diseases

- This transforms raw data into actionable knowledge.

Pattern Discovery

- Types of patterns that can be discovered:
  - Association patterns
    - These show relationships between items that frequently occur together.
    - Example: Customers who buy bread often buy butter.
    - This relationship is expressed like Bread→Butter
    - Retailers use this information for:
      - product recommendations
      - targeted marketing
  - Sequential patterns
    - These capture order relationships in time.
    - Example: User visits product page → reads reviews → buys product
    - Companies like streaming platforms or e-commerce websites use sequential patterns to understand customer journeys.
  - Frequent patterns
    - These identify items or combinations that appear very often in datasets.
    - Example: A supermarket might discover that {milk, cereal} appears in 25% of transactions.
    - Frequent pattern mining helps identify common behaviors in large populations.

Pattern Discovery

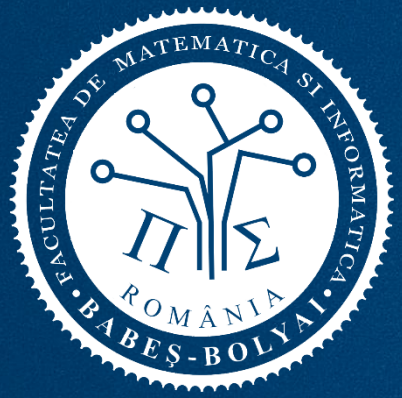
- Example: Market Basket Analysis

Transaction	Item purchased
1	Milk, bread
2	Milk, diapers, beer
3	Bread, butter
4	Milk, diapers, beer

- A pattern discovery algorithm may find diapers→ beer

- This means customers buying diapers frequently also buy beer.

Pattern Discovery



Association Rule Mining

Association Rule Mining

- Association Rule Mining is a data mining technique used to discover relationships between variables in large datasets.
- It identifies items that frequently occur together in transactional data.
- The goal is to uncover rules of the form: $A \rightarrow B$
  - where:
    - **A** = antecedent (if part)
    - **B** = consequent (then part)
- Interpretation: If event **A** occurs, event **B** is likely to occur as well.

Association Rule Mining

- Association rule mining is widely used in **transactional datasets**, where each record represents a **set of items**.

Transaction	Item purchased
1	Milk, bread
2	Milk, diapers, beer
3	Bread, butter
4	Milk, diapers, beer

Example rule

diapers→ beer

- Customers who buy diapers often also buy beer.

- This pattern was discovered in retail data and became a famous example in data mining literature.

Association Rule Mining

- In industry
  - Retail (Market Basket Analysis)
    - Companies analyse shopping baskets to discover product relationships.
    - Laptop → Laptop bag
      Laptop → Mouse
  - E-commerce
    - Platforms such as Amazon use association rules to power recommendations
  - Healthcare
    - Association rules can detect relationships between:
      - symptoms
      - diseases
      - Treatments
  - Web Usage Mining
    - Association rules can identify user behaviour patterns.
    - homepage → product page → checkout

Association Rule Mining

- Before mining rules, we define item sets.
- Item→ A single product or attribute. (example: milk, bread)
- Item set → A group of items appearing together.
  - example: {milk, bread}
- Frequent item set →An itemset that appears frequently in the dataset.
  - {milk, bread} appears in 40% of transactions.
- Once frequent item sets are discovered, we generate rules.
  - Example rule: $milk \rightarrow bread$
  - Interpretation: Customers buying milk are likely to buy bread.

Association Rule Mining- Evaluation metrics

- **Support** measures how frequently the rule appears in the dataset.

$$support(A \rightarrow B) = \frac{transactions\ containing\ A\ and\ B}{total\ transactions}$$

Transactions	Count
Total transactions	100
Transactions containing milk and bread	30

- Support (milk→bread)= 30/100=0.30

- 30% of all transactions contain both items.

Association Rule Mining- Evaluation metrics

- **Confidence** measures how often B occurs when A occurs.

- $confidence(A \rightarrow B) = \frac{support(A \cup B)}{support(A)}$

- Example:
  - transactions with milk and bread : 30
  - Transactions with milk : 50

- Confidence(milk→bread)= 30/50 = 0.60

- 60% of customers who buy milk also buy bread.

Association Rule Mining- Evaluation metrics

- **Lift** measures how strong the relationship is compared to random chance.
- $lift(A \rightarrow B) = \frac{confidence(A \rightarrow B)}{support(B)}$

- Lift=1 → no relationship
- Lift >1 → positive association
- Lift < 1→ negative association

- Example: lift=2 (Means the items occur together twice as often as expected by chance.)

Algorithms for Association Rule Mining

- Mining association rules directly is computationally expensive because the number of possible item sets grows exponentially.

- Example:
  - If we have **100 items**, the number of possible item combinations is:

$$2^{100}$$

- Therefore, specialized algorithms are needed.

Algorithms for Association Rule Mining-Apriori Algorithm

- Apriori is one of the earliest and most famous algorithms for association rule mining.
- Key Idea: If an itemset is frequent, then all of its subsets must also be frequent.
- Example: if {milk, bread, butter} is frequent, then:
  - {milk, bread}, {milk, butter}, and {bread, butter} must also be frequent.
- Steps
  - Find frequent individual items.
  - Generate candidate item sets.
  - Remove those below the support threshold.
  - Repeat for larger item sets.

Algorithms for Association Rule Mining-FP-Growth Algorithm

- FP-Growth improves Apriori by avoiding candidate generation.
- Instead of scanning the database many times, it builds a compressed structure called an FP-tree.
- Advantages:
  - faster than Apriori
  - fewer database scans
  - scalable for larger datasets

Algorithms for Association Rule ECLAT Algorithm

- ECLAT Algorithm (Equivalence Class Clustering and bottom-up Lattice Traversal)
- ECLAT is another algorithm for frequent itemset mining, but it uses a different data representation.
- Instead of using transaction lists, ECLAT uses vertical data format.

Item	Transaction IDs
Bread	T1,T3
Milk	T1,T2
beer	T2,T3

Association Rule Mining

- Association rule mining faces several challenges.
- Combinatorial Explosion→ Number of possible item sets grows exponentially.
- Spurious Patterns→ Some patterns occur due to random chance.
- Interpretability→ Large datasets may produce thousands of rules, many of which are not meaningful.

Software Tools for Knowledge Discovery

Knowledge Discovery in Databases

- Knowledge Discovery in Databases (KDD) is the comprehensive process of extracting valuable insights from extensive information.

- It includes multiple phases such as data selection, cleaning, preprocessing, transformation, data mining, and the interpretation/evaluation of identified patterns.

- KDD encompasses more than mere algorithm application; it is an iterative process that converts raw data into significant and usable knowledge.

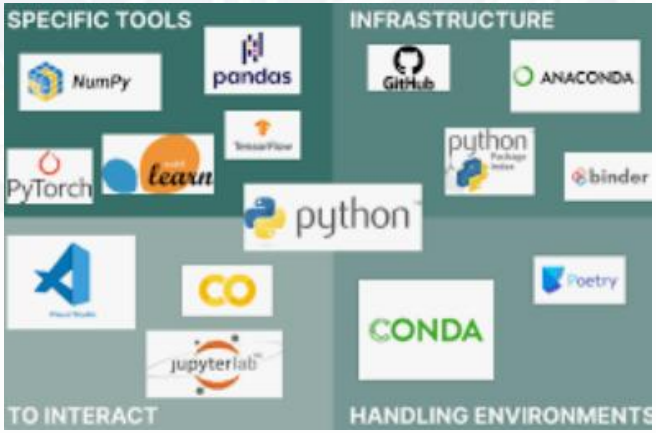
Features of Knowledge Discovery Tools

- Interoperability
  - The capability to assimilate many data sources, including databases, data warehouses, and big data platforms.
  - Compatibility with several data formats (CSV, JSON, ARFF, SQL, etc.).
- Functionality
  - Graphical User Interfaces (GUIs) for user-friendliness.
  - Advanced scripting functionalities for proficient users.
  - Environments for visual programming based on workflows.
- Modularity and Extensibility
  - Support for plug-ins to accommodate supplementary algorithms.
  - Integration with programming languages such as R and Python.
- Scalability
  - Management of extensive datasets (Big Data facilitation through distributed computing frameworks such as Hadoop and Spark).
- Extensive Functionality
  - Utilities for preprocessing and data cleansing.
  - Assistance for diverse data mining techniques: classification, clustering, regression, association rules, and anomaly detection.
  - Tools for model evaluation, visualization, and interpretation.

Python Ecosystem (scikit-learn, pandas, TensorFlow, etc.)

- **Strengths**:
  - Flexibility and wide adoption.
  - Rich ecosystem for machine learning (scikit-learn), deep learning (TensorFlow, PyTorch), and data manipulation (pandas, numpy).

- **Research Applications**:
  - Used in both industrial and academic projects, including real-time analytics, natural language processing, and recommender systems (*Pedregosa et al., 2011*).



WEKA (Waikato Environment for Knowledge Analysis)

- **Origin**: University of Waikato, New Zealand.
- **Type**: Open-source, Java-based.

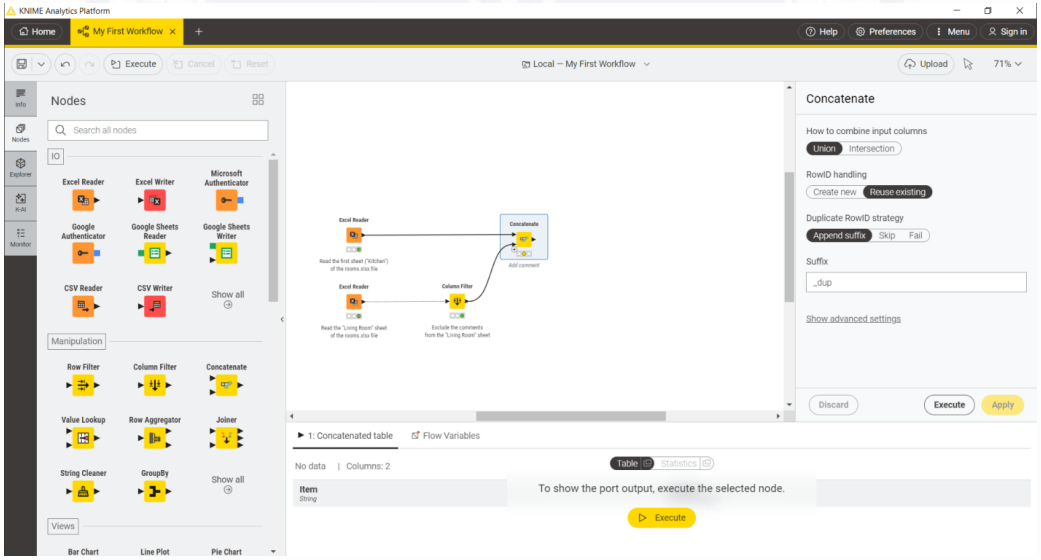
- **Strengths**:
  - Rich collection of algorithms for classification, regression, clustering, and association rule mining.
  - Visualization tools and preprocessing filters.
  - Easy to use GUI, as well as command-line interface.
- **Use Case in Literature**:
  - Widely used in academic experiments (*Manevitz & Yousef, 2001*) for benchmarking machine learning algorithms.
  - https://ml.cms.waikato.ac.nz/weka/Witten_et_al_2016_appendix.pdf

RapidMiner

- **Type**: Open-source with commercial options.
- **Strengths**:
  - Drag-and-drop GUI with support for complex workflows.
  - Broad algorithm support and data handling.
  - Integration with R, Python, and big data platforms.
- **Use in Research**:
  - Popular in healthcare and marketing research (*Hofmann & Klinkenberg, 2013*).
  - Used for customer segmentation and predictive analytics.
  - https://academy.rapidminer.com/learning-paths/get-started-with-rapidminer-and-machine-learning

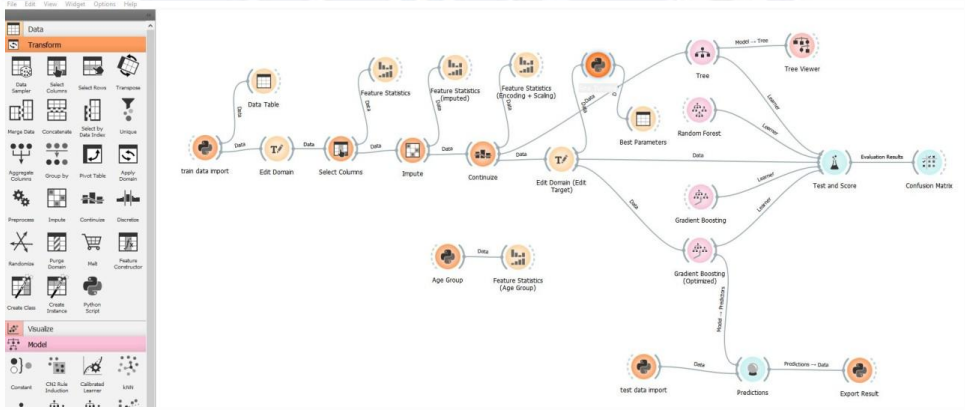
KNIME (Konstanz Information Miner)

- **Strengths**:
  - Visual workflow editor.
  - Large collection of data mining, text mining, and image mining components.
  - Strong integration with R, Python, Weka, and TensorFlow.
- **Use in Scientific Literature**:
  - Frequently used in bioinformatics and chemoinformatics (*Berthold et al., 2008*).
  - https://www.knime.com/get-started



Orange

- **Type**: Open-source, Python-based.
- **Strengths**:
  - Visual programming and component-based design.
  - Emphasis on educational use and prototyping.
  - Interactive data visualizations.
- **Academic Use**:
  - Extensively used in teaching and introductory data mining research projects.
- https://orangedatamining.com/



R and RStudio

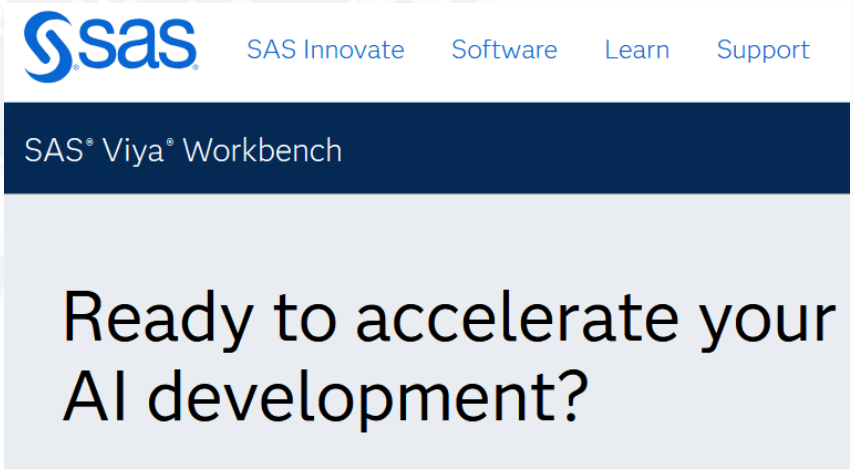
- **Strengths**:
  - Extensive statistical and data mining packages: caret, randomForest, e1071, arules, etc.
  - Powerful for statistical modeling, visualization, and report generation (e.g., using R Markdown).
- **Use in Literature**:
  - Used in environmental science, epidemiology, and financial data mining studies.

- https://rstudio-education.github.io/hopr/starting.html

SAS

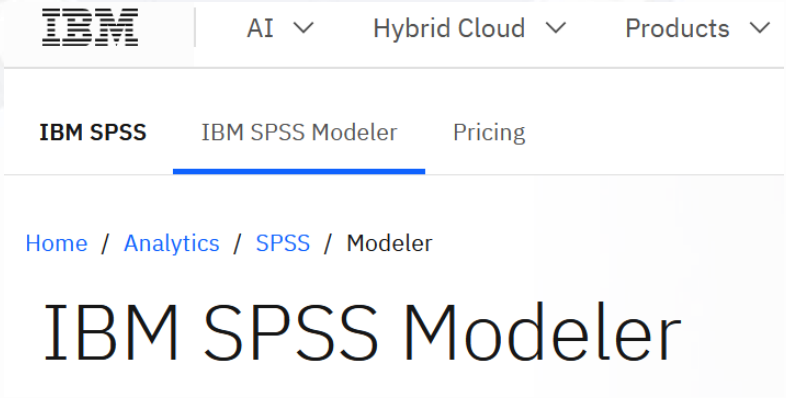
- **Type**: Commercial.
- **Strengths**:
  - Comprehensive, scalable, and secure enterprise-level KDD tool.
  - Strong support for automated model building and deployment.
- **Use in Industry**:
  - Extensively used in sectors such as finance, insurance, and healthcare for risk modeling, fraud detection, and customer analytics.

https://www.sas.com/



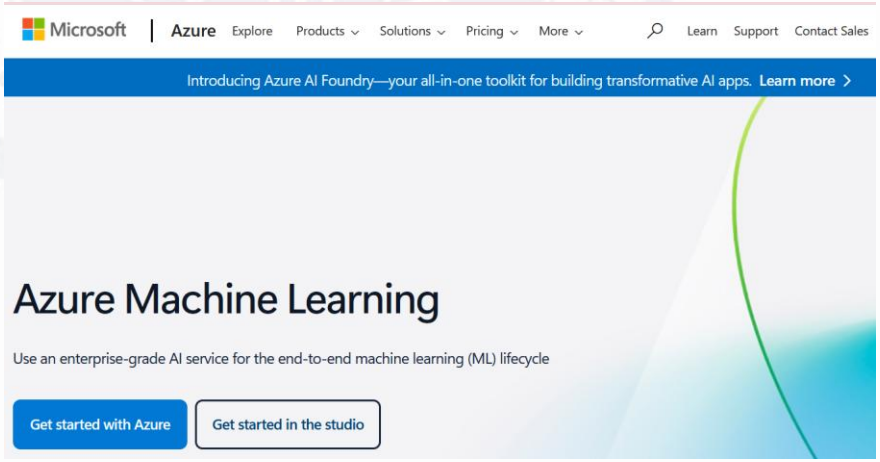
IBM SPSS Modeler

- **Type**: Commercial.
- **Strengths**:
  - GUI-based, suited for business users and analysts.
  - Emphasis on predictive modeling and text analytics.
- **Use Cases**:
  - Applied in social sciences and marketing analytics.
  - https://www.ibm.com/products/spss-modeler



Microsoft Azure Machine Learning Studio

- Microsoft Azure Machine Learning Studio, a cloud-based application, provides a drag-and-drop interface for creating, training, and deploying machine learning models.
- Azure offers a wide range of machine learning and deep learning techniques and works seamlessly with other Microsoft services, making it an excellent choice for enterprise-level applications.



Comparative Analysis

Tool	Ease of Use	Algorithm Variety	Scalability	Community Support
WEKA	High	Medium	Low	High
RapidMiner	Very High	High	Medium	High
KNIME	High	High	High	High
R	Medium	Very High	Medium	Very High
Python	Medium	Very High	High	Very High
SAS Enterprise Miner	Medium	High	High	Medium
IBM SPSS Modeler	High	Medium-High	Medium	Medium
Microsoft Azure ML Studio	Very High	High	Very High (Cloud)	High

Trends in KDD Tools

- AutoML Integration: Automated machine learning functionalities designed to minimize human involvement in model selection and optimization.
- Explainable AI (XAI): Emphasis on model interpretability and transparency.
- Cloud-based Knowledge Discovery in Databases Tools: Scalability and accessibility through systems such as Google Cloud AutoML and Azure ML Studio.
- Specialized Tools: Customized KDD frameworks for fields like as biology, finance, and cybersecurity.

How to choose?

- Software solutions for knowledge discovery are essential for facilitating data-driven decision-making.
- The scientific literature emphasizes the significance of tools that are scalable, extendable, and user-friendly, while also facilitating advanced analytics.
- Open-source platforms such as KNIME, RapidMiner, and Python have democratized access to data mining capabilities, although commercial solutions like SAS and IBM SPSS maintain dominance in enterprise settings.
- The choice of a KDD tool should be determined by:
  - The complex nature and scale of data.
  - The necessary level of analysis.
  - The scope of application.
  - The users' technological proficiency.
  - The integration of methodological principles with advanced software tools results in more dependable, informative, and actionable knowledge discovery outcomes.

Standards in Data Mining

Standards in Data Mining

- Standards in data mining define established methods, rules, and best practices that guarantee consistency, reproducibility, and quality in the data mining process.

- These standards include data management, algorithm execution, model assessment, and outcome documentation.

Areas of Standards in Data Mining

- Data Representation and Format Standards:
  - A fundamental element of data mining is the assurance that data is represented in a manner conducive to efficient processing and analysis.
  - Various data formats are frequently employed in data mining, including CSV (Comma-Separated Values), ARFF (Attribute-Relation File Format), and XML.
  - The standardization of these formats facilitates interoperability among various software applications and systems.
- Data Preprocessing Standards:
  - Data preprocessing consists of cleaning, manipulating, and preparing data for analysis.
  - Standards in this area specify how missing values, outliers, and inconsistencies should be handled.
  - For example, data normalization or standardization is frequently standardized to ensure dataset comparability.

Areas of Standards in Data Mining

- Algorithm Standardization:
  - In order to assure reproducibility and accuracy, algorithms must be used consistently.
  - Certain techniques, such as k-means clustering, decision trees, and association rule mining, have been standardized in a variety of software applications.
  - However, there are established guidelines for measuring algorithm performance, such as cross-validation, confusion matrices, and ROC curves.
- Model Evaluation and Validation:
  - Standards for determining the quality and accuracy of data mining models are critical.
  - Precision, recall, F1 score, AUC (Area Under the Curve), and RMSE (Root Mean Squared Error) are all standardized metrics that serve as uniform benchmarks for model evaluation.

Areas of Standards in Data Mining

- Data Mining Process Standards: Data mining frequently adheres to a set of standard steps (CRISP-DM, for example).
- Cross-industry Standard Process for Data Mining (CRISP-DM): A commonly used methodology for data mining that outlines the following steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.
- SEMMA (Sample, Explore, Modify, Model, and Assess): SAS Institute developed this process, which is widely used for statistical data mining.
- These approaches ensure that a structured and methodical approach is taken during data mining.

Exam Details

Exam Details

- Written Exam – Students must obtain a minimum grade of 5 in order to pass the exam.

- Exam Dates:
  - Data Science: the 17th of June 2026, 17:00- Room C335
  - AI4CI & Erasmus: the 18th of June 2026, 17:00- Room C335

- Secondary dates:
  - Data Science: the 18th of June 2026, 17:00- Room C335
  - AI4CI & Erasmus: the 17th of June 2026, 17:00- Room C335

- Retake Session (all programs)
  - The 7th of July 2026, 17:00-Room C335
  - The evaluation consists exclusively of the written exam; no project presentation will take place, as the project was considered part of the semester work.

- **Important**
  - Participation in the **secondary date** is allowed **only in special cases and with my prior approval, otherwise you cannot attend the secondary date!**
  - To successfully complete the course, students must obtain a minimum grade of 5 in both the project and in the written exam.

EXAM OVERVIEW

DURATION
1h 30 min

COVERAGE
All lectures are required for the exam, except the invited guest sessions.

TYPES OF EXERCISES

1. MULTIPLE CHOICE
Test your understanding of key concepts and methods.

2. MINI CASE STUDIES
Analyze a scenario and design an appropriate solution.

3. INTERPRETATION EXERCISES
Interpret results, tables, and outputs from analytical models.

4. SMALL PROBLEMS
Solve practical problems (e.g., clustering, anomaly detection, recommendation algorithms, association rules, compute metrics, etc.).

5. PLOT INTERPRETATION
Analyze clustering, data quality, data distribution, or other EDA steps.

6. RESEARCH-ORIENTED CRITICAL THINKING QUESTIONS
Apply concepts, evaluate approaches, and justify decisions.

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