Data Analysis and Knowledge Discovery

Lecture 2

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

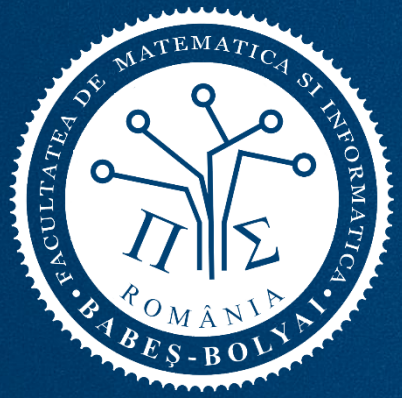
Sergiu Limboi, PhD Teaching Assistant

Motto: From Raw Data to Analysis-Ready Data.

Data Quality & Data Preprocessing

AGENDA

- Warm-Up
- Transition: Reality of Data
- Why Preprocessing Exists
- Data Quality Dimensions
- Teamwork Time 1
- The Data Preprocessing Pipeline
- Raw Data
- Data Cleaning
- Missing Data
- Outliers
- Teamwork Time 2
- Encoding Categorical Data
- Feature Scaling
- Analysis-Ready Dataset
- Reflection Time
- Key Takeaways



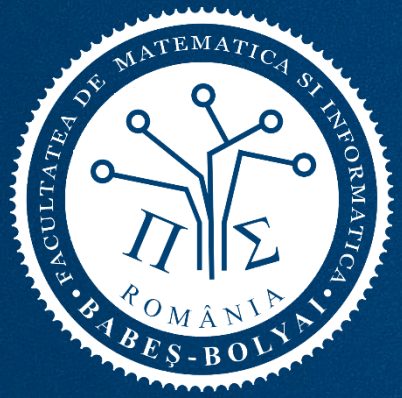
Warm-Up

Faculty of Mathematics and Computer Science

Warm-Up

Go to www.menti.com and enter the code **4149 7336**

or use the QR code



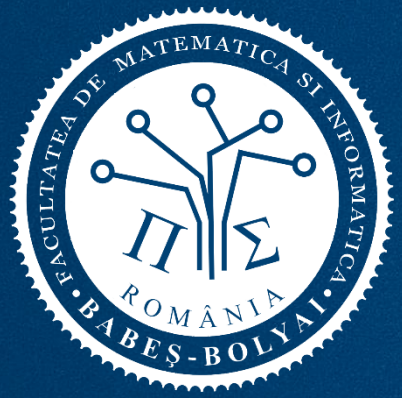
Transition: Reality of Data

Faculty of Mathematics and Computer Science

Transition: Reality of Data

- Real-world data is messy.
- Real datasets suffer from:
  - measurement errors
  - human input mistakes
  - system failures
  - inconsistent standards
- Academic datasets are clean because someone already cleaned them.
- Industry Practice (Consulting / Banking / Healthcare): **60–80% of total project time** is spent on data collection & cleaning
- Large Enterprise / Big Data Systems: **70–80% of total time** is spent on data cleaning & feature engineering

Why Preprocessing Exists



Faculty of Mathematics and Computer Science

Why Preprocessing Exists

- Algorithms assume:
  - numerical consistency
  - comparable scales
  - valid observations
  - independent samples
- Raw data violates these assumptions.
- When we apply an algorithm, we implicitly assume that data behaves nicely.
- But real-world data is collected by humans, sensors, software systems, and business processes — all of which introduce errors.

Why Preprocessing Exists

- Preprocessing converts reality into mathematical representation.
- Completeness assumption

Age	Salary
25	3000
30	NULL

- Many algorithms cannot compute with NULL.
- Consistency assumption: Male, male, M, man
- Numerical representation: gender feature (male/female) must be encoded numerically, because algorithms works with numbers.

Why Preprocessing Exists

- Comparable scale assumption

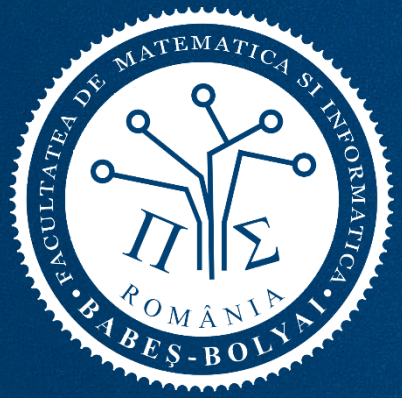
Feature	Range
Age	18-70
Income	1.000-10.000 $

- Validity Assumption (e.g., age=250, temperature =-5000)

What Happens Without Preprocessing?

- **Model Problems**
  - unstable learning
  - incorrect patterns
  - biased predictions
  - poor generalization

- Companies invest heavily in:
  - data engineers
  - data pipelines
  - governance
  - Not only data scientists.

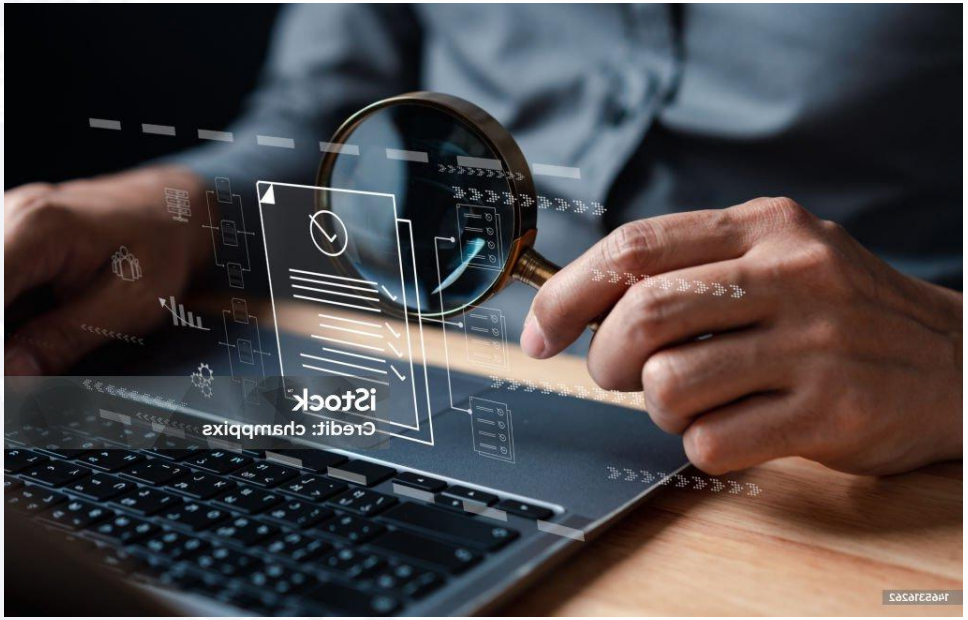


Data Quality Dimensions

Faculty of Mathematics and Computer Science

Data Quality Dimensions

- Completeness
- Accuracy
- Consistency
- Validity
- Uniqueness
- Timeliness



Completeness

- Are values missing?

- Example:
  Income missing for many customers.

Accuracy

- Are values correct?

- GPS location = airport while user at home.
  - Real case: Food delivery optimization failures.

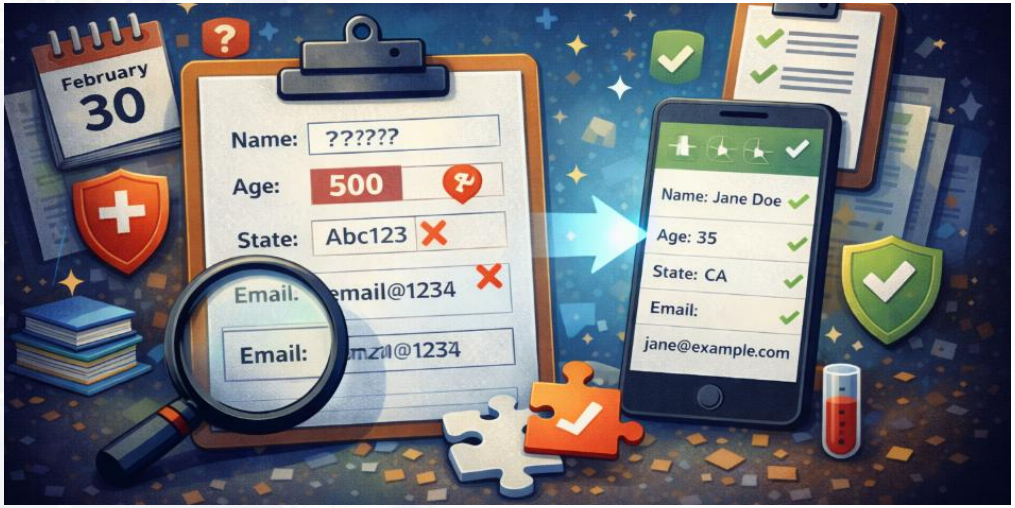
Consistency

- Same information stored uniformly?

- Example : Male/ M/ Man/ male

Validity

- Does data respect rules?



- Example: Salary < 0 → invalid.

Uniqueness

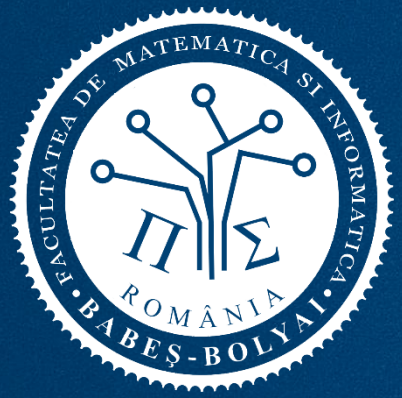
- Duplicates exist?

- Example: Same customer twice.

Timeliness

- Is data outdated?

- Example: Old medical records.



Teamwork Time 1

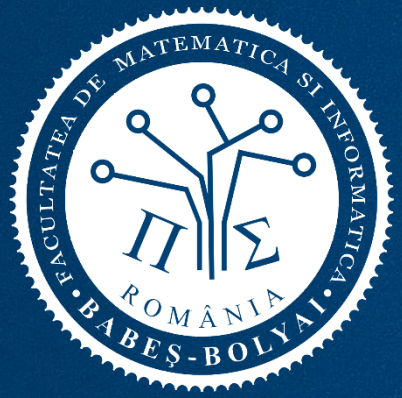
Faculty of Mathematics and Computer Science

Teamwork Time 1

- A hospital dataset contains:
  - missing diagnoses
  - duplicated patients
  - inconsistent gender labels
  - outdated records

- Which data quality problems exist?

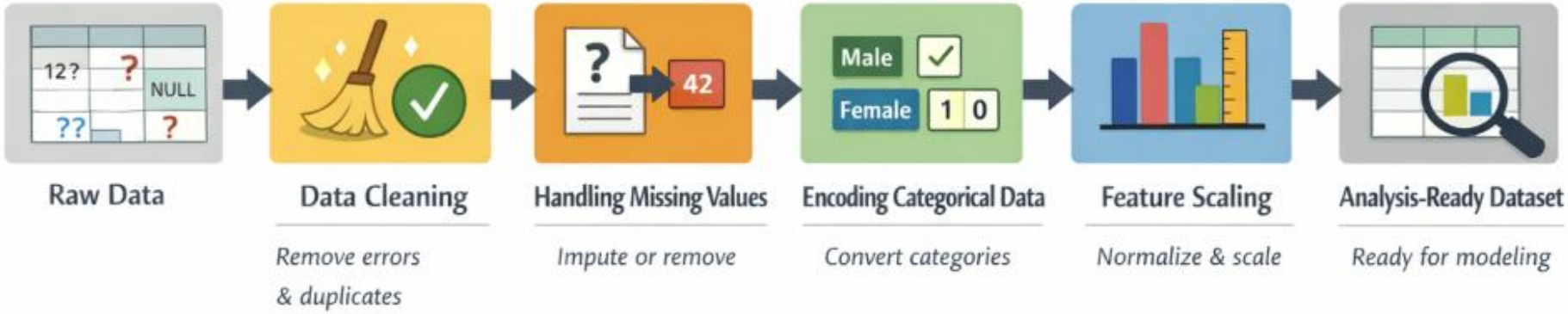
- Time: 5-7 minutes



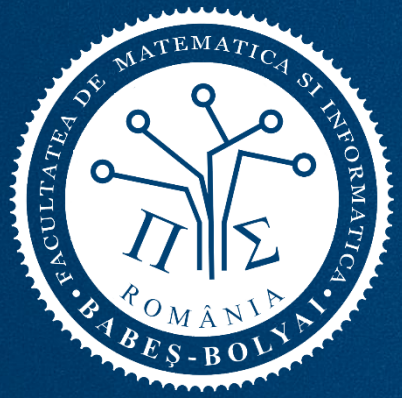
The Data Preprocessing Pipeline

Faculty of Mathematics and Computer Science

The Data Preprocessing Pipeline



- **Preprocessing converts reality into computation.**



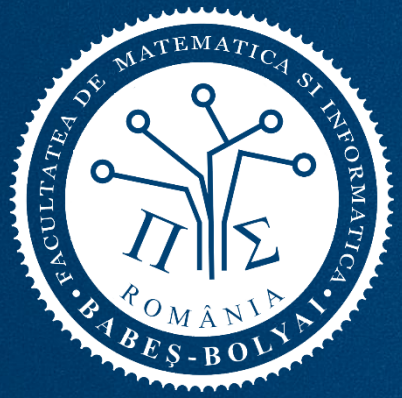
Raw Data

Faculty of Mathematics and Computer Science

Raw Data

- Data directly collected from reality.
- Sources:
  - databases
  - sensors
  - surveys
  - logs
  - APIs
  - human input
- Raw data is usually:
  - Incomplete
  - Inconsistent
  - Noisy
  - Duplicated
  - Incorrectly formatted

Age	Gender	Salary
25	Male	3000
NULL	M	?
250	male	4000



Data Cleaning

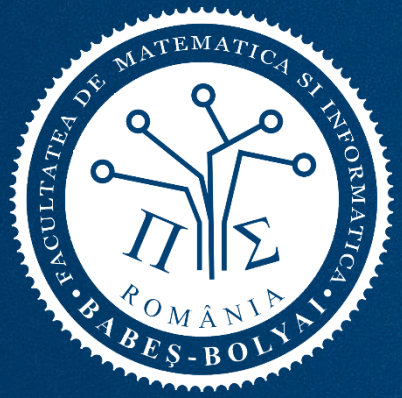
Faculty of Mathematics and Computer Science

Data Cleaning

- Remove **errors and inconsistencies**.

- Operations:
  - Removes duplicates
  - Fix inconsistent labels (e.g., male/M/Male → MALE) → Standardization
  - Correct invalid values (e.g., age=250 → error)

- Cleaning improves **data validity** and **consistency**.



Missing Data

Faculty of Mathematics and Computer Science

Missing Data

- Missing data ≠ random accident.

- Types of Missing Data:
  - MCAR → Missing Completely At Random
  - MAR → Missing At Random
  - MNAR → Missing Not At Random

- MCAR — Missing Completely At Random
  - Probability of missing independent of data.
  - Example: sensor malfunction.
  - Safe handling possible.

- MAR — Missing At Random
  - Depends on observed variables.
  - Example: young people skip income questions.
  - Bias exists but manageable.

- MNAR — Missing Not At Random
  - Depends on hidden value itself.
  - Example: high earners hide salary.
  - Most dangerous case.

- Hardest missing data type? MNAR

Handling Missing Data

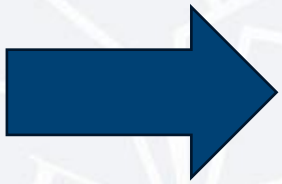
- Method 1 : Deletion
- Method 2 — Mean / Median Imputation
- Method 3 — Model-Based Imputation
- Method 4- Use most frequent value

Age
20
25
30
NULL
35

Missing Data: deletion

- Remove rows
- Problem: loss of information.

Age
20
25
30
NULL
35



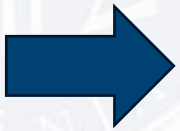
Age
20
25
30
~~NULL~~
35

Mean/median Imputation

- Replace missing values with the **average (mean)** of observed values.
- Replace missing values using the **median** (middle value).

- (20+25+30+35)/4=27.5

Age
20
25
30
NULL
35



Age
20
25
30
27.5
35

Model-based Imputation

- Separate rows with known values
- Train prediction model
- Predict missing values
- Replace missing entries

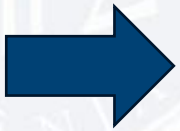
Age	Education	Salary
25	Bachelor	3000
30	Master	4500
40	PhD	?

- We use a Machine learning model to predict the salary.

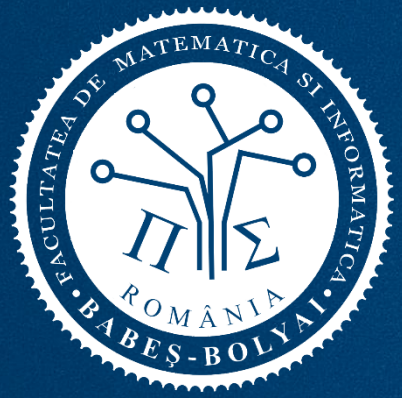
Most frequent value Imputation

- standard method for categorical variables.
- we cannot compute a mean or median.
- Most Frequent Value Imputation replaces missing values using the most common category in the dataset.

Gender
Male
Female
Male
NULL
Male
Male



Gender
Male
Female
Male
Male
Male
Male



Outliers

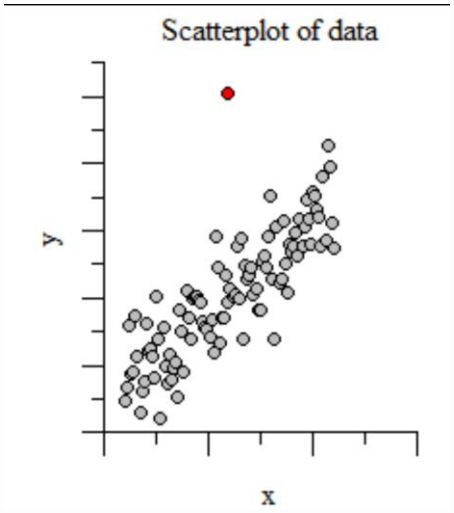
Faculty of Mathematics and Computer Science

Outliers

- Outliers are values located **far away from the typical data distribution**.

- Example: passenger fares→7, 8, 10, 9, 11, 8, 7, 512
- Value **512** behaves very differently.

- Outliers are observations that do not follow the general pattern of the data

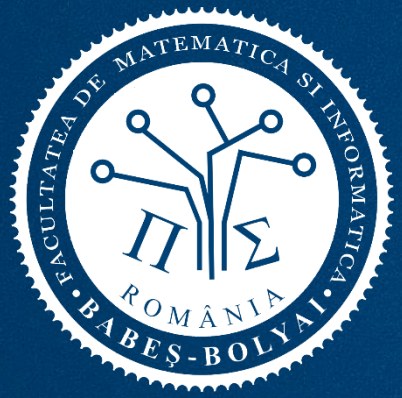


- Should we remove all outliers? NO, they may contain valuable information.

- Vizualization: boxplot, scatterplot, etc.

Why Do Outliers Appear?

- Measurement Errors
  - Sensor malfunction or human mistake.
  - Example: Age = 999
- Data Entry Errors
  - Typing mistake.
  - 1000000 instead of 10000
- Natural Rare Events
  - Natural Rare Events
  - Example: Very wealthy passenger on Titanic.
- Important Anomalies
  - Sometimes the **most valuable observations**.
  - Example: Fraudulent bank transaction.



Teamwork Time 2

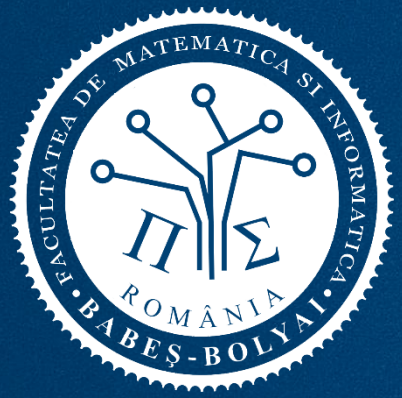
Faculty of Mathematics and Computer Science

Teamwork Time 2

- Bank transaction dataset contains:
  - €1 transactions
  - €10,000,000 transaction

- Teams discuss:
  - error OR fraud OR VIP client?
  - Give context and decisions

- 5-7 minutes



Encoding Categorical Data

Faculty of Mathematics and Computer Science

Encoding Categorical Data

- Algorithms process numbers.

- But reality contains categories:
  - gender
  - country
  - product type

- Label encoding
  - Assign integer values to categories.
  - Example: sex feature: male→ 0 ; female → 1

Encoding Categorical Data

- One-hot encoding
  - Create one binary column per category.
  - Disadvantage: Many categories → many columns.

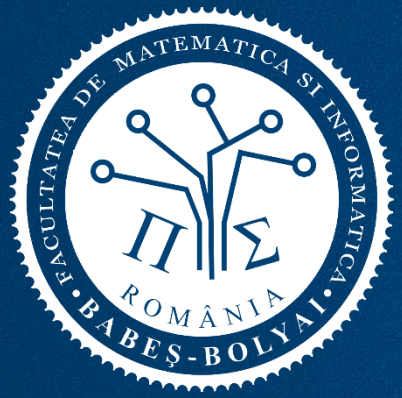
Color
Red
Blue

Red	Blue
1	0
0	1

- Ordinal encoding
  - Used when order exists.

Education
High School
Bachelor
Master
PhD

Education
0
1
2
3



Feature Scaling

Faculty of Mathematics and Computer Science

Feature Scaling

- Features measured on different scales can distort learning algorithms.

- Types of Features Scaling/ Data Normalization:
  - Min-max
  - Log
  - Clipping
  - Z-score standardization

Min-Max Normalization

- Rescales values into fixed interval: [0,1]
- Formula:

$$x_i^{new} = \frac{x_i - min(x)}{max(x) - min(x)}$$

- When to use:
  - scaling to a range is a good choice when both of the following conditions are met:
  - You know the approximate upper and lower bounds on your data with few or no outliers (we will end up with smaller standard deviations, which can suppress the effect of outliers)
  - Your data is approximately uniformly distributed across that range.

- A_{min} = 10, B_{min}=80
- A_{max}=25, B_{max}=200
- Column A, row 1 normalized:

- $(10 - 10)/15 = 0/15 = 0.0000$

Row	Feature A	Feature B
1	10	100
2	12	80
3	18	120
4	20	200
5	25	160

Log Scaling

- Log scaling is helpful when a handful of your values have many points, while most other values have few points.

- Formula:

$$x_i^{new} = \log(x_i)$$

- Examples: Movie ratings are a good example. Most movies have very few ratings (the data in the tail), while a few have lots of ratings (the data in the head).

Clipping Scaling

- Thresholding the data which caps all feature values above (or below) a certain value to fixed value.
- Formula:

$$x_i^{new} = \begin{cases} x_i, & x_i < threshold \\ threshold, & x_i \geq threshold \end{cases}$$

- When to use: if your data set contains extreme outliers

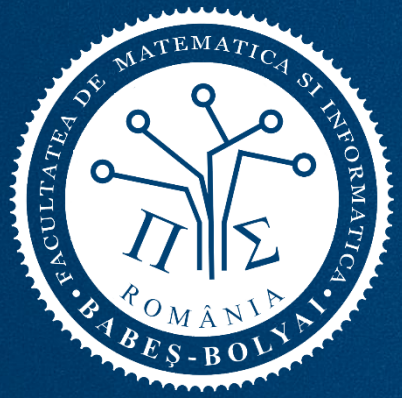
Z-score normalization

- The result of standardization (or Z-score normalization) is that the features will be rescaled so that they will have the properties of a standard normal distribution with $\mu=0$ and $\sigma=1$ (where μ is the mean (average) and σ is the standard deviation from the mean).

- Formula:

$$x_i = (x_i - \mu)/\sigma$$

$$stdev = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n-1}}$$

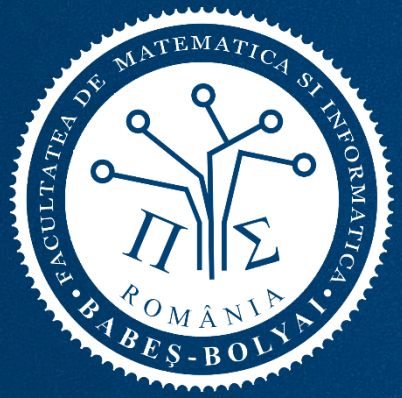


Analysis-Ready Dataset

Faculty of Mathematics and Computer Science

Analysis-Ready Dataset

- Now data is:
  - Complete
  - Consistent
  - Numerical
  - Comparable
  - Valid
- This dataset can safely enter:
  - statistical analysis
  - data mining
  - machine learning



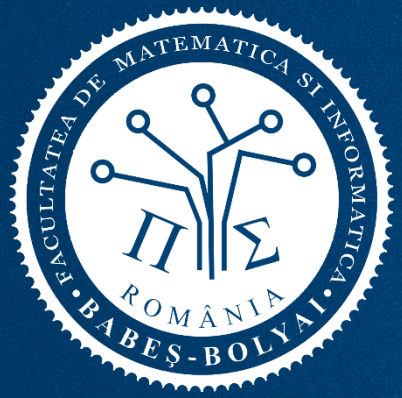
Reflection Time

Faculty of Mathematics and Computer Science

Reflection Time

Go to www.menti.com and enter the code **1887 3853**

or use the QR code



Key Takeaways

Faculty of Mathematics and Computer Science

Key Takeaways

- Real data is imperfect
- Quality determines reliability
- Missing data has mechanisms
- Outliers require reasoning
- Preprocessing prevents failure

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