# *The Data Analysis Lifecycle:* *Project Management, Methodologies, and Decision Making*

01.04.2025
Cluj-Napoca

# About me

## Experience

2014 ----------------2017 ----------------2018 ----------------present

business analyst--- team leader --------delivery manager----

## Studies

2013 ----------------2015 ----------------2024-------present

BSc Statistics &
Economic Forecasting
BBU

MSc
Econometrics
BBU

Executive MBA
Maastricht University

## Other activities

- Founder member of ASECU Youth -Students Association of South and Eastern Europe and the Black Sea Region Economic Universities 2011 → present
- Founder Member and Volunteer - The Student College for Academic Performance BBU (2013-2015)
- Volunteer - The Organization of Economics Students (2010 – 2014)

# Structure of the presentation

- Methodologies
- Types of Data Analysis
- Questions and Debate

# You will learn/ find:

- **Structured Approach to Data Mining**

- **Problem-Solving Framework:** CRISP-DM helps students break down complex data problems into manageable parts. They'll learn how to define a problem, analyze it from different angles, and choose the right methods and tools.

- **Collaboration Between Business and Technical Teams:** CRISP-DM emphasizes clear communication between data scientists and business stakeholders. Students will learn how data science should align with business goals and how cross-functional teams work together.

- **Data-Driven Decision-Making:** Through CRISP-DM, students will see how data analysis informs strategic decisions. They'll learn how to use data to make actionable insights that impact business operations.

- **Hands-On Data Preparation and Modeling:** As students progress through the stages, they'll gain experience in data cleaning, transformation, and using machine learning algorithms for modeling, which are vital skills in the industry.

- **Evaluation and Improvement:** They'll learn how to evaluate models and results, refine processes, and adjust strategies to improve outcomes. This will help them understand iterative work in data science.

- **Critical Thinking and Problem Exploration:** The methodology encourages exploration, questioning assumptions, and iterative improvement, fostering critical thinking when working with data

# Data analysis

- **Descriptive Analytics**: Summarizing past data.
- **Diagnostic Analytics**: Understanding why something happened.
- **Predictive Analytics**: Forecasting future trends using ML models.
- **Prescriptive Analytics**: Recommending actions based on predictions.
- **Big Data & AI**: Role of machine learning in automating insights.

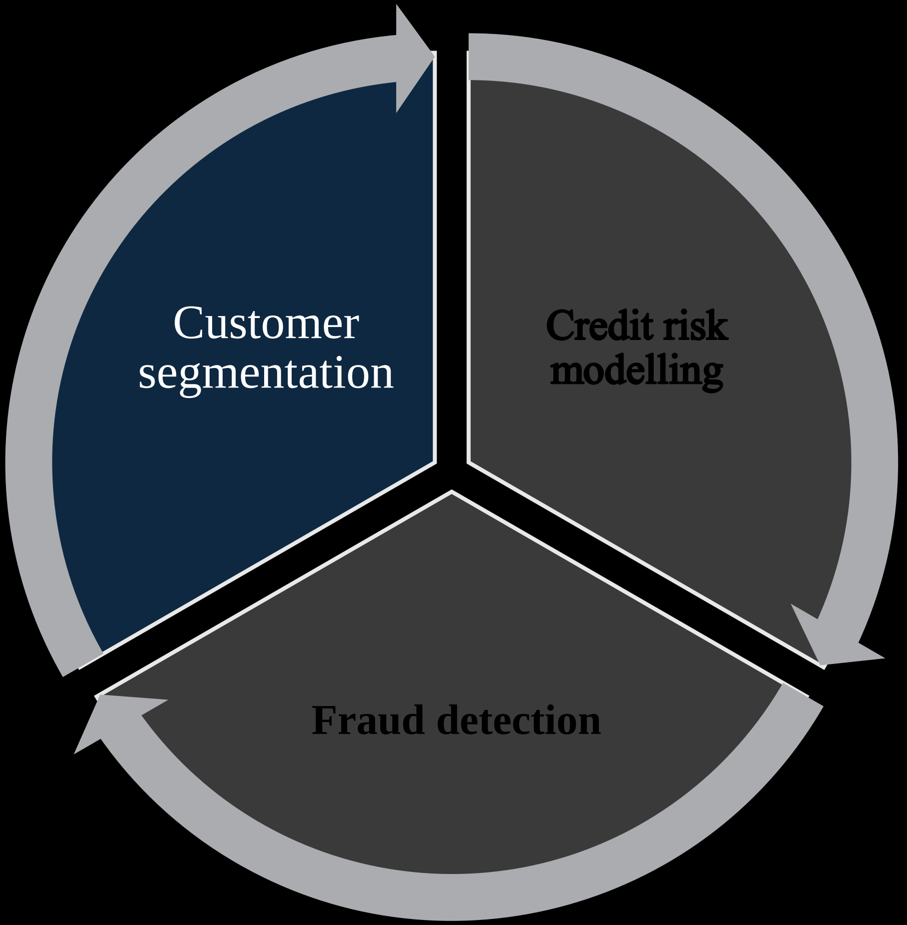
✓ Batch data

✓ T-1 or T-n

✓ Near Real – Time

✓ Real - time

# I. Applications of Data Science in Banking



- BI & reports  analysis
- Market studies
-  Statistical models
- Actions based in data
  - Stakeholders
  - Customers

# II. Methodologies in Data Science

| Aspect | CRISP-DM | SEMMA | KDD (Knowledge Discovery in Databases) | TDSP (Team Data Science Process) | Agile Data Science |
|---|---|---|---|---|---|
| **Developed By** | IBM, Daimler-Benz, NCR (1996) | SAS Institute | Academia (1990s) | Microsoft | Industry (modern approach) |
| **Best For** | Business-focused, structured projects | Statistical modeling, Data Mining | Research and large-scale data processing | Enterprise AI/ML projects, MLOps | Rapid experimentation & startup environments |
| **Used In Banking?** | For credit scoring, fraud detection, compliance, risk modeling and customer segmentation | | Limited, more academic use cases | Yes, for scalable banking AI solutions | Yes, for real-time analytics, fintech solutions |
| Phases | Business understanding, Data understanding, Data preparation, Modeling, Evaluation, Deployment | Sample, Explore, Modify, Model, Assess | Selection, Preprocessing, Transformation, Data mining, Interpretation | Business understanding, Data acquisition, Modeling, Deployment, Feedback | Continuous cycles of data gathering, modeling, deployment, improvement |
| Focus | End-to-end business-driven process | Technical ML modeling | Pattern discovery from large datasets | Collaborative, scalable DS projects | Iterative, fast-paced modeling |
| Iterative? | Partially (some feedback loops) | No (linear, stepwise approach) | No (sequential discovery) | Yes, supports iteration & feedback | Yes, follows Agile methodology |
| Deployment Focus | Explicit deployment phase | Focuses more on modeling | Mainly research-based | Integrates MLOps & DevOps | Continuous, cloud-based deployment |

# II. A. **CRISP-DM phases**

### 1. Business Understanding

- Define the project objectives and business goals.
- Identify success criteria and constraints.

*Distribution of new customers per month*

| | | | A | B | C | D |
|---|---|---|---|---|---|---|
| 1 | total new customers | | 100 | 100 | 100 | 100 |
| 2 | branch | | 80 | 70 | 70 | 60 |
| 3 | online assisted | | 0 | 10 | 20 | 30 |
| 4 | online non-assisted | | 20 | 20 | 10 | 10 |
| 5 | % online | = (3+4)/1 | 20.00% | 30.00% | 30.00% | 40.00% |
| 6 | % online assisted | = 3/1 | 0.00% | 10.00% | 20.00% | 30.00% |
| 7 | % online non-assisted | = 4/1 | 20.00% | 20.00% | 10.00% | 10.00% |

**MIT**

“participatory sensemaking”

•Listening
•Analyze
•Synthesize

→ **Which bank has the most online customer onboarding?**

# II. A. CRISP-DM phases

## 2. Data Understanding

### Data Collection

- Gather the relevant datasets from different sources.
- Verify that the data aligns with business objectives.

### Data Description

- Summarize key characteristics such as:
  - Number of records (rows) and attributes (columns).
  - Data types (numeric, categorical, text, etc.).
  - Basic statistics (mean, median, standard deviation).

### Data Exploration

- Perform exploratory data analysis (EDA) using:
  - Visualizations (histograms, box plots, scatter plots).
  - Correlations and relationships between variables.
  - Identification of outliers or anomalies.

### Data Quality Assessment

- Identify missing values and inconsistencies.
- Check for duplicates and redundant information.
- Detect potential biases or errors in data collection.

most common aspects of the
**Data Understanding**

✓ Missing Data
✓ Outliers
✓ Data Inconsistencies
✓ Data Bias & Class Imbalance
✓ Data Distribution & Skewness

**How can data understanding issues be addressed?**

# II. A. CRISP-DM phases

## 3. Data Preparation

**Data Cleaning**

- **Handling Missing Values**
  - Remove records with too many missing values.
  - Impute missing values (mean, median, mode, or predictive methods).
- **Handling Outliers**
  - Detect using box plots, z-scores, or the IQR method.
  - Remove or transform extreme values.
- **Fixing Inconsistencies**
  - Standardize formats (e.g., date formats, text case).
  - Remove duplicates.

**Data Transformation**

- **Feature Engineering**
  - Create new variables.
  - Convert categorical data into numerical (one-hot encoding, label encoding).
- **Scaling & Normalization**
  - Normalize features for models sensitive to magnitude differences.
  - Log transformations for skewed data.
- **Reducing Dimensionality**
  - Remove redundant information.

**Data Integration**

- **Merging multiple data sources**
  - Join datasets using common keys (e.g., customer ID, transaction ID).
  - Ensure consistency across sources.
- **Dealing with Schema Mismatches**
  - Align column names and data types across datasets.

**Data Formatting & Structuring**

- Convert data into a structured format (CSV, SQL tables, JSON).
- Ensure correct data types (integers, floats, dates).
- Optimize storage by reducing memory usage.

**What is the recommended order for applying data preparation techniques?**

# II. A. CRISP-DM phases

### 4. Modelling

- Selecting,
- Building and
- fine-tuning machine learning or statistical models

to analyze the prepared data and generate predictions or insights.

### 5. Evaluation

- assesses the performance of the trained models to ensure they align with business objectives and deliver reliable results.

### 6. Deployment

- the trained and evaluated model is implemented in a real-world environment to generate actionable insights or automate decision-making.

*This topic will be explored in greater detail in the afternoon.*

# What are the deliverables of the CRISP-DM methodology?

**1. Business Understanding Phase**

- Project Goals and Objectives: Clear definition of the business problem and objectives.
- Success Criteria: Criteria to measure the success of the data mining project.
- Business Requirements: A detailed specification of the business needs and data-related requirements.

**2. Data Understanding Phase**

- Data Collection Report: Summary of the data sources and the data collected.
- Initial Data Exploration: Descriptive statistics, visualizations, and basic insights from the data.
- Data Quality Report: Identifying issues like missing values, inconsistencies, and outliers.
- Data Summary: A thorough understanding of the dataset’s attributes and structure.

**3. Data Preparation Phase**

- Data Cleaning Report: Detailed account of how missing values, outliers, and inconsistencies were handled.
- Data Transformation Documentation: Information about feature engineering, normalization, encoding, etc.
- Prepared Dataset: A clean and well-organized dataset ready for modeling.

**4. Modeling Phase**

- Modeling Approach and Algorithms: Documentation of selected models, algorithms, and techniques.
- Trained Models: The actual models built and their respective hyperparameters.
- Performance Metrics: Evaluation of model performance based on training and testing data.

**5. Evaluation Phase**

- Model Evaluation Report: A comprehensive analysis of model performance and effectiveness.
- Comparison of Models: If multiple models were tested, this document compares their strengths and weaknesses.
- Assessment of Business Goals: A report on how well the models meet the business objectives.
- Recommendations: Suggestions on the next steps, model adjustments, or business strategies.

**6. Deployment Phase**

- Deployment Plan: Detailed steps for integrating the model into the production environment.
- Final Model: The final, fully trained and tested model ready for deployment.
- Performance Monitoring Plan: Strategy for tracking the model's performance post-deployment.
- User Documentation and Training Materials: Guides for stakeholders or end-users on how to use the model.

- Stakelders
- Business owner
- Data analyst
- Statistician
- Data analyst’s boss
- GDPR officer
- IT

Which step of the CRISP-DM methodology do you think takes the longest, and why?

Order of the steps in the **CRISP-DM** methodology based on their relative duration, from longest to shortest:

1. Data preparation
2. Data Understanding
3. Modelling
4. Evaluation
5. Business Understanding
6. Deployment

# Business impact

- **Strategic Decision-Making**: Cost reduction, efficiency, customer targeting.
- **Risk Management**: Fraud detection in banking.
- **Personalization**: Customer experience improvements.
- **Data Governance** & Compliance: Regulations (GDPR, CCPA, DORA)

# Useful resources

**Data Methodologies**

- **Data Science Central** – A great resource for data science methodologies, including best practices, tutorials, and case studies. https://www.datasciencecentral.com/
- **Towards Data Science** – A Medium publication that shares articles on data methodologies, data science techniques, and case studies. https://towardsdatascience.com/
- **KDnuggets** – Offers articles on data science, machine learning, and best practices. https://www.kdnuggets.com/

**CRISP-DM**

- **CRISP-DM Website** – Official site for the CRISP-DM methodology, including detailed guides and case studies. https://www.crisp-dm.org/
- **CRISP-DM 1.0 Handbook** – A complete guide to CRISP-DM, downloadable from various educational sites like Springer. https://link.springer.com/chapter/10.1007/978-3-642-15505-8_6
- **Data Mining: Practical Machine Learning Tools and Techniques** – Book by Ian Witten and Eibe Frank. Covers CRISP-DM in detail. Available on Amazon or through academic libraries.

# ***Feedback | Q & A***

vasile.anton.25@gmail.com

01.04.2025
Cluj-Napoca

# Assignment

## The Paradox of Data Quality and Data Science: When Accuracy Clashes with Practicality

**Format Requirements:**

The written assignment must be 3-4 pages long, excluding title page, table of contents, introductory notes, and appendices.

The document should be formatted with 1.5 line spacing, using Times New Roman or Arial, size 12.

If many diagrams, figures, tables, or charts are included in the main text, students must compensate by increasing the overall page length accordingly.

Proper APA referencing is mandatory.

**Assignment Overview:**

This paper explores the controversial aspects of data quality techniques, particularly the trade-offs between accuracy, consistency, timeliness, and cost. While organizations strive for high-quality data, achieving perfection is often impractical. Students will analyze real-world cases where data quality improvements led to unintended consequences, such as increased costs, system inefficiencies, or ethical dilemmas.

# Assignment

## The Paradox of Data Quality and Data Science: When Accuracy Clashes with Practicality

**1. Data Cleaning vs. Data Integrity** – Examining cases where aggressive cleaning removes valuable outliers or introduces bias.

**2. Timeliness vs. Accuracy** – Analyzing the trade-off between real-time data processing and potential errors.

**3. Subjectivity in Data Standards** – Discussing how different organizations define "high-quality data" differently, leading to inconsistencies.

**4. The Cost of Data Quality** – Evaluating how data quality investments can lead to diminishing returns or pose challenges for smaller organizations.

**5. Ethical Concerns in Data Quality** – Investigating cases where data modifications skew results, such as in predictive policing or healthcare analytics.

**6. The Accuracy vs. Explainability Paradox -**
Complex machine learning models (e.g., deep learning) are often more accurate but less interpretable, while simpler models (e.g., linear regression) are more interpretable but less accurate

**7. The Automation vs. Human Oversight Paradox -** Automating data science tasks (e.g., automated feature selection, AI-driven decisions) improves efficiency, but full automation can introduce biases or errors that humans could catch

# Assignment

## The Paradox of Data Quality and Data Science: When Accuracy Clashes with Practicality

**Students are required to:**

- Analyze real-world case studies and present possible solutions.
- Conduct a small experiment comparing different data cleaning techniques.
- Write a short essay (maximum 4 pages) reflecting their critical thinking, personal opinions, comparisons, and examples. The essay must include at least 5 references.
- Use of AI Tools (e.g., ChatGPT):
    - Students are allowed to use AI tools such as ChatGPT responsibly for research, idea generation, and writing assistance. However, they must:
    - Ensure their work is original and critically analyzed—AI-generated content should not be copied directly.
    - Properly fact-check AI-generated information before including it in their paper.
    - Cite sources appropriately, especially when using AI-generated references.
    - Maintain their own critical thinking and writing style rather than relying solely on AI.

# Assignment

## The Paradox of Data Quality and Data Science: When Accuracy Clashes with Practicality

**Assessment Criteria:**

Submissions will be evaluated based on:

- Fulfillment of all requirements outlined above.
- Originality and depth of analysis.
- Clarity and readability of the content.
- Practical applicability of the ideas presented.

**Presentation & Debate:**

- At the end of the course, three to four students will be randomly selected to present their work. They may use:
- A PowerPoint presentation (maximum 5 slides) OR
- A verbal presentation summarizing their key findings.

**Presentation format:**

- 10 minutes for presentation
- 10 minutes for debate and discussion with classmates

All students are expected to actively participate in the discussion and debate.