Intelligent techniques for processing large and structured data

Lecture 1

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

Sergiu Limboi, PhD Teaching Assistant

Motto: From Raw Data to Actionable Knowledge

Introduction into Data Analysis, Data Mining, and Knowledge Discovery

AGENDA

- Course organization
- Why this course exists?
- Let’s get to know the audience
- Evaluation
- What is Data?
- What is Data Analysis?
- What is Data Mining?
- What is Knowledge Discovery?
- Types of Data
- Industry Pipeline for Large & Structured Data
- Key Takers

Course organization

Faculty of Mathematics and Computer Science

Course organization

- Instructors
  - Course: Teaching Assistant PhD Sergiu Limboi
  - Laboratory: Associate Teacher Andreea Gabrian

- Team channel code: j8zn8u3

- Course structure:
  - Lectures 2 hours per week
  - Laboratories 2 hours per week

- Additional information
  - All materials will be posted on GitHub: https://github.com/SergiuLimboi/Intelligent-techniques-for-processing-structured-and-large-data/tree/main
  - MS Teams will be used for announcements (e.g., exam dates, invited guests, etc.)

Why this course exists?

Faculty of Mathematics and Computer Science

Why this course exists?

- “Do all real datasets look like a Pandas table?”
- “What breaks first when data becomes large: the algorithm or the infrastructure?”
- “If a model works on 10,000 rows, will it work on 100 million?”
- “If your data is clean but your pipeline is slow, where is the real problem?”
- “Is structured data really simple?”
- “In large-scale systems, what is more important: accuracy, speed, or reliability?”

Why this course exists?

- This course assumes you already know Machine Learning (ML).

- We will focus on what ML *cannot* naturally handle

- Most AI students can:
  - Train a model;
  - Tune hyperparameters;
  - Get a good accuracy score.

- But real projects fail because:
  - The data is wrong;
  - The problem is wrongly defined;
  - The evaluation is misleading;
  - The dataset is biased or incomplete.

- This course is NOT about “better models”. It is about getting the right knowledge from data.

Why this course exists?

- Typical student mindset
  - “ I just load the data and try Random Forest/ ANN/ etc. and compute some evaluation measures.”
- Professional mindset
  - “What does the data represent, how was it generated, and what decision will be made based on it?”

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 **7361 1112**

or use the QR code

Evaluation

Faculty of Mathematics and Computer Science

Evaluation

- Evaluation structure:
  - Written exam (during the official exam session): 40%
  - Laboratory assignments (3 assignments during the semester): **10%**
  - Project development (during the semester): 50%

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

- To be eligible for the exam session, students must meet the following requirements:
  - Number of laboratory attendances: 8 out of 12
  - Number of turned laboratory assignments 2 out of 3
  - A minimum grade of 5 in the project

Evaluation

- 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.

- Project timeline:
  - Project development will start in **Week 5** of the semester.

- The project will consist of the following components:
  - Project development (an application with a minimal UI is required)
  - Research report
  - Presentation during the final laboratory session

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?

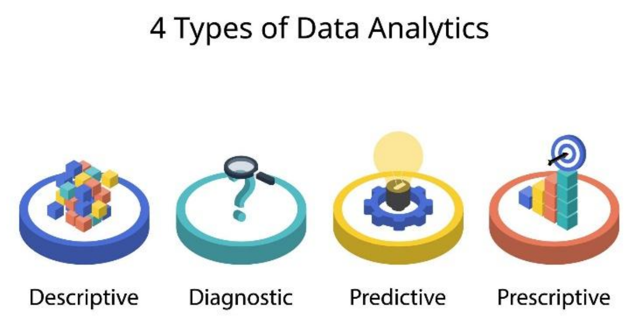
- Key concepts of Data Analysis:
  - Data collection
  - Data cleaning
  - Data transformation
  - Exploratory Data Analysis (EDA)
  - Interpretation

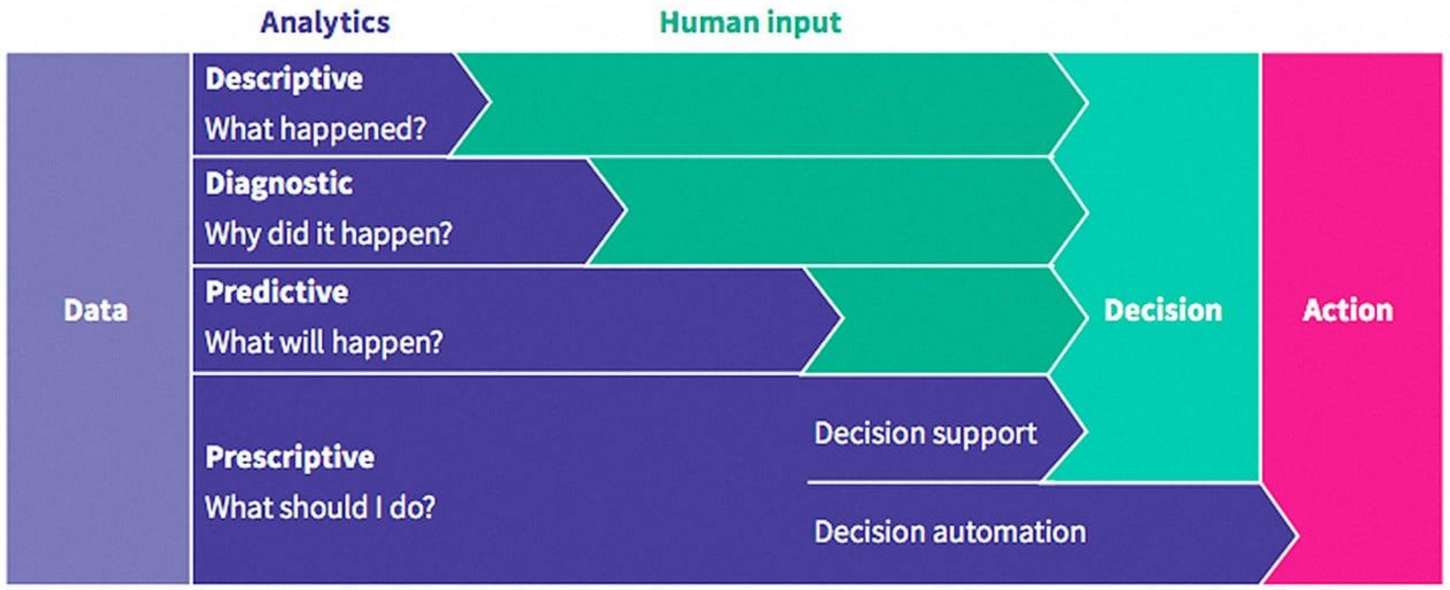
What is Data Analysis?

- 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 Data Mining?

Faculty of Mathematics and Computer Science

What is Data Mining?

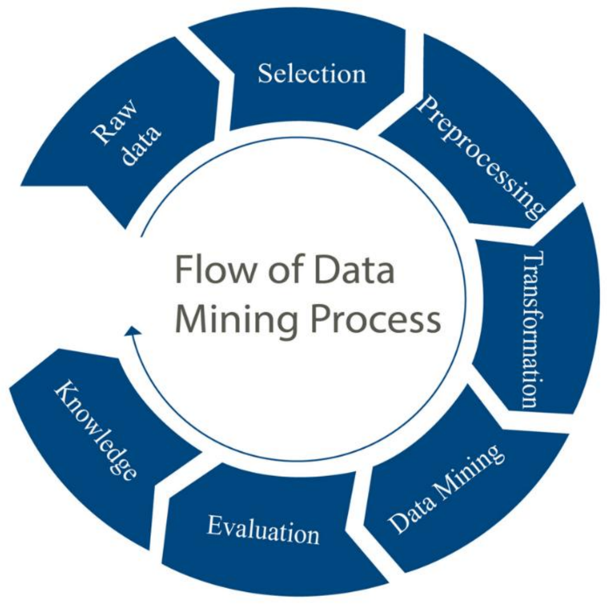
- **Data Mining** is the process of extracting valuable information from large databases that was previously unknown and using it to make informed business decisions.

- 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

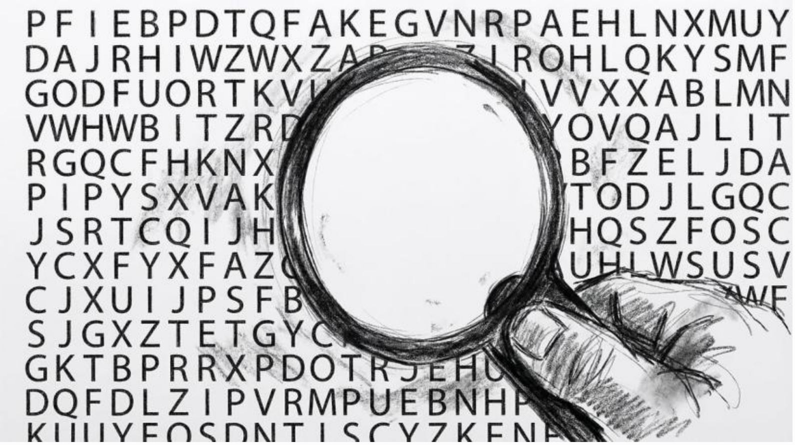
What is Knowledge Discovery?

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.

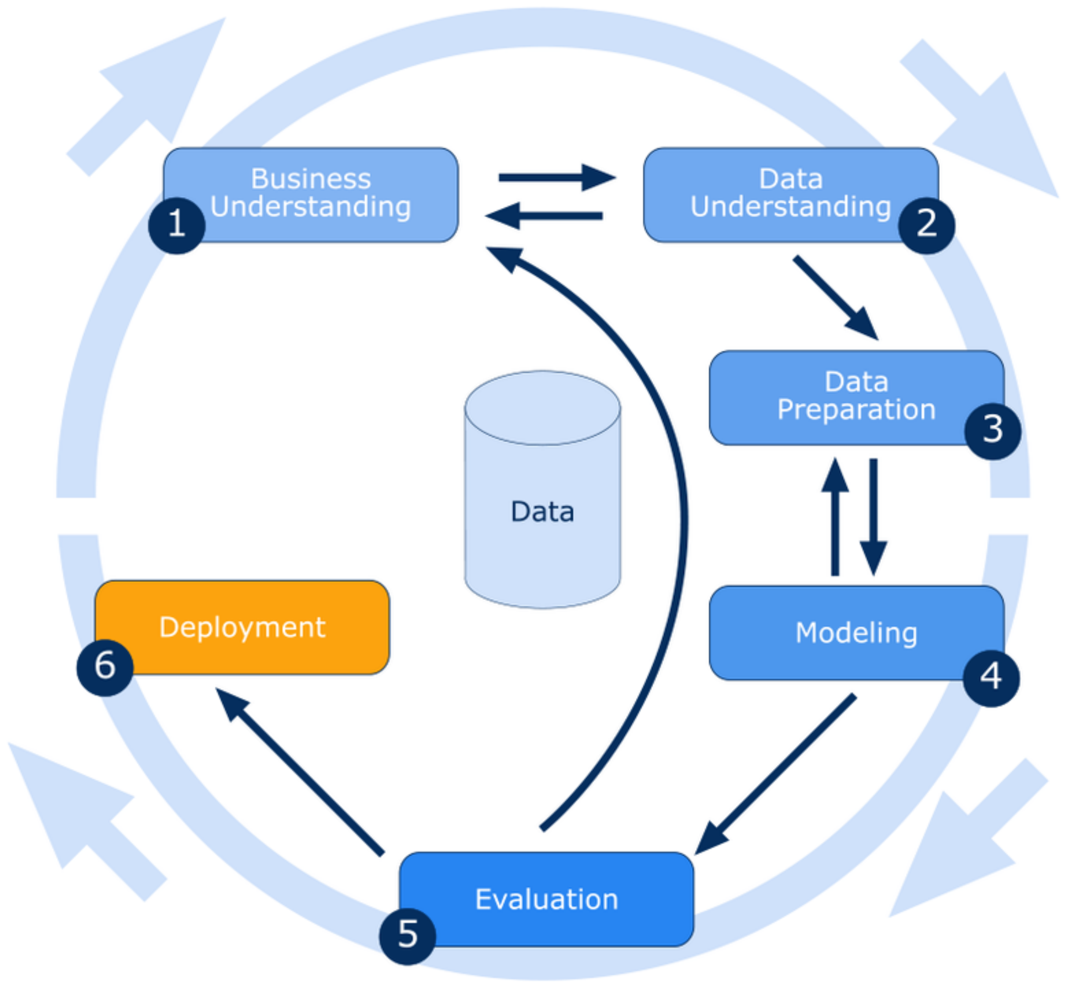


What is Knowledge Discovery?

- 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.

The Knowledge Discovery Process



The Knowledge Discovery Process

- Step 1- Business Understanding
  - What question are we answering?
  - Who uses the result?
  - What decision depends on it?

- Step 2 – Data Understanding
  - What data do we have?
  - Where does it come from?
  - What does each column mean?
  - What is missing and why?
  - Observation: Many datasets are **not random samples**.

The Knowledge Discovery Process

- Step 3: Data Preparation (Most Time-Consuming)
  - Cleaning
  - Filtering
  - Encoding
  - Feature engineering
  - Aggregations
  - Etc.
  - In real projects: 60%-70% of the effort
  - This is where **domain knowledge matters more than algorithms**.

- Step 4: Data Mining/ Modelling Pattern mining
  - Classification
  - Anomaly detection
  - Regression
  - Clustering
  - This step is often less than 20% of total effort

The Knowledge Discovery Process

- Step 5- Evaluation
  - Is the result correct?
  - Is it useful?
  - Is it stable?
  - Includes: metrics, validation strategies, error analysis, sanity checks

- Step 6- Deployment = Interpretation & use (knowledge extraction)
  - Reports
  - Dashboards
  - Decisions
  - Monitoring

Data Analysis vs. Data Mining vs. Machine Learning vs. Knowledge Discovery

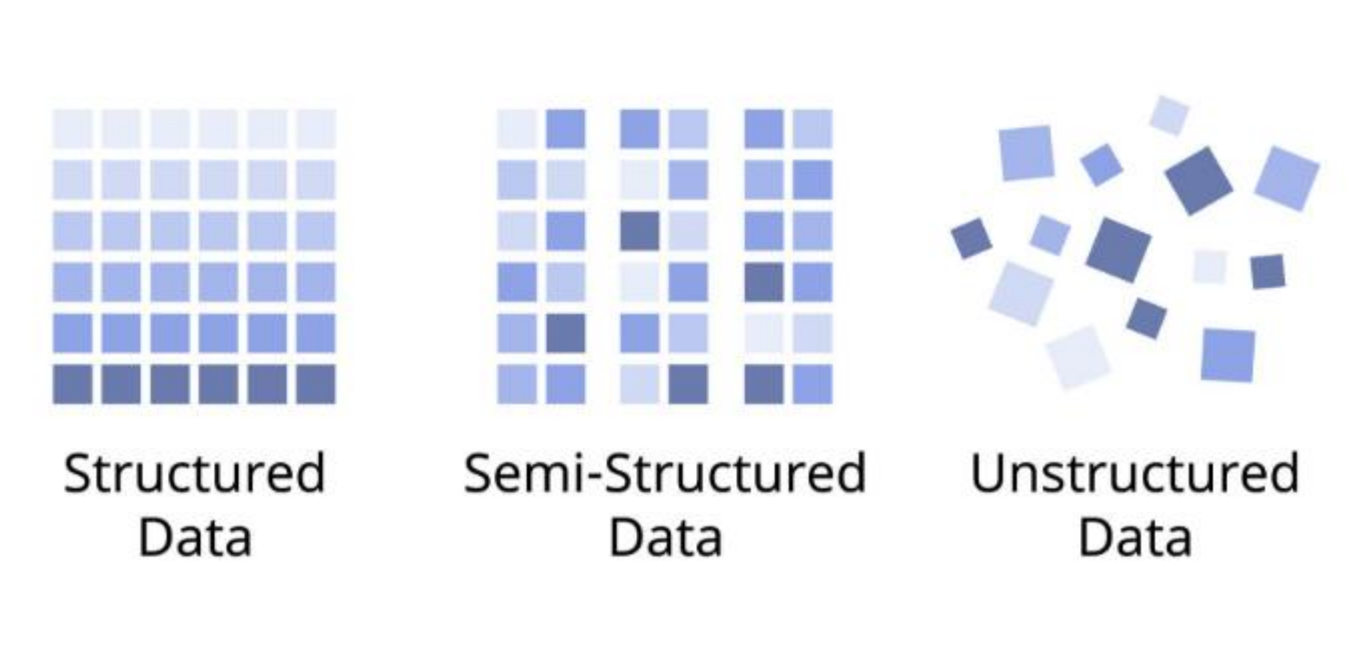
- **Data Analysis:** understand & describe data
- **Data Mining:** discover hidden patterns
- **Machine Learning:** algorithms that learn patterns
- **Knowledge Discovery** focuses on interpretation and meaning.

- **Knowledge Discovery**→ the **whole journey** from raw data to insight
- **Data Mining**→ the **core step** where patterns are extracted
- **Machine Learning**→ the **toolbox of algorithms** used inside data mining
- **Data Analysis**→ supports **every step** (before, during, after)

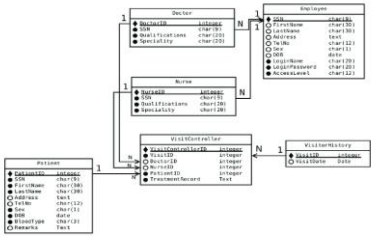
Types of Data

Faculty of Mathematics and Computer Science

Types of Data

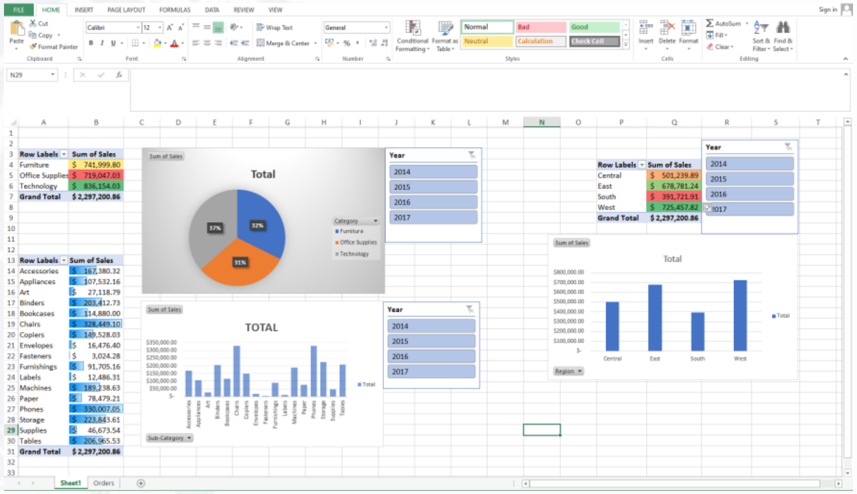


Structured data



- Data organized in a **fixed schema**, where every instance has the same attributes.

- Characteristics:
  - Rows and columns
  - Well-defined data types
  - Easy to store in databases
- Examples:
  - SQL tables
  - CSV/ Excel files
  - Transaction records



Row Labels	Sum of Sales
Furniture	$ 741,999.80
Office Supplies	$ 719,047.03
Technology	$ 836,154.03
Grand Total	$ 2,297,200.86

Row Labels	Sum of Sales
Accessories	$ 167,380.32
Appliances	$ 107,532.16
Art	$ 27,118.79
Binders	$ 203,412.73
Bookcases	$ 114,880.00
Chairs	$ 328,449.10
Copiers	$ 149,528.03
Envelopes	$ 16,476.40
Fasteners	$ 3,024.28
Furnishings	$ 91,705.16
Labels	$ 12,486.31
Machines	$ 189,238.63
Paper	$ 78,479.21
Phones	$ 330,007.05
Storage	$ 223,843.61
Supplies	$ 46,673.54
Tables	$ 206,965.53
Grand Total	$ 2,297,200.86

Row Labels	Sum of Sales
Central	$ 501,239.89
East	$ 678,781.24
South	$ 391,721.91
West	$ 725,457.82
Grand Total	$ 2,297,200.86

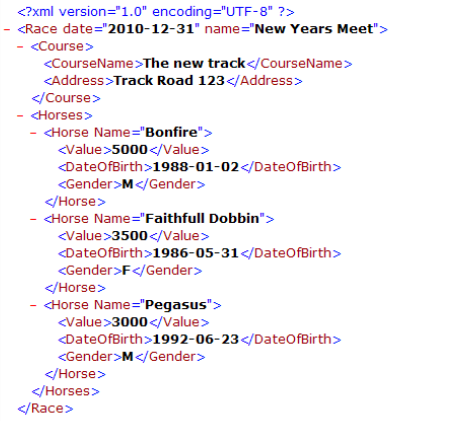
- Structured data is **easy to query**, but **hard to design correctly**.

Semi-structured data

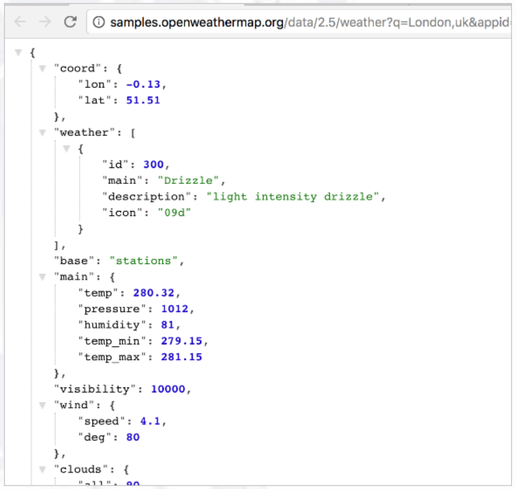
- Data with **some structure**, but not a fixed schema.

- Characteristics:
  - flexible attributes
  - nested fields
  - schema may vary per instance

- Examples:
  - JSON
  - XML
  - API responses



```
<?xml version="1.0" encoding="UTF-8" ?>
- <Race date="2010-12-31" name="New Years Meet">
  - <Course>
      <CourseName>The new track</CourseName>
      <Address>Track Road 123</Address>
    </Course>
  - <Horses>
    - <Horse Name="Bonfire">
        <Value>5000</Value>
        <DateOfBirth>1988-01-02</DateOfBirth>
        <Gender>M</Gender>
      </Horse>
    - <Horse Name="Faithfull Dobbin">
        <Value>3500</Value>
        <DateOfBirth>1986-05-31</DateOfBirth>
        <Gender>F</Gender>
      </Horse>
    - <Horse Name="Pegasus">
        <Value>3000</Value>
        <DateOfBirth>1992-06-23</DateOfBirth>
        <Gender>M</Gender>
      </Horse>
    </Horses>
</Race>
```

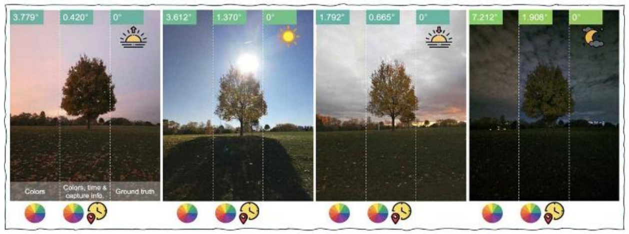


samples.openweathermap.org/data/2.5/weather?q=London,uk&appid

```
{
  "coord": {
     "lon": -0.13,
     "lat": 51.51
  },
  "weather": [
     {
        "id": 300,
        "main": "Drizzle",
        "description": "light intensity drizzle",
        "icon": "09d"
     }
  ],
  "base": "stations",
  "main": {
     "temp": 280.32,
     "pressure": 1012,
     "humidity": 81,
     "temp_min": 279.15,
     "temp_max": 281.15
  },
  "visibility": 10000,
  "wind": {
     "speed": 4.1,
     "deg": 80
  },
  "clouds": {
     "all": 90
```

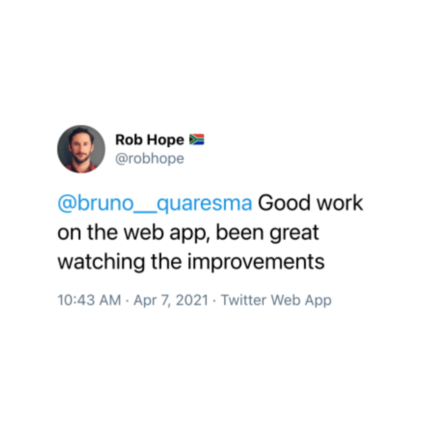
- Harder to query, requires parsing and normalization before mining.

Unstructured data



- Data without predefined organization.

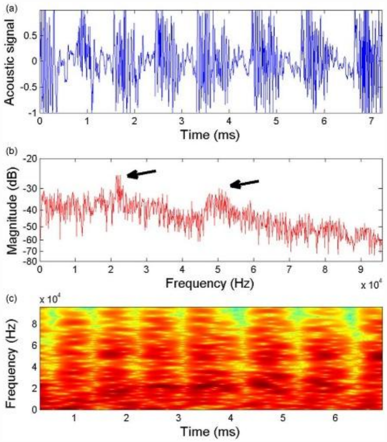
- Examples:
  - text documents
  - images
  - audio
  - video
  - social media posts



Rob Hope 🇿🇦
@robhope

@bruno__quaresma Good work on the web app, been great watching the improvements

10:43 AM · Apr 7, 2021 · Twitter Web App



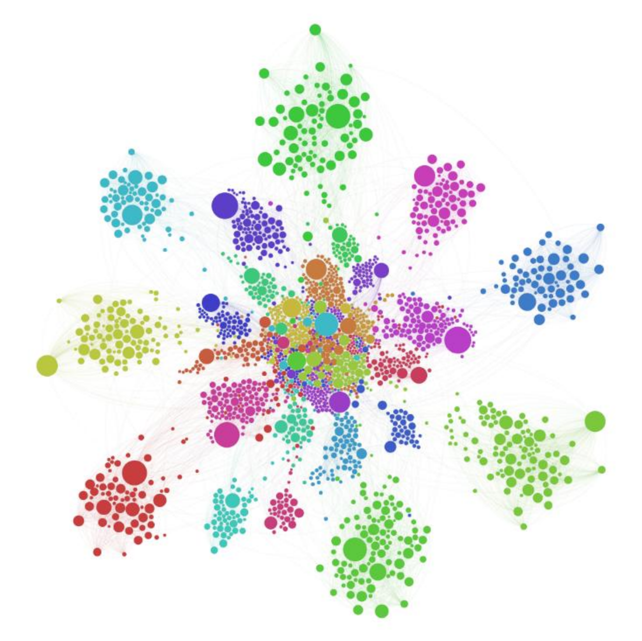
- Even unstructured data is often **converted into structured form** before mining (e.g., features, embeddings).

Why this course focuses on structured data?

- Most business data is structured
- Decisions rely on tables
- Pipelines are more complex than models
- Scalability issues appear here first

Large Data

- **Large data** refers to datasets that:
  - Do **not** fit in memory on a single machine
  - Require distributed storage & processing
  - Need specialized frameworks
- Typical characteristics:
  - Millions to billions of records
  - Tens of GBs → TBs → PBs
  - High velocity (data arrives continuously)
  - Often semi-structured and unstructured
- Examples:
  - Social media stream (tweets)
  - Transactional data at scale
  - Logs



Large vs. Small Data

Small	**Large**
Pandas works	Pandas struggles
Quick experiments	Need planning
Few joins	Complex joins
Fast feedback	Performance matters

- Large ≠ Big Data buzzword
- Large = **practical processing difficulty**

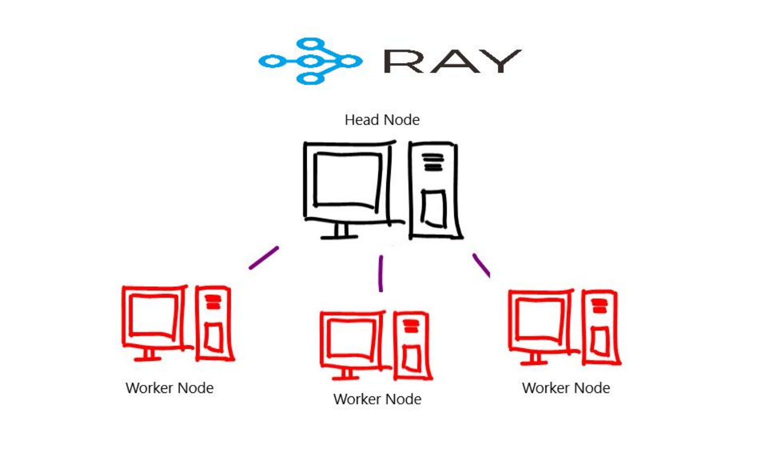
Why Large Structured Data?

- Common issues:
  - Memory limits
  - Slow joins
  - Slow processing
  - Imbalanced data
  - Wrong aggregation logic
  - Hidden bias.

- Using future information when predicting the past
  → model looks “amazing”
  → completely useless in reality

Tools & Mindset

- Tools:
  - Pandas (baseline)
  - Polars/ PySpark/ Dask/ Ray (for scale)
  - ML libraries



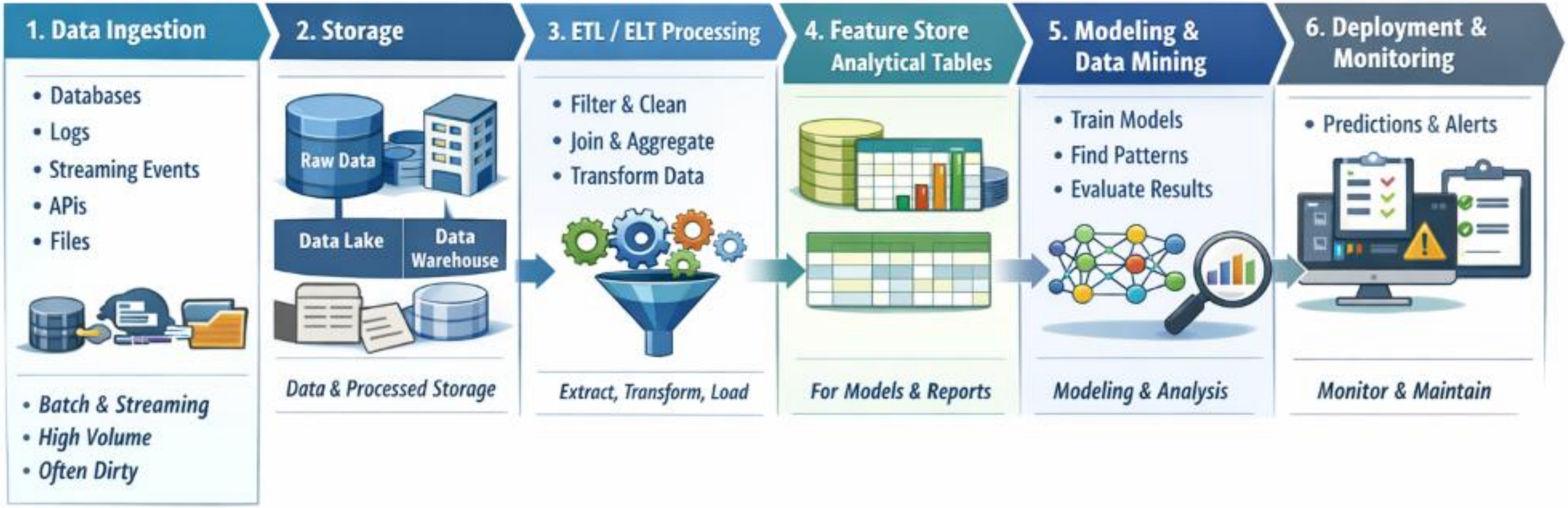
- Mindset:
  - Thinking
  - Reasoning
  - Correctness

- **All industrial systems aim to produce structured data** for decision-making.

Industry Pipeline for Large & Structured Data

Faculty of Mathematics and Computer Science

Industry Pipeline for Large & Structured Data



Step 1: Data Ingestion

- **Goal:** Bring data into the system.
- **Sources:**
  - databases
  - logs
  - streaming events
  - APIs
  - Files

- **Key properties:**
  - batch or streaming
  - high volume
  - often dirty

Step 2: Storage

- Storage is not just “where data lives”
- Storage defines how data can be used
- Wrong storage → slow pipelines, wrong results
- Common layers:
  - raw data storage (data lake)
  - processed data storage (warehouse)
- Raw data is almost never used directly for modelling.
- What goes in a data lake?
  - Database dumps, logs, CSV files, JSON files, etc.
- What goes in a data warehouse?
  - Cleaned and transformed data (e.g., aggregated tables)

Step 3: ETL/ ELT Processing

- ETL = Extract, Transform, Load
- Operations include:
  - filtering
  - joining tables
  - aggregations
  - feature computation
  - data validation

- This step creates **analysis-ready structured datasets**.

Step 4: Feature Store/ Analytical Tables

- At this stage:
  - Data is clean
  - Schema is stables
  - Features are well-defined

- This data feeds:
  - Machine Learning models
  - Dashboards
  - Reports

- Feature computation is done at step 3 (e.g., compute number of purchases per user, average transaction values, etc.)

- Feature engineering is done at step 4 (e.g., selecting useful features for a ML model, define feature semantics,

Step 5: Modelling & Data Mining

- **At this stage, we can:**
  - Train predictive models
  - Discover patterns in data
  - Evaluate results and performance

- Models are **consumers**, not owners of data.

Step 6: Deployment & Monitoring

- Deployment = using results in real systems

- Outputs are:
  - Predictions;
  - Scores;
  - Clusters;
  - Alerts – triggered when conditions are met.

- Monitoring checks:
  - data drift (e.g., feature no longer look the same, changes in input data)
  - model degradation (e.g., performance decreases over time)
  - pipeline failures (e.g., missing data, delayed pipelines)

Key Takers

Faculty of Mathematics and Computer Science

Key Takers

- Data mining is a **process**, not an algorithm
- Machine Learning is a **tool**, not the goal
- Structured data drives real-world decisions
- Raw data is almost never used directly
- Pipelines matter more than models
- Knowledge must be **useful and explainable**

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