Intelligent techniques for processing large and structured data

Lecture 4

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

Sergiu Limboi, PhD Teaching Assistant

Motto:“Patterns are easy to find in small data.
Finding reliable patterns in massive data is the real challenge.”

Large Scale Data Mining

AGENDA

- Warm-Up
- Why Large-Scale Data Mining Exists?
- Fundamental Data Mining Tasks
- Computational Complexity in Data Mining
- Techniques for Scalable Data Mining
- Teamwork Time 1
- Association Rule Mining
- Industry Case Studies
- Teamwork Time 2
- Key Takeaways

Warm-Up

Faculty of Mathematics and Computer Science

Warm-Up

Go to www.menti.com and enter the code **7598 6394**

or use the QR code

Why Large-Scale Data Mining Exists?

Why Large-Scale Data Mining Exists?

- Expectation: Data mining finds truth
- Reality: Data mining finds weird correlations

Why Large-Scale Data Mining Exists?

- Small datasets allow us to easily find patterns.

DATASET
10.000
customers

DATASET
1 billion users
100 billion
transactions

- We can easily detect
  - correlations
  - clusters
  - patterns.

- Computational cost
- Noise
- False patterns

Why Large-Scale Data Mining Exists?

- Large datasets contain:
  - measurement errors
  - missing values
  - irrelevant features
  - Duplicates
- More data does not mean more knowledge.
- Sometimes: **More data = more noise (noise amplification)**

Why Large-Scale Data Mining Exists?

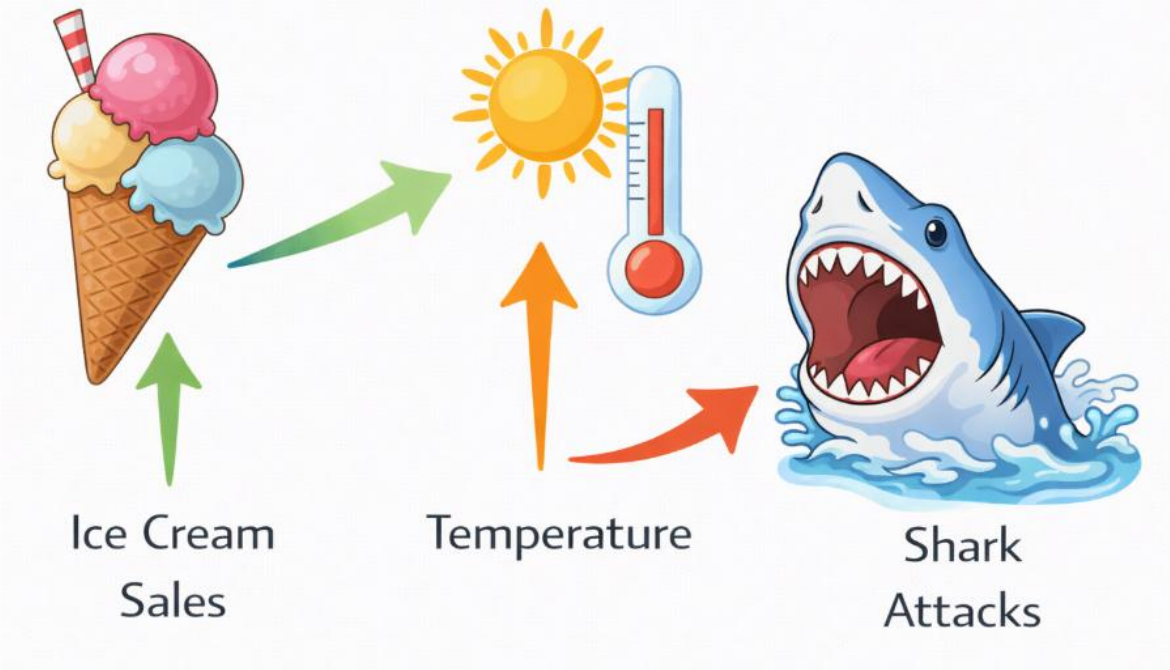
- False patterns
  - With billions of records, random correlations appear.

- Research dataset found correlation between:
  - number of people who drowned in swimming pools
  - number of films Nicolas Cage appeared in

spurious correlation

Why Large-Scale Data Mining Exists?

- Two variables may appear correlated because a third variable affects both.



Why Large-Scale Data Mining Exists?

- Hidden bias
  - Large datasets often contain bias.
  - Hiring dataset : 90% male engineers
  - Model learns: male → better engineer

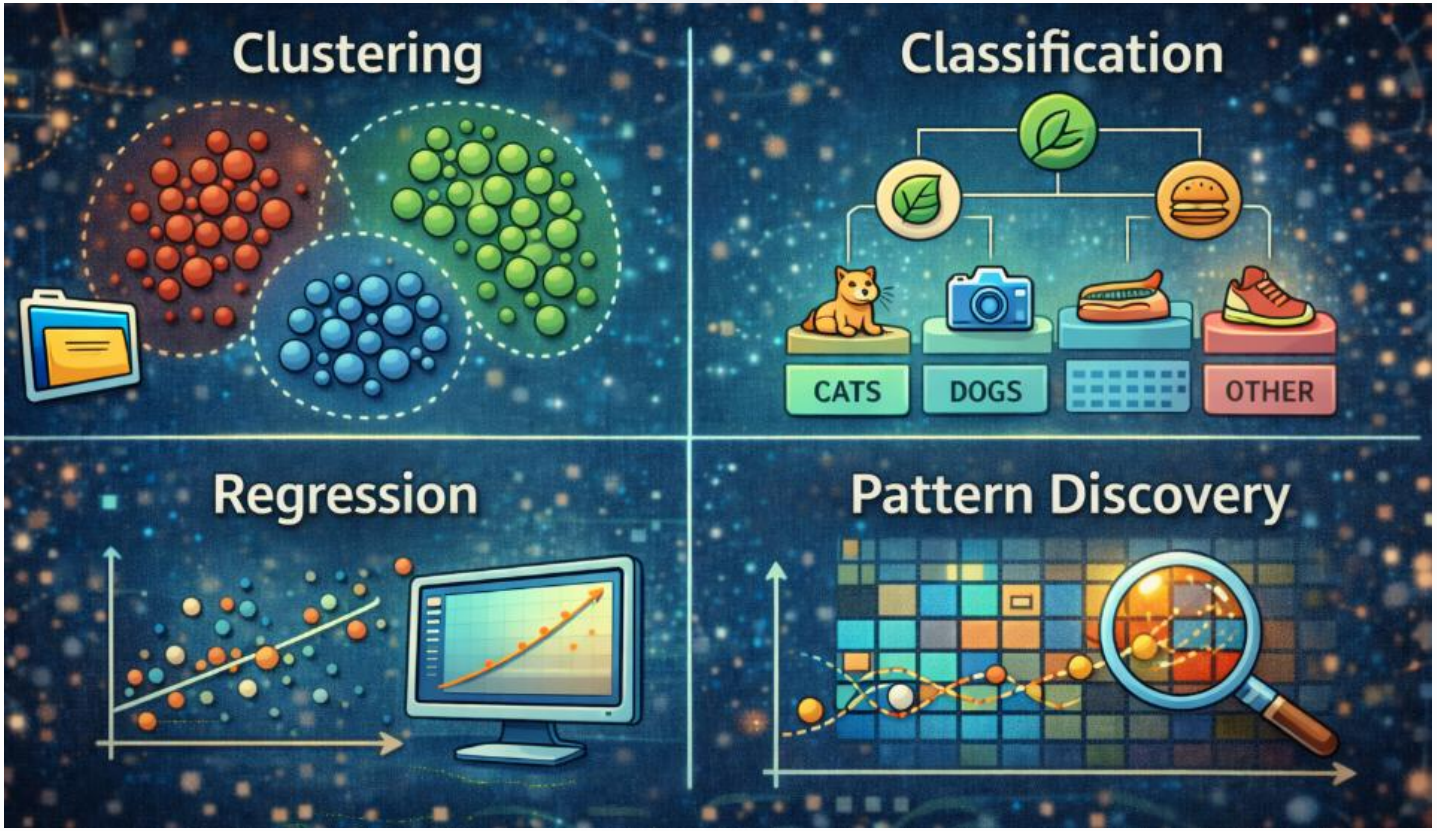
Why Large-Scale Data Mining Exists?

- The goal is not just prediction.
- The goal is discovering **patterns that influence decisions**.

- Examples:
  - Which products are bought together
  - Which users will churn
  - Which ad will generate revenue

Fundamental Data Mining Tasks

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

Pattern Discovery

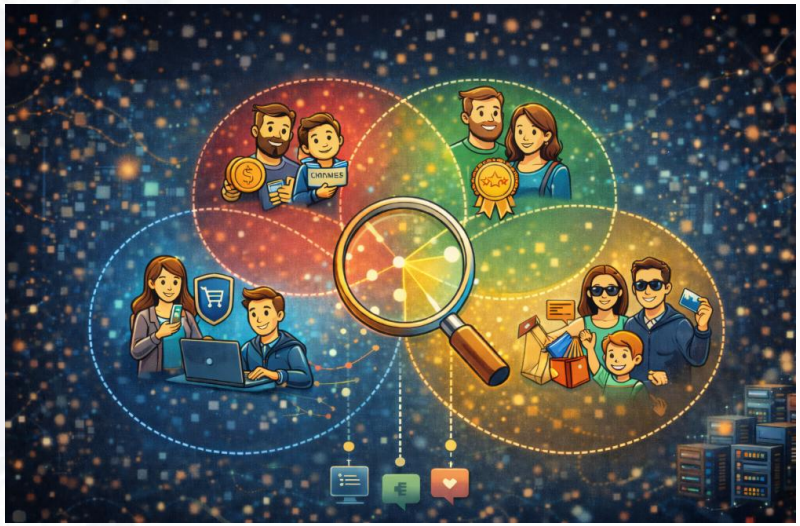
- Association rule mining is one of the most important technique to discover interesting relationships or structures in data.

- Algorithms for association rules:
  - Apriori
  - FP-Growth
  - ECLAT

- Evaluation metrics:
  - Support
  - Confidence
  - Lift

Clustering

- Goal: Group similar observations without labels.
- Example: customer segmentation
- Algorithms:
  - K-Means
  - Hierarchical clustering
  - K-Medoids
  - DBSCAN
  - Etc.



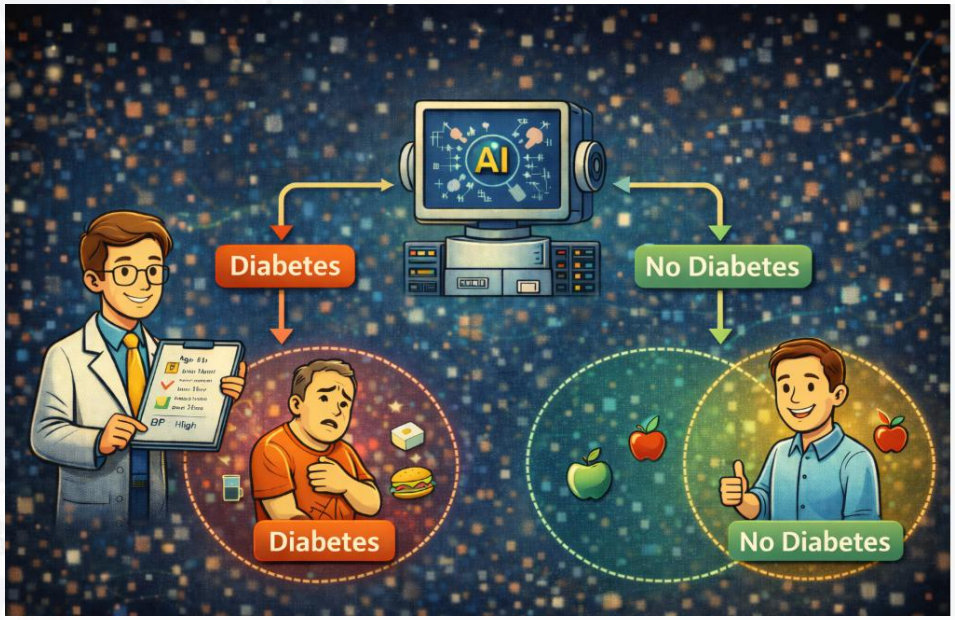
- Challenge: Distance computation becomes expensive.

Classification

- Goal: Predict categorical labels.

- Examples
  - spam detection
  - fraud detection
  - disease prediction

- Algorithms
  - logistic regression
  - random forest
  - decision trees
  - gradient boosting
  - Etc.



Regression

- Goal: Predict continuous values.

- Examples:
  - house price prediction
  - energy consumption
  - demand forecasting

- Algorithms:
  - Linear regression
  - Gradient boosting
  - Neural networks
  - Etc.

Computational Complexity in Data Mining

Computational Complexity in Data Mining

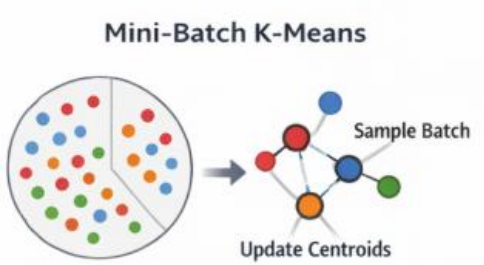
- A critical concept in large-scale mining is algorithm complexity.
- Classic algorithms fail on massive data.
- We usually express complexity using **Big-O notation**.

Complexity	Interpretation
O(n)	Linear growth
O(n log n)	Efficient for large data
$O(n^2)$	Expensive for large datasets
$O(2^n)$	Practically impossible

  - where n is the number of data points

Complexity in Common Data Mining Tasks

- Clustering→ classic K-Means
- Steps:
  - Initialize centroids
  - Compute distance
  - Update clusters



- Problem: Distance calculation expensive.

- Solution? Mini-batch K-Means ✔
  - Instead of full dataset, use small batches (e.g., 1000 samples)
  - Benefits: faster, scalable, similar accuracy

Complexity in Common Data Mining Tasks

- k-Nearest Neighbours
  - To classify a sample: compute distance to every point
  - Complexity: O(n x d), where n is the number of points, d number of features/dimensions
  - Example: 10 million users, 100 features →Now a single prediction requires 1 billion operations.

- Solution? ✔
  - KD-Trees (Space Partitioning)
    - Instead of comparing the query point with all points, the algorithm searches only in relevant regions.
  - Ball Trees
    - Instead of splitting by feature values, Ball Trees divide data into hyperspheres.
    - Each node contains: a centre, a radius, all points inside the sphere
  - Approximate Nearest Neighbour (ANN)
    - We only need very close neighbours.
    - Approximate Nearest Neighbour algorithms find neighbours much faster with small error.
    - Instead of searching the entire dataset, the algorithm searches only a subset of candidate points.

Techniques for Scalable Data Mining

Techniques for Scalable Data Mining

- Sampling
  - Instead of analysing the entire dataset, we use a representative subset
  - Example:
    - Dataset with 1 billion rows
    - Sample: 1 million rows
    - Benefits:
      - Faster training
      - Reduced memory
      - Often similar results
- Distributed Computing
  - Instead of one machine , we use cluster of machines
  - Example:
    - 10 nodes
    - Each node processes 10% of the dataset
    - Results are then aggregated
    - This paradigm follows the MapReduce model

Techniques for Scalable Data Mining

- Feature Reduction
  - High dimensionality increases complexity.
  - Example: 1.000 features → distance computation expensive
  - Solution? Dimensionality reduction
    - PCA
    - Feature selection
    - Autoencoders
- Approximate Algorithms
  - Instead of finding the exact best solution, they find a solution that is very close, and much faster to compute
  - Example: Instead of finding the exact nearest neighbour, the algorithm finds a very close neighbour.

Techniques for Scalable Data Mining

- Approximate Algorithms
  - Approximate Nearest Neighbour (ANN)→Find points that are very close to the query without comparing with all data points.
    - ANN methods search only a small subset of candidate points.
  - Graph-based ANN
    - Some modern ANN algorithms build graphs of neighbours.
    - Each data point connects to its nearest neighbours.
    - Searching becomes a graph traversal problem.
  - Locality Sensitive Hashing (LSH)
    - LSH hashes similar items into the same buckets.
    - Objects that are close in the feature space are likely to fall into the same hash bucket.
    - When searching neighbours, the algorithm only looks inside the same bucket.

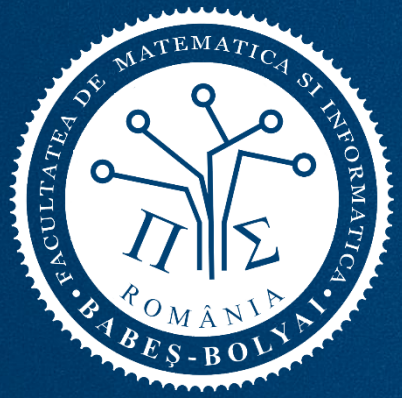
Teamwork Time 1

Teamwork time 1-Spotify's Secret Clusters

- You work as data scientists at Spotify.
- Spotify wants to automatically discover groups of similar users in order to improve music recommendations.
- The platform has:
  - 500 million users
  - billions of listening events
  - millions of songs
- Each user has features such as:
  - favourite genres
  - listening time per day
  - average song popularity
  - skip rate
  - playlist creation frequency
  - number of artists listened to
  - mood of songs
- Your task is to design a clustering solution.

Teamwork time 1-Spotify's Secret Clusters

- In teams of **3–4 students**, discuss the following questions.
- You have **10 minutes**.
- Mission:
  - What patterns could exist in this dataset?
    - What kinds of **user groups** might appear?
  - Which clustering algorithm would you use?
    - Questions to think about:
      - Do we know the number of clusters?
      - Could there be noise users?
      - Are clusters well separated?
  - What infrastructure would you need?
    - Spotify has hundreds of millions of users.
    - Think about scalability.
  - What insights could Spotify gain?
    - What patterns could help the company?
  - Each team must invent **some cluster names**.



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

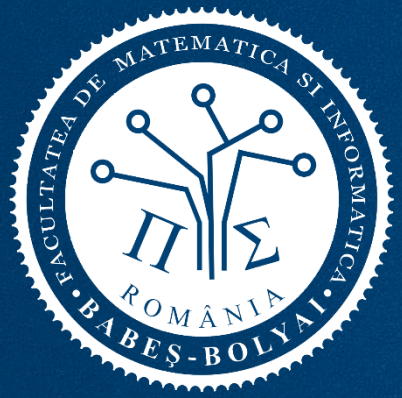
Algorithms for Association Rule ECLAT Algorithm

- 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.



Industry Case Studies

Industry Case Studies-Amazon

- Amazon uses large-scale data mining for:
  - Product Recommendations
  - Dynamic Pricing
  - Supply Chain Optimization
- Users buying iPhone also buy charger, screen protector.
- Amazon computes these patterns using massive transaction datasets.
- Amazon adjusts prices based on demand, competitor prices, user behaviour.
- Data mining identifies patterns like: high demand → increase price
- Amazon predicts future demand and pre-positions inventory.
  - This reduces delivery time.

Industry Case Study - Facebook

- Facebook mines social graphs.
  - Nodes→users
  - Edges→friendships
- Mining tasks
  - community detection
  - recommendation
  - ad targeting
- Example
  - If many friends like travel pages, the system recommends travel ads

Industry Case Study – Tik-Tok

- TikTok analyses billions of user interactions to recommend videos in the “For You” feed.

- The goal is to predict: Which video a user is most likely to watch next.

Teamwork Time 2

Teamwork time 2- Design a Data Mining System

- A large online platform similar to Amazon or TikTok wants to improve its recommendation system.

- The platform has:
  - 200 million users
  - 50 million products or videos
  - billions of interactions per day
- Examples of interactions:
  - views
  - likes
  - purchases
  - comments
  - search queries

- The company wants to build a data mining system to recommend content or products to users.

Teamwork time 2- Design a Data Mining System

- Time: 10 minutes
- Groups of students
- Discuss:
  - What data would you use?
  - What data mining task would you apply?
  - What algorithm could be used?
  - What challenge might appear at large scale?

Teamwork time 2- Design a Data Mining System

Why do recommendation systems often feel so accurate?

- In a future lecture, we will study recommendation systems in detail and see how companies like Amazon, Netflix, and Spotify build large-scale personalized recommendations using data mining and machine learning.

Key Takeaways

Key takeaways

- Large-scale data mining introduces computational challenges.
- Classical algorithms often do not scale.
- Large datasets create statistical risks
  - spurious correlations
  - bias
  - noise.
- Successful systems combine
  - scalable infrastructure
  - efficient algorithms
  - domain knowledge.
- The goal of data mining is not accuracy. The goal is reliable knowledge.

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