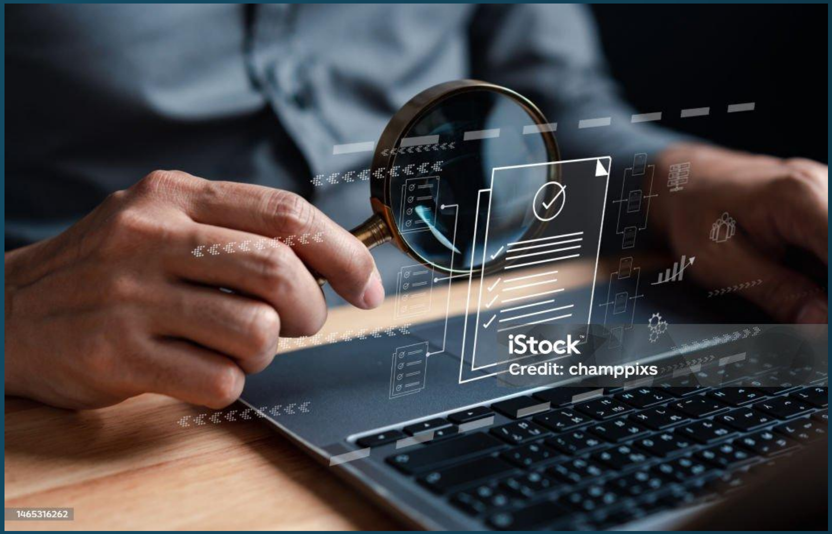
Intelligent techniques for processing large and structured data

Lecture 2

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

Sergiu Limboi, PhD Teaching Assistant

Motto:Understanding Real-World Data Before Machine Learning



Data Understanding and Data Quality in Large-Scale Systems

AGENDA

- Warm-Up
- Industry reality
- Data understanding
- Data quality
- Teamwork time 1
- Data bias
- Industry case: Amazon Hiring AI Failure
- Data leakage
- Why large data makes everything worse?
- Key industry mindset shift
- Teamwork time 2
- Key takeaways

Warm-Up

Faculty of Mathematics and Computer Science

Warm-Up

Go to www.menti.com and enter the code **5904 2936**

or use the QR code

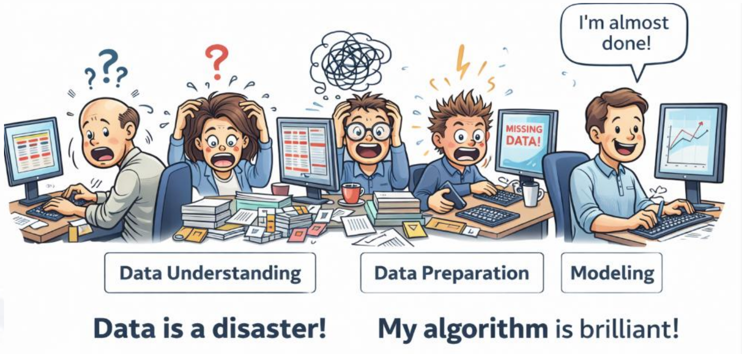
Industry reality

Faculty of Mathematics and Computer Science

Industry reality

- If two teams use the SAME model, why does one succeed and the other fail?

Industry reality



- Based on **Gartner Research, Google ML Systems, IBM &New Vantage Surveys**, most AI/Data Science project failures originate from:

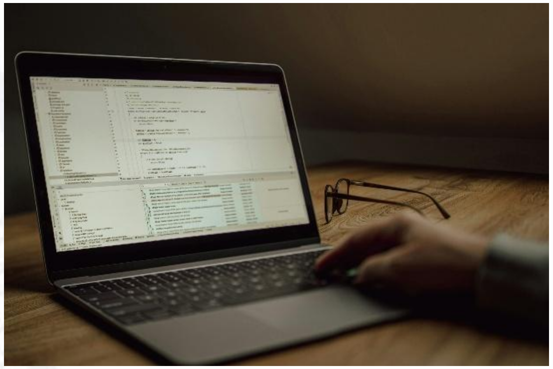
Problem source	Failure rate
Wrong data	~60%
Bad problem definition	~20%
Model choice	<10%
Infrastructure	~10%

Data understanding

Faculty of Mathematics and Computer Science

Data understanding

- Data understanding means knowing what your data represents before analysing or modelling it.



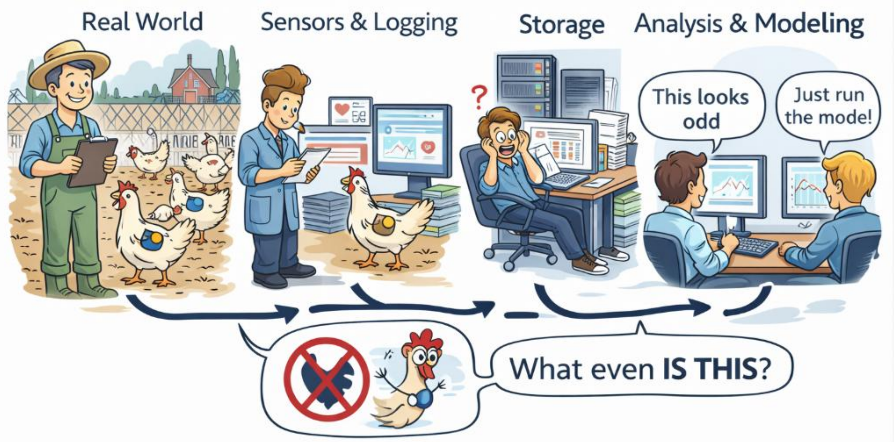
- Data understanding means answering:
  - What does data represent?
  - How was it generated?
  - What process created it?
  - What decision depends on it?

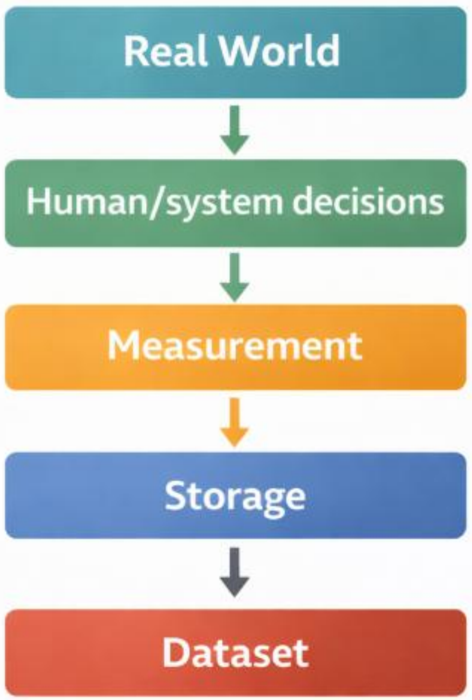
- Datasets are **observations of processes**, not reality.

- Example: Bank transactions ≠ customer behaviour

Data Generating Process

- Every dataset comes from:





- Errors appear everywhere.

Data quality

Faculty of Mathematics and Computer Science

Data quality

- Data quality measures how suitable data is for decision-making.
- Data quality dimensions:
  - Completeness
  - Accuracy
  - Consistency
  - Validity
  - Uniqueness
  - Timeliness

Completeness

- Do we have all required data?

- We talk about a missing values problem.

- Example:
  - Medical datasets missing tests for poor patients. Impact: Bias introduced automatically. ❌
  - Customer income missing for 40% users. Impact: Credit scoring becomes biased. ❌

Accuracy

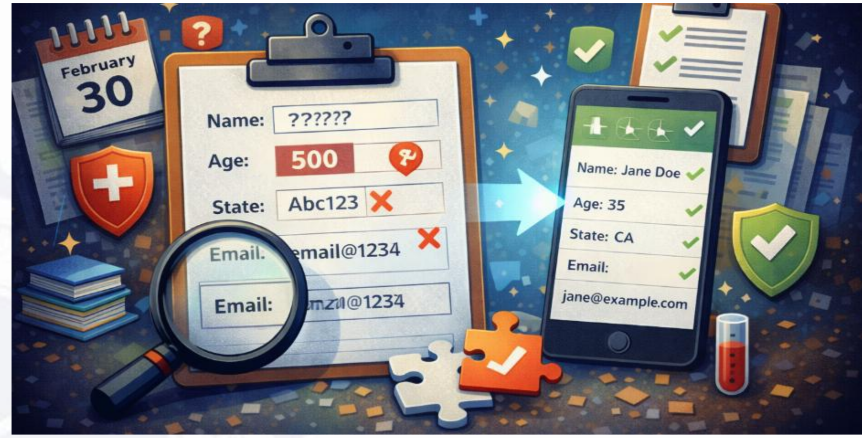
- Is data correct?

- Examples:
  - GPS location = airport while user at home.
    - Real case: Food delivery optimization failures. ❌
  - IoT temperature sensors shift slowly.
    - Model degradation occurs silently. ❌

Consistency

- Same entity → same value everywhere.
- The information should be stored uniformly.
- Example: Male/M/Man/male ❌

Validity



- Does data respect constraints?

- Examples:
  - Age = −5 ❌
  - Transaction date in future ❌
  - Transaction before account creation. ❌

Uniqueness

- Duplicates destroy aggregation.

- Do we have duplicates?

- Examples:
  - Customer counted twice → revenue inflated. ❌

Timeliness

- Is data outdated?
- Freshness matters.
- Examples:
  - Fraud detection using 24h delayed data = useless. ❌
  - System using old medical records. ❌

Data quality

- Is more data always better?

Data Quality Pyramid

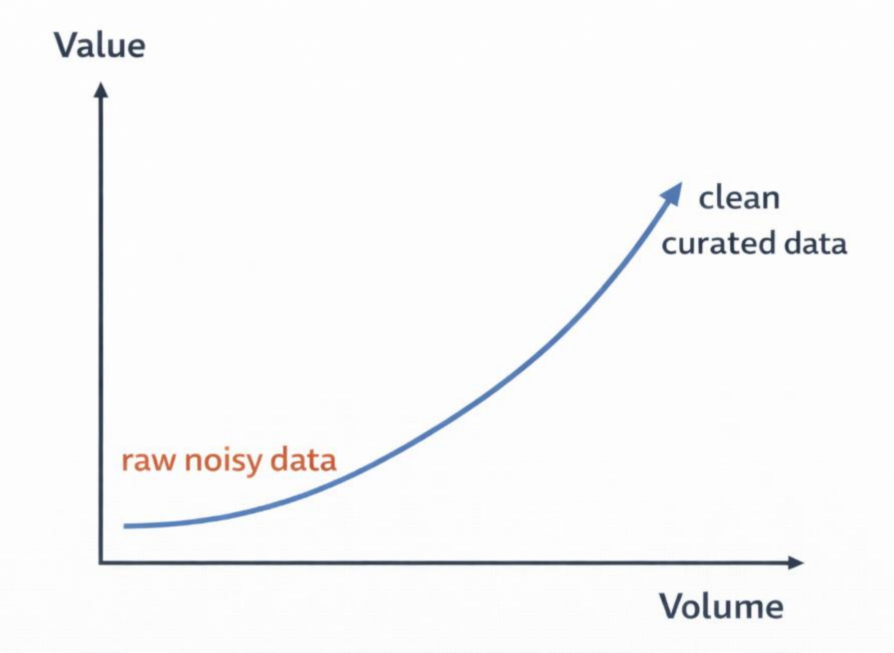
The biggest industry problem

- DIRTY DATA

- Typical dataset problems:
  - Missing values
  - Duplicates
  - Outliers
  - Schema drift
  - Unit mismatch
  - Encoding errors
  - Etc.

- Example: Banking dataset where balance is stored as :
  - 1000
  - 1.000
  - 1k
  - 1000.00

Industry data



- More volume ≠ more knowledge.

Teamwork time 1

Teamwork time 1- “Trust This Dataset?”

- Dataset description: Online retail dataset with missing prices, duplicated customers, inconsistent timestamps, currency inconsistencies (prices stored in RON, USD, EUR), some quantities are negative, product name variation (e.g., iPhone, Iphone, Apple Iphone).

- Task:
  1. Identify 3-5 risks
  2. Rank them by severity
  3. Propose fixes.

- 10 minutes

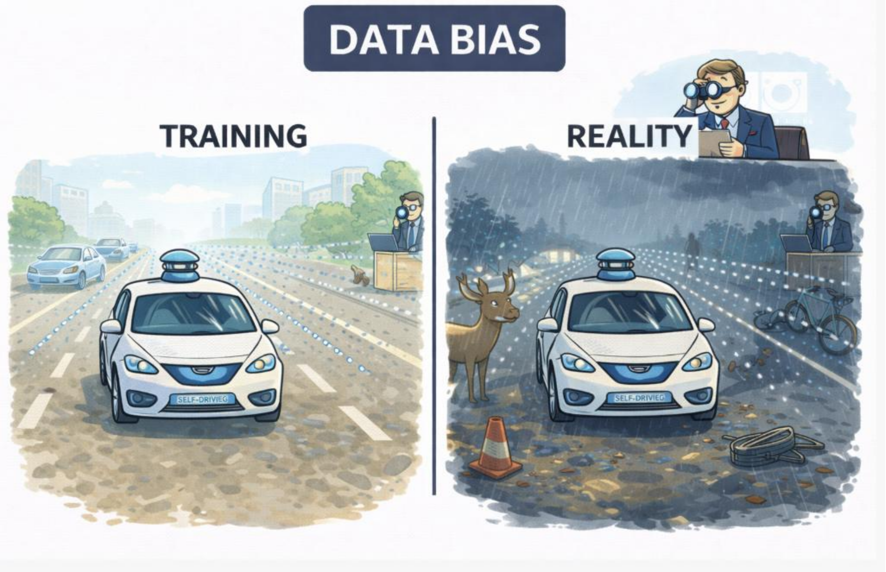
- Teams of 4-6 students

Data bias

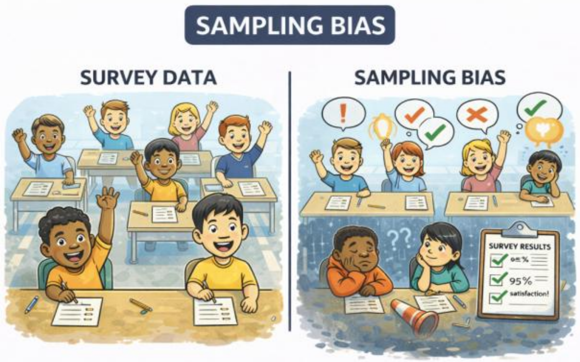
Data bias

- What is a bias?
  - Systematic distortion introduced during data creation.

- Models do not create bias — they learn it.
- Types of bias:
  - Sampling
  - Historical
  - Measurement
  - Temporal
  - Selection



Sampling Bias

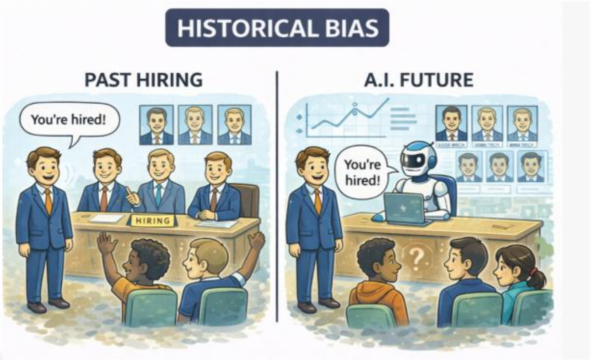


- Sampling bias occurs when the collected dataset is not representative of the entire population because some groups or observations are systematically overrepresented or underrepresented during data collection.

- Dataset not representative.

- Examples:
  - Only urban customers sampled. Rural predictions fail.
  - Predict salaries using LinkedIn users only. Overestimated salaries.

Historical Bias

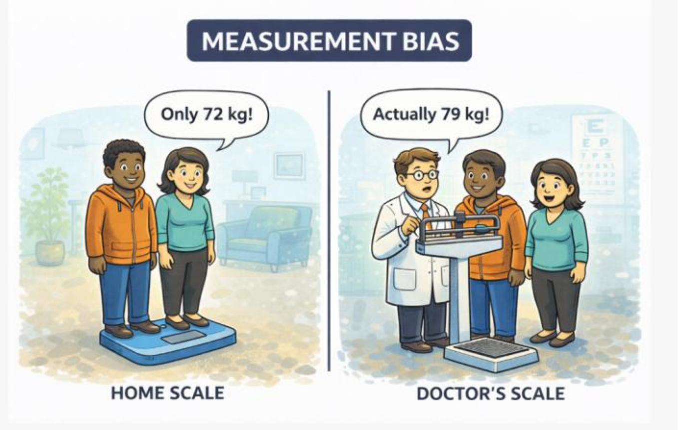


- Historical bias occurs when past human decisions or societal patterns are embedded in data, and models learn to reproduce them.

- Model learns past unfair decisions.

- Examples:
  - Loan approvals historically unequal. Model reproduces discrimination.
  - Company hired mostly men for technical roles.
  - Police historically patrolled certain neighbourhoods more.
  - Autonomous Driving Data collected mostly daytime and sunny weather.

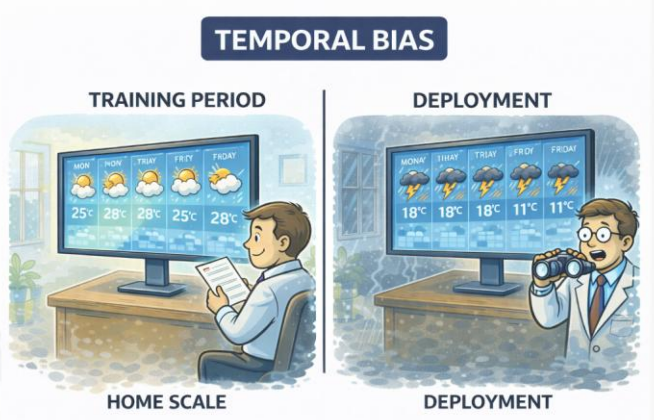
Measurements Bias



Measurement bias occurs when the process used to collect or measure data systematically produces incorrect or distorted values compared to the true underlying phenomenon.

- Sensors or labels are incorrect.
- Example:
  - Manual annotation fatigue. (e.g., Human annotators label tweets.)
  - Different hospitals measure differently.
  - Temperature sensor slowly drifts.
  - Using **job title** as proxy for skill.

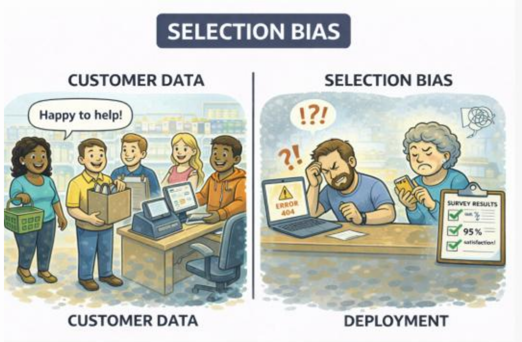
Temporal Bias



- Temporal bias occurs when data collected during one time period does not correctly represent future or past conditions, causing models to learn patterns that are no longer valid over time.

- Examples:
  - Online shopping behavior during COVID lockdown.
  - Ride-sharing models→ Traffic patterns learned before remote work adoption.

Selection Bias



- Selection bias occurs when the process used to select observations for a dataset systematically favours certain individuals, groups, or outcomes, causing the sample to differ from the intended population.

- Examples:
  - Company analyses only **active users**. Conclusion: Customers highly engaged.
  - Goal: Measure student satisfaction. Survey sent only to students attending lectures.
  - Hospital dataset contains only patients who visited hospital. Missing: people without healthcare access
  - Training data collected mostly in safe driving situations.

Industry case: Amazon Hiring AI Failure

Industry case: Amazon Hiring AI Failure

- Around **2014**, Amazon started developing an internal **AI recruiting system**.
- Goal: Automatically review resumes and rank job candidates. The system used **machine learning** to score candidates from **1 to 5 stars**.
- Motivation:
  - Amazon received **thousands of applications daily**
  - recruiters spent enormous manual effort
  - automation promised efficiency
- The AI was trained using:
  - **10 years of historical hiring data**
  - resumes submitted to Amazon
  - past successful employees
- The assumption was: Past successful hires → future successful hires.

Industry case: Amazon Hiring AI Failure

- Amazon's historical technical workforce was predominantly **male**.
  - Training data contained mostly:
    - male resumes
    - male career patterns
    - male wording styles

- What the AI learned?
  - **Penalize resumes containing:**
    - the word **"women's"**
    - participation in women-only organizations
    - graduates from women's colleges

- Example reported behavior: "women's chess club captain" → lower score

- The algorithm effectively learned that **female-associated signals correlated with rejection**.

Industry case: Amazon Hiring AI Failure

- Amazon engineers **never programmed gender discrimination**.

- Why this happened?
  - HISTORICAL BIAS

- Amazon tried to:
  - remove gender indicators
  - adjust feature weights
  - However:
    - New hidden correlations kept appearing.
    - Bias could not be reliably removed.

- Final outcome: By 2017-2018 Amazon abandoned the project completely.

Industry case: Amazon Hiring AI Failure

- https://www.hubert.ai/insights/why-amazons-ai-driven-high-volume-hiring-project-failed

- https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Data leakage

Data leakage

- Data leakage occurs when information that would not be available at prediction time is used during model training or evaluation, leading to unrealistically optimistic performance.

- Golden rule: Any information unavailable at prediction time is illegal.

- **Why Data Leakage Is Dangerous?**
  - Accuracy looks excellent
  - Model fails immediately in production

Data leakage

- Examples:
  - Predict if customer will leave→ Feature used: account_closed_date. Accuracy ≈ 99%, but impossible in reality. This variable exists only after churn happens.
  - Delivery Delay Prediction → Feature used: customer_refund_issued. Refund happens after delay.
  - Credit risk→Feature used: loan_restructuring_status. Occurs after repayment problems.

Data leakage

- If performance looks amazing → suspect leakage.

Why large data makes everything worse?

Why large data makes everything worse?

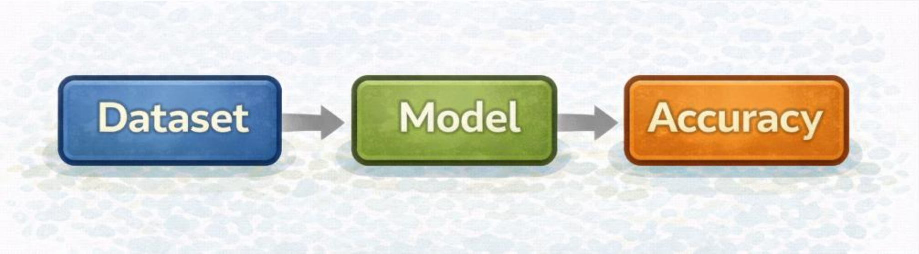
- Scaling amplifies the errors
- Large data = small problems at massive scale.

Problem	Small data	Large data
Missing values	Visible	Hidden
Bias	Manageable	Amplified
Errors	Detectable	Systemic
Debugging	Easy	Nightmare

Key industry mindset shift

Key industry mindset shift

Student mindset



Professional mindset

Teamwork time 2

Teamwork time 2

- A machine learning model has been successfully deployed in production.

- For several months:
  - accuracy was high
  - business decisions worked well
  - stakeholders trusted the system

- Suddenly:
  - Predictions become unreliable
  - Business KPIs drop
  - Complaints appear
  - Management asks for explanation

- **What happened?**

Teamwork time 2

- Team structure: each team contains 3 roles.

1. Data scientist- responsibility: model performance
   - Did accuracy drop?
   - Is there data drift?
   - Was retraining performed?

2. Business manager- responsibility: business impact
   - When did problems start?
   - Which customers are affected?
   - Revenue impact?

3. Data engineer- responsibility: pipeline & infrastructure
   - Did data schema change?
   - Missing columns?
   - Pipeline delay?

Teamwork time 2

- Present possible causes (e.g., data drift, missing features, etc)
- Decide the root cause category (model issue, data issue, pipeline issue, business issue)
- Propose a fix/plan (e.g., monitoring)
- Time: 10-15 minutes

Key takeaways

Key takeaways

- Data quality determines model quality
- Bias originates in data collection
- Leakage destroys evaluation
- Large data amplifies mistakes

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